ID N. 006

UNIVERSITÀ CAMPUS BIO-MEDICO DI ROMA

DEPARTMENT OF ENGINEERING

ISTITUTO PER LA RICERCA E L'INNOVAZIONE BIOMEDICA
(IRIB) - CNR

Italian National Ph.D. in Artificial Intelligence
Health and Life Sciences
XXXVII Cycle

Multimodal AI to Unravel the Neurodevelopmental Cascade: From Early Movements to Advanced Communicative and Social Skills

Supervisors

Giovanni Pioggia

Gennaro Tartarisco

Candidate

Roberta Bruschetta

June, 2025

Acknowledgements

Completing this thesis would not have been possible without the support and guidance of many individuals and institutions.

First and foremost, I would like to express my deepest gratitude to my supervisors, Giovanni Pioggia and Gennaro Tartarisco, for their invaluable mentorship, insightful advice, and continuous encouragement throughout this research. Their expertise and dedication have been a cornerstone of this project.

I am also deeply thankful to Professors Salvatore Anzalone and Cristiano De Marchis, who generously dedicated their time to review my dissertation, providing me with important and constructive feedback that enriched this work.

I am also profoundly grateful to my colleagues, Simona Campisi, Liliana Ruta, Marilina Mastrogiuseppe, and the entire Early Change Team, for their stimulating discussions, constructive feedback, and constant collaboration, which have enriched this experience.

This research has been made possible by the financial and institutional support of IRIB CNR and the Campus Bio-Medico of Rome, to which I am deeply indebted.

I would also like to extend my gratitude to the research groups and institutions with whom I had the privilege to collaborate:

- Maria Luisa Scattoni and her team at the *Istituto Superiore di Sanità*, for their contribution to the research on newborns.

- Jana Iverson and her team at the *Sargent College of Health and Rehabilitation Sciences*, for welcoming me as a Visiting Scholar in her lab at Boston University for six months. This experience enriched my research and deepened my understanding of neurodevelopmental cascade theory.

- Giuseppe Di Cesare and his teams at the *Cognitive Architecture for Collaborative Technologies Unit, Italian Institute of Technology, Genoa*, and the *Department of Medicine and Surgery, University of Parma, Parma, Italy*, for involving me in his innovative research on vitality forms.

- Andrea De Gaetano and the *IASI Biomatlab team* of *CNR*, for their technical and mathematical expertise, which provided a fresh perspective on the analysis of visual attention.

- Olga Capirci of the *ISTC CNR*, for her clinical expertise and significant input in developing the coding scheme for children's gestural production.

- Bhismadev Chakrabarti of the *University of Cambridge*, for his contributions to the study of visual attention.

- Gessica Vasco and her team at the *Ospedale Bambino Gesù in Rome*, for their support and partnership, which enabled the successful completion of the study on dysarthria.

Abstract

This thesis explores the application of multimodal Artificial Intelligence (AI) to advance early diagnosis and understanding of Neurodevelopmental Disorders (NDD) within the framework of the neurodevelopmental cascade theory. NDD present significant diagnostic and therapeutic challenges due to their heterogeneity and overlapping symptoms, requiring new scalable and precise methodologies. By integrating AI across multiple domains and examining how motor, communicative, attentional, and socio-emotional behaviors dynamically interact throughout development, this research identifies early biomarkers, uncovers novel patterns, and addresses traditional assessment limitations.

A marker-less analysis of newborns' spontaneous movements, using Deep Learning (DL), identified kinematic patterns of delays in foot motor development in 10-day-old infants, predicting adverse outcomes with 85% accuracy. Furthermore, longitudinal analyses of motor behaviors during social engagement and object-reaching tasks revealed cascading effects of early hand movements at six months, highlighting how the emergence of reaching serves as a crucial precursor by enabling active exploration. This early motor foundation plays a pivotal role in shaping later communicative and social development.

A microanalytic study of gestures, gaze, and language coordination in naturalistic parent-child interactions characterized socio-communicative behaviors, highlighting reduced complexity in neurodivergent toddlers. Building upon this, an automatic coding system based on a transformer architecture was developed to identify deictic gestures from videos with high performance, demonstrating the feasibility of scalable gesture recognition. Gaze behaviors, modeled through a novel eye-tracking paradigm and Markov chains, revealed divergences in atten-

tional dynamics in preschool-aged children with NDD, such as increased gaze aversion and repetitive nonsocial focus, reflecting sensory coping strategies and their downstream effects on social cognition. Additionally, the analysis of the expression of vitality forms linked motor behaviors with socio-communicative skills, emphasizing how emotional expressions differ in social contexts, further reinforcing the cascading nature of developmental processes.

Finally, a hierarchical AI model combining Machine Learning (ML) and DL was developed to support automatic speech analysis in school-aged children, identifying dysarthria with 90% accuracy and stratifying its severity with 80% accuracy. This final stage of the research exemplifies how initial sensory-motor and attentional variations contribute to later expressive and linguistic differences, reinforcing the longitudinal approach of this study.

These findings demonstrate the dynamic potential of AI to integrate multimodal dimensions and unravel the complex interplay underlying neurodiverse trajectories. By adopting a developmental perspective, this study underscores how disruptions in foundational skills propagate through interconnected domains, influencing later functional outcomes. This approach paves the way for improved diagnostic accuracy, enabling earlier and more personalized interventions that align with individual neurodevelopmental profiles.

Contents

List of Figures

List of Tables

Abbreviations

10FCV 10-Fold Cross-Validation. 7, 97, 99, 100

AAC Augmentative and Alternative Communication. 45

ABA Applied Behavior Analysis. 21

ADHD Attention-Deficit/Hyperactivity Disorder. 18, 19, 21, 25, 26, 28, 42, 44, 138

ADI-R Autism Diagnostic Interview-Revised. 28

ADOS Autism Diagnostic Observation Schedule. 9, 28, 74, 75, 103, 113, 115, 138, 139, 150, 152, 231, 232

AI Artificial Intelligence. 3, 4, 9, 22, 27, 28, 31, 33, 39–45, 47–52, 54, 55, 71, 74, 81–83, 86–90, 103, 106, 110, 136, 137, 152, 154, 174, 176–181, 183, 230

AIMS Alberta Infant Motor Scale. 29

Ama Area differing from Moving Average. 59

ANN Artificial Neural Network. 37

AOIs Areas of Interest. 7–10, 113, 115–122, 129, 131, 133, 135, 182, 236

ASC Autism Spectrum Condition. 7–10, 18, 19, 22, 25, 26, 28, 30, 31, 33–35, 37, 39–45, 52, 74, 75, 90, 92, 96–103, 112–115, 119–123, 126, 129–135, 137–139, 141, 142, 146, 149–152, 181–183, 185, 231, 232, 236, 239–244

AUC Area Under the Curve. 37

C-SVC C-Support Vector Classifier. 64

CA Congenital Non-Progressive Ataxia. 155

CC Cross-Correlation. 58

CELF Clinical Evaluation of Language Fundamentals. 31, 43

CIMA Computer-based Infant Movement Assessment. 35

CNN Convolutional Neural Network. 36, 38, 39, 42–45, 54, 55, 143, 168, 169, 183

CP Cerebral Palsy. 20, 29, 33–35, 37, 38, 113, 154, 232, 233

CTMC Continuous-Time Markov Chain. 113, 116–118, 121, 133, 135, 182

DCD Developmental Coordination Disorder. 139

DCNN Deep Convolutional Neural Network. 40

DeepLabCut DeepLabCut Framework. 37, 38

DenseNet121 Dense Convolutional Network Model with 121 layers. 106

dfB degrees of freedom for between-groups. 63

dfW degrees of freedom for within-groups. 63

DL Deep Learning. 3, 4, 33, 34, 36, 39, 42, 52, 66, 110, 143, 151, 152, 154, 156, 167, 170, 172, 174–176, 180, 181, 183

DLD Developmental Language Disorder. 25, 26

DQ Developmental Quotient. 9, 115, 231

DSM-4 Diagnostic and Statistical Manual of Mental Disorders, 4th Edition. 74

DSM-5 Diagnostic and Statistical Manual of Mental Disorders, 5th Edition. 18, 103, 113, 232

DTMC Discrete-Time Markov Chains. 116

EfficientHourglass EfficientHourglass Model. 36

EfficientNet-B6 EfficientNet Model, Version B6. 38

EfficientPose EfficientPose Framework. 36

EL Elevated Likelihood. 25

ELAN Eudico Linguistic Annotator. 94–96

ESDM Early Start Denver Model. 93

FMs Fidgety Movements. 29, 35, 36

FRI Fluid Reasoning Index. 138

GMA General Movements Assessment. 29, 33, 35, 37

GMs General Movements. 37, 51, 52, 59, 60, 71, 180

GMT General Movements Toolbox. 34, 35

GRUs Gated Recurrent Units. 43

GTCCs Gammatone Cepstral Coefficient. 160, 162

H Healthy. 155

H-Ratio Harmonic Ratio. 164

HMLM Hierarchical Machine Learning Model. 8, 154, 170–176, 183

HR High-Risk. 35–37, 51, 52, 74, 75

ICC Intraclass Correlation Coefficient. 84

IQ Intellectual Quotient. 8, 138, 139

IS Infantile Spasms. 35

ISS Istituto Superiore di Sanità. 19, 53

k-NN k-Nearest Neighbors. 8, 9, 170, 173, 174

LCD Liquid Crystal Display. 114

LD language development delays. 75, 78, 81

Log-Reg Logistic Regression. 7, 97, 99, 100

LOOCV Leave-One-Out Cross-Validation. 7, 64, 97, 99, 100

LR Low-Risk. 74, 75

LSA Language Sample Analyses. 31

LSTM Long Short-Term Memory Network. 37–39, 42, 43

MFCCs Mel-Frequency Cepstral Coefficient. 160–162

MI Mutual Information. 97, 99

MIMAS Markerless Infant Movement Analysis System. 35

ML Machine Learning. 4, 8, 9, 34, 42–44, 54, 73, 101, 102, 112, 154, 156, 168–170, 172–175, 183

MRI Magnetic Resonance Imaging. 155

MSB Mean Squares Between-groups. 62

MSW Mean Squares within Groups. 62, 63

NB Naïve Bayes. 8, 170, 173

NDBI Naturalistic Developmental Behavioral Intervention. 93

NDD Neurodevelopmental Disorders. 3, 4, 7, 18–23, 25–28, 30, 33–35, 43–53, 56, 58, 59, 61–64, 67–73, 89, 90, 112, 136, 154, 175–180, 182–184, 227–229

NICU Neonatal Intensive Care Unit. 37, 38

NIDA Network for Early Detection of ASC. 52

NLP Natural Language Processing. 31, 43, 44

NS no symptoms. 75, 78, 81

NT Neurotypical. 7–10, 44, 48, 52, 53, 58, 61–64, 67–72, 88, 92, 96–102, 112–115, 118–123, 126, 129–135, 137–139, 141, 142, 146, 149–152, 181, 183, 227, 228, 231, 232, 236, 239–244

OP Object Presentation. 76, 81, 83, 84

P Periodicity. 60

PA Progressive Ataxia. 155

PBDEs Polybrominated Diphenyl Ethers. 20

PCA Principal Component Analysis. 96, 100, 113, 118, 119, 121, 122, 133–135, 182

PCBs Polychlorinated Biphenyls. 20

PCI Parent-Child Interaction. 93

PDMS Peabody Developmental Motor Scales. 29

PPVT Peabody Picture Vocabulary Test. 31

PSI Processing Speed Index. 138

R Pearson's Correlation Coefficient. 58

RAVEN Raven's Progressive Matrices. 150

REML Restricted Maximum Likelihood. 64

ResNet-152 Residual Network Model with 152 layers. 38

RMS Root Mean Square. 83

RMSE Root Mean Square Error. 58, 65, 66

RNN Recurrent Neural Network. 39

RPM Raven's Standard Progressive Matrices. 139

SARA Scale for the Assessment and Rating of Ataxia. 154, 156, 170

SCQ Social Communication Questionnaire. 74

SE Standard Error. 68, 69, 86, 87

SHAP SHapley Additive exPlanations. 97, 98, 100, 101

SIDS Sudden Infant Death Syndrome. 37

SLI Specific Language Impairment. 26, 28, 30, 44

SM Spontaneous Movements. 7, 76, 78, 81, 83

SMI SensoMotoric Instruments. 114, 115

SSB Sum of Squares Between Groups. 62

SSBD Self-Stimulatory Behavior Dataset. 39

SSRIs Selective Serotonin Reuptake Inhibitors. 21

SSRs Sensory Social Routines. 113, 115, 118, 119, 121, 133–135, 182

SSW Sum of Squares Within Groups. 63

STFT Short-Time Fourier Transform. 168

SVM Support Vector Machine. 7–9, 35, 37, 40, 64, 65, 70, 168, 170, 173, 174

tDCS Transcranial Direct Current Stimulation. 21

TEC Test of Emotion Comprehension. 8, 138, 139, 150

TMS Transcranial Magnetic Stimulation. 21

VCI Verbal Comprehension Index. 138

VF Vitality Form. 8, 10, 49, 136, 137, 139–142, 149–153, 183, 239

VGGish Visual Geometry Group CNN. 168, 169

VideoMAE Video Masked Autoencoder. 39

VSI Visual Spatial Index. 138

WHO World Health Organization. 18, 22

WISC-IV Wechsler Intelligence Scale for Children, 4th Edition. 138, 150

WMI Working Memory Index. 138

Chapter 1

Introduction

1.1 Overview of Neurodevelopmental Disorders

1.1.1 Definition

Neurodevelopmental Disorders (NDD) include a range of conditions that originate during the developmental period, affecting brain and nervous system functions. These conditions disrupt typical neurodevelopment, leading to varying levels of impairment in social, cognitive, and emotional functioning [1]. According to the Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-5) [2], NDD include conditions such as intellectual disabilities, communication disorders, Autism Spectrum Condition (ASC), Attention-Deficit/Hyperactivity Disorder (ADHD), specific learning disorders, and motor disorders. Understanding their diverse presentations and ensuring early and accurate diagnosis are essential for tailoring effective interventions that improve developmental outcomes [3]. Globally, NDD represent a significant public health issue. The World Health Organization (WHO) estimates that ASC alone affects approximately 1 in 160 children worldwide, though prevalence rates vary, with estimates ranging from 0.5% to 2.5% depending on the region [4]. In the United States, about 15% of children aged 3 to 17 are reported to be affected by NDD, with ADHD, learning disabilities, and ASC being the most diagnosed conditions [5]. These disorders are notably more prevalent among males, with gender ratios of approximately 4:1 for ADHD and 3:1

for autism [4, 6]. In Europe, neuropsychiatric conditions, including NDD, are the third leading cause of premature mortality, underscoring their impact on public health and quality of life. In Italy, recent data from the Istituto Superiore di Sanità (ISS) reveal that approximately 1 in 77 children are diagnosed with ASC, further emphasizing the importance of ensuring accessible services and early detection mechanisms to mitigate long-term impacts [7].

1.1.2 Clinical Presentation and Diagnostic Challenges

NDD are typically diagnosed in early childhood, with symptoms often becoming apparent by the age of five [1]. While these disorders primarily emerge in childhood, they frequently persist into adolescence and adulthood, with symptoms that may evolve over time. This dynamic progression can delay recognition and diagnosis, as certain manifestations become more noticeable with age [8].

Diagnosing NDD presents significant challenges due to the overlap of symptoms across conditions and the frequent presence of comorbidities. For instance, individuals with ASC often exhibit intellectual disabilities, while those with ADHD commonly present specific learning disorders. This symptom overlap, combined with the wide variability in severity and impact across domains such as language, learning, memory, motor coordination, and social functioning, complicates differential diagnosis [3].

The heterogeneity of NDD further adds to this complexity. Symptoms can vary widely among individuals, ranging from specific learning difficulties to impairments in executive function and social skills. Early signs may be identified through screening tools, but more complex cases often require comprehensive clinical evaluations to ensure an accurate diagnosis [6].

NDD are characterized by delays or atypical development in key areas such as language, social skills, and motor coordination [9]. Affected children may face significant communication challenges, exhibit atypical behaviors, and struggle to form friendships or adhere to social norms, such as taking turns in conversations or respecting personal boundaries. For example, children with ADHD often display hyperactivity, impulsivity, or difficulty maintaining focus on tasks. Additionally, many NDD involve altered sensory processing, where children may be hypersensi-

tive to stimuli, such as sounds, lights, or textures, or seek sensory input through repetitive behaviors like spinning or rocking [3].

Emotional and behavioral difficulties are also prevalent in NDD. Many individuals experience heightened levels of anxiety, depression, or aggression, often coupled with challenges in emotional regulation. These issues can significantly impact personal, social, academic, and occupational functioning, leading to difficulties in areas such as emotional control, learning processes, and memory retention. For instance, children with specific learning disorders may struggle with tasks requiring sustained attention and organizational skills, despite having intact intelligence and strong perceptiveness, factors that can profoundly affect their self-esteem [10].

1.1.3 Causes and Risk Factors

Understanding the causes of NDD is essential for developing effective interventions and advancing research. Although the exact etiology remains largely unknown, NDD are believed to result from a complex interplay of genetic, environmental, neurological, and prenatal factors. Genetic predisposition plays a central role, with numerous mutations and genetic variations identified as contributors [11]. These include both inherited alterations and de novo mutations that arise spontaneously during fetal development.

Environmental factors also significantly influence NDD risk. Adverse early experiences, such as exposure to toxic chemicals, chronic stress, or neglect, can exacerbate vulnerability to these conditions [6]. Additionally, environmental contaminants like lead, methylmercury, and Polychlorinated Biphenyls (PCBs) are well-established risk factors for neurodevelopmental impairment. For instance, lead exposure has been strongly associated with cognitive deficits and attention difficulties, while prenatal exposure to methylmercury has been linked to delays in cognitive development [12]. Emerging research is further exploring the effects of other chemicals, such as pesticides, flame retardants (e.g., Polybrominated Diphenyl Ethers (PBDEs)), and phthalates, highlighting their potential associations with behavioral and learning challenges [6].

Neurological factors are also implicated in the pathogenesis of NDD. Conditions such as Cerebral Palsy (CP) or epilepsy can disrupt typical brain development

and function [13]. The prenatal period is particularly critical, as maternal factors, including substance use, malnutrition, and high stress levels during pregnancy, can adversely affect fetal brain development and increase the likelihood of NDD [14].

1.1.4 Current Treatment Strategies

Due to the complex and varied nature of NDD, treatment must be highly individualized to address each person's specific needs [15]. A multidisciplinary approach is often required to manage symptoms effectively and improve quality of life. Key strategies include behavioral interventions, pharmacotherapy, and neuromodulation techniques [16, 17, 18]:

- **Behavioral Interventions:** These therapies aim to enhance social, communication, and behavioral skills, often involving structured programs such as Applied Behavior Analysis (ABA) or social skills training. Family participation plays a crucial role in reinforcing therapeutic goals, providing emotional support, and maintaining consistency across settings.

- **Pharmacotherapy:** Medications are used to manage core symptoms, such as hyperactivity and inattention, or to treat co-occurring conditions like anxiety, depression, or mood dysregulation. For example, stimulants such as methylphenidate are commonly prescribed for ADHD, while Selective Serotonin Reuptake Inhibitors (SSRIs) may be used to address anxiety or depression.

- **Neuromodulation Therapies:** Non-invasive techniques such as Transcranial Magnetic Stimulation (TMS), Transcranial Direct Current Stimulation (tDCS), and Neurofeedback are emerging as promising tools for managing symptoms with minimal side effects. These techniques have shown potential in modulating neural activity and improving attention and behavior in targeted brain regions, while Neurofeedback offers real-time brain activity monitoring to train self-regulation skills.

1.1.5 The Importance of Early Diagnosis

Early diagnosis is critical in shaping the developmental trajectories of children with NDD. The WHO emphasizes that prompt identification enables timely interventions, providing children with tailored strategies to improve their developmental outcomes and daily functioning [4]. Evidence strongly supports that early interventions lead to substantial gains in cognitive, social, and adaptive skills, key areas often impacted by NDD. These interventions, such as behavioral therapy, speech therapy, and family-centered support, can significantly enhance a child's quality of life [19, 20]. However, barriers such as variability in symptom presentation and limited access to specialized services often delay diagnosis. For instance, subtle early signs of conditions like ASC may go unnoticed, particularly in areas with fewer diagnostic resources or trained professionals [21]. To address these challenges, structured screening processes and accessible diagnostic resources are essential. Innovative tools like AI-driven diagnostic platforms and standardized behavioral checklists can enhance early identification and reduce diagnostic delays [22]. Additionally, increasing awareness among healthcare providers through training programs and expanding early diagnostic resources are fundamental to ensuring timely and effective care for children with NDD [3, 7].

1.2 The Neurodevelopmental Cascade Theory for Understanding Developmental Pathways in Children

1.2.1 The Developmental Cascades Framework

As highlighted in the previous section, NDD constitute a diverse and heterogeneous group, characterized by significant variability in the manifestation, severity and progression of core diagnostic behaviors and functional capabilities across multiple developmental domains. This diversity necessitates a conceptual framework that flexibly examines cross-domain interactions while accounting for the modulatory influence of environmental factors [1]. A multimodal perspective is essential to understand the reciprocal and evolving relationships among motor, cognitive, linguistic, and socio-emotional domains. Both the timing of key developmental milestones and real-time interactions within everyday contexts are pivotal in shaping the developmental trajectories of children with NDD [23, 24].

Recent research underscores the interdependence of developmental domains, demonstrating that advancements or delays in one area can trigger cascading effects that significantly influence functioning in other domains, ultimately shaping a child's overall developmental trajectory [25, 26, 27]. Importantly, subtle early deviations can be amplified through these cascading processes, culminating in pronounced long-term developmental outcomes [28].

The developmental cascades framework integrates foundational theoretical models, including Gottlieb's developmental behavioral genetics framework [29], Smith and Thelen's dynamic systems theory [30], Sameroff's transactional model [31], and Karmiloff-Smith's neuroconstructivist approach [32]. Collectively, these paradigms conceptualize development as the cumulative result of continuous, bidirectional interactions among biological substrates, environmental contexts, and cross-domain influences [28]. By synthesizing these perspectives, the developmental cascades framework offers a comprehensive approach to investigating how these complex interactions unfold over time, ultimately shaping individual developmental pathways [33].

1.2.2 Applying Developmental Cascades Theory to Infancy

Extending the developmental cascades framework to infancy, a period marked by rapid neurodevelopment and heightened neural plasticity [34], provides critical insights into the mechanisms driving both typical and atypical developmental trajectories. During this phase, the attainment of foundational milestones is shaped by complex, reciprocal interactions between biological predispositions and environmental influences. Understanding how developmental cascades unfold in infancy is essential to elucidate how early differences or delays can progressively reshape developmental pathways, potentially magnifying disparities across cognitive, linguistic, motor, and social-emotional domains [35, 36]. Even subtle perturbations in early development can initiate cascading effects, where initial delays compound over time, hindering the acquisition of more complex skills.

This framework offers a nuanced understanding of neurodiversity by emphasizing the temporal dynamics and interdependence of developmental domains. It highlights how early behaviors and bioregulatory processes unfold over time, influencing subsequent developmental trajectories. The timing, quality, and consistency of these early processes are pivotal in shaping long-term outcomes, as they contribute to cascading effects across multiple functional domains [37]. By investigating these early mechanisms, researchers can identify sensitive periods in development—windows of heightened plasticity—during which targeted interventions can have the greatest impact, ultimately refining early detection and prevention strategies [36].

The Role of Early Motor Development Motor development plays a foundational role in infancy, acting as a gateway for environmental exploration and multimodal learning [38]. As infants acquire motor skills such as independent sitting or crawling, their capacity to engage with their surroundings expands, which in turn stimulates growth in cognitive, communicative, and social domains [25]. For example, head and neck control facilitates torso stability, enabling infants to engage in sustained visual exploration. Achieving independent sitting provides new vantage points, enhancing object manipulation and social referencing, while crawling allows infants to actively seek social partners and explore novel environ-

ments [28].

These motor milestones promote physical autonomy, catalyzing advancements in language and social cognition. Sitting upright expands the visual field, enabling infants to track caregivers' gestures and gaze more effectively. This enhanced visual access supports the emergence of joint attention, a foundational skill for effective communication [39, 40]. Furthermore, the shift in posture associated with sitting alters the vocal tract's configuration, facilitating the production of speech-like sounds, including canonical babbling [41]. Conversely, delays in motor development can disrupt these cascading processes, limiting infants' exposure to crucial social and linguistic inputs. Infants at Elevated Likelihood (EL) for neurodevelopmental disorders frequently exhibit delayed motor milestones, which may dampen subsequent advancements across multiple developmental domains [42, 43, 44].

Communicative Gestures in NDD Gestures form a crucial bridge between motor and communicative development, serving as early indicators of both linguistic and social competencies [45]. They emerge from an infant's ability to coordinate motor actions with social attention, exemplifying how cascading developmental processes link early motor skills to later communicative abilities. Appearing shortly after foundational motor milestones, communicative gestures play a pivotal role in language acquisition. Actions such as pointing, showing, and waving are fundamental to early communication, and their production serves as a key diagnostic marker for NDD, including ASC, ADHD, and Developmental Language Disorder (DLD). Delays or atypical patterns in gesture production often arise as some of the earliest indicators of neurodevelopmental divergence [46, 47], offering a critical window for early detection.

Toddlers at EL for NDD, such as siblings of individuals with ASC or those born preterm, frequently exhibit a reduced frequency and diversity of gestures [48]. By two years of age, toddlers with ASC typically produce significantly fewer joint attention gestures than their neurotypical peers [46], highlighting difficulties in integrating nonverbal communicative cues with gaze. Given that gestures often precede and predict verbal language development, early deficits in this domain can hinder vocabulary expansion, pragmatic language use, and social engagement [49].

Importance of Visual Attention and Gaze Behavior Visual attention and gaze behavior exemplify the cascading nature of neurodevelopment, shaping cognitive and social trajectories from infancy onward. Visual attention is integral to cognitive and social development, influencing how infants engage with their environment and extract information from social cues. From early infancy, gaze-following and sustained attention to social stimuli establish the foundation for reciprocal social interactions. Disruptions in attentional processes can cascade into broader challenges in learning, communication, and social interaction.

Infants and toddlers with NDD frequently exhibit atypical gaze patterns, such as reduced eye contact or a preference for non-social stimuli, potentially restricting their opportunities to learn from social exchanges [50, 51]. For example, toddlers with ASC may focus on object features rather than faces, hindering the development of theory of mind and social cognition [52]. Similarly, children with ASC often show diminished attention to faces and social stimuli, whereas those with ADHD typically struggle with sustained attention and task persistence [53, 50].

Language Development and its Cross-Domain Impact Language development, deeply interconnected with earlier motor, gestural, and attentional processes, represents a fundamental pillar of neurodevelopment. Preverbal behaviors, including vocalizations, eye contact, and gesture use, support the development of subsequent speech and literacy skills [54, 55]. Language development is a complex, multidimensional process that interacts with cognitive, motor, and social domains. Disruptions in early language acquisition often cascade into broader difficulties in academic achievement, social interaction, and adaptive functioning. Early language abilities in toddlers are strong predictors of later academic performance, influencing literacy, executive functions, and social cognition [33]. However, language deficits associated with NDD vary widely across conditions such as ASC, SLI, and Down syndrome, necessitating individualized assessment approaches. Delays in this domain, as observed in DLD, can have long-lasting effects on reading proficiency, expressive language, and peer interactions, highlighting the importance of early intervention [33].

Symbolic gestures serve as a crucial mechanism in word learning, facilitating the transition from nonverbal to verbal communication. Given their predictive

role in language acquisition, gestures act as early markers of linguistic and social competencies. Notably, cross-linguistic research indicates that bilingual toddlers can transfer lexical and conceptual knowledge across languages, reinforcing linguistic development through interdependent processing. This phenomenon exemplifies how cascading effects extend beyond single-domain boundaries, shaping cognitive flexibility and enhancing metalinguistic awareness [33].

Implications for Early Intervention and Clinical Practice The developmental cascade framework emphasizes the necessity of identifying early indicators, such as motor milestones, gesture use, and gaze behavior, to inform the design of personalized, cross-domain intervention strategies [56]. Leveraging technologies like eye-tracking and sensor-based assessments offers clinicians and researchers objective, continuous data streams, enhancing diagnostic precision and tailoring intervention timing to individual developmental profiles [57]. Early, multimodal interventions that harness the interconnectivity of developmental domains can effectively alter trajectories, reducing long-term functional impairments and improving quality of life for infants and toddlers with NDD [56].

Integrating cutting-edge methodologies, including AI, expands the potential for scalable and accessible early assessments. Integrating these innovations within the developmental cascades framework enables practitioners to move beyond static, domain-specific assessments, shifting toward comprehensive models that reflect the dynamic interplay among biological, behavioral, and environmental factors [58]. This holistic paradigm enhances early identification efforts and guides the design of interventions that are developmentally appropriate, family-centered, and outcome-oriented, ultimately promoting more favorable long-term trajectories.

1.3 Limitations and Challenges of Traditional Assessment Methods

As described in the previous section on the neurodevelopmental cascade, diagnosing NDD involves navigating a complex interplay of developmental processes, where early disruptions in one domain can cascade into subsequent difficulties across multiple functional areas. This inherent complexity is amplified by the wide variability in symptom presentation, severity, and the frequent occurrence of comorbidities. Disorders such as ASC, ADHD, and Specific Language Impairment (SLI) are typically identified during early childhood but can emerge along different developmental trajectories, with manifestations ranging from subtle delays to profound impairments affecting cognitive, linguistic, motor, and social domains. Accurate and timely diagnosis is essential, as early interventions significantly enhance outcomes across cognitive, social, and daily adaptive functioning, ultimately improving the child's quality of life [21]. However, capturing the nuanced progression of symptoms within and across these domains remains a significant challenge.

Traditional assessments rely heavily on developmental history, observational methods, and standardized diagnostic tools. Instruments like the Autism Diagnostic Observation Schedule (ADOS) [59] and Autism Diagnostic Interview-Revised (ADI-R) [60] are considered "gold standards" for diagnosing ASC, using structured observations and comprehensive parental interviews to assess social communication and restricted, repetitive behaviors. Despite their clinical value, these methods face notable limitations. Comorbidities such as motor or language impairments often overlap with core NDD symptoms, complicating differential diagnosis and potentially delaying intervention [21]. Additionally, socio-economic and cultural factors can influence access to healthcare services and the interpretation of developmental milestones, further widening diagnostic disparities [21].

This chapter explores these challenges through an examination of assessment methods across key developmental domains, motor skills, gesture production, visual attention, and language, highlighting their limitations and the potential of emerging technologies to address them. By integrating advances in AI into clinical practice, it is possible to enrich traditional methods with more objective, scalable,

and continuous assessments that better capture the dynamic nature of neurodevelopmental trajectories.

Early Motor Skills Clinical assessments of motor performance traditionally employ standardized tools such as the General Movements Assessment (GMA) [61], Alberta Infant Motor Scale (AIMS) [62], and Peabody Developmental Motor Scales (PDMS) [63].

The GMA is widely recognized for its predictive validity, particularly in detecting conditions like CP and other motor disorders in preterm infants. By analyzing spontaneous movement patterns, especially Fidgety Movements (FMs) between 9 and 16 weeks post-term, the GMA provides early insights into potential motor dysfunctions [64]. The AIMS focuses on gross motor maturation through postural assessments in various positions, helping map an infant's developmental trajectory [62], while the PDMS evaluates both fine and gross motor skills, offering a comprehensive overview of motor abilities [63].

Despite their clinical utility, these methods have significant limitations. Assessments often rely on subjective interpretations of observable behaviors, making them vulnerable to inter-rater variability [65]. External factors, such as the child's mood, fatigue, or the testing environment, can further influence outcomes, reducing reliability. Moreover, these tools provide only momentary snapshots of development rather than continuous insights, which are essential for tracking the cascading effects of early motor delays on subsequent cognitive and social functioning. Their administration is time-consuming, requires specialized training, and is often inaccessible in resource-limited settings [65]. These constraints highlight the need for scalable alternatives capable of providing objective, longitudinal data. Digital tools and sensor-based methods, offering automated and real-time assessments, are emerging as promising solutions, enabling earlier identification of motor delays and facilitating targeted interventions that consider broader developmental trajectories.

Gesture Production Conventional assessments of gestures involve clinician-led observations or video coding to evaluate gesture frequency, variety, and communicative function. While these methods are clinically informative, they are

labor-intensive, subject to observer bias, and often fail to capture subtle but diagnostically significant features such as movement fluidity, temporal coordination, and spatial accuracy [66, 67].

Children with NDD frequently exhibit atypical gesture patterns. For instance, those with ASC may display gestures that are abrupt or lack synchrony with speech and gaze, reflecting broader deficits in motor planning and social engagement [45]. Similarly, children with SLI might rely on gestures to compensate for verbal limitations, though these gestures are often less complex and less fluid [68]. Delays in joint attention gestures, such as pointing or showing, are particularly salient markers of ASC and are critical for subsequent language acquisition and social development [47]. Traditional methods may overlook these nuanced deviations, limiting early diagnostic sensitivity.

Emerging automated approaches, including motion capture and computer vision technologies, offer objective and scalable alternatives. These systems can quantify gesture kinematics, such as trajectory, velocity, and temporal coordination, with greater precision, enhancing the detection of atypical patterns even in naturalistic settings [66, 69]. These tools allow for real-time feedback and longitudinal monitoring, supporting the assessment of how early motor-communicative interactions influence later social and linguistic development.

Visual Attention Eye-tracking technology has revolutionized the assessment of visual attention, providing non-invasive and quantitative measures of gaze patterns and fixation metrics. These methodologies offer critical insights into attentional preferences and processing strategies, playing a pivotal role in identifying early attentional divergences in infants and toddlers at risk for NDD. By capturing real-time data on gaze allocation, eye-tracking facilitates the early detection of atypical social attention, allowing for preemptive interventions designed to enhance social engagement and mitigate downstream effects on language, cognition, and adaptive functioning [70, 57].

However, traditional eye-tracking systems face several limitations, including calibration challenges, sensitivity to environmental conditions, and the need for specialized equipment and expertise. These constraints hinder their applicability in ecologically valid settings and reduce scalability [71]. Moreover, conventional

approaches often capture attentional behavior at isolated time points rather than continuously, making it difficult to assess how early attentional deficits evolve and impact later socio-communicative and cognitive development. To maximize the potential of eye-tracking in developmental research and clinical practice, future advancements should prioritize increased adaptability, integration with naturalistic environments, and enhanced longitudinal monitoring capabilities.

Advances in AI and portable eye-tracking devices now enable more accessible, real-time assessments in naturalistic environments. AI-driven models enhance the analysis of gaze patterns by extracting and interpreting complex eye-tracking signals, integrating them with facial expression and speech metrics to provide a more comprehensive understanding of attentional and social behaviors [71]. These innovations improve early detection and characterization of attentional atypicalities, supporting the development of targeted interventions aimed at mitigating long-term developmental challenges.

Language Assessment Traditional methods, such as the Clinical Evaluation of Language Fundamentals (CELF) [72], Peabody Picture Vocabulary Test (PPVT) [73], and Language Sample Analyses (LSA) [74, 75], focus on evaluating core linguistic components, syntax, vocabulary, and comprehension.

While standardized assessments provide valuable normative benchmarks, they often fail to capture pragmatic and contextual aspects of language use, particularly in children with ASC, who may perform well in structured tasks but struggle in spontaneous conversation [76, 77]. LSA, which involves detailed analysis of spontaneous speech, offers deeper insights into pragmatic competence but is time-consuming and requires significant manual effort, limiting its scalability [78].

Recent advances in AI and Natural Language Processing (NLP) have enabled the development of automated tools capable of efficiently analyzing large volumes of language data. These technologies can assess syntactic complexity, lexical diversity, and discourse coherence, providing objective, reproducible metrics that enhance the precision of language evaluation [79]. Crucially, they facilitate longitudinal monitoring of language development, offering insights into how early communicative delays shape broader developmental trajectories. By minimizing the reliance on labor-intensive transcription and manual coding, these tools im-

prove the feasibility of comprehensive, ecologically valid assessments that more accurately capture a child's everyday communicative functioning.

1.4 AI Applications to Advance Diagnosis and Assessment of Neurodevelopmental Disorders: Related Works

Building upon the limitations of traditional assessment methods discussed in the previous section, recent advancements in artificial intelligence (AI) have opened new avenues for investigating and assessing NDD. Unlike conventional approaches, often constrained by manual coding, observer bias, and limited scalability, AI-driven tools enable rapid data processing, advanced feature extraction, and real-time analysis across multiple functional domains, including motor skills, visual attention, social communication, and language development [58]. By integrating multimodal data, these technologies offer a richer, more holistic understanding of individual developmental trajectories, enhancing both diagnostic accuracy and intervention planning [80, 81].

This section reviews recent works that leverage AI to address diagnostic challenges, focusing on applications across key developmental domains.

1.4.1 Assessment of Early Motor Skills

Early detection of motor delays, particularly in conditions such as cerebral palsy (CP) and autism spectrum conditions (ASC), has greatly benefited from AI-based methods. These technologies enable precise and scalable tracking of motor behaviors, surpassing the limitations of manual assessments. In particular, video-based deep learning (DL) methods have emerged as a leading solution for analyzing infants' general movements, facilitating the early identification of subtle motor delays [82, 83].

These AI methods enhance tools like the General Movement Assessment (GMA), which captures spontaneous movements, key indicators of neurological development [84]. Unlike traditional methods, AI-powered tools can analyze movement patterns with exceptional precision and track the trajectory of an infant's limbs in real-time, enabling longitudinal monitoring across uncalibrated, natural environments. This approach ensures a more accurate reflection of developmental

milestones and reduces the need for controlled settings [85].

1.4.1.1 Marker-less Analysis of Infants' Movements

Non-intrusive video-based methods for assessing infants' movements have gained increasing attention for their ability to capture naturalistic behavior in a variety of settings [82]. These approaches can be broadly divided into two main groups: those employing classical computer vision algorithms and those using DL techniques. Classical methods typically rely on predefined models or manually designed features to track and analyze motion, which, while effective in controlled environments, often struggle with the complexity and variability of human movements. In contrast, DL approaches leverage advancements in neural networks to automatically extract and analyze intricate motion patterns directly from video data, offering greater robustness and scalability [84].

The key strength of these marker-less systems lies in their unobtrusive design, which allows infants to move freely without the need for physical sensors or markers. This ensures more natural and representative assessments of motor behavior, enhancing both the ecological validity and practical utility of the data. Recent innovations, such as the integration of depth-sensing cameras and advanced ML algorithms, have further improved the accuracy and feasibility of these systems, enabling real-time analysis and expanding their applicability to natural environments, including homes and clinics [82].

Despite these advancements, challenges persist, particularly in standardizing methodologies and validating these systems across diverse populations and settings. Addressing these issues will be crucial to ensuring the widespread adoption of marker-less approaches in clinical and research contexts, as they hold great promise for advancing the understanding of infants' neurodevelopmental trajectories [82].

Traditional Computer Vision Methods Classical computer vision techniques have been widely employed to analyze infants' movements, providing valuable tools for the early detection of NDD such as CP and ASC. Among these, the General Movements Toolbox (GMT) developed by Adde et al. [86] represents a pivotal contribution. The GMT utilizes change detection algorithms to quantify motion

parameters, including variability, velocity, and acceleration, derived from pixel differences between consecutive video frames. By automating the assessment of FMs, a key marker for CP, the GMT provides an objective and accessible alternative to the traditional GMA, which relies on expert interpretation.

Similarly, Tacchino et al. [87] introduced the Markerless Infant Movement Analysis System (MIMAS), a non-invasive tool that analyzes spontaneous movements captured via RGB video recordings. By extracting 39 motion parameters, MIMAS demonstrated its ability to distinguish between typical and atypical development, particularly in preterm infants, making it a scalable and cost-effective tool for clinical use.

Optical flow techniques have also been extensively utilized. Stahl et al. [88] combined optical flow with wavelet analysis to track motion features and classify CP risk with high accuracy using Support Vector Machines (SVMs). Ihlen et al. [89] advanced this approach with the Computer-based Infant Movement Assessment (CIMA) model, which tracks the movements of six body parts and analyzes their trajectories, frequency, and amplitude using ML algorithms, achieving reliable predictions of CP risk. Rahmati et al. [90] refined these methods further by incorporating graph-based segmentation and semi-supervised particle matching. This refinement allowed for robust tracking of dense motion trajectories, reducing the need for manual intervention even in complex scenarios.

These classical techniques have also been adapted to ASC detection. For instance, Caruso et al. [91] developed MOVIDEA, a semi-automatic software that extracts kinematic features from both 2D and 3D video recordings, enabling differentiation between High-Risk (HR) and typically developing infants. Baccinelli et al. [92] contributed with a graphical interface designed to track hand and foot movements, usable in both clinical and home-video settings. Additionally, Das et al. [93] focused on infant kicking movements, employing KAZE point detection and SVM classifiers to identify neuromotor risks such as CP and Infantile Spasms (IS).

Collectively, these methods highlight the adaptability and potential of classical computer vision techniques as accessible, non-invasive tools for the early detection of NDD. They provide a foundation for integrating automated motion analysis into clinical and research contexts.

Deep Learning Techniques DL techniques provide advanced and highly automated methods for analyzing infants' movements from video recordings. These approaches are increasingly employed for tasks such as pose estimation and movement classification, offering enhanced accuracy and scalability in detecting and interpreting motor patterns. By leveraging the power of neural networks, DL enables the identification of complex and subtle movement dynamics, pushing the boundaries of traditional motion analysis techniques [82].

An example is the work by Shin et al. [94], where the AlphaPose [95] model was applied to track spontaneous movements in HR preterm infants, focusing on kinematic parameters like joint angles and angular velocities. By quantifying movement complexity with sample entropy and assessing inter-limb synchronization, they identified that lower complexity patterns were associated with a higher risk of developmental delays, highlighting the utility of automated kinematic analysis for neurological assessment.

Building on this concept, Reich et al. [96] used OpenPose [97] to detect FMs in typically developing infants. Their approach relied on full-body pose extraction to generate a 25-point skeleton, which served as input for a neural network trained on 2,800 annotated video frames. This method achieved an 88% classification accuracy, demonstrating that non-invasive motion tracking can facilitate naturalistic movement analysis without the need for markers or devices.

OpenPose was further explored by Chambers et al. [98], who retrained the algorithm on labeled infants' videos to enhance pose estimation for neuromotor risk assessment. By extracting kinematic features such as velocity and postural symmetry and applying a Bayesian Surprise metric, they identified significant deviations in movement patterns among infants at higher neuromotor risk, emphasizing the accessibility and cost-effectiveness of such tools.

To achieve greater accuracy in pose estimation, Groos et al. [99] utilized Convolutional Neural Networks (CNNs) such as EfficientHourglass Model (EfficientHourglass) [100] and EfficientPose Framework (EfficientPose) [101]. Their study trained models on a large dataset of clinical and home videos, reaching precision levels comparable to those of human experts, particularly with the EfficientHourglass B4 model. This work highlights the scalability and potential of ConvNet-based methods for early motor development screening.

Another marker-less approach was implemented by Moro et al. [102], who used DeepLabCut Framework (DeepLabCut) [103] to analyze preterm infants' spontaneous movements. Their pipeline extracted quantitative kinematic features and utilized classifiers such as SVMs and Long Short-Term Memory Networks (LSTMs) to distinguish between normal and abnormal patterns, achieving an accuracy of 85.7%. This highlights its applicability for early neuromotor disorder detection in clinical settings.

Focusing on General Movements (GMs), Passmore et al. [104] developed a DeepLabCut-based framework to automate GMA. Using 503 smartphone videos of infants aged 12 to 18 weeks, their system achieved an Area Under the Curve (AUC) of 0.80 and 76% sensitivity for detecting abnormal movements. This demonstrates the practicality of smartphone-based, non-invasive assessments for early CP screening.

The analysis of GMs during the writhing stage was further refined by Letzkus et al. [105], who developed a Neonatal Intensive Care Unit (NICU)-trained pose estimation model to detect anatomical key points with high accuracy. Their movement analysis identified significant differences in lower limb dynamics between normal and cramped-synchronized GMs, providing early insights into CP risk.

Similarly, Doi et al. [106] applied ML to explore the relationship between 4-month-old infants' spontaneous movements and ASC-like behaviors at 18 months. Their analysis revealed reduced movement strength and atypical body center dynamics in HR infants, supporting the potential of early video-based ASC risk detection. Extending their work, Doi et al. [107] also investigated neonatal movements during sleep and awake states, finding that motor patterns during sleep provided better predictive markers for ASC risk, emphasizing the role of subtle early indicators.

Shifting focus to infant safety, Singh et al. [108] introduced a smart monitoring system combining OpenPose [97] and Dlib [109] for real-time infant posture and sleep analysis. By detecting unsafe conditions such as face-down sleeping, the system offered caregivers immediate alerts, reducing the risk of sudden Sudden Infant Death Syndrome (SIDS) and enhancing infant safety.

Tsuji et al. [110] developed a marker-less system incorporating an artificial Artificial Neural Network (ANN) to analyze 25 motion indices from video recordings.

Their system achieved 90.2% classification accuracy, aligning closely with expert evaluations and paving the way for early non-invasive diagnoses of disorders like CP.

The potential of DeepLabCut was also explored by Abbasi et al. [111], who combined it with deep neural networks to analyze spontaneous movements in infants. By fine-tuning models such as Residual Network Model with 152 layers (ResNet-152) [112] and EfficientNet Model, Version B6 (EfficientNet-B6) [113] and integrating a Kalman filter, their system achieved over 94% accuracy in pose estimation, enhancing the precision of automated assessments.

Pose-based features were central to the work of McCay et al. [114], who used OpenPose to classify abnormal movements with up to 91.67% accuracy, emphasizing the robustness of ML for CP detection. Their earlier study in 2019 laid the foundation for this approach by introducing innovative pose-based features for movement classification, achieving remarkable accuracy on both synthetic and real-world datasets.

Advancing these efforts, Sakkos et al. [115] developed a system combining CNNs and LSTMs to model limb movements. By integrating visualization modules to highlight influential body parts, they improved interpretability for healthcare applications while addressing class imbalance with data augmentation.

Further exploring early movement detection, Doroniewicz et al. [116] employed OpenPose to analyze writhing movements in newborns, extracting features that described movement scope, shape, and location. Their ML classifiers demonstrated high sensitivity and specificity, reinforcing the value of pose-based analysis for early neurological diagnoses.

Lastly, Moccia et al. [117] utilized depth videos from NICUs to propose a 3D convolutional neural network framework. Their approach achieved precise pose localization, offering a non-invasive method for monitoring preterm infants' development and identifying early signs of disorders like CP.

1.4.2 Gesture Recognition in Neurodevelopmental Assessment

The assessment of gesture production is essential for understanding social communication skills, particularly in neurodevelopmental disorders such as ASC. Gestures, including pointing, waving, and reaching, serve as early markers of social and communicative development. However, despite their well-established diagnostic relevance, traditional methods often struggle to capture the full complexity and subtlety of developmental trajectories. This limitation is particularly evident in longitudinal studies, where early variations in gesture use may provide critical insights into emerging social or communication delays [45, 58]. In this context, artificial intelligence (AI) has emerged as a powerful tool, enabling automated, objective, and real-time assessments that enhance the precision and scalability of neurodevelopmental evaluations.

1.4.2.1 AI for Gesture Detection and Classification

Deep Learning Approaches DL models, particularly CNNs and Recurrent Neural Networks (RNNs), have demonstrated exceptional capabilities in detecting and classifying complex gestural patterns. Alkahtani et al. [118] developed a CNN-LSTM model to analyze video recordings of children exhibiting repetitive behaviors, such as hand flapping, an established early marker of ASC. Their model, trained on the Self-Stimulatory Behavior Dataset (SSBD) [58], achieved an impressive 96% accuracy, underscoring the potential of AI in automating the recognition of diagnostic behaviors [119].

Similarly, Singh et al. [120] utilized a Video Masked Autoencoder (VideoMAE)-based model [121] to detect stereotypical behaviors such as arm flapping and head banging in children with ASC. Excelling in feature extraction from video data, their system achieved a 97.7% success rate. These advancements highlight the robustness of DL methods in processing high-dimensional video data, making them particularly suitable for large-scale screening initiatives and remote assessments.

Transformer architectures [122], with their capacity to model long-range dependencies in sequential data, have further revolutionized gesture recognition. Floris et al. [123] iintroduced a transformer network capable of accurately identifying

hand gestures with a 91.67% accuracy rate, demonstrating enhanced adaptability to diverse gesture types beyond strictly clinical contexts. Extending this approach, Song et al. [124] developed a multimodal screening method combining pose tracking, head pose estimation, and speech recognition, achieving 93.3% accuracy in detecting responses to names. These multimodal systems leverage the complementary strengths of various data streams, providing a comprehensive profile of a child's communicative and motor behaviors.

In addition, McDonald et al. [125] explored head movement patterns during dyadic interactions using computer vision. Their approach proved more accurate in predicting autism status than monadic analysis, highlighting the importance of social cues in understanding social communication difficulties in autism.

While manual coding remains a gold standard in behavioral research, it is inherently time-consuming and prone to subjectivity. AI integration can streamline this process without compromising accuracy. Samanta et al. [126] demonstrated this by combining a Deep Convolutional Neural Network (DCNN) with a SVM to classify 12 distinct gestures during infant-caregiver interactions. The strong correlations between machine-coded and human-coded gestures underscore the potential of AI to enhance traditional coding methods, particularly in large-scale studies of language acquisition and cognitive development.

Applications in Clinical and Naturalistic Settings The deployment of AI-powered systems in real-time gesture recognition has transformative implications for clinical diagnostics and early intervention. Gopi et al. [127] illustrated how AI models can track subtle motor patterns and detect gestural anomalies indicative of developmental delays. Unlike traditional assessments conducted in clinical settings, these systems can be seamlessly integrated into everyday environments, enabling unobtrusive, continuous monitoring that enhances ecological validity.

Hashemi et al. [128] validated mobile applications capable of analyzing engagement, head movements, and emotional responses in both home and clinical contexts. Such tools offer scalable solutions for ASC screening, reducing the reliance on specialized clinical visits and thereby increasing accessibility for families in remote or resource-limited settings.

Beyond diagnostic applications, gesture recognition technologies have been in-

corporated into therapeutic contexts through socially assistive robotics. Ivani et al. [129] demonstrated how integrating AI with robotic platforms can enhance therapy for children with ASC, alleviating the cognitive load on therapists while improving gesture recognition accuracy. These systems not only facilitate individualized intervention plans but also promote engagement and motivation in children, making therapy sessions more interactive and effective.

The combination of AI with wearable sensor platforms further expands the possibilities for continuous, real-time monitoring. Siddiqui et al. [130] developed a system using accelerometers and gyroscopes to capture movement data in children with ASC, achieving approximately 91% accuracy through ML models. Such devices allow for prolonged data collection in naturalistic settings, providing clinicians with rich datasets to inform early intervention strategies and track developmental progress over time.

The integration of advanced AI methodologies, ranging from deep learning architectures to transformer-based and multimodal models, has significantly enhanced the accuracy, scalability, and ecological validity of gesture recognition systems. These technological advancements enable comprehensive assessments of cognitive, motor, and social behaviors, with profound implications for early detection and intervention in neurodevelopmental disorders like ASC.

However, despite these promising developments, several challenges remain. Ensuring the generalizability of AI models across diverse populations requires access to large, representative datasets. Privacy concerns associated with continuous monitoring, especially in home environments, must be addressed through robust data protection protocols. Future research should focus on refining multimodal integration, improving model transparency, and exploring ethical considerations related to the deployment of AI-driven assessments in vulnerable populations.

By addressing these challenges, the field can move toward more precise, scalable, and effective interventions, ultimately enhancing the quality of care and developmental outcomes for children with neurodevelopmental disorders.

1.4.3 Gaze and Visual Attention Pattern Analysis

AI-driven technologies have revolutionized gaze and visual attention assessments by enabling automated quantification of gaze metrics, such as fixation duration and gaze shifts, even in complex and dynamic environments. These tools provide refined analyses of disorder-specific attentional patterns, including reduced face-directed gaze in ASC and heightened distractibility in ADHD, which serve as early behavioral markers [53, 50]. Unlike traditional methods, which often require controlled laboratory settings, AI-powered systems facilitate reliable data processing in naturalistic environments, thereby enhancing ecological validity and clinical applicability [131, 80].

A significant advancement lies in the integration of visual attention data with other behavioral indicators, such as gestures and vocalizations, to create comprehensive multimodal profiles. This approach captures the intricate interplay between visual attention, joint attention, and social cues, supporting more accurate diagnostics and personalized intervention planning [132].

1.4.3.1 AI Techniques for Gaze Analysis in Neurodevelopmental Disorders

Advanced ML and DL algorithms, including CNNs and LSTMs, offer detailed insights into the temporal and spatial dynamics of gaze behaviors. Li et al. [133] utilized an LSTM-based model to analyze raw video data, capturing sequential gaze patterns during social interactions and achieving high classification accuracy for ASC. Similarly, Vabalas et al. [134] integrated kinematic and gaze metrics, underscoring the advantages of multimodal approaches in profiling fixation duration, saccadic movements, and gaze shifts.

Standardizing gaze analysis methods has become increasingly feasible through AI-based quantitative approaches. For instance, Jeyarani et al. [135] stressed the importance of consistent protocols to identify autism markers reliably, while Akter et al. [136] demonstrated how gaze metrics could inform personalized treatment plans. These studies collectively highlight how AI methodologies address variability issues inherent to traditional assessments.

Clinical and Diagnostic Applications of Gaze Analysis Atypical gaze behavior is a well-established marker of neurodevelopmental disorders, particularly ASC. Clinical applications of AI-driven gaze analysis have yielded high diagnostic accuracy rates. Zhang et al. [137] employed CNNs to analyze gaze during interactive tasks, achieving a 92% classification accuracy. Complementarily, Ahmed et al. [138] leveraged LSTMs and GRUss architectures to model temporal gaze dynamics, reporting an impressive 98.33% accuracy.

Beyond diagnostics, gaze metrics provide insights into broader cognitive and social processes. Sasson et al. [139] found that children with ASC demonstrate reduced fixation on faces and heightened attention to non-social stimuli, patterns linked to social communication challenges. Kang et al. [140] further elaborated on how gaze deviations impact emotion recognition and social cue processing.

Emerging technologies extend gaze analysis beyond clinical settings. Remote monitoring solutions, like those proposed by Bidwe et al. [141], enable real-time, home-based assessments, bridging gaps in traditional diagnostic practices. Stuart et al. [142] demonstrated the utility of gaze tracking in monitoring therapeutic progress, while Cho et al. [143] and Moradizeyveh et al. [144] emphasized the scalability and accessibility of AI-driven tools.

Integrating AI into gaze and visual attention assessments enhances the precision, scalability, and accessibility of neurodevelopmental evaluations. By improving diagnostic accuracy and enabling continuous monitoring, these advancements support personalized interventions and better outcomes for individuals with NDD.

1.4.4 Language Assessment

AI has transformed language assessment through advanced tools such as NLP and ML, enabling detailed analyses of syntactic complexity, prosody, and lexical diversity. These technologies facilitate the identification of language delays often associated with NDD [145]. Traditional assessments, such as the CELF, frequently fail to capture pragmatic language use in naturalistic contexts. AI-based methods address this limitation by providing objective and continuous data on conversational patterns, including turn-taking, topic maintenance, and responsiveness, offering a more comprehensive understanding of communicative abilities [75]. This approach

enables interventions to be tailored to the specific linguistic and communicative needs of each child.

1.4.4.1 AI Techniques for Language Analysis in Neurodevelopmental Disorders

This section explores the technical methods through which AI identifies linguistic patterns in children with NDD. Children with ASC, SLI, and ADHD present distinct linguistic profiles that impair effective communication. Leveraging ML, advanced voice recognition models, and behavioral analyses, AI allows for the detection of subtle linguistic features often overlooked by traditional methods, thus enhancing diagnostic accuracy and facilitating early identification of language difficulties.

Numerous studies have demonstrated how various AI techniques support the analysis of specific linguistic domains. Villasanti et al. [146] and Sharma et al. [147] applied ML models to identify phonetic, syntactic, and semantic features in children with SLI, revealing linguistic patterns typically missed by standardized evaluations. Gale et al. [148] utilized NLP on natural speech data to differentiate between ASC and SLI, highlighting linguistic markers that enhance diagnostic processes. Acoustic analyses have also advanced significantly; Pahwa et al. [149] demonstrated that AI-based models outperform manual methods in distinguishing NT speech from autistic speech, while Radha et al. [150] employed CNNs, achieving an accuracy rate of 86.6%.

These findings highlight how AI, through targeted techniques and multimodal analyses, enables comprehensive linguistic assessments, paving the way for more precise interventions.

Clinical Applications of AI in Language Processing While the previous section focused on technical aspects, this section examines the clinical applications of AI technologies in diagnosis and intervention. The integration of linguistic and behavioral data enables tools capable of monitoring the evolution of communicative skills and personalizing treatment strategies.

Wang et al. [151] proposed a multimodal approach that integrates auditory and behavioral data, such as responses to one's name, to detect developmental delays.

Similarly, Bhardwaj et al. [145] developed individualized phonetic profiles to guide targeted interventions. Narala et al. [152] combined CNNs with architectures like EfficientNet for simultaneous analysis of vocal signals and facial expressions, enhancing ASC diagnostics by capturing multimodal markers.

Assistive and Augmentative Communication (Augmentative and Alternative Communication (AAC)) technologies have particularly benefited from AI-driven advancements. Costanzo et al. [153] developed Talkitt, a system that translates unintelligible vocalizations into words, promoting social integration for children with severe language impairments. Similarly, mobile applications such as Fluency SIS [154] provide continuous, accessible therapy that can be seamlessly integrated into daily routines, supporting therapeutic goals beyond clinical environments.

These applications demonstrate how AI enhances not only diagnostic precision but also long-term therapeutic strategies tailored to the evolving communicative needs of children with NDD.

Innovative Applications and Emerging Trends Innovation in AI-based language assessment is continuously evolving, with emerging technologies expanding diagnostic and therapeutic possibilities beyond current methods. Unlike the previous sections, this part focuses on future directions and applications not yet fully integrated into standard clinical practice.

Multilingual voice recognition models, such as those developed by Ashwini et al. [155], have achieved an accuracy rate of 91.3%, demonstrating the adaptability of AI across diverse linguistic contexts. Trayvick et al. [156] introduced virtual reality environments that allow children to practice language skills in controlled, immersive scenarios, offering real-time adaptive feedback. Educational platforms like SmartSpeech [157] gamify the diagnostic process, increasing engagement while providing clinicians with real-time progress data.

Future research aims to integrate linguistic, cognitive, and behavioral data to develop comprehensive AI models. Almutairi et al. [158] proposed wearable sensors coupled with AI algorithms to monitor vocalizations and social interactions in real time, enabling more responsive therapy adjustments. Donolato et al. [159], Sindhu and Sainin [160], and Hu et al. [161] emphasize the importance of multimodal frameworks to improve both diagnostic precision and therapeutic efficacy.

These trends hold promise for overcoming current limitations, providing increasingly accurate, accessible, and personalized tools that foster a holistic approach to the assessment and treatment of language impairments in NDD.

Chapter 2

Research Motivation and Objectives

The primary objective of this research is to advance the understanding of NDD by investigating multiple developmental domains throughout childhood using an AI-driven, multimodal approach grounded in the theoretical principles of the neurodevelopmental cascade. As discussed in the previous chapter, this model assumes that early deficits, such as motor impairments, can trigger cascading effects, impacting higher-order cognitive, communicative, and social functions. Adopting an evolutionary perspective, the study explores how motor, communicative, and emotional domains dynamically interact across distinct childhood stages. The ultimate goal is to refine the characterization of neurodevelopmental trajectories and identify novel biomarkers to improve both detection and intervention strategies.

Leveraging advanced AI techniques, this research focuses on automating and enhancing the identification of early sensory, cognitive, and behavioral indicators, including communicative gestures, speech capabilities, social engagement patterns, visual attention processes, and emotional expressions. This integrative, data-driven approach enables a comprehensive assessment of neurodevelopmental pathways, facilitating more precise diagnoses and personalized interventions. By optimizing diagnostic precision and tailoring therapeutic strategies, the study aims to contribute to more effective support for children with NDD.

Methodology

1. **Motor Development in Early Infancy: From Spontaneous Movements to Reaching Behaviors**
   The research initially focused on analyzing spontaneous movements in newborns to explore early motor patterns and their correlation with the later onset of NDD. This phase aimed to identify potential risks of developmental divergence by assessing general movements during the first weeks of life. As infants reach six months of age, a period typically characterized by the emergence of object reaching and the development of postural control, the scope expanded to include the analysis of more complex hand movements, particularly during social interactions and object presentation tasks. This progression, from basic spontaneous movements to goal-directed reaching behaviors, follows the developmental trajectory of motor skills and their implications for neurodevelopmental risk, thereby deepening the understanding of how early motor patterns may relate to the later onset of NDD.

2. **Emerging Communicative Skills in Toddlers: Gestures, Gaze, and Vocalizations in Naturalistic Interactions**
   As children transitioned into the toddler stage, the study shifted its focus to the production of gestures during naturalistic parent-child interactions, capturing a wide range of communicative behaviors in ecologically valid contexts. Particular attention was given to deictic gestures, which are known to be crucial precursors for the development of early social and communicative skills. The study compared these behaviors between Neurotypical (NT) children and those at risk for NDD and introduced an AI-driven tool for the automatic recognition of gestures from naturalistic video data. This tool provides an efficient and scalable method for assessing gesture production, advancing the study of early communication in real-life contexts and enhancing the understanding of developmental trajectories across various neurodevelopmental profiles.

3. **Advancing Visual Attention to Social Cues in Preschoolers: Exploring Gaze Patterns Development**

The investigation of visual attention in preschool-aged children employed an innovative eye-tracking paradigm enhanced by AI technologies. This approach enabled a precise analysis of gaze patterns and attentional allocation in social settings, revealing deficits commonly associated with NDD. The study specifically examined how children distribute attention to social cues, providing insights into the role of visual attention divergence in social cognition and communication difficulties in children with NDD.

4. **Neurodivergent Trajectories in Middle Childhood: Effects on Expressive Kinematics and Speech**
   Further investigations were conducted with school-aged children to explore more complex manifestations of neurodivergence. Specifically, we analyzed the kinematics of "Vitality Forms" (VFs), which describe how children convey emotions and affect through actions and body movements during social interactions. By comparing the dynamic expressions of children with NDD to those of their typically developing peers, the study provided valuable insights into differences in emotional communication and social cue processing in NDD. Additionally, we examined language development using AI-driven tools to detect and assess dysarthria, facilitating the early identification of speech-related impairments.

Innovation, Impact, and Future Directions This research represents an innovative integration of technological advancements with the neurodevelopmental cascade model. By employing a multimodal AI-based approach across different developmental stages, the study explores how early disruptions in motor or communicative skills may cascade into broader cognitive and social impairments. Identifying early biomarkers not only distinguishes NDD from neurotypical development but also enables the stratification of disorders based on specific functional domains. This process supports targeted interventions tailored to individual neurodevelopmental profiles, ultimately improving long-term outcomes.

Integrating data across motor, communicative, and emotional domains provides a more comprehensive perspective on neurodevelopmental trajectories, shifting from isolated assessments to holistic evaluations. This approach enhances clinical

practice by facilitating more personalized and timely interventions.

Future research will focus on embedding AI-based tools into clinical workflows, ensuring accessibility and effectiveness. Long-term studies will be essential for tracking developmental trajectories, allowing clinicians to monitor progress and refine interventions accordingly. Further exploration of specific biomarkers will enhance diagnostic precision and early detection. Beyond clinical applications, this paradigm has implications for personalized intervention strategies, educational support, and policy decisions related to early childhood development. By prioritizing practical applications, the overarching goal is to improve the quality of life for children with NDD and provide tailored resources for families.

Chapter 3

Motor Development in Early Infancy: From Spontaneous Movements to Reaching Behaviors

Early motor development is a fundamental building block in children's growth, exerting profound influence on the acquisition of cognitive, communicative, and social competencies. As highlighted in the preceding chapters, motor skills enable infants to actively engage with their environment, facilitating multimodal learning experiences that shape subsequent developmental trajectories. Disruptions in early motor behaviors can initiate cascading effects, amplifying delays across domains such as language, social interaction, and executive functions. The present chapter employs AI to investigate motor development in infants at HR for NDD from the earliest days of life. The overarching aim is to leverage the predictive potential of AI-based models to detect subtle motor patterns that may signal divergent neurodevelopmental pathways. To this end, two studies are presented: the first focuses on the marker-less analysis of spontaneous movements in newborns, emphasizing the characterization of GMs and kinematic features; the second examines the emergence of goal-directed hand movements, with particular attention to reaching behaviors during social engagement and object-presentation tasks.

3.1 Marker-less Analysis of Spontaneous Movements in Newborns

Spontaneous movements during the early postnatal period are fundamental indicators of the integrity and maturation of the developing nervous system. Deviations in the quality or complexity of these movements can provide early signs of potential neurodevelopmental challenges, highlighting the importance of timely identification to facilitate targeted interventions. This section employs a non-intrusive, marker-less AI approach to examine spontaneous motor patterns in newborns at HR for NDD. By focusing on a detailed analysis of GMs and kinematic features, the study aims to capture subtle alterations that traditional observational methods may overlook. Through automated video processing, the proposed method offers an objective, scalable solution for detecting early deviations in motor behavior, which may be indicative of adverse outcomes.

3.1.1 Participants

For this analysis, 74 HR infants were recruited from the Italian Network for Early Detection of ASC, also known as the NIDA network [162]. The NIDA network represents the largest Italian cohort of infants at risk for NDD and includes siblings of children diagnosed with ASC, preterm newborns, and small-for-gestational-age newborns. For each participant, we recorded and analyzed five videos capturing the infant in a state of natural and unrestricted movements at five specific time points corresponding to 10 days, 6, 12, 18, and 24 weeks of age. These time points were chosen to capture crucial moments in motor development from birth onward.

Furthermore, a comprehensive clinical/diagnostic assessment of the infants and toddlers using standardized tools/tests and parental structured interviews were conducted to confirm the presence/absence of NDD. Within this cohort, 5 infants dropped out of the study, 28 received a diagnosis of NDD, while 33 have been assessed as NT. Additionally, 8 infants are still pending evaluation since they have not yet reached the appropriate age for a stable diagnosis. Videos of no-label or drop-out infants were used exclusively for validating our DL-based automatic tracking approach and not for the subsequent analysis steps aimed at identifying

predictive variables for NDD.

The dataset (summarized in Table 3.1) lacks complete sets of five videos for each participant, as not all individuals could be recorded at every designated time point. Furthermore, in addition to the videos reported in the dataset and recorded as described in the following paragraph, we used four additional amateur-recorded videos to test our tracking algorithm under completely unstructured conditions. Further details about the dataset can be found in the supplementary Table A.1.

Time Point	**NDD**	**NT**	**No Label**	**Drop-out**
10 days	11	15	2	4
6 weeks	18	22	6	4
12 weeks	22	25	6	2
18 weeks	14	26	5	2
24 weeks	18	16	6	1

Table 3.1: Overall dataset with the number of videos for each timepoint. Subsequently, for analyses aimed at identifying early motor features predictive of clinical outcomes, only data from children who have received a diagnosis (NDD and NT) were included.

The study was approved by the ethics committee of the ISS (Approval Number: Pre 469/2016), all the families that voluntarily participated in the study signed a written informed consent and all methods were performed in accordance with the relevant guidelines and regulations. Informed consent for publication of identifying information/images was also obtained from the parents of the subjects whose images appear in this section.

3.1.2 Methods

3.1.2.1 Experimental Setup

Infants were recorded in a home setting, lying on a green blanket supplied by the NIDA network. The camera was positioned at a distance of 50 cm above the infant's chest. Each recording lasted for a minimum of 5 minutes, aiming to capture spontaneous movements of the infant's entire body.

3.1.2.2 Video Editing

Recorded videos were reviewed revealing that in most cases, segments of high quality (i.e., without external disturbances such as operator intervention to soothe the infant) did not exceed 3 minutes in duration. Consequently, the decision to retain a 3-minute video segment representing the highest quality portion for each recording was made. To ensure consistency in the dataset, videos were manually edited to meet the following specific criteria: a duration of three minutes and the infant in a supine position, displaying a state of well-being and spontaneous motor activity, and without any episodes of crying. In the cases where the videos contained more than 3 minutes of high-quality footage, we opted to analyze the initial 3 minutes of high-quality content. Any video frames that exhibited operator or parental interferences, as well as accidental camera movements, were excluded from the subsequent analysis. All the analysis presented in the following paragraphs were performed using Python [163] and Matlab [164].

3.1.2.3 Tracking Procedure

The trajectories of infants' body landmarks during their spontaneous movements were automatically extracted from all the collected videos using a customized application of the Mediapipe Pose Landmark Detection solution [165]. MediaPipe is an open-source framework developed by Google for creating AI and ML pipelines, offering a suite of libraries and tools for multimedia processing across multiple platforms. The MediaPipe Pose Landmarker, a key component of this framework, enables real-time human pose estimation by detecting and tracking 33 body landmarks, including significant points such as joints and limb endpoints like hands and feet, in images and videos.

The architecture is based on BlazePose [166], a specifically modified version of the MobileNet [167] CNN, which processes input data to extract features and generate heatmaps representing the likelihood of each landmark's presence. A regression model then refines the coordinates of the detected landmarks to improve accuracy. The system outputs these landmarks as both image coordinates (pixel values) and 3-dimensional world coordinates, facilitating applications such as posture analysis, movement categorization, and body point identification. Me-

diaPipe's lightweight and efficient architecture is optimized for use across various platforms, including mobile devices and web applications.

The comprehensive list of landmarks is provided in Figure 3.1. For each reference point, we stored a $N \times 2$ matrix containing the x and y coordinates in the image for each of the N frames in the respective video.

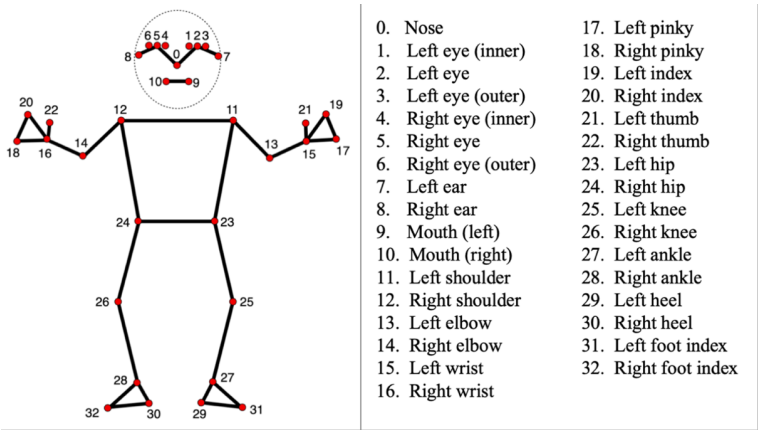


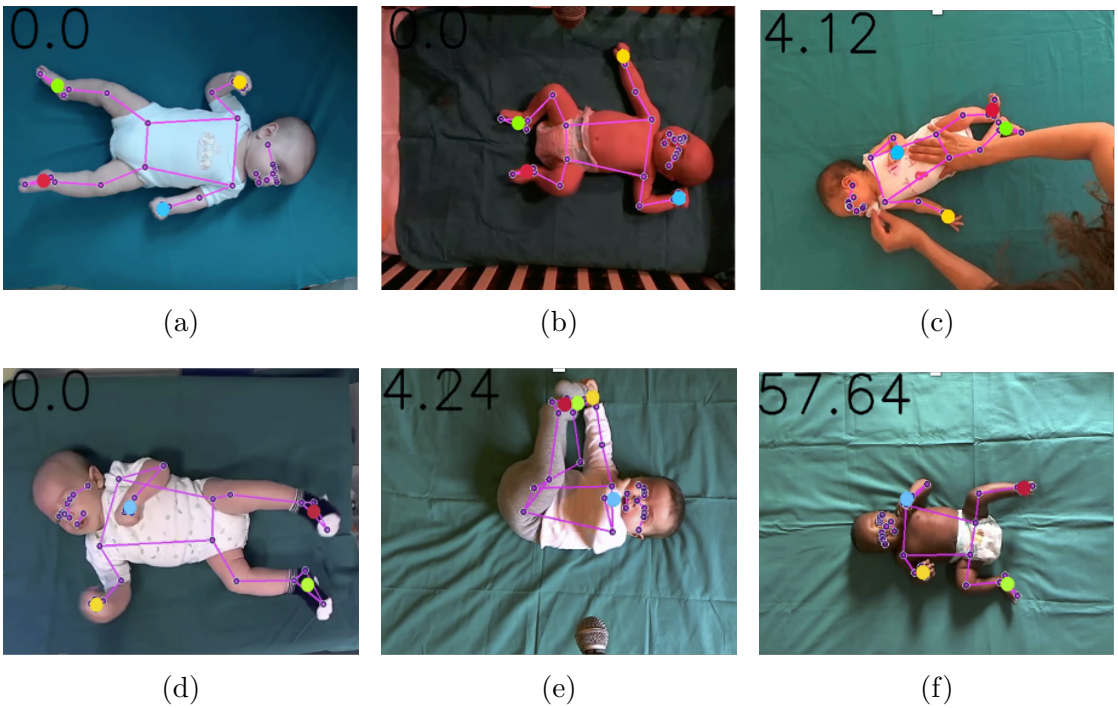
Figure 3.1: List of the 33 body landmarks locations tracked using Mediapipe Pose solution [165].

We selected the MediaPipe Pose Landmark Detection as the body pose estimation solution for our study due to several advantages. Notably, it has undergone an extensive training on large datasets and fine-tuning to ensure precise landmarks detection. Additionally, it offers real-time pose estimation with minimal latency. Furthermore, the CNN architecture is easily customizable to meet specific project requirements.

In our work, we fine-tuned the hyperparameters of the CNN during the training process to optimize the performance of pose landmark detection for our specific task.

This AI approach in our study has demonstrated remarkable robustness also in conditions such as low resolution (Figure 3.2a), inadequate lighting (Figure 3.2b), inadvertent operator intrusion into the frame (Figure 3.2c), infants wearing socks or clothing that covered their limbs (Figures 3.2d and 3.2e), or variations in skin tones (Figure 3.2f). In all these scenarios, motion tracking was performed accurately, as further confirmed by the model validation described in the next para-

graphs.



(a) (b) (c)

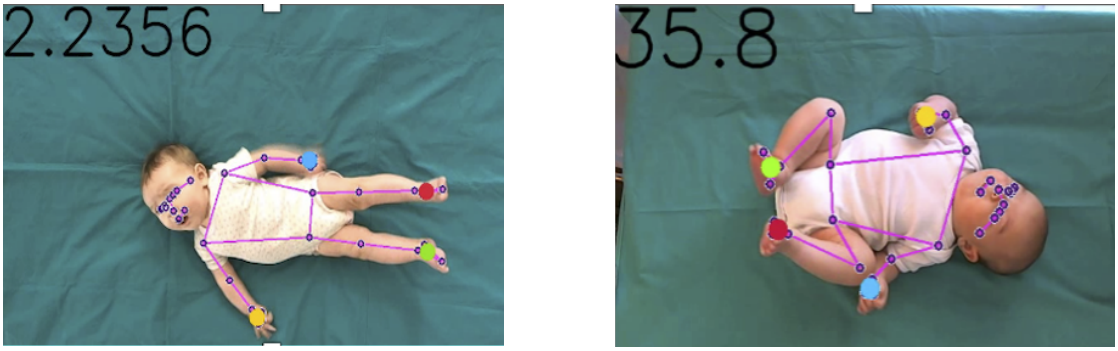
(d) (e) (f)

Figure 3.2: Example of videos with low resolution (a), poor lighting (b), hand operator intrusion in the frame (c), and children wearing socks and clothes that covered their limbs (d and e) or had different skin tones (f).

3.1.2.4 Signal Preprocessing

Using the outlined methodology, we identified the x and y coordinates of 33 body landmarks for each frame of each video in our dataset. The AI model ensures consistency across diverse frame sizes by normalizing coordinate values within the range of 0 to 1. The resultant output for each video undergoing processing was a file that contained, for each landmark, the corresponding x and y coordinates across every frame in which the model successfully performed body pose estimation. For subsequent analyses aimed at identifying early motor predictors of NDD, we specifically focused on infants' limbs. The centroid coordinates of each end-effector were computed for each frame mediating the x and y coordinates of wrist, pinky, and index for the hands and of ankle, heel, and foot index for the feet (Figure 3.3). Each signal was then preprocessed by applying a zero-phase moving

average filter with a selected window size of 5 samples in order to smooth noisy fluctuations.



(a) (b)

Figure 3.3: Example of two frames with the detected reference points and the corresponding skeleton overlaid. The hands and feet centroids are indicated by larger dots. Figure (a) shows the overlaid skeleton when the limbs are fully extended, whereas Figure (b) illustrates the case where the limbs are bent.

3.1.2.5 Model Validation: Comparison with Movidea

To evaluate the performance of our automatic tracking approach, we conducted a comparative analysis between the trajectories derived from our procedure, described in the previous paragraphs, and those obtained using Movidea [92], a validated software package purposefully designed for the semi-automatic extraction of kinematic features in assessing infants' motor skills. Movidea is a software which facilitates the tracking of infants' end-effectors during free movement, from single-camera video footage. However, it requires the presence of an operator for the selection of the infant's reference measures, specifically the head length and the symmetry line of the body. The operator is also responsible for selecting the central point of each limb's end-effector to initialize the tracking algorithm, and they must oversee the entire process, intervening to reset the points in the event of errors and to skip frames in cases where a limb is not visible.

During the validation procedure, in addition to the videos listed in Table 3.1 acquired through the method described in the paragraphs *Experimental Setup* and *Video Editing*, we also tested four additional amateur videos. These videos were recorded under conditions different from the structured procedure established for

the study. To assess the similarity between our tracking results and those from the operator-dependent software Movidea, we computed the Pearson's Correlation Coefficient (R), the p-value testing the hypothesis of no relationship between the two signals, and the Root Mean Square Error (RMSE). For each signal, similarity measures were computed within 10-second windows and then averaged across all these windows. Subsequently, we checked the normality of the data distribution for each measure, using the Shapiro-Wilk statistical test. Upon identifying non-normal distributions ($p < 0.001$), we expressed the aggregated results across all the signals using the median value (Q1-Q3).

3.1.2.6 Feature Extraction

To characterize the spontaneous movements of infants in terms of synchronicity, smoothness, repetitive movements, and symmetry, we used the trajectories of each end-effector's centroid extracted from videos of NDD and NT children at each time point to compute the following features and kinematic parameters:

Velocity and Acceleration: The velocity of each limb was calculated as the Euclidean distance between the location of the corresponding reference point in every two subsequent frames. The obtained signal was then multiplied by the frame rate to normalize any difference due to videos captured using different devices. To reduce the fast oscillations in the velocity profiles, a third-order low-pass Butterworth filter with a cutoff frequency equal to 95% of the Nyquist frequency was then applied. Analogously, the acceleration of each limb was calculated as the difference between two consecutive velocity samples. Descriptive statistics, including mean, median, standard deviation, variance, mode, skewness, kurtosis, maximum, and minimum were then computed for both the velocity and acceleration of each limb. Furthermore, the features obtained for the right and left hands, as well as for the right and left feet, were averaged to aggregate the kinematics of the upper body and lower body, respectively.

Cross-Correlation (CC): The calculation of the zero-lag cross-correlation between the velocity of each pair of limbs was performed applying the following equation:

$$CC_{v_1 v_2} = \frac{\sigma_{v_1 v_2}}{\sqrt{\sigma_{v_1}^2 \cdot \sigma_{v_2}^2}} \tag{3.1}$$

where $CC_{v_1 v_2}$ is the cross-correlation between the velocity v_1 and the velocity v_2, $\sigma_{v_1 v_2}$ is the covariance of v_1 and v_2, $\sigma_{v_1}^2$ is the variance of v_1, and $\sigma_{v_2}^2$ is the variance of v_2. CC serves as a metric to evaluate the synchronization of limbs movements, a fundamental information for describing the distinct motor patterns exhibited by infants. Consequently, it holds significant importance in the early assessment of NDD [168].

Area differing from Moving Average (Ama): The smoothness is another crucial parameter for the assessment of infants' proper development, as typical GMs are characterized by continuous, flowing patterns without jerky or abrupt transitions. To quantify this aspect, we computed the moving average for both the x and y components of each limb's trajectory across the entire recording, using a 2-second window size [168], as described by the following equation:

$$\bar{x}_i = \frac{1}{k} \sum_{j=i-k/2}^{i+k/2} x_j \tag{3.2}$$

where $\bar{x}_i$ is the moving average computed at the i-th frame, k is the window's size, and x_j is the point position in the j-th frame.

Successively, for every sample in the trajectory, we calculated the deviation from the moving average by subtracting the trajectory value from the corresponding moving average value, as shown in the following equation:

$$A_{max} = \sum_{i=k/2}^{l-k/2} |x_i - \bar{x}_i| \tag{3.3}$$

where A_{max} is the area differing from the moving average of the x component and l is the total number of frames of the recording. Finally, the total Ama was computed for both the lower and upper limbs. This was accomplished by summing the areas that deviated from the moving average for the two components of the feet and the two components of the hands respectively.

Periodicity (P): Periodicity, as defined by Meinecke, L. et al. [168], is a parameter designed to assess the presence of repetitive movements in limbs motion. To measure the periodicity of infants' spontaneous movements, we computed a high-order moving average for both the x and y components of each limb's trajectory across the entire recording, we segmented the recordings using a 1000-samples window size. This window length was chosen to capture the overall movement trend without closely tracking the trajectory, allowing us to detect fast movements with high amplitude [168]. Within each window, we computed the mean trajectory for each limb's movement component, and we then identified the points where the trajectory intersected with the mean. Subsequently, we calculated the mean distance (d) and standard deviation (σ_d) between consecutive intersections. Finally, the periodicity (P) was determined by combining the aforementioned parameters using the following equation:

$$P = \frac{1}{d + \sigma_d} \tag{3.4}$$

Maximum displacement along the x and y axes: It was calculated for each limb by subtracting the minimum coordinate from the maximum coordinate occupied by the reference point during the recording. This measurement was employed to quantify motion amplitude, another important characteristic given that typical GMs exhibit high range and extent.

The smallest and largest eigenvalues of the 95% error ellipse between the x and y components of the body centroid trajectory were computed: A covariance error ellipse in two dimensions is a graphical representation of the spread or dispersion of a bivariate distribution. It is centered on the mean value of the two variables and its shape and orientation depend on the covariance between them. The major and minor axes of the ellipse correspond to the directions of greatest variation in the data (eigenvectors), as determined by the covariance matrix. The orientation of the ellipse indicates the correlation between the variables. A longer major axis implies greater variability in that direction, and vice versa. The eigenvalues of the covariance matrix represent the variance of the data along the eigenvectors. In the case of correlated data, the eigenvectors indicate

the direction of the largest spread, while the eigenvalues, the magnitude of that spread. Therefore, covariance error ellipses are helpful in understanding the distribution of data points and within our specific application can be used to quantify the magnitude of motion [169].

Percentage of covered space: For each limb we computed the ratio between the sum of all the different positions occupied by that limb and the total number of pixels in the frame to quantify the area of movement.

Mean Pearson Correlation Coefficient: It was computed between the x and y components of left and right hands and between left and right feet trajectories to also measure the symmetry between the two sides.

The Difference between mean velocity of upper body and mean velocity of lower body: It has been calculated with the aim of comparing the amount of movement between the upper and lower parts of the body according to the following formula:

$$v_{diff} = v_{lefthand} + v_{righthand} - v_{leftfoot} + v_{rightfoot} \tag{3.5}$$

Where v_{diff} is the difference between mean velocity of upper body and lower body, $v_{lefthand}$ is the mean velocity of left hand, $v_{righthand}$ is the mean velocity of right hand, $v_{leftfoot}$ is the mean velocity of left foot and $v_{rightfoot}$ is the mean velocity of right foot.

Features were normalized by the infant's bust size to remove differences due to children size. Infants' bust measure was computed by calculating the Euclidean distance between the midpoint of the shoulders and the midpoint of the hips for each frame, and subsequently extracting the median of these values.

3.1.2.7 Feature Selection

Applying the feature extraction procedure detailed in the preceding paragraph, we derived a comprehensive set of parameters that characterize the spontaneous movements of each infant in our dataset across both the NDD and NT groups

and throughout all time points. These features comprised 45 parameters, encompassing a range of kinematic variables tailored to assess the gross motor skills of infants. Specifically, they included 18 statistical descriptors related to velocity and acceleration for both lower and upper limbs, 6 cross-correlations among limb pairs, 2 parameters capturing the periodicity of upper and lower body movements, and 2 metrics representing deviations from the moving average for the upper and lower limbs. Additionally, our extraction process encompassed 8 features associated with the maximum displacement of each limb along the x and y axes, 2 eigenvalues of the centroid, 4 indicators measuring the percentage of space occupied by each limb, 2 Pearson correlation coefficients for upper and lower body movements, and 1 feature expressing the difference in velocity between hands and feet. All these variables were examined to identify potential early predictors of clinical outcomes, specifically distinguishing between NT children and those diagnosed with NDD.

The following procedure was independently executed for each time point after applying min-max scaling normalization, which accounted for differences in measurement units and value ranges of features. As the initial step in reducing the high number of variables, we conducted a correlation analysis using the Pearson method. When variable pairs showed a correlation exceeding 80%, we systematically removed the one displaying lower variability within our sample. Subsequently, we utilized the ANOVA F-value between the labels (NDD versus NT) and each of the remaining features to rank them and identify the optimal subset of variables for each time point [170, 171]. In the context of feature selection, the one-way ANOVA assesses the significance of a feature in explaining the variance in the data and distinguishing between the NDD and NT groups. Specifically, the F-value used in this test quantifies the ratio of the variance between groups to the variance within groups. A high F-value indicates that the feature is more relevant for distinguishing between groups, making it a strong candidate for feature selection [172]. The F-value is calculated as the ratio of the Mean Squares Between-groups (MSB) to the Mean Squares within Groups (MSW):

$$F = \frac{MSB}{MSW} \tag{3.6}$$

The variance between the groups' means MSB is calculated by dividing the Sum

of Squares Between Groups (SSB) by the degrees of freedom for between-groups (dfB), equal to $k-1$, where k is the number of groups:

$$SSB = \sum n_i(\mu_i - \bar{\mu})^2 \tag{3.7}$$

where n_i is the number of observations in the i-th group, μ_i is the mean of the i-th group, and $\bar{\mu}$ is the overall mean.

The variance within each group MSW is measured by dividing the Sum of Squares Within Groups (SSW) by the degrees of freedom for within-groups (dfW), equal to $N-k$, where N is the total number of observations:

$$SSW_i = \sum (x_i - \mu_i)^2 \tag{3.8}$$

Here, SSW_i is the sum of squared differences between individual data points x_i and their group mean μ_i for each group i. The total SSW is the sum of these SSW_i values across all groups.

To select the optimal number of features for the subset at each time point, we employed a recursive algorithm. Beginning with the subset comprising the two features that obtained the highest-ranking scores, the algorithm computed classification performance at each iteration while sequentially adding features in descending order of their ranking. Ultimately, we chose the optimal subset that achieved the best performance with the fewest features. The two groups were compared for the selected variables at each time point using the non-parametric unpaired two-samples Wilcoxon statistical test. We used the feature selection procedure to assess, for each early time point, the existence of a subset within all features capable of significantly discriminating between children later assessed as NT or NDD. Then, only for the time points where this subset was identified, we proceeded by applying the min-max scaling technique for normalization, given the differences in measurement units and value ranges of features.

3.1.2.8 Mixed Effects Model for Selected Features

Given the longitudinal nature of our dataset, with repeated measurements for each subject at different time points, we decided to conduct an additional analysis focused only on the features selected as described in the preceding paragraph, to

address the potential correlation among repeated measures within the same individual. We used a mixed-effects model, which effectively handles this complex data structure, including missing data [173]. This approach enabled us to thoroughly examine our variables by incorporating fixed effects for Group (NT vs. NDD), Time Point, and the Group:Time Point interaction. Random effects were also included to account for within-subject variability (1 | Subject). We used Restricted Maximum Likelihood (REML) as the estimation method. The model applied to analyze each feature was defined as follows:

$$\text{Selected Features} \sim 1 + \text{Time Point} + \text{Group} + \text{Time Point:Group} + (1 \mid \text{Subject}) \tag{3.9}$$

When significant results were found, we performed post hoc analyses to interpret them.

3.1.2.9 Classification Model

The selected and normalized features were employed as input to train a C-Support Vector Classifier (C-SVC) [174] for the early identification of NDD infants at the specific time point. C-SVC is a specific type of SVM designed to find an optimal decision boundary or hyperplane for effectively separating different classes of data in a dataset, ensuring the maximum margin between the classes while still allowing for some classification errors based on the chosen value of C. C represents, indeed, a regularization parameter that controls the trade-off between maximizing the margin (the distance between the decision boundary and the nearest data points of each class) and minimizing classification errors. A smaller value of C will result in a larger margin but may allow some misclassification of data points, while a larger value of C will lead to a narrower margin but fewer misclassifications.

This algorithm is particularly useful when dealing with non-linearly separable data, as it can map the data into a higher-dimensional space to find a linearly separable hyperplane. Moreover, several studies have demonstrated the effectiveness of this classification model even with small-scale datasets. In our work, we set $C = 1$ and chose a radial basis function kernel since data are not linearly separable [174]. Leave-One-Out Cross-Validation (LOOCV) [175] was employed to train and test

the SVM classifier considering the limited size of the dataset. The performances were assessed using the following metrics [176]:

- **Accuracy:** This metric evaluates the ratio of correct predictions to the total number of instances.
- **Precision:** Precision represents the ratio of true positive instances to the total instances classified as positive. It assesses the classifier's capability to refrain from mislabeling a negative sample as positive. A high precision value indicates a low rate of false positives, suggesting that when the model predicts a positive result, it is likely correct.
- **Recall or Sensitivity:** This metric measures the proportion of positive instances that are accurately classified. Recall reflects the classifier's ability in identifying all positive samples.
- **Specificity:** Specificity gauges the proportion of true negatives correctly identified by a classification model out of the total number of actual negatives. Essentially, it quantifies the model's ability to avoid false positives.
- **F1 Score:** The F1 score is the harmonic mean of precision and recall. It provides a balanced assessment, where the relative contributions of precision and recall are equal.

3.1.3 Results

Model Validation with Movidea The Pearson correlation coefficient between each signal extracted using our automatic tracking approach and its corresponding signal obtained from Movidea is 93.96 (88.61-96.60)%. The RMSE was 9.52 (7.29-12.37) pixels. The p-value < 0.0001 evidenced a significant relationship between the two signals. An example of a trajectory extracted using our deep-learning-based automatic approach with the same trajectory obtained from Movidea is provided in Figure 3.4.

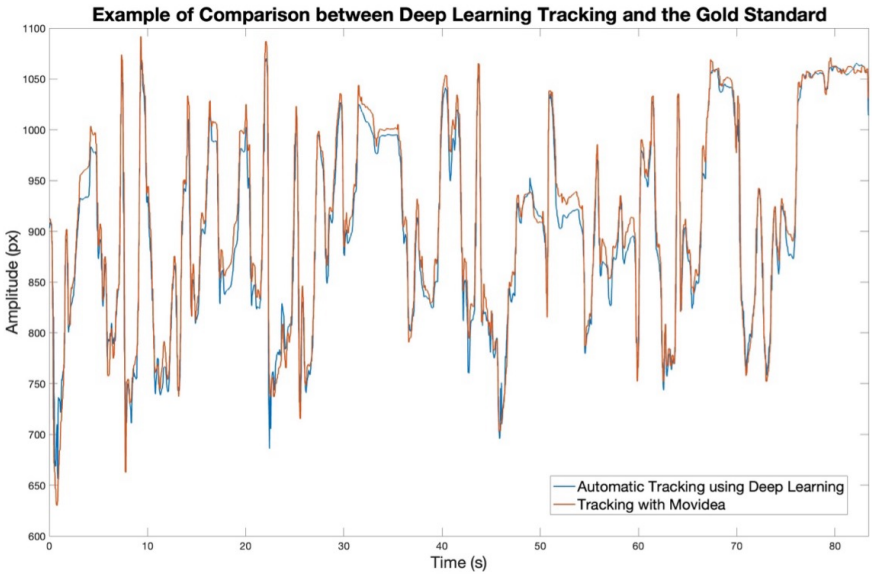


Figure 3.4: Overlay of a foot trajectory extracted using our DL-based automatic approach (blue line) and the same trajectory obtained with the Movidea software, used as the gold standard for validating our approach (red line).

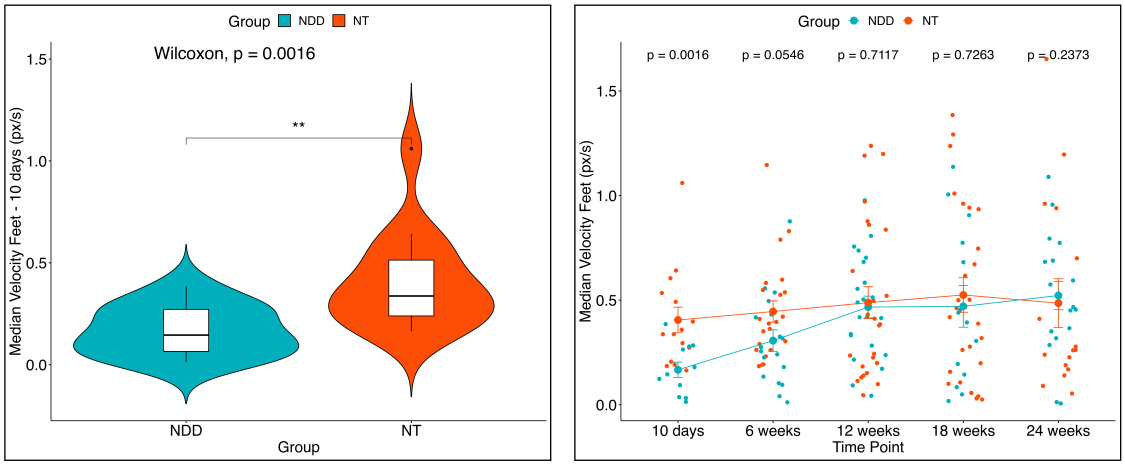
Testing the four amateur videos, we achieved an R of 95.75% and an RMSE of 8.31 px, even when the background was not green (Figure 3.5a). Our approach also demonstrated robust tracking capabilities, achieving an R of 91.44% and an RMSE of 10.62 px, even when the child was not in a supine position (Figure 3.5b). However, for accurate tracking, it was essential that the entire body of the infant remained within the camera's view, as also evidenced by the evaluation of missing data. The algorithm's performance when the infant moved outside the frame was $R = 42.56\%$ and RMSE=19.7 px (Figure 3.5c). Additionally, if the infants were dressed, their clothing needed to have a distinct color from the bedsheet or surface they were lying on. Tests that did not adhere to these constraints resulted in poorer performance, with an R of 64.58% and an RMSE of 11.69 px (Figure 3.5d).

(a) (b) (c) (d)

Figure 3.5: Examples of amateur-recorded videos in unstructured conditions: (a) no green background, (b) no supine position, (c) infant outside the camera's view, (d) infant being clothed in a color similar to the bedsheet.

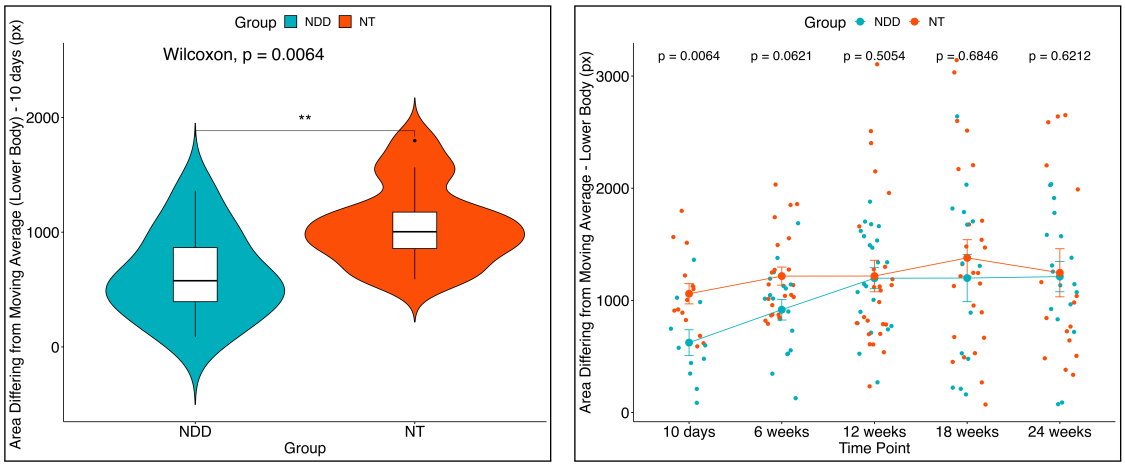
Feature Selected and Mixed Effects Model The most important features identified through the feature selection procedure from the total set of 45 include the 'Median Velocity of the Feet', and the 'Area differing from moving average' and 'Periodicity', all within the lower body domain.

Non-parametric unpaired two-sample Wilcoxon statistical tests between NT and NDD showed that the median values of these three variables were significantly lower in NDD infants than in NT at 10 days. No significant effects were observed for the remaining time points (Figures 3.6, 3.7 and 3.8). Descriptive statistics for these three variables are reported in Table 3.2.



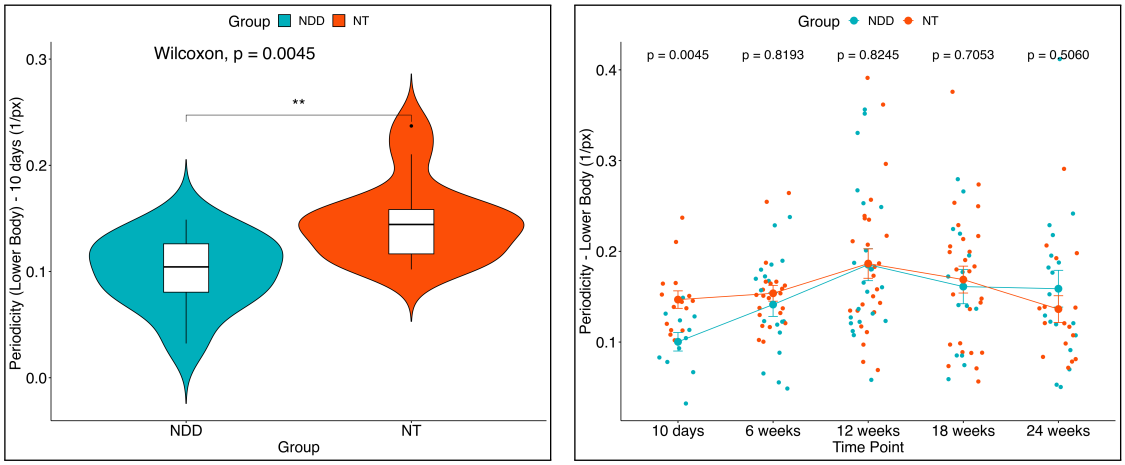
(a) (b)

Figure 3.6: Median Velocity of Feet: (a) Violin plot for the first time point (10 days) and (b) Trend across the 5 time points with mean values and the SE for the two groups: NDD (light blue) and NT (red). The p-values related to the comparison between the two groups were computed using the unpaired two-samples Wilcoxon test. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.



(a) (b)

Figure 3.7: Area Differing from Moving Average of Lower Body: (a) Violin plot for the first time point (10 days) and (b) Trend across the 5 time points with the mean values and the SE for the two groups: NDD (light blue) and NT (red). The p-values related to the comparison between the two groups were computed using the unpaired two-samples Wilcoxon test. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.



(a) (b)

Figure 3.8: Periodicity of Lower Body: (a) Violin plot for the first time point (10 days) and (b) Trend across the 5 time points with the mean values and the SE for the two groups: NDD (light blue) and NT (red). The p-values related to the comparison between the two groups were computed using the unpaired two-samples Wilcoxon test. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Variable	Group	Median	Min	Max	95% CI (Lower)	95% CI (Upper)	U / p-value
Median Velocity of the Feet [px/s]	NDD	0.145	0.0140	0.385	0.0848	0.249	24 / 0.002
	NT	0.337	0.1641	1.060	0.2742	0.536	
Area differing from Moving Average (Lower body) [px]	NDD	576.866	86.1931	1361.686	368.1953	878.608	31 / 0.006
	NT	1003.886	589.6773	1798.358	864.5623	1253.916	
Periodicity (Lower body) [1/px]	NDD	0.104	0.0323	0.149	0.0776	0.123	29 / 0.004
	NT	0.144	0.1020	0.237	0.1261	0.167	

Table 3.2: Descriptive statistics and results of the non-parametric unpaired two-sample Wilcoxon statistical test between the two groups for the three selected variables at 10 days.

The mixed-effects model applied across the multiple time points (repeated measures) revealed a significant impact of both time and group ($p < 0.05$) for all three variables. However, the interaction between group and time did not reach statistical significance. Post-hoc analyses confirmed that these three variables are significantly different between the two groups only at the initial time point (10 days). Results of post-hoc tests are reported in Table 3.3. To verify that the assumptions were met, we examined the residual plots to confirm the absence of deviations from homoscedasticity or normality.

Variable	Time Point	Group 1	Time Point	Group 2	Difference	SE	t	Df	p-value
Median Velocity of Feet	10 days	NDD	10 days	NT	-0.25532	0.1262	-2.0234	177	0.045
	6 weeks	NDD	6 weeks	NT	-0.14762	0.1019	-1.4484	173	0.149
	12 weeks	NDD	12 weeks	NT	-0.02627	0.0942	-0.2790	169	0.781
	18 weeks	NDD	18 weeks	NT	-0.07217	0.1061	-0.6801	175	0.497
	24 weeks	NDD	24 weeks	NT	0.00852	0.1098	0.0776	176	0.938
Area differing from Moving Average (Lower body)	10 days	NDD	10 days	NT	-488.51	232	-2.1036	177	0.037
	6 weeks	NDD	6 weeks	NT	-329.69	190	-1.7363	167	0.084
	12 weeks	NDD	12 weeks	NT	-51.00	177	-0.2888	158	0.773
	18 weeks	NDD	18 weeks	NT	-240.54	197	-1.2201	170	0.224
	24 weeks	NDD	24 weeks	NT	-89.90	203	-0.4418	173	0.659
Periodicity (Lower body)	10 days	NDD	10 days	NT	-0.04338	0.0260	-1.6703	177	0.097
	6 weeks	NDD	6 weeks	NT	-0.01417	0.0211	-0.6709	170	0.503
	12 weeks	NDD	12 weeks	NT	-0.00121	0.0196	-0.0616	163	0.951
	18 weeks	NDD	18 weeks	NT	-0.01011	0.0219	-0.4605	173	0.646
	24 weeks	NDD	24 weeks	NT	0.01251	0.0227	0.5515	174	0.582

Table 3.3: Post hoc test of the mixed-effects models.

Classification The combination of these parameters, once tested, was used as input for the SVM classifier to discriminate NDD infants vs NT with an accuracy of 84.6%. Details of metrics are reported in Table 3.4 and the confusion matrix is shown in Figure 3.9.

Time Point	Accuracy	Precision	Sensitivity	F1 Score	Specificity
10 days	84.62%	100%	63.64%	77.78%	100%

Table 3.4: Performance metrics achieved for the first time point (10 days) by the SVM classifier.

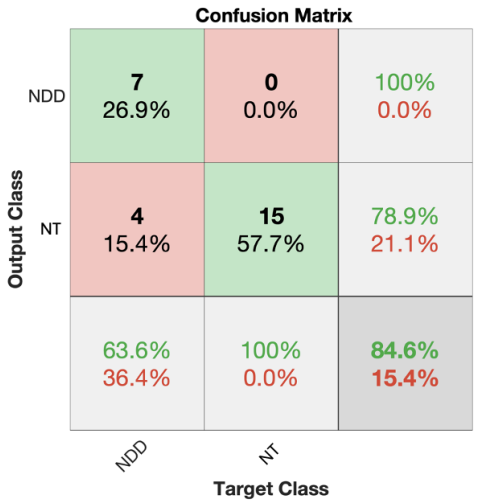


Figure 3.9: Confusion Matrix showing the percentages of subjects accurately and mistakenly associated with each class (children who developed NDD and NT children) for the first time point (10 days). The rightmost column, denoted as Precision, indicates how many infants, assigned to a particular group by the classifier, truly belong to that class. Similarly, the bottom row of the matrix shows the Recall or Sensitivity, indicating for each class how many of the total subjects of that class were correctly recognized by the classifier, providing valuable insights into the accuracy and reliability of the classification process. The total accuracy is displayed in the lower right corner.

3.1.4 Discussion

In the first part of our research, we employed AI to automatically track newborns' spontaneous movements and analyze their trajectories, aiming to characterize both kinematic behaviors and GMs. This stage focused on assessing how early motor patterns might be related to the later onset of NDD, with the overarching goal of identifying specific biomarkers capable of efficiently distinguishing infants at risk from those following a NT developmental trajectory. Such early identification would provide a significant boost in enabling timely interventions that could improve long-term outcomes.

3.1.4.1 Interpretation of Results

At the first assessment, conducted 10 days post-birth, three key parameters related to lower-limb movements emerged as strongly correlated with clinical outcomes: "Median Velocity of Feet," "Deviation from the Moving Average" and "Periodicity" of lower-limb trajectories. These features allowed the classification model to differentiate between infants with NDD and those with NT development with an accuracy of approximately 85%. Notably, all significant parameters pertained to lower-limb activity—particularly feet movements—highlighting the importance of specifically focusing on these limbs during this critical window in development. These findings align with previous work by Moro et al. [102], which highlighted foot acceleration as a crucial early indicator, and reinforce the importance of assessments within the critical window from birth to six weeks [91].

Longitudinal analysis across five time points revealed that differences in lower-limb movements are most pronounced shortly after birth, progressively diminishing in subsequent assessments. This pattern indicates that while early spontaneous motor activity is delayed in infants who later develop NDD, these differences tend to resolve as the infants grow older. However, the initial divergence in motor patterns, although attenuating over time, warrants further investigation into how these early deviations evolve and whether they can provide predictive insights into the final diagnosis. The predictive value of these early motor alterations supports their potential role in shaping other developmental domains, in line with the "neurodevelopmental cascade" theory [25, 177, 56, 178]. While the broader cascading effects of these early deviations are still being explored at this preliminary stage, the strong correlations observed early on emphasize the critical role of motor performance in influencing subsequent developmental trajectories. Future studies should focus on understanding how these early motor differences may impact cognitive and communicative development in the long term.

3.1.4.2 Significance and Value

This study introduces a novel, marker-less, deep-learning-based method for analyzing newborns' spontaneous movements. The fully automated approach, requiring no operator input, facilitates the assessment of infants even when clothed and

supports the use of video recordings taken in non-ideal conditions, making it particularly suitable for home-based monitoring. By eliminating the need for physical markers, it ensures unobtrusive and ecologically valid observations, preserving the natural quality of movements and minimizing potential discomfort or distraction for the infant.

Such accessibility is particularly valuable in underserved regions where clinical resources are limited, enabling caregivers to perform developmental screenings using widely available devices like smartphones. The established link between early motor abilities and later social and communicative development highlights the importance of further exploring this aspect to enhance NDD screening, especially in high-risk populations, such as preterm infants or siblings of children with autism. Furthermore, the integration of ML techniques offers the potential to detect subtle motor anomalies that are often missed by traditional assessments, providing a scalable and efficient tool for early detection.

3.1.4.3 Limitations & Future Improvements

The main limitation of this study is the relatively small sample size, which may restrict the generalizability of the model. Ongoing efforts are focused on expanding the dataset to include a broader participant pool with confirmed diagnoses, which will allow for the exploration of additional movement features, such as rotational patterns and tremors to increase clinical relevance.

Future work will target time points that have not yet revealed significant variables, leveraging the longitudinal nature of the data to capture evolving motor patterns. Investigating how early foot movements relate to later milestones, including reaching and grasping, could provide further insight into the developmental cascade and refine early identification strategies. Building upon the findings concerning spontaneous lower-limb movements, subsequent sections will examine the role of more complex motor behaviors, such as hand use during object exploration, and their implications for language development trajectories.

3.2 AI Analysis of Infants' Hands Movements in Social Engagement and Object-Reaching Tasks

As infants transition from spontaneous to goal-directed movements, the emergence of reaching behaviors marks a critical milestone in neurodevelopment. Reaching facilitates exploration and manipulation of the environment, playing a key role in the development of communication and social engagement. This section aims to investigate hand movements in infants aged 2.5 to 6 months, with a specific focus on reaching behaviors during social and object-directed interactions. Using AI-based video analysis, the study examines movement trajectories between HR and Low-Risk (LR) infants, seeking to explore the relationship between the development of these skills and typical growth across various domains. Additionally, by correlating longitudinal motor data with clinical assessments at 36 months, the research emphasizes the importance of early movements in shaping cognitive and socio-communicative development. Finally, the study highlights the potential of AI-based video analysis as a valuable tool for early assessment of motor skills, offering a scalable solution for detecting developmental delays.

3.2.1 Participants

The study included 43 full-term infants (15 female, 28 male) with no birth complications, defects, sensory impairments, or known genetic syndromes, all from English-speaking families. Infants were followed longitudinally from 2.5 to 36 months of age within a study investigating the development of reaching, sitting, and object exploration in infants with LR and HR for autism. Among the total sample, 20 infants (8 female, 12 male) were considered at HR due to having at least one older sibling diagnosed with ASC. The sibling's diagnosis was verified prior to enrollment using the ADOS [179], the Social Communication Questionnaire (SCQ) [180], and the DSM-4 Clinical Best Estimate [181]. HR infants were enrolled through a university-based autism research program (Pittsburgh Early Autism Study), parent support groups, local agencies, and schools for children

with ASC.

The remaining 23 infants (7 female, 16 male) had no first-degree relatives diagnosed with ASC, and their older siblings had no history of developmental delays or referrals for intervention, so they were considered LR for ASC. These infants were recruited from the Magee-Women's Hospital of Pittsburgh birth registry, local parent-infant programs, and daycare centers.

Each participant from both groups underwent an outcome assessment visit at 36 months using the ADOS scale, with no one receiving an ASC diagnosis. Infants were classified as language delayed if they did not receive an ASC diagnosis and met at least one of the following criteria:

1. Standardized scores on the MacArthur–Bates Communicative Development Inventory, Words and Sentences (CDI-II; [182]) or CDI-III that fell at or below the 10th percentile at two or more time points between 18 and 36 months.

2. A standardized score on the CDI-III at or below the 10th percentile, accompanied by a standardized score on the Receptive Language and/or Expressive Language subscales of the Mullen Scale of Early Learning (MSEL; [183]) that was 1.5 or more standard deviations below the mean at 36 months.

Among all participants, 13 children exhibited signs of language development delays (LD), 24 showed no symptoms (NS), and 6 were missing because they dropped out of the study or were lost to follow-up. Specifically, 10 out of the 20 infants classified as HR and 3 out of the 23 LR infants exhibited language delays.

The ethics committee of Boston University approved the study (Approval Number: 7613E). All families who voluntarily participated provided written informed consent, and the study procedures adhered strictly to the applicable guidelines and regulations. Additionally, informed consent for the publication of identifying information and images was obtained from the parents of the subjects whose images are included in this section.

3.2.2 Methods

3.2.2.1 Procedure & Data Collection

To participate in the trials, infants were visited at home every two weeks, starting at 2.5 months of age and continuing until 6 months. All sessions were recorded using a standard camera. The analyzed video segments included a Spontaneous Movements (SM) session followed by four Object Presentation (OP) trials [184]. During each trial, the infant lay supine on a mat. To minimize frustration, spontaneous movements were always filmed first, followed by the object presentation tasks, as infants may find it challenging to transition from a toy condition to one without toys. During the SM session, the infant faced an experimenter who spoke quietly to maintain engagement for a duration of 2 minutes (Figure 3.10a). In the OP segments, the infant was presented with four different toys serially, each for 30 seconds (Figure 3.10b). Toys were positioned at midline and chest height, ensuring they were within the infant's reachable area, based on the wrist position when the arm was fully extended [185]. If the infants did not reach for a toy within 15 seconds, it was gently placed in their hand. If the infants dropped the toy before the trial was completed, it was promptly replaced in their hand. The toys used during the four object presentation tasks, a rattle, a double shaker, a toy with keys, and a chain of rings, were chosen with the aim of providing sensory and motor stimulation for the children at different levels.

During each trial, participants wore an APDM Opal wearable inertial sensor on both wrists to capture the kinematic parameters of the upper limbs. These wireless sensors, weighing less than 25 g (including the battery), are equipped with two triaxial accelerometers, a triaxial magnetometer, and a triaxial gyroscope, all sampling at 128 Hz. Data were transmitted to a nearby laptop via USB connection through an access point, with device synchronization achieved using a custom-built trigger box. Data loss occurred for specific trials, leading to their exclusion from the analysis. Details about the available sensor data are provided in Table 3.5.

(a) Spontaneous Movement Trial

(b) Object-Presentation Task Trial

Figure 3.10: Example frames from each type of trial.

3.2.2.2 Video Editing

Trials were initially recorded to preserve data for later review, as the original project focused on the manual coding of motor behaviors rather than the application of artificial intelligence for automatic movement tracking. Recordings were captured in a naturalistic setting, reflecting real-life conditions without strict control over filming parameters. While features such as high resolution, consistent lighting, uniform backgrounds, and fixed camera positioning were not prioritized, this approach provided a more ecological representation of infant behaviors.

Camera angles and distances were adapted to the naturalistic context, which

may have occasionally limited the visibility of certain movement details, particularly the hands, which are the primary focus of the current analysis. For these reasons, all recordings were carefully reviewed and edited, precisely trimming the videos so that each resulting clip aligned exactly with the start and end of the trial, as indicated by the experimenter's voice conducting the recordings and a specific sound signal from the sensors used during the trials to mark the beginning and end of data collection. Additionally, only clips in which the camera remained sufficiently stable throughout the trial, without significant changes in angle, were retained, ensuring that the child's hands were clearly visible and that the lighting was adequate. In this way, 5 out of 185 videos were excluded (2.7 %).

Not all participants completed the trials at each of the scheduled time points due to family availability, and this data cleaning process, necessary to ensure accurate analyses, resulted in further data loss. Details regarding the available data for each time point are provided in Table 3.5 .

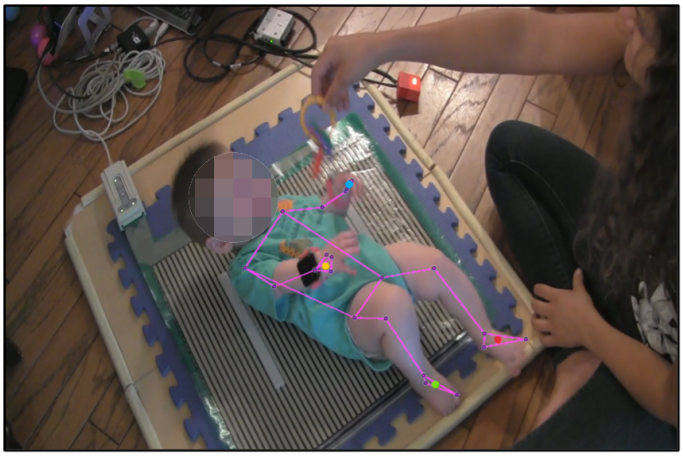
Timepoint (months)	**Total number video data (LD/NS)**	**Total number APDM sensor data (LD/NS)**
2.5	25 (7/14)	18 (5/10)
3	22 (6/13)	20 (7/10)
3.5	21 (7/11)	21 (7/11)
4	29 (9/18)	21 (8/11)
4.5	28 (8/18)	16 (4/12)
5	27 (8/16)	21 (8/12)
5.5	20 (8/11)	17 (7/10)
6	13 (5/8)	13 (7/6)

Table 3.5: Data from SM trials included in the analysis in this study. The missing data from the total are those of infants who dropped out or were lost to follow-up.

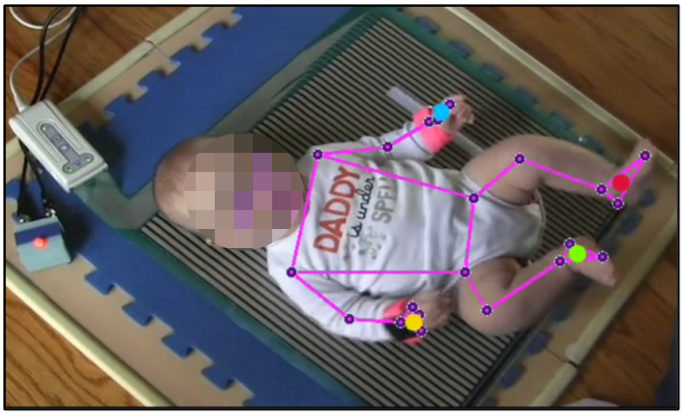
3.2.2.3 Tracking Procedure

A customized version of the MediaPipe Pose Solution [165] was applied to each of the obtained clips to extract the trajectories of the children's movements during the trials at each of the specified time points. This procedure yielded the 3D coordinates of 33 body landmarks for each participant at each time point during the execution of all trials (Figure 3.11). Consequently, for each video, a matrix of size $n \times 99$ was generated, where n represents the number of frames and 99 comes from 3×33 (three coordinates for each of the 33 landmarks) plus an additional

$n \times 1$ column indicating the specific frame. From this dataset, we focused on the x and y coordinates of the right and left wrists for subsequent analyses, aiming to study hands movements to evaluate the development of early object exploration skills in the first months of life. The wrist was chosen as the reference point for the hand due to its relative stability.



(a) Spontaneous Movement Trial



(b) Object-Presentation Task Trial

Figure 3.11: Examples of skeletons extracted through automatic tracking for each type of trial.

3.2.2.4 Trajectories Processing & Feature Extraction

Each wrist trajectory underwent a preprocessing phase before feature extraction. First, normalization was performed based on the child's torso size to reduce vari-

ability in the data caused by different camera distances during recording. The torso size was calculated for each video by determining the Euclidean distance between the midpoint of the shoulders landmarks and the midpoint of the hips landmarks for each frame, followed by computing the median value. Additionally, each trajectory was smoothed using a moving average filter with a window of 10 samples to reduce fluctuations and noise in the data.

Furthermore, in the event of losing a maximum of 5 consecutive samples during automatic tracking, cubic spline interpolation was performed to estimate the missing data. This threshold of 5 samples was selected based on literature recommendations that suggest not exceeding this limit to ensure the reliability of the interpolation results [186]. Furthermore, it is worth noting that there was never a total loss of more than 10% of samples in any video during automatic tracking, which aligns with the acceptable threshold indicated in the literature [187]. Thus, all data were deemed adequate for analysis.

Subsequently, for each video, we extracted a set of kinematic features from the processed hands trajectories. Kinematics are often among the first aspects assessed in motor development evaluations, providing crucial insights into motor skills development [188]. Early analysis of kinematics can help quickly identify any deviations from typical movement patterns.

Specifically, the velocity magnitude of each hand was calculated using the Euclidean distance between the positions of wrist reference points in consecutive frames. This resulting signal was then multiplied by the frame rate to normalize differences from videos recorded on various devices. To smooth out rapid fluctuations in the velocity profiles, a third-order low-pass Butterworth filter with a cutoff frequency set to 95% of the Nyquist frequency was applied. The direction of the velocity was determined by calculating the arctangent of the ratio of the y-component to the x-component of the velocity. Similarly, the acceleration of each hand was derived from the difference between two consecutive velocity samples, with its direction calculated from the ratio of the respective components. Descriptive statistics, including mean, median, standard deviation, variance, mode, skewness, kurtosis, maximum, and minimum, were computed for both the velocity and acceleration magnitudes and directions of each hand. Additionally, the median of the features from the right and left hands was calculated to summarize the

kinematics of the upper body.

For the videos of the OP tasks, conducted at the 6-month time point, the 'reaching time' feature was manually extracted by measuring the interval from the beginning of the trial, when the therapist presented the object, until the child reached for it.

3.2.2.5 Statistical Analysis

We began our analysis by comparing the kinematic parameters described in the previous paragraphs between the two groups of infants at each time point during SM trials. To determine significant differences in the distributions of these parameters between the groups, we applied the non-parametric Mann-Whitney U test at each time point.

The analysis of OP tasks focused specifically on trials conducted at six months. This time point was chosen because significant behavioral differences had already been observed during SM trials only at this age. Furthermore, many infants had not yet acquired object-reaching skills before six months. Due to the substantial time required for video preprocessing, our analysis prioritized this critical time point. The available data for the analyses related to the OP tasks at the 6-months time point included 18 videos (6 LD, 11 NS and 1 missing) and 12 recordings from APDM sensors (6 LD and 6 NS).

At six-month, we extended our analysis to include both the kinematic parameters quantified by the automated AI model and the reaching time. Both parameters were analyzed using the non-parametric Mann-Whitney U test to compare differences between the two groups.

For the analysis of the features extracted from the OP tasks, where each child performed four trials with four different objects, two approaches were used. First, the average value across the four trials was calculated for each participant to obtain a single value for that participant at the given time point, and the Mann-Whitney U test was applied to these data. This provided a simple and interpretable representative value, allowing exploration of macro differences between groups while reducing variability due to noise. However, this approach may lead to a loss of information regarding the variability between different trials and objects, as any

systematic differences between them would not be captured by averaging.

To address this, the analyses were also performed using generalized mixed-effects models to account for the repeated nature of the measurements and the potential variability due to both the subject and the object. Generalized models were chosen because, with fewer than 25 subjects, the assumption of normal distribution could not be made [189, 190, 191, 192]. Specifically, the model considered the object as a fixed effect and the individual variability (the subject) as a random effect. The model used was:

$$\text{Output Variable} \sim 1 + \text{Type of Toy} + \text{Group} + \text{Type of Toy:Group} + (1|\text{Subject}) \tag{3.10}$$

The analyses were conducted after removing outliers using the interquartile range method.

To determine whether motor differences identified at the six-month time point were predictive of clinical outcomes at 36 months (No symptoms vs. Language delays), we tested for the absence of linguistic differences between the two groups at six months. This was done by applying the Mann-Whitney U test to the receptive and expressive language scores on the Mullen Scales of Early Learning [183].

3.2.2.6 Analysis of sensors data

To qualitatively validate the results obtained through AI, we analyzed the data collected using APDM sensors. After internal filtering and processing, the APDM opal sensor output for each hand and sample included temperature (°C), angular velocity (rad/s), linear acceleration (m/s^2), magnetic field strength (μT), and quaternion orientation data. For our analyses, we focused specifically on angular velocity and linear acceleration due to their direct relevance to movement dynamics such as motor control, movement quality, and coordination. Additionally, these measures are more interpretable for evaluating movement kinetics in this context and align with established research methodologies in neurodevelopmental studies, where they are commonly used to assess movement variability, amplitude, and organization.

Previous research [24, 193, 194] has shown that hand movements in infants up

to 6 months of age typically exhibit frequencies of up to approximately 3 Hz. To account for this, we applied a low-pass Butterworth filter with a cutoff frequency of 3 Hz to the angular velocity and linear acceleration data for both hands in all three dimensions (x, y, z).

For each video clip, we extracted a comprehensive set of descriptive statistics from the filtered signals for each hand, averaging the parameters across the three dimensions. The parameters included mean, median, maximum, minimum, standard deviation, variance, maximum autocorrelation value, energy, integral, skewness, kurtosis, Root Mean Square (RMS), dominant frequency, and the mean and standard deviation of the derivative of both angular velocity and linear acceleration.

The resulting data were then compared between the two groups across different time points, using the Mann-Whitney U test.

We used Python 3 [163] for automatic movement tracking with Mediapipe [165], Matlab 2023b [195] for trajectory processing and parameter extraction, and R [196] for subsequent statistical analyses.

3.2.3 Results

The analysis of data extracted through AI-based automatic tracking revealed a significant difference between the two groups in the median velocity of hands during SM trials performed at six months ($p = 0.0062$, $U = 2.00$). Specifically, children who exhibited language development delays at 36 months showed significantly lower hands movement velocity at 6 months. No significant differences were found at other time points in these trials (Figure 3.12a, Supplementary Table A.2).

Similarly, the analysis of data collected using APDM sensors during the same trials highlighted a significant difference between the groups in median angular hands velocity at six months ($p = 0.041$, $U = 5.00$), with lower velocity observed in the group of children with delays. No significant differences were identified at the other time points (Figure 3.12b, Supplementary Table A.3). Moreover, the analysis of linear hands acceleration during these trials showed no significant differences (Figure 3.13).

Analysis of the data collected during OP trials using AI revealed a significant

difference between the groups at six months in hands movement velocity, averaged across the four OP tasks ($p = 0.015$, $U = 9.00$). Again, the group of children with typical development exhibited higher hands velocity compared to those later diagnosed with communicative delays.

Reaching time calculated during OP trials at six months was significantly longer in the group of children with language delays compared to those with typical development ($p = 0.048$, $U = 13.00$).

The Mann-Whitney U test conducted on data extracted from the sensors showed no significant differences in angular velocity, averaged across the four OP tasks, between the two groups ($p > 0.05$). However, the average angular velocity was higher in the group of children with typical development compared to those later diagnosed with communicative delays.

The omnibus test of the mixed-effects model for Reaching Time revealed a significant effect of Group (Language delays vs No symptoms) ($F(1, 14.7) = 4.78, p = 0.045$), although the parameter estimate did not reach statistical significance ($\beta = -4.36$, SE $= 2.24$, $t = -1.945, p = 0.071$). This discrepancy may be attributed to the small sample size, which limits the statistical power to detect effects at the individual parameter level. The analysis did not reveal a significant main effect of the Type of Toy on Reaching Time, nor was the interaction between Group and Type of Toy significant. These results suggest that the Type of Toy does not modify the effect of Group on Reaching Time.

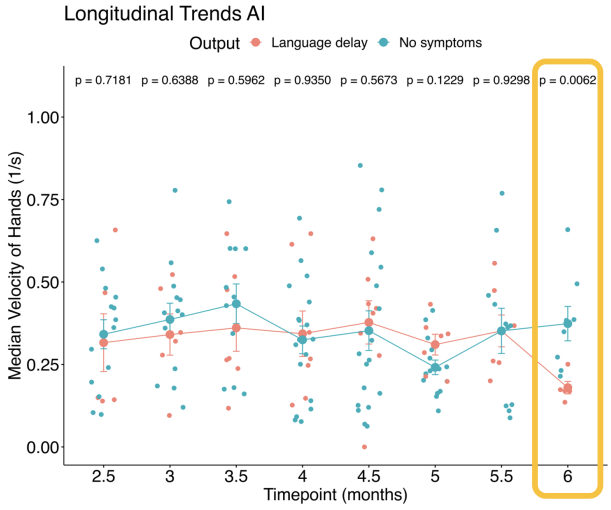
Random intercepts for participants indicated substantial between-subject variability ($\sigma^2 = 17.02, \text{SD} = 4.13$), with an Intraclass Correlation Coefficient (ICC) of 0.67, suggesting that 67% of the total variance was attributable to between-subject differences. The remaining variability was accounted for by residual variance ($\sigma^2 = 8.36$).

For Hands Velocity, also the mixed-effects model revealed a significant main effect of Group in both the omnibus test ($F(3, 16.5) = 9.73, p = 0.006$) and the parameter estimate ($\beta = 0.26$, SE $= 0.02$, $t = 11.49, p < 0.001$). In contrast, the Type of Toy did not exhibit a significant main effect ($F(4, 107.3) = 0.19, p = 0.942$), with parameter estimates ranging from $\beta = 0.00717$ to $\beta = 0.0279$, all $p > 0.10$. Similarly, the interaction between Group and Type of Toy was not significant ($F(3, 106.5) = 0.282, p = 0.838$), with parameter estimates for interaction terms

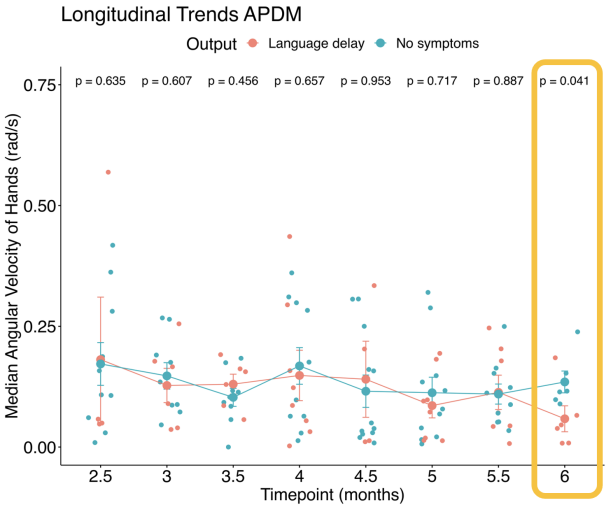
ranging from $\beta = -0.03122$ to $\beta = -0.01060$, all $p > 0.10$. These findings indicate that the Type of Toy does not directly affect Hands Velocity, nor does it moderate the effect of Group.

Random intercepts for participants showed considerable between-subject variability ($\sigma^2 = 0.005, \text{SD} = 0.07$), with an ICC of 0.54, indicating that 54% of the total variance in Hands Velocity was attributable to differences between subjects. The residual variance ($\sigma^2 = 0.004$) accounted for the remaining variability.

No significant differences were identified between the two groups at six months in either the receptive language or expressive language scores on the Mullen Scales of Early Learning.



(a)



(b)

Figure 3.12: Trends of the *Median Velocity of Hands* calculated from trajectories extracted using AI (a) and of the *Median Angular Velocity of Hands* obtained using APDM Opal sensors (b) across the five time points with mean values and SE for the two groups: Language Delays (pink) and No Symptoms (light blue). Data were normalized using the min-max scaling method to ensure comparability between the values extracted by AI and those extracted by the sensors. The p-values for the comparison between the two groups were computed using the Mann-Whitney U test. Significance levels are indicated as follows: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. The yellow square highlights the six-month time point where a significant difference between the two groups was observed with both methods.

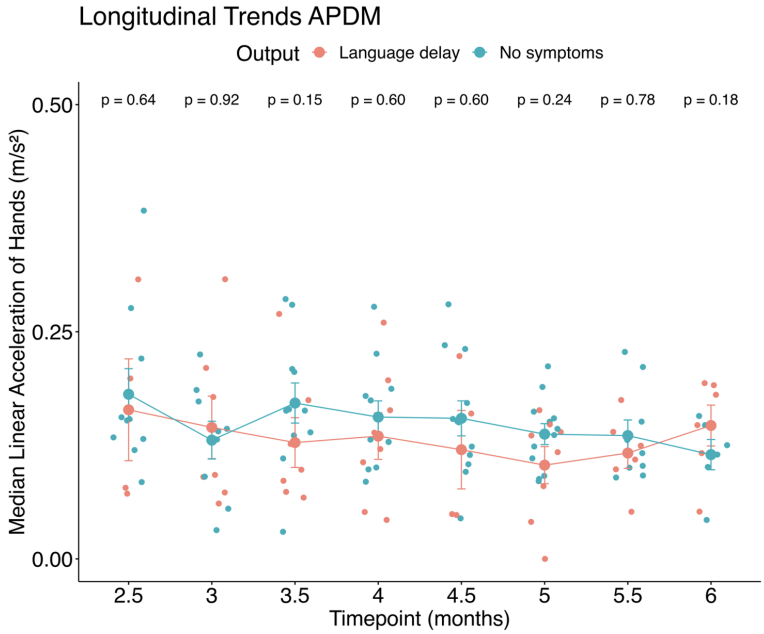


Figure 3.13: Trend of the *Median Linear Acceleration of Hands* (m/s^2) obtained using APDM Opal sensors across the five time points with mean values and SE for the two groups. The p-values for the comparison between the two groups were computed using the Mann-Whitney U test. Significance levels are indicated as follows: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

3.2.4 Discussion

This section continues the longitudinal investigation of motor development, focusing on the emergence of key motor milestones. Specifically, we examined differences in hand movement characteristics during social engagement with a therapist and object-directed reaching tasks among infants who later exhibited language delays, compared to those with typical development. By integrating longitudinal data, advanced AI-based tracking, and kinematic analysis from wearable sensors, this study enhances our understanding of how early motor delays impact exploration, communication, and social development. The multi-timepoint design, spanning from 2.5 to 6 months of age, captures the transition from early spontaneous movements to goal-directed behaviors, such as reaching and object manipulation, facilitating the identification of the optimal time point for evaluating these skills.

3.2.4.1 Interpretation of Results

Our analysis revealed significant differences in hand movement characteristics between 6-month-old infants who later exhibited language delays at 36 months and those with typical development. These findings are consistent with previous research by Iverson et al. [197], which demonstrated that fine motor differences become evident by six months in high-risk infants.

During object-presentation tasks, infants with language delays exhibited longer reaching times compared to their NT peers. Kinematic data from sensors confirmed these delays, highlighting reduced angular hand velocity as a distinguishing factor, while linear acceleration remained similar across both groups. This suggests that rotational movements may play a more critical role than linear movements in differentiating motor performance. Additionally, manually calculated reaching times further support the notion of less efficient spatial exploration in infants with delays, potentially limiting their opportunities for environmental interaction.

It is important to note that AI captured both linear and rotational movements, while sensor data specifically focused on angular velocity and linear acceleration. The consistent outcomes from this dual approach, despite methodological differences, further support the findings.

3.2.4.2 Significance and Value

The results once again highlight the crucial role of early motor behaviors in shaping developmental trajectories. Significant differences in hand movements emerged at six months, aligning with previous studies that have shown how motor development progresses from basic, spontaneous movements to more complex, goal-directed actions over time [198, 199]. The development of skills such as reaching typically follows the maturation of more foundational spontaneous movements, underscoring the importance of early motor milestones in predicting later developmental outcomes, including communication delays.

The longitudinal design of this research adds significant value, as it tracks the progression from spontaneous movements at 10 days to goal-directed behaviors like reaching by six months. This approach offers a deeper understanding of how motor skills evolve, with earlier disruptions potentially signaling future developmental

challenges.

By integrating AI-based tracking with sensor-derived kinematic data, this study demonstrates the practicality of non-invasive movement monitoring. The AI method, validated against wearable sensor data, effectively captures meaningful motor patterns while maintaining a naturalistic environment for the infant. This was particularly evident in the successful analysis of unstructured video recordings, which still provided reliable insights despite variations in recording conditions.

3.2.4.3 Limitations & Future Improvements

It is important to note that variability in video quality occasionally hindered AI-based analyses, highlighting the need for standardized recording protocols to improve data consistency. Additionally, the relatively small sample size limits the generalizability of the findings. Expanding the cohort in future studies will enhance the statistical power and robustness of the conclusions.

Further investigations should examine the longitudinal relationship between early spontaneous movements and later-reaching behaviors to better understand the developmental pathways underlying the neurodevelopmental cascade. Moreover, the observed significance of angular velocity as a distinguishing factor between groups presents a promising avenue for refining early screening methods.

Beyond motor development, early social communication plays a crucial role in understanding Neurodevelopmental Disorders. The following chapters explore how children with NDD integrate multimodal behaviors, such as gestures, gaze, and speech, during naturalistic social interactions, offering a broader perspective on developmental trajectories.

Chapter 4

Emerging Communicative Skills in Toddlers: Gestures, Gaze and Vocalizations in Naturalistic Interactions

Emerging communicative skills in early childhood, including gestures, gaze, and vocalizations, are fundamental to the development of social interactions and language acquisition. During this critical developmental window, toddlers increasingly rely on multimodal coordination to engage with their environment, triangulate attention, and convey intentional communication. Disruptions in these early communicative behaviors have been identified as early markers of NDD, particularly ASC, underscoring the importance of timely and ecologically valid assessment methods. Traditional structured evaluations, while informative, often fail to capture the spontaneous nature of social communication, potentially overlooking subtle but significant behavioral cues [200, 201]. To address this limitation, the research presented in this chapter adopts a naturalistic approach, focusing on behaviors exhibited during parent-child interactions in free play contexts. By integrating microanalytic behavioral analysis with advanced AI models, this work aims to enhance early diagnosis of NDD by detecting nuanced communicative patterns that may not emerge in traditional clinical settings.

Two complementary studies are presented. The first investigates multimodal communicative behaviors through detailed microanalytic coding of naturalistic video recordings. The second presents the development of a transformer-based deep learning model to automatically recognize deictic gestures from the same video data, aiming to improve diagnostic efficiency while preserving the richness of naturalistic observation.

4.1 Early Multimodal Behavioral Cues in Autism: a Microanalytic Exploration of Actions, Gestures and Speech during Naturalistic Parent-Child Interactions

Multimodal integration of communicative behaviors is a fundamental aspect of early socio-communicative development. In toddlers, these integrated behaviors underpin the emergence of joint attention, intentional communication, and early language skills. Alterations in the frequency, diversity, and coordination of these behaviors are commonly observed in children with ASC, often manifesting as reduced gesture production, diminished positive affect, and atypical gesture-gaze coordination [202, 203, 204, 205, 206]. Early identification of these differences is essential, as they can lead to more significant communicative and social difficulties if left unaddressed.

This section presents a microanalytic exploration of multimodal communicative behaviors in toddlers during naturalistic parent-child interactions. Using a detailed second-by-second coding scheme, the study systematically analyzed gestures, motor actions, and speech to capture the intricacies of real-world communicative exchanges. By leveraging naturalistic play contexts, this study provides ecologically valid insights into early socio-communicative development, emphasizing behavioral patterns that may be overlooked in structured assessments [207, 208].

4.1.1 Participants

17 autistic (ASC) and 15 NT toddlers (7 ASC and 8 NT females, respectively), participated in this study. ASC and NT children were individually matched according to the ASC child's non-verbal developmental age (ASC: mean= 24 months, sd=8.5; NT: mean= 22 months, sd= 5.8). ASC children and their families were recruited through and participated in naturalistic caregiver-child interactions at the clinical facilities of the National Research Council of Italy, Institute for Biomedical Research and Innovation (CNR-IRIB) in Messina and at the clinical and territorial service in Catania. NT children were recruited via mainstream nursery schools in

the local territory of Messina and Catania. Inclusion and exclusion criteria together with demographic and clinical characteristics of the sample are reported in the *Supplementary Information*. The study received ethical clearance by the local health ethics committee (protocol number 08/2021), and all the caregivers provided informed consent to participate in the study.

4.1.2 Methods

4.1.2.1 Parent-child interaction protocol

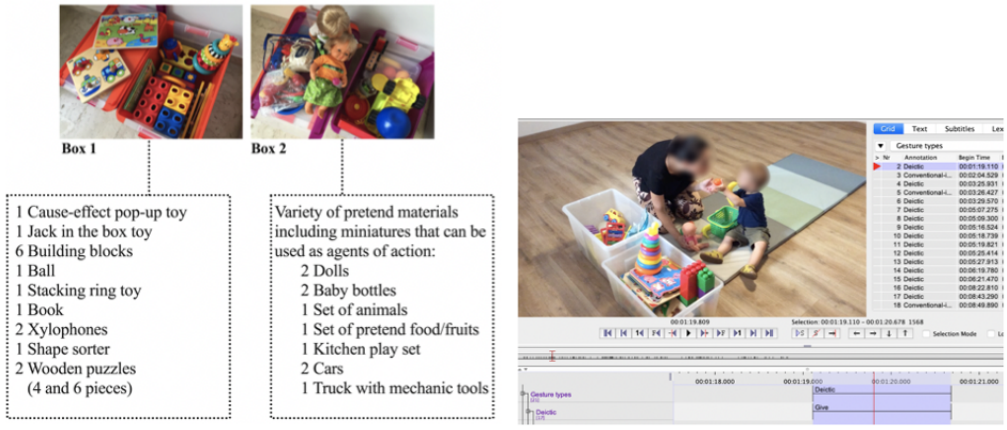
All the mothers and children in the study were involved in a Parent-Child Interaction (PCI) protocol conducted before they started a Naturalistic Developmental Behavioral Intervention (NDBI) following the principles of the Early Start Denver Model (ESDM). The PCI protocol consists of a 10-minute free play interaction, during which mothers are given a standardized set of age-appropriate toys and instructed to engage with their child as they would at home. The PCI toy set is organized into two boxes providing a diverse array of play materials and duplicates of toys to encourage interactive play (Figure 4.1a). The PCI interactions took place at the CNR-IRIB facilities, in a quiet room without any other play material or distractions. A square carpet approximately 2x2 m was placed on the floor to allow the participants to sit naturally and comfortably. The two boxes, without lids, were placed on one side of the carpet to be easily accessible by both child and parent. The interactions were recorded through a high-resolution digital camera, which was operated by the researcher. The researcher maintained an appropriate distance to ensure a frontal perspective of the scene. Under no circumstances did the researcher engage with either the parent or the child to offer information or participate in the activity. Special care was taken to ensure that the researcher's presence did not interrupt the natural flow of the interaction.

4.1.2.2 Behavioral Microanalytic coding

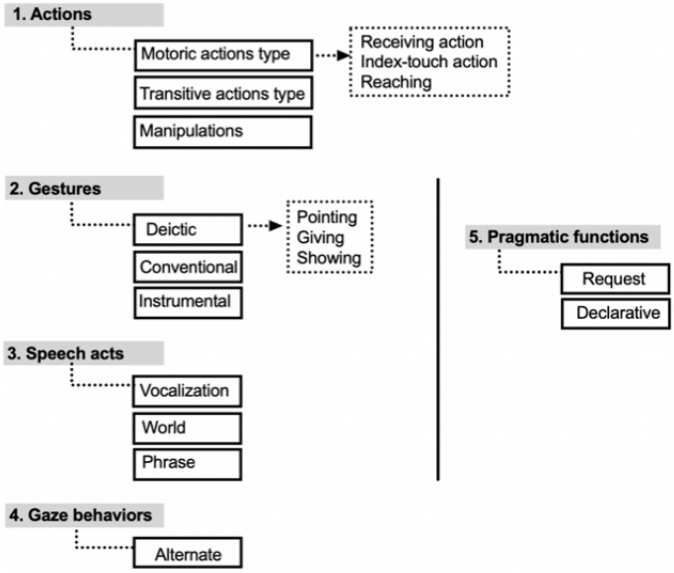
Video footage of all the mother-child interactions was carefully reviewed to be transcribed for detailed analysis of actions and multimodal communication. Relevant behavioral information from video contents were extracted using an open-source

behavioral annotation tool for audio and video recordings, which allows an unlimited number of textual annotations, supports creation of multiple tiers and tier hierarchies Eudico Linguistic Annotator (ELAN); [209]. See Figure 4.1b for an overview of the ELAN annotation system.

A novel microanalytic hierarchical coding scheme, specifically designed for the study, was implemented. It examined various motor schemes and communication modalities for social orienting, social responsiveness, and social initiative. The analysis included the frequency and/or duration of target behaviors, such as motor actions, gestures, and speech. Subcategories of these behaviors, such as reaching behavior within motoric actions or pointing, giving, and showing within deictic gestures, were also explored. Additionally, the study investigated chained behaviors, such as object-partner alternate gaze associated with gesture, along with the pragmatic functions of gestures (i.e., request of objects or help, versus declarative). Figure 4.1c presents a comprehensive depiction of the coding scheme, and all the details of the coding manual are also provided in the *Supplementary Information*.



(a) (b)



(c)

Figure 4.1: Composite figure showing various elements: (a) List of parent-child interaction toys and materials; (b) Screenshot of the ELAN annotation system (written consent obtained); (c) Microanalytic coding scheme hierarchical architecture.

4.1.2.3 Inter-coder reliability

During an initial training phase, four coders underwent formal training, supervised by an expert in the coding procedure. Organized sessions were conducted to familiarize coders with the ELAN software, offering guidance on its usage and facilitating the exploration of selected target behaviors. This involved coding four videos featuring both NT and ASC children. After the training phase, each coder was tasked with independently coding six randomly selected video clips (3 ASC, 3 NT), constituting 20% of the total observations. The coding was conducted in a blinded manner to the autism condition. The expert independently coded the same video segments and computed Cohen's Kappa [210] as a reliability metric. Cohen's Kappa was determined through pairwise comparisons between the expert and each independent coder individually. Subsequently, an average Kappa was calculated for each of the macro categories considered in the analysis. Cohen's kappa values were .72 for actions, .80 for gestures, .71 for speech, reaching a level considered "substantial" [211]. Disagreement locations were pinpointed and collaboratively reviewed and resolved by both coders and the expert.

To ensure intercoder reliability, weekly meetings were held among the experts and coders. These meetings aimed to sustain agreement on macro categories, establish consensus on hierarchical categories, validate the coding process, and address any uncertainties or concerns about the coding procedure.

4.1.2.4 Data Analysis

All statistical analyses and graphical visualizations were implemented in R (version 4.2.1) [196]. To comprehensively explore multiple target behaviors, we conducted a Principal Component Analysis (PCA) as a first step. The aim was to extract the most relevant information from our rich dataset by efficiently reducing the number of variables. This reduction facilitated clearer visualization and interpretation of the data, revealing its inherent patterns. Specifically, we projected the data onto a condensed set of principal components to explore their effectiveness in discriminating between ASC and NT children. A scatterplot was created for the participants using the first two principal components (PC1 and PC2), and different colors were assigned to children in each group. Additionally, to evaluate

the influence of each original variable on the first two principal components, we overlaid a correlation plot of the variables onto PC1 and PC2. Variables that have stronger correlations with PC1 and PC2 have greater relevance in explaining variability within the dataset. Consequently, these variables were utilized as input to train the Logistic Regression (Log-Reg) classifier for automated differentiation between the ASC and NT groups.

Specifically, we employed a Mutual Information (MI) estimation to identify, from the comprehensive original set of behavioral features, those demonstrating the highest predictive power. MI quantifies the dependence between two random variables, in our case, a feature and a discrete target variable. A MI score of zero indicates complete independence, whereas higher values indicate a higher level of dependency [212, 213]. After ranking the variables according to their MI values, we identified the top four by analyzing the knee point in the corresponding bar chart. Then, using the selected features as input, we proceeded to train and test the classifier using two different cross-validation methods: 10-Fold Cross-Validation (10FCV) and LOOCV.

To report the outcome of the classifier, we applied the following performance metrics [176], considering ASC as a positive class:

- **Accuracy**: Calculates the ratio of correct predictions to the total number of instances, providing an overall measure of predictive correctness.
- **Precision**: Is the ratio of true positive instances to the total instances classified as positive; a higher precision indicates a lower rate of false positives.
- **Recall**: Measuring the proportion of positive instances correctly classified, it signifies the classifier's ability to identify all positive samples.
- **F1 Score**: Being the harmonic mean of precision and recall, provides a balanced measure where the contributions of precision and recall are equally considered.

To enhance the interpretability of our model results, we employed SHapley Additive exPlanations (SHAP) values [214]. This game-theoretic approach provides valuable insights into how the essential features of a dataset influence the

model's output. In the beeswarm plot, each sample is represented as a dot, with its x-coordinate indicating the SHAP value for the respective feature. The color of each dot reflects the original value of the feature. Features are organized based on their predictive power, offering a visual representation of each feature's impact on the model's predictions.

4.1.3 Results

Upon closer examination of the variables' correlation with the principal components, and observing sample projections in the new space, a distinctive pattern became apparent. Specifically, NT children predominantly occupy quadrants characterized by positive correlations with: (a) greater gesture use: this not only involves a higher frequency of gestures like pointing, showing, giving, and conversational-interactive gestures but also encompasses multimodal integration and social intention, such as gesture-gaze integration and gestures with declarative intentions; (b) higher verbal communication: NT children in these quadrants exhibit a higher number of phrases and words to communicate with their partner; (c) increased engagement in functional play schemes: this involves a greater involvement in functional play schemes with objects.

Conversely, ASC children predominantly exhibited a positive correlation with variables indicating: (a) lower gesture production: less showing, giving, pointing, and conventional gestures and more instrumental gestures. These gestures are mainly used for expressing behavioral regulation and requesting functions rather than declarative intentions; (b) presence of motor actions, including reaching behaviors, hand opening to receive an object from the partner as well as the use of the index finger mainly to carry out motor actions on objects rather than pointing for communication purposes; (c) lower presence of partner-object alternate gaze during gesturing; (d) greater use of vocalizations both intentional and unintentional, rather than words or phrases; (e) greater use of nonfunctional object manipulation compared to engagement in functional play schemes.

Taken together, these findings emphasize unique behavioral patterns that play a role in distinguishing between NT and ASC children. Figure 4.2 shows the behavioral characteristics of the ASC and NT children projected onto the PC1

and PC2 plane.

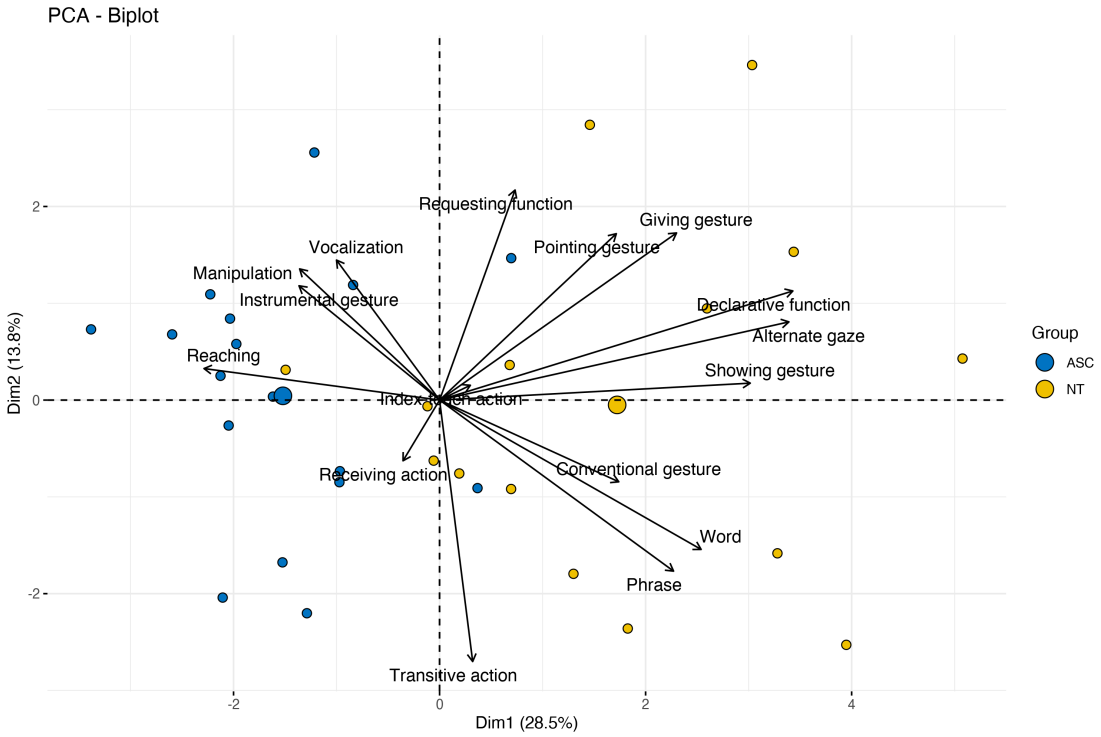


Figure 4.2: Two-dimensional scatter plot of ASC and NT children's behavioral features, projected onto PC1 and PC2 plane.

Applying the MI algorithm on behavioral variables to pinpoint the most influential subset, the features that emerged as robust predictors of an autism condition encompassed 'Declarative Function', 'Manipulation', 'Alternate Gaze', and 'Reaching'. Using these identified features as input for the Log-Reg model, which aimed to classify between ASC and NT children, we attained an accuracy of around 93%. A comprehensive overview of the model's performance metrics for both LOOCV and 10FCV and LOOCV accuracy trend increasing the number of features are provided in Figure 4.3 and Table 4.1.

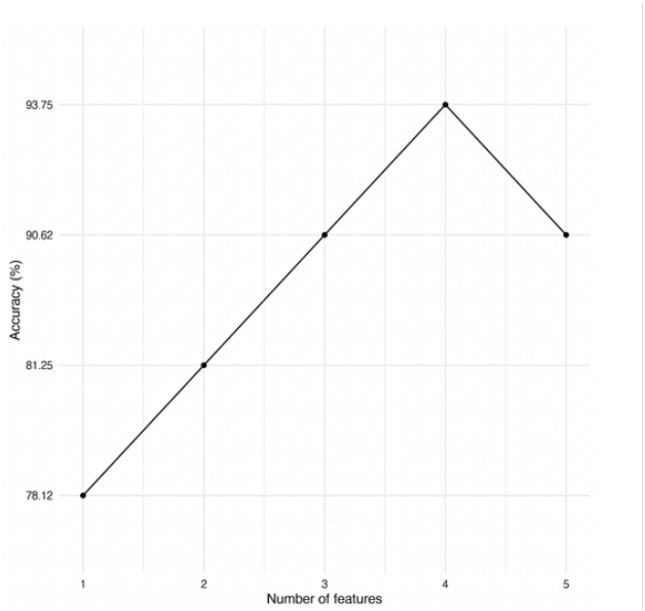


Figure 4.3: Leave-one-out accuracy increasing the number of features.

Measure	**Leave-one-out Cross-Validation (average)**	**10-folds Cross-Validation (average, SD)**
Accuracy	93.75%	93.33% ± 13.33%
Precision	94.12%	95% ± 15%
Recall	94.12%	95% ± 15%
F1 Score	94.12%	93.33% ± 13.33%

Table 4.1: Performance metrics of the Log-Reg model for classifying ASC vs NT children for both LOOCV and 10FCV.

SHAP values associated with the key features validated our initial observations from PCA. As depicted in the beeswarm plot presented in Figure 4.4, elevated values for the features 'Declarative Gesture' and 'Alternate Gaze' exert a positive influence on the model output, leaning it towards the NT group. Conversely, for the remaining two features, namely 'Manipulation' and 'Reaching', higher values guide the model toward the classification of ASC children. This analysis provides a nuanced insight into the distinct contributions of individual features in shaping the model's classification decisions.

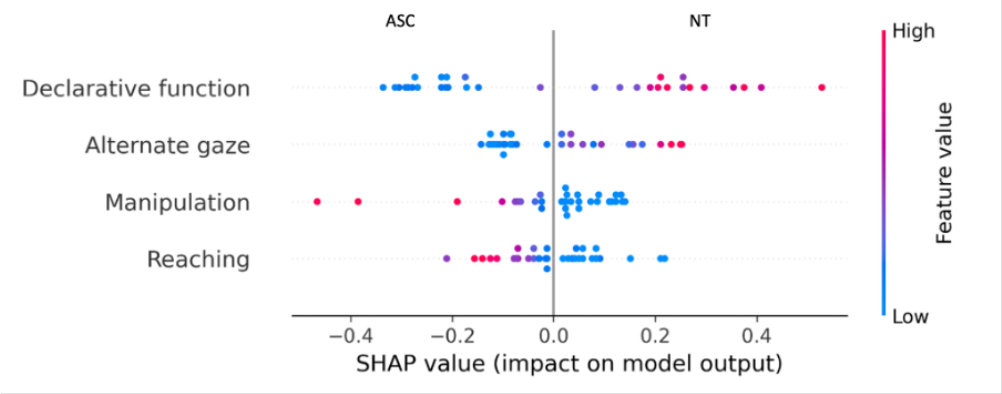


Figure 4.4: Beeswarm plot displaying SHAP values for the top four features.

4.1.4 Discussion

Continuing our investigation into early markers of atypical neurodevelopment, this study focused on a microanalytic examination of multimodal communicative behaviors during naturalistic parent-child interactions. We analyzed a broad range of behavioral variables, including manipulation, transitive actions, gaze, gestures, and speech patterns, among young children with ASC and NT peers. By employing an innovative second-by-second coding scheme, we captured detailed social–communicative profiles for each child. Subsequent ML analysis revealed distinct behavioral clusters, emphasizing significant differences in gesture production, action patterns, and speech integration between ASC and NT children [202, 203, 204, 205, 206].

4.1.4.1 Interpretation of Results

Behavioral clustering demonstrated that NT children exhibit greater complexity and diversity in their social–communicative behaviors compared to children with ASC. Specifically, NT children were observed to use declarative gestures more frequently, integrate eye contact more consistently during interactions, and participate actively in functional play schemes. In contrast, children with ASC tended to rely predominantly on request-based gestures, showed reduced gaze alternation, and preferred object manipulation over more functional action patterns. Our ML analysis identified four pivotal features, declarative function, alternate gaze, ma-

nipulations, and reaching, that together achieved over 93% classification accuracy and 94% precision in predicting autism.

4.1.4.2 Significance and Value

The microanalytic approach adopted in this study offers several innovations. By analyzing naturalistic parent-child interactions, we obtained ecologically valid insights into early communicative behaviors that are often missed in structured clinical assessments [200, 201]. The integration of multimodal data (gestures, vocalizations, and gaze) provided a comprehensive understanding of the social–communicative profiles that differentiate ASC from NT children. These findings underscore the diagnostic importance of subtle behavioral markers and support the use of ML techniques in early autism research.

4.1.4.3 Limitations & Future Improvements

The relatively small sample size of children with ASC limits our ability to explore potential autism subtypes and may affect the generalizability of the findings. Furthermore, the absence of a comparative analysis with other clinical groups restricts the specificity of our results to autism. Future research should aim to expand the sample size, include additional clinical groups for comparison, and extend the longitudinal tracking of individual behavioral trajectories to further elucidate how early multimodal communicative behaviors influence later social and language outcomes. Incorporating additional multimodal features, such as the interplay between gesture, gaze, and speech, could also provide a richer understanding of early autism markers.

4.2 A Deep Learning Approach for Automatic Video Coding of Deictic Gestures in Children with Autism

Building upon the behavioral insights gained from the microanalytic exploration, which highlighted the importance of assessing gestures, the second study addresses the challenges of manual gesture coding, specifically its time-consuming nature and susceptibility to observer bias. Deictic gestures, such as pointing, showing, giving, and requesting, are central to early communication, serving as key indicators of joint attention and intentionality [205, 206]. Given their diagnostic significance, automating the detection of these gestures from naturalistic videos has the potential to streamline early assessment processes while preserving ecological validity.

This section introduces the development of an AI-driven tool based on a transformer architecture to automatically recognize deictic gestures in toddlers during parent-child free play interactions. By training the model on annotated video data from the first study, this work aims to create a tool that simplifies the characterization and assessment of communicative gestures in children. Ultimately, the study seeks to demonstrate the potential of deep learning for the automatic coding of gestures.

4.2.1 Participants

This study was conducted on a group of 6 young children between the ages of 23 and 63 months who had been clinically diagnosed with ASC according to the DSM-5 [2] and established by expert clinicians using the ADOS [215]. Children with autism were recruited at the clinical facilities of the Institute for Biomedical Research and Innovation of the National Research Council of Italy (IRIB-CNR) in Messina. The study received ethical clearance by the Ethics Committees of Azienda Ospedaliera Universitaria P. Giaccone, Palermo, Italy, approval ID 07/2022 12/07/2022.

4.2.2 Methods

4.2.2.1 Data collection

For data collection and processing, we employed the detailed strategy described in the preceding section. In summary, we recorded and analyzed 10-minute videos of naturalistic mother-child interactions for each participant. Parents were instructed to play spontaneously with their children using a standardized set of age-appropriate toys, designed to elicit a broad range of play behaviors, from basic exploration to symbolic play. Each of the six sessions was captured with a fixed video camera and processed using the described moment-by-moment coding procedure, enabling both quantitative and qualitative analysis of gestural production. Specifically, here we focused on targeting and detecting four categories of deictic gestures (Figure 4.5).

Deictic gestures are a type of nonverbal communication that refers to an object or event by directly pointing or touching it. Their meaning is dependent on the context in which they are used. Mastrogiuseppe et al. [216] proposed the classification of children's deictic gestures into four categories:

- **Pointing:** it involves the child using distinctly the index finger to direct attention towards a specific object, place, or event.
- **Showing:** when the child grasps an object towards the adult to show it.
- **Giving:** it involves the child offering an object to the partner.
- **Requesting:** when the child extends the arm with the palm facing up to ask for something, usually accompanied by opening and closing the hand.

From six video processing, 37 repetitions of deictic gestures were identified resulting in 37 separated video clips to use for model development. Detected gestures are divided as follows:

- **Pointing:** 9 repetitions.
- **Showing:** 3 repetitions.
- **Giving:** 11 repetitions.

- **Requesting:** 14 repetitions.

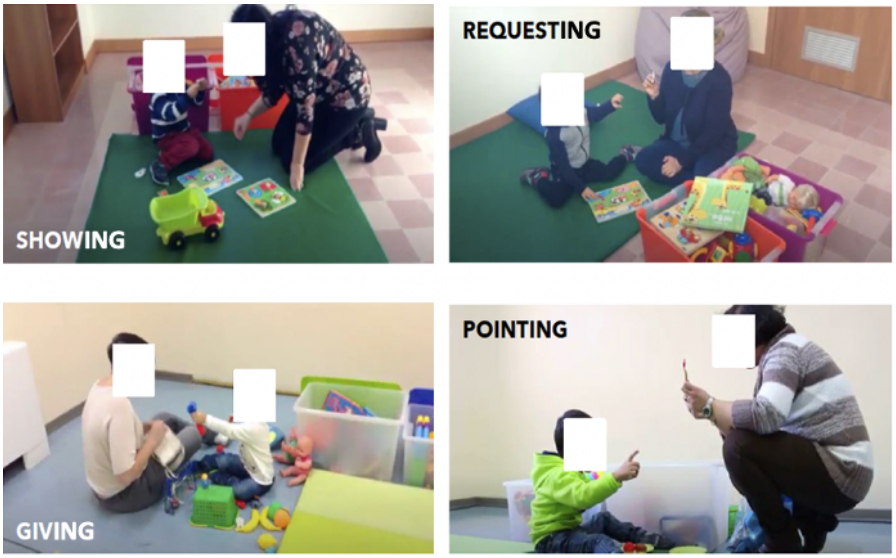


Figure 4.5: Examples of the four deictic gestures.

4.2.2.2 Deep learning Classifier: The Transformer Architecture

Transformers are simple models that exploit the attention mechanism to transform input sequences into output sequences using an encoder-decoder architecture. Positional encodings are computed using sine and cosine functions and summed to input embeddings before encoder and decoder stacks in order to include information about the relative position of the elements in the sequence [122].

Both the encoder and the decoder consist of 6 identical blocks. In the encoder, each block includes a multi-head self-attention network and a fully connected feed-forward network. Instead, decoder blocks include a further multi-head attention sublayer applied to the outputs of the corresponding encoder block.

The attention mechanism proposed in [122], called *Scaled Dot-Product Attention*, applies an attention function for mapping a query and a set of key-value pairs (with dimension d_k and d_v respectively) to an output. The output is computed as the weighted sum of the values, where the weight of each value is calculated by the dot product between queries of each position and keys of other positions in

the sentence dividing by d_k and applying a softmax function:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{4.1}$$

In particular, in self-attention layers of the encoder, all the queries, keys and values come from the output of the previous encoder and in the same way for self-attention layers of the decoder that represents an autoregressive model. In *encoder-decoder attention* layers instead, the queries come from the previous decoder layer, and the memory keys and values come from the output of the encoder.

At the end, the decoder output is converted through a linear layer followed by a softmax function to predict next token probabilities [122].

To automatically identify the four deictic gestures from selected video clips, we developed a multi-frame approach based on AI models considering each video clip as a whole, rather than training and classifying single frames.

In more detail, as a first step, each video clip was preprocessed by resizing its frames to the input dimensions of the Dense Convolutional Network Model with 121 layers (DenseNet121) [217] exploited for the feature extraction procedure: $128 \times 128 \times 3$. In addition, a fixed length of 2 seconds (25 fps) was set as duration for each clip. Shorter videos were padded to 50 frames by adding empty frames at the end, while longer videos were centered on the action of interest cutting frames in excess.

Feature extraction was performed by removing the fully-connected layer at the top of the DenseNet121 pretrained on ImageNet, and by applying a global average pooling to the output of the last convolutional block. A total of 1024 features was thus obtained for each video frame. Extracted features from frames of each video clip were employed for training a multi-frame classification model based on the transformer architecture proposed in [122]. Information about the relative positions of the frames was incorporated into the model exploiting the positional encoding. Positions of frames within videos were encoded using an `Embedding` layer preceding the transformer encoder and then added to the pre-computed features.

The described architecture is followed by a maximum pooling operation (`Global MaxPooling1D` layer) and a dropout layer (0.5 `Dropout` layer) to prevent overfitting. Model output consists of a densely-connected layer with Softmax activation

function that assigns output probabilities for the four possible categories and performs classification. The resulted model includes ∼4.27 million trainable parameters. Details about model architecture and trainable parameters are reported in Table 4.2 and Figure 4.6.

Layer type	**Output shape**	**Parameters**
Input Layer	-	0
Positional Embedding	(None, None, 1024)	51200
Transformer Encoder	(None, None, 1024)	4211716
Global Max Pooling 1D	(None, 1024)	0
Dropout	(None, 1024)	0
Dense	(None, 4)	4100
Total parameters		4267016
Trainable parameters		4267016
Non-trainable parameters		0

Table 4.2: Details about model architecture.

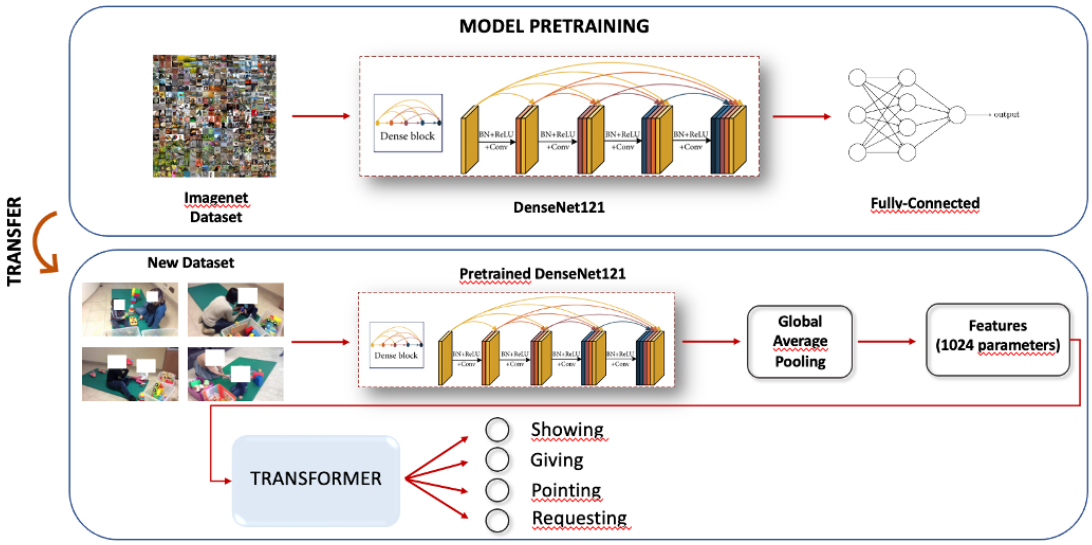


Figure 4.6: Architecture Overview.

80% of available video clips were employed for model training and validation

(validation split was set to 0.15), while the remaining 20% was used for internal testing. In addition, 5 video clips recorded in a different environmental setting were used as external-testing set. Details about training, validation and testing sets are reported in Table 4.3.

Model training and validation were performed for 50 epochs using Adam optimization and sparse categorical cross-entropy as loss function.

All the steps of video pre-processing, feature extraction, training, validation and testing of the model were performed using Python 3 [163] with the following libraries: Tensorflow 2.9.2, Keras 2.9.0, Pandas, Numpy, ImageIO, CV2 and VisualKeras.

Set	**Total**	**Showing**	**Requesting**	**Giving**	**Pointing**
Training + Validation	31	2	12	10	7
Internal Testing	6	1	2	1	2
External Testing	5	1	-	2	2

Table 4.3: Description of Training, Validation and Testing Sets.

4.2.3 Results

Classification performance of training, validation and internal testing are reported in Table 4.4. Results indicate that the model was able to achieve an overall accuracy of 67% in classifying the video clips into four actions of interest.

A detailed analysis of the internal testing results revealed that the model was highly accurate in identifying video clips depicting the gesture of "Requesting" with 100% accuracy (2 out of 2 clips correctly classified), and similarly for the gesture of "Giving" and "Showing" with 100% accuracy (1 out of 1 clip correctly classified for each gesture). However, the model struggled in correctly identifying video clips depicting the gesture of "Pointing", with 0% accuracy (both clips were incorrectly classified as "Giving").

When the model's performance was evaluated on an external testing set, it was found that it was able to accurately classify video clips depicting the gesture of "Pointing" and "Giving" with 100% accuracy (2 out of 2 clips correctly classified

for each gesture). However, the model incorrectly classified a video clip depicting the gesture of "Showing" as "Giving" (0 out of 1 clip correctly classified). There were no video clips depicting the gesture of "Requesting" in the external testing set, hence no evaluation for this gesture was possible. More detailed information on the results of internal and external testing can be found in Table 4.5.

Metric	**Training**	**Validation**	**Internal Testing**
Accuracy	100%	80%	67%
Loss	0.01	2.02	1.17

Table 4.4: Overall model performance.

Set	**Instance**	**True Label**	**Showing**	**Requesting**	**Giving**	**Pointing**
Internal Testing	1	Showing	59.74%	4.59%	34.49%	1.19%
	2	Requesting	7.59%	82.17%	3.45%	6.80%
	3	Requesting	0.57%	72.46%	18.46%	8.51%
	4	Giving	1.20%	13.85%	63.40%	21.55%
	5	Pointing	0.72%	0.71%	97.83%	0.74%
	6	Pointing	10.95%	0.63%	85.32%	3.10%
	Single Class Accuracy		100%	100%	100%	0%
External Testing	1	Showing	6.87%	7.97%	78.27%	6.89%
	2	Giving	18.52%	0.69%	51.52%	29.27%
	3	Giving	12.64%	14.51%	71.79%	1.06%
	4	Pointing	1.15%	10.98%	24.24%	63.62%
	5	Pointing	0.11%	1.02%	4.94%	93.93%
	Single Class Accuracy		0%	-	100%	100%

Table 4.5: Details about Internal and External testing results.

4.2.4 Discussion

The second phase of our research on communicative gestures focused on automating the detection and coding of deictic gestures from naturalistic video recordings. Deictic gestures, such as pointing, showing, requesting, and giving, are critical for establishing joint attention and intentional communication in early development

[205, 206]. To address the time-consuming and labor-intensive nature of manual coding, we developed an AI-based system leveraging a transformer-based DL model. This model was trained on annotated video data to analyze entire clips, thereby capturing the continuity of social interactions.

4.2.4.1 Interpretation of Results

The AI-based system demonstrated the capability to automatically detect deictic gestures from video sequences with internal and external accuracies of 67% and 80%, respectively. By processing entire video clips rather than individual frames, the system was able to capture the flow of dynamic social interactions and the contextual dependencies between gestures, enhancing its ability to correctly interpret actions. However, misclassifications, particularly in the "Giving" category, were observed, which we attribute to imbalances in the training dataset. Despite these challenges, the model's performance underscores the feasibility of using transformer-based deep learning methods for the automated analysis of naturalistic socio-communicative behaviors.

4.2.4.2 Significance and Value

The automated coding model represents a significant advancement over traditional manual coding methods by offering a scalable, efficient, and objective approach to gesture analysis in naturalistic settings. The transformer-based DL framework, with its ability to analyze entire video clips rather than individual frames, captures the continuity of social interactions, enabling more accurate characterization of gestures and actions. This approach shows promise for large-scale neurodevelopmental research, providing a tool for rapid, high-specificity assessments that could be integrated into clinical evaluations and personalized intervention strategies. By reducing the need for manual coding, the system could expand applications in early autism detection and contribute to more precise, neurodiversity-affirming diagnostic practices.

4.2.4.3 Limitations & Future Improvements

The transformer-based model exhibited classification biases, most notably in the "Giving" gesture, due to class imbalances within the training dataset. Refining the training process by incorporating a more balanced and diverse dataset is essential to enhance the model's generalizability and overall performance. Data heterogeneity, arising from variations in video quality, camera angles, and zoom levels, also presented challenges; implementing a standardized video-recording protocol would likely enhance consistency and reliability. Additionally, further improvements in preprocessing techniques and model architecture may reduce misclassifications and improve accuracy. Future work should also consider integrating additional multimodal cues (e.g., combining gesture with concurrent gaze and speech data) to further refine the automated detection of deictic gestures. Finally, validating the system on larger and more heterogeneous samples will be critical for ensuring its applicability in varied clinical and naturalistic contexts.

Chapter 5

Advancing Visual Attention to Social Cues in Preschoolers: Exploring Gaze Patterns Development

Visual attention to social cues is fundamental for developing social learning and communication [218, 219]. Evidence suggests that NT children exhibit a natural preference for socially salient stimuli, such as faces and voices, from early infancy [220, 221, 222]. In contrast, children with ASC often display reduced attentional engagement with these cues [223], which may contribute to later difficulties in social interaction and communication. These differences highlight the importance of attentioning gaze patterns as early markers of NDD and as key targets for early intervention. In this chapter, we try to extend current knowledge on visual attention in preschoolers by leveraging a novel ML-based approach to analyze gaze behavior in dynamic, naturalistic social contexts.

5.1 Decoding Social Attention in Preschoolers: A New Eye-Tracking Paradigm with Markov Chain Analysis

In this study, we examine gaze behavior in children with ASC and NT while they view video stimuli depicting naturalistic social interactions. These videos, designed to reflect real-world social dynamics, include structured tasks such as Sensory Social Routines (SSRs) with songs and object-based play involving musical instruments. The scenarios incorporate key interactive components—mutual imitation, turn-taking, and shared attention—providing ecologically valid contexts for capturing diverse gaze behaviors. Using eye-tracking technology, we precisely measure gaze shifts across social and nonsocial elements. To analyze these patterns, we employ advanced statistical techniques: Continuous-Time Markov Chains (CTMCs) model probabilistic transitions between Areas of Interest (AOIs), while PCA helps identify the most informative patterns of gaze transitions. This methodological approach addresses the limitations of previous research, which often relied on static images or brief stimuli [224, 225, 226], offering a more detailed and dynamic characterization of social attention in preschoolers.

5.1.1 Participants

The sample consisted of 55 preschoolers, aged between 29 and 93 months. Of these, 24 children were clinically diagnosed with ASC, based on the DSM-5 criteria [2]. Expert clinicians performed the diagnostic assessments, using the ADOS-2 [227] as a supporting tool. These assessments took place at the Institute for Biomedical Research and Innovation of the National Research Council of Italy (IRIB-CNR) in Messina. The NT group included 31 children, who were recruited from two mainstream nursery schools in Messina.

The inclusion criteria for the ASC group required the absence of known genetic syndromes (e.g., fragile X syndrome, tuberous sclerosis), inborn metabolic disorders (e.g., aminoaciduria, peroxisomal disorders), epilepsy with uncontrolled seizures, movement disorders, or CP. For the NT group, exclusion criteria included

any clinical diagnosis of neurodevelopmental conditions (e.g., language and/or motor delays) and a family history of autism. The two groups were matched for age, with no significant differences ($p > 0.05$).

All participants had normal or corrected-to-normal vision and no history of auditory impairment. Ethical approval for the study was granted by the Ethics Committee of CNR and informed consent was obtained from all caregivers for their children's participation in the study.

5.1.2 Methods

5.1.2.1 Task procedure

During the eye-tracking experiment, children sat in a small chair within a quiet, controlled environment, positioned 80 cm from a high-resolution 24" widescreen LCD monitor (1024 x 768 pixels). A research team member (S.L.) ensured the children's engagement and promptly repositioned them if they moved outside the trackable range. Gaze patterns were recorded with the SMI iView XTM RED dark-pupil 250Hz eye-tracking system and exported using SMI BeGaze 2.4 software.

Before starting the task, a nine-point calibration grid featuring dynamic targets, like a cartoon cat with meowing sounds, was used. Participants then performed the task, and those with a calibration error exceeding approximately 1.2 visual degrees were excluded from the analysis [228].

As a result, two children (one ASC and one NT) were excluded due to calibration errors. This resulted in a final analyzable sample of 53 children: 23 ASC children and 30 NT children. Table 5.1 summarizes the demographic and clinical characteristics of the final sample.

Variable	**ASC**	**NT**	**Statistic**	**p-value**
N (female, %)	23 (4, 17.4%)	30 (19, 63.3%)	Chisq.	< 0.01
Age (months) (mean ± SD)	56.64 ± 19.77	59.43 ± 8.84	t.test	> 0.05
Total DQ (mean ± SD)	76.53 ± 24.21	120.16 ± 14	t.test	< 0.0001
ADOS 2 – SA (mean ± SD)	9 ± 4.84	-	-	-
ADOS 2 – RRB (mean ± SD)	5.25 ± 2.67	-	-	-
ADOS 2 – Total (mean ± SD)	14.18 ± 7.40	-	-	-

Table 5.1: Demographic and clinical characteristics of the participants.

Participants watched four 25-second videos in random order, each depicting a naturalistic interaction between a child and an adult seated at a small table. These interactions emphasized turn-taking and mutual imitation with four actions performed by each individual per clip. Distractor toys were placed on a shelf behind the table, positioned in the left and right corners of the scene. Two videos involved joint musical activities with instruments (*Drums* and *Xylophones*), and the other two featured SSRs with songs and body gestures (*Sheriff* and *Witches*). In one musical activity (*Drums*) and one song routine (*Sheriff*), the child initiated the action, with the adult imitating after 2–3 seconds; in the others (*Xylophones* and *Witches* respectively), the roles were reversed. The task was carefully balanced across several conditions, including activity type, role counterbalancing, sensory consistency, and distractor placement, to capture varied interactional and sensory contexts within a naturalistic yet systematically varied setting. Each video began with drift correction, followed by a 5-second black screen.

5.1.2.2 Data Pre-Processing

The SMI system was used to define seven distinct AOIs for a detailed analysis of gaze patterns. These AOIs were: (1) Adult Face, (2) Child Face, (3) Adult Activity, (4) Child Activity, (5) Left Distractor Object, and (6) Right Distractor Object. The Adult and Child Face AOIs included the head, neck, and upper shoulders of each actor. The Adult and Child Activity AOIs covered the upper body, arms, hands, and any object being used in the object-based play videos (Figure 5.1). AOIs were manually adjusted each second to account for natural

movement during interactions, ensuring precise data capture. Trials with over 25% track loss were excluded from the eye-tracking analysis, resulting in the omission of 9 out of 212 trials (4.2%) (Supplementary Table A.5). To handle missing data, imputation was performed using the substitution technique with the median value.

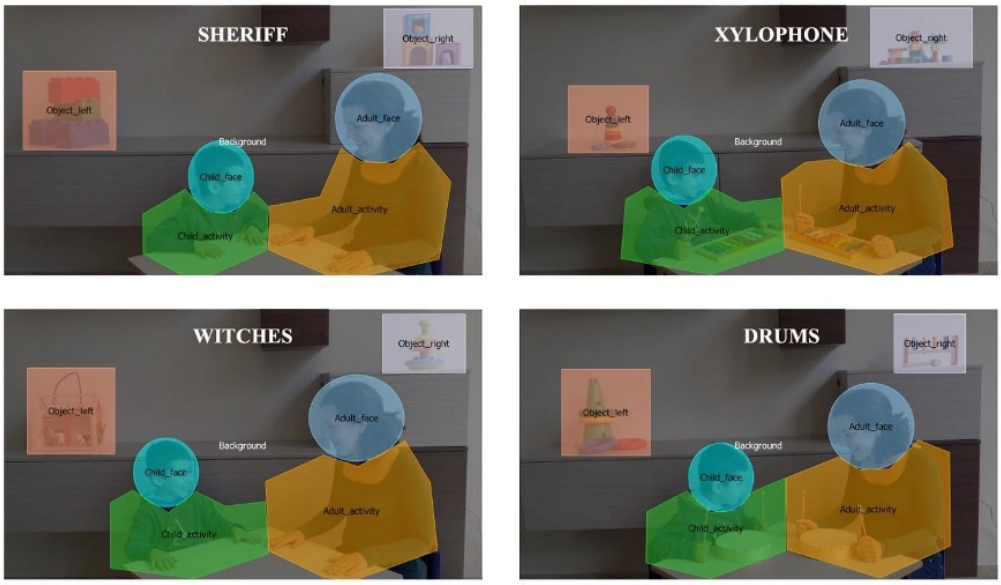


Figure 5.1: Frame from each video clip with overlapped AOIs. The defined regions include: Adult Face, Child Face, Adult Activity, Child Activity, Left Distractor Object and Right Distractor Object.

5.1.2.3 Markov Chain Data Modeling

In this section, we outline the CTMCs used to model gaze patterns within the defined AOIs as exhibited by children while watching the videos. A CTMC is a stochastic process characterized by a state space S, namely a finite or countable set representing all possible states, which in our case correspond to different AOIs, and by the propensities (or transition rates) that describe how quickly transitions occur between states per unit of time. Unlike the more conventional Discrete-Time Markov Chains (DTMC), which evolves in fixed time steps and is defined by transition probabilities, the CTMC framework allows transitions to occur at any continuous point in time. Although our experiment used a digital system with a theoretically fixed sampling interval, which might suggest that a DTMC approach

could suffice, employing CTMC models offers an advantage. Specifically, they can account for uncertainties or slight delays in the data capture process, phenomena that we have indeed observed.

The evolution of a finite time-homogeneous CTMC (with n states) is characterized by a transition matrix Q, whose general element q_{ij} denotes the propensity of a transition from state i to j and q_{ii} represents the opposite of the probability of leaving the state i, which ensures that each column of Q sums to zero.

The dynamics of the probability $P(X_t = i)$ that the CTMC at a general time t is found in the general state i is governed by the following system of n ordinary differential equations:

$$\frac{d}{dt}P(X_t = i) = \sum_{j \neq i} P(X_t = j) \cdot q_{ji} - P(X_t = i) \cdot \sum_{j \neq i} q_{ij} \tag{5.1}$$

The above equation system, called the *Master Equation*, captures the continuous-time evolution of the CTMC probability distribution, where transitions among states are probabilistic and occur with propensities specified by the elements of the transition matrix Q.

In our modeling framework, each AOI is represented as a distinct state within the Markov chain, denoted as X_t at time t, indicating the specific AOI capturing the child's visual attention. This approach offers a sophisticated means to capture and analyze the nuanced transitions in gaze behavior over time, facilitating a detailed examination and prediction of gaze patterns across the defined AOIs. For each participant and each of the four trials, we computed a 6-state CTMC ($n = 6$) using the following AOIs: Child Face, Adult Face, Child Activity, Adult Activity, Left Distractor Object, and Right Distractor Object. As a result, for each child and video stimulus, we obtained a 6×6 transition matrix, from which $n \cdot (n-1) = 30$ transition propensities were estimated for each trial, following the procedure outlined below.

According to the theory of continuous-time Markov processes, the expected waiting time in a state j before transitioning to state i is an exponential random variable with the mean equal to the inverse of the propensity from j to i. Therefore, for each child and trial, we estimated the general element q_{ij} of the transition matrix Q as the statistical mean of the observed waiting times from state j to

state i. In cases where a state j was not visited for a given participant and trial, the corresponding column j of matrix Q would be a zero vector. This would lead to undesirable properties of the underlying behavior, particularly the non-uniqueness of the CTMC equilibrium distribution. To prevent such issues, when a state was not observed, we imputed the transition propensities from the non-visited state j by calculating the median value of the corresponding transition propensities from state j across all other subjects within the same group and trial. As a result, the unique equilibrium probabilities from the Master Equation were computed for each individual and trial, and these probabilities were analyzed as described in the next paragraphs.

5.1.2.4 PCA

To investigate whether NT gaze patterns show preferences for specific areas of the video stimuli, we analyzed the equilibrium probabilities for the six AOIs across the four trials in the NT group. Specifically, we compared the median values of the equilibrium probabilities for each AOI, using the non-parametric Kruskal-Wallis test. For further examination of significant differences, pairwise comparisons were conducted using the Dwass-Steel-Critchlow-Fligner method. The analysis revealed that SSRs drew NT children's attention to faces, while musical activities with instruments directed their gaze to the activity areas. Specifically, in the *Sheriff* video stimulus, the Child Face AOI garnered significantly more attention than other areas, whereas in the *Witches* video, the Adult Face AOI was the primary focus. This pattern aligns with our expectations, as the *Sheriff* song is initiated by the child, and the *Witches* song by the adult. In contrast, during the *Drums* and *Xylophone* videos, the activity areas for both the child and adult were significantly more attended to than the facial areas (see Supplementary Tables A.6 and A.7). This result is consistent with our expectations, as object-based activities naturally draw attention to the materials and the actions involved, directing gaze toward the activity areas. Building on these observations, subsequent analyses were aimed at investigating gaze transitions in relation to the facial AOIs during the SSRs trials, and the activity-related AOIs during the musical instrument trials.

To deepen our understanding of these patterns, we conducted a PCA, a stan-

dard technique for dimensionality reduction. PCA identifies the most important features in a dataset and projects them into a lower-dimensional space, while retaining as much of the original information as possible. In our analysis, we focused on the transition propensities to and from the AOIs associated with faces for the SSRs trials (*Sheriff* and *Witches*), and the transition propensities to and from the AOIs related to activities for the trials involving musical instruments (*Drums* and *Xylophone*).

This approach resulted in 18 distinct features for each trial, totaling 36 features for each stimulus group (SSRs and musical activities with instruments). To further explore the data, we visually represented the PCA outcomes for both video groups by creating a scatter plot that displayed NT children and ASC children based on the first two principal components. We used the PCA to simplify the high-dimensional dataset, allowing us to uncover underlying patterns and explore the distinctive gaze scanning behaviors between NT children and ASC children.

In addition to the scatter plot, we included a correlation plot, depicted with arrows, to illustrate the relationship between the input features and the principal components. Features with higher correlations to PC1 and PC2 were considered the most significant in explaining the variability within the dataset. We categorized the features into five groups and color-coded them based on their type. For the SSRs trials, these groups included:

1. Transition propensities between the Child Face and the Adult Face (and vice versa), shown in green.

2. Transition propensities from the Adult Face to other areas excluding the Child Face, shown in blue.

3. Transition propensities from the Child Face to other areas excluding the Adult Face, shown in red.

4. Transition propensities towards the Adult Face from other areas excluding the Child Face, shown in pink.

5. Transition propensities towards the Child Face from other areas excluding the Adult Face, shown in gray.

The same approach was applied to the activity areas in the musical instrument trials. Details about the feature groups are provided in the Supplementary Table A.8.

5.1.2.5 Data Visualization with Chord Diagrams

We included chord diagrams to visually represent the complex patterns of gaze transitions, providing a clearer and more intuitive understanding of the differences in gaze behavior between the ASC and NT groups. Transition propensities between AOIs are depicted through a simplified arrow-based design. Each AOI is represented as a segment along the circumference of the diagram, with directed transitions illustrated by curved arrows connecting the segments. The thickness of the arrows reflects the strength of the transition, enabling an immediate visual comparison of gaze shifts between AOIs.

The arrow's direction indicates the direction of the gaze shift, moving from the starting AOI to the target AOI. The color of the arrow corresponds to the starting AOI and matches the color of the outer circle segment representing that AOI. At the base of the arrow, a colored segment represents the target AOI and matches the color of the corresponding inner circle segment. This design helps to visually distinguish the origin and destination of each transition.

We generated these chord diagrams for each group (ASC and NT) and for each trial (*Sheriff*, *Witches*, *Drums*, and *Xylophone*). Self-transitions (i.e., transitions where the gaze remains within the same AOI) were excluded to reduce visual clutter, as their high propensity could dominate the visualization. Additionally, we applied a scaled version of the chord diagram, in which all AOI segments on the circumference were set to have equal size. Within each segment, however, the arrows representing transitions were proportionally scaled to reflect the fraction of interactions directed toward other AOIs. This scaling approach normalized the sector sizes while preserving the relative strength of transitions, providing a clear, intuitive comparison of gaze transition patterns across groups and trials.

5.1.3 Results

Using CTMCs to model gaze transition patterns within defined AOIs, we observe distinct group differences in the SSRs with songs and musical activities with instruments. For SSRs, NT children display significantly higher transition propensities between face-related AOIs (Adult Face and Child Face) when compared to ASC children, as shown by green arrows in the PCA visualization. This greater propensity for gaze shifts between faces is also evident in the chord diagrams, where the blue arrow from Adult Face to Child Face is larger for NT children in the *Sheriff* trial, and the green arrow from Child Face to Adult Face is more prominent for NT children in the *Witches* trial. NT children also show a significantly higher propensity to shift attention from distractor objects or the activity areas toward faces, as indicated by the pink and gray arrows in the PCA visualization. This behavior is further confirmed by the chord diagrams, where the light gray arrow from Object Left to Child Face, the dark gray arrow from Object Right to Adult Face, and the red arrow from Adult Activity to Adult Face are all larger in NT children compared to the ASC group. These patterns indicate a stronger inclination to maintain attention on social elements, as seen in the frequent gaze shifts between faces, and to reorient attention toward faces when initially directed toward non-social elements during visual exploration. In contrast, ASC children exhibited a higher propensity to avert their gaze from faces, shifting attention instead to non-social elements in the scene, such as distractor objects and activity areas, as suggested by the red and blue arrows in the PCA visualization (Figure 5.2). This tendency is also clearly reflected in the chord diagrams, where the green arrows from Child Face to Child Activity and from Child Face to Object Left, as well as the blue arrows from Adult Face to Adult Activity and from Adult Face to Object Right, are larger in the ASC group (Figures 5.3 and 5.4). Table 5.2 presents the descriptive statistics and results of the Mann-Whitney U test for between-group comparisons of transition propensities across the selected AOIs during the SSRs trials.

During musical activities with instruments, NT children, compared to ASC children, display a gaze pattern that redirects attention to the activity areas, either by triangulating their gaze from the face of a social partner to the activity area

of the other social partner (e.g., from the Adult Face to the Child's Activity AOI in the *Drums* trial) or by shifting focus from Distractor Objects (Left or Right) back to the Activity areas. This pattern is also reflected in the chord diagrams, particularly in the green arrow from Child Face to Adult Activity, which is more prominent for NT children, indicating their higher tendency to engage in gaze triangulation. Additionally, the dark gray arrow from Object Right to Adult Face and the light gray arrow from Object Left to Child Face are larger for NT children, confirming their greater propensity to reorient their gaze from distractor objects back to faces. In contrast, ASC children show significantly higher gaze transition propensities between the Activity-related AOIs (Adult Activity and Child Activity) compared to NT children, as well as from Activity-related AOIs to distractor objects (specifically, from the Adult Activity AOI to the Object on the Right). This behavior is clearly visible in the chord diagrams, where the red arrow from Adult Activity to Child Activity and the yellow arrow from Child Activity to Adult Activity are both larger in the *Drums* trial, indicating stronger transitions between activity areas. ASC children also show a higher propensity to shift their gaze from activity-related AOIs to distractor objects, as reflected in the chord diagrams by the red arrow from Adult Activity to Object Right and the yellow arrow from Child Activity to Object Left, which are more prominent in this group. Additionally, ASC children tend to shift gaze from the face to the Activity-related AOI within the same individual, without directing attention to the other social partner in the scene, thus showing less of the gaze triangulation pattern observed in NT children (Figure 5.5).

Furthermore, the PCA visualization indicates that gaze shifts in the ASC group are more diffusely distributed along the main dimensions, suggesting greater variability in gaze-shifting behavior among children in this group and highlighting increased heterogeneity in attentional patterns within the ASC group. This behavior is confirmed by the chord diagrams, where the green arrow from Child Face to Child Activity and the blue arrow from Adult Face to Adult Activity are significantly larger in the ASC group (Figure 5.6 and 5.7). Table 5.3 presents the descriptive statistics and results of the Mann-Whitney U test for between-group comparisons of transition propensities across the selected AOIs during the musical activities with instrument trials.

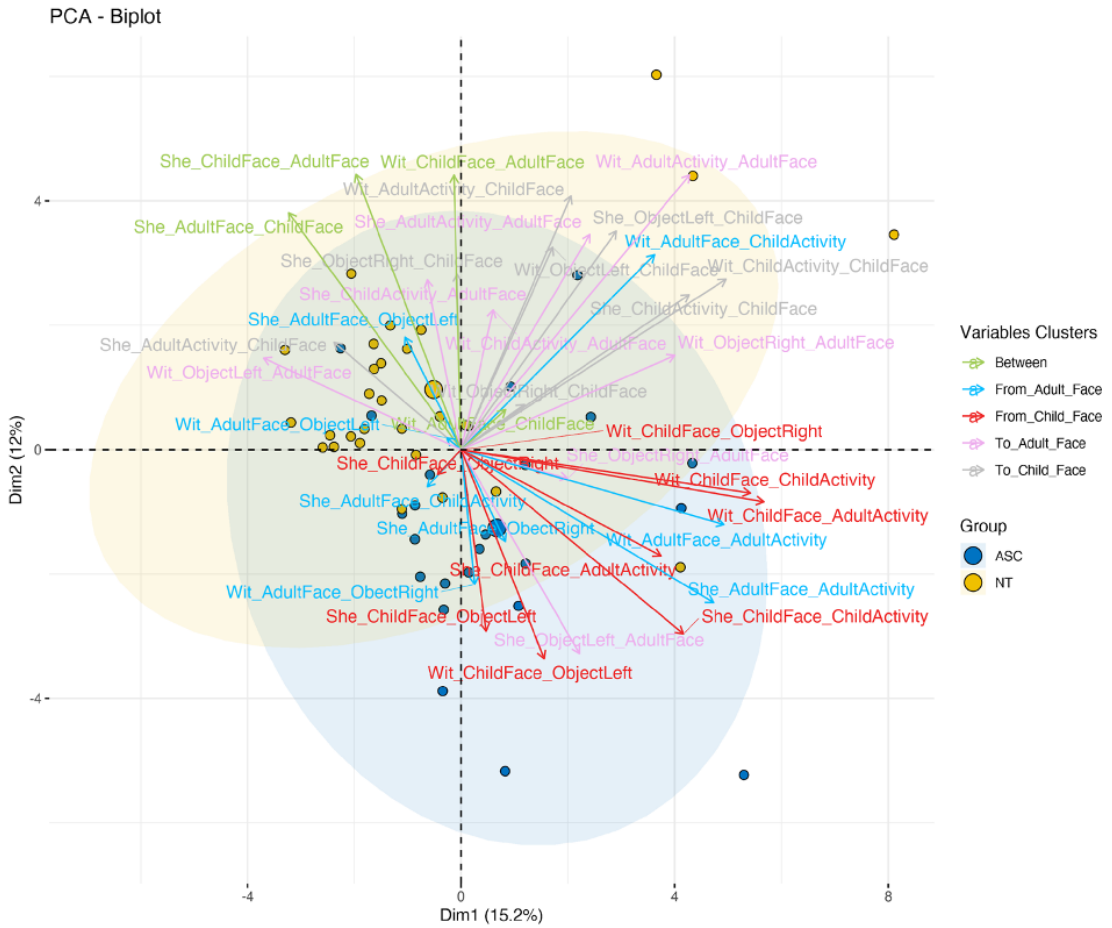
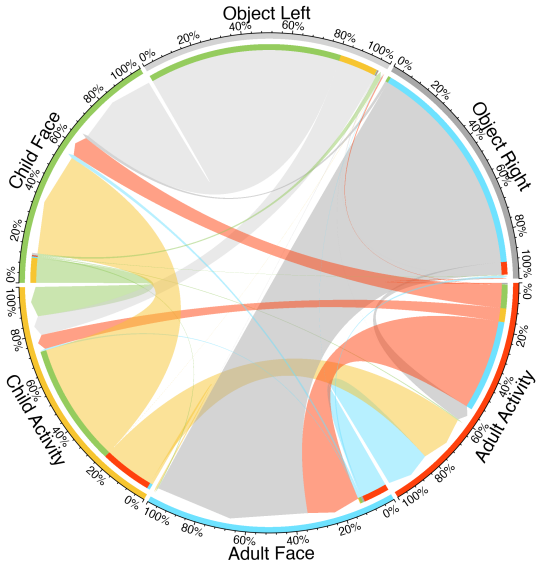
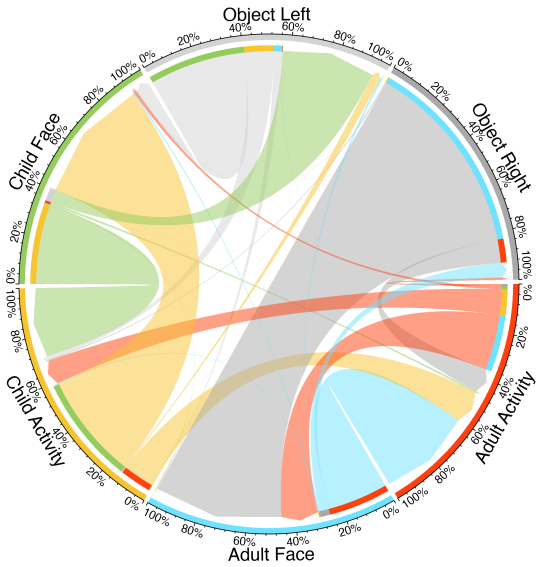


Figure 5.2: Scatter plot showing individuals based on the first and second principal components, with concentration ellipses around each group (ASC and NT). Overlaid, a correlation plot displaying the correlation between the first and second principal components, with variables (transition propensities) color-coded based on the categorization. Stimulus videos: *Sheriff* and *Witches*.

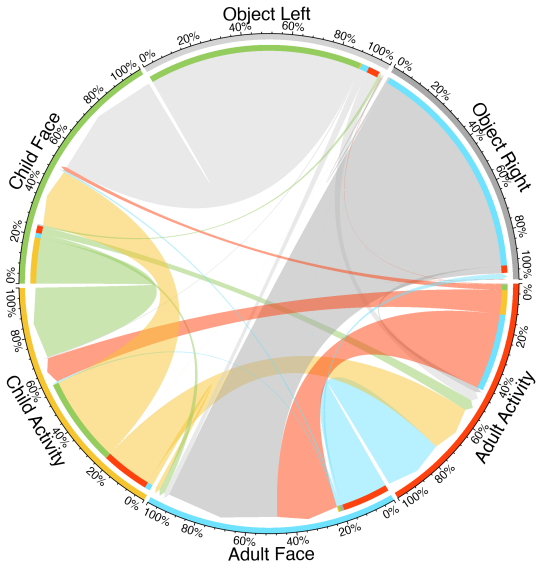


(a) Sheriff - NT Group

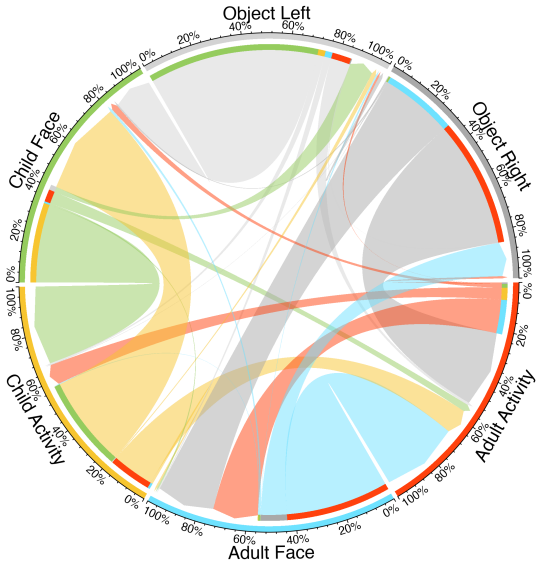


(b) Sheriff - ASC Group

Figure 5.3: Chord diagrams representing gaze transition propensities between areas of interest (AOIs) during the "Sheriff" trial. The diagram on the top illustrates transition patterns for the NT group, while the diagram on the bottom shows those for the ASC group. AOIs are color-coded as follows: Child Face (green), Child Activity (yellow), Adult Face (blue), Adult Activity (red), Object Right (dark gray), and Object Left (light gray). Arrow thickness indicates the strength of the transitions between AOIs.



(a) Witches - NT Group



(b) Witches - ASC Group

Figure 5.4: Chord diagrams representing gaze transition propensities between areas of interest (AOIs) during the "Witches" trial. The diagram on the top illustrates transition patterns for the NT group, while the diagram on the bottom shows those for the ASC group. AOIs are color-coded as follows: Child Face (green), Child Activity (yellow), Adult Face (blue), Adult Activity (red), Object Right (dark gray), and Object Left (light gray). Arrow thickness indicates the strength of the transitions between AOIs.

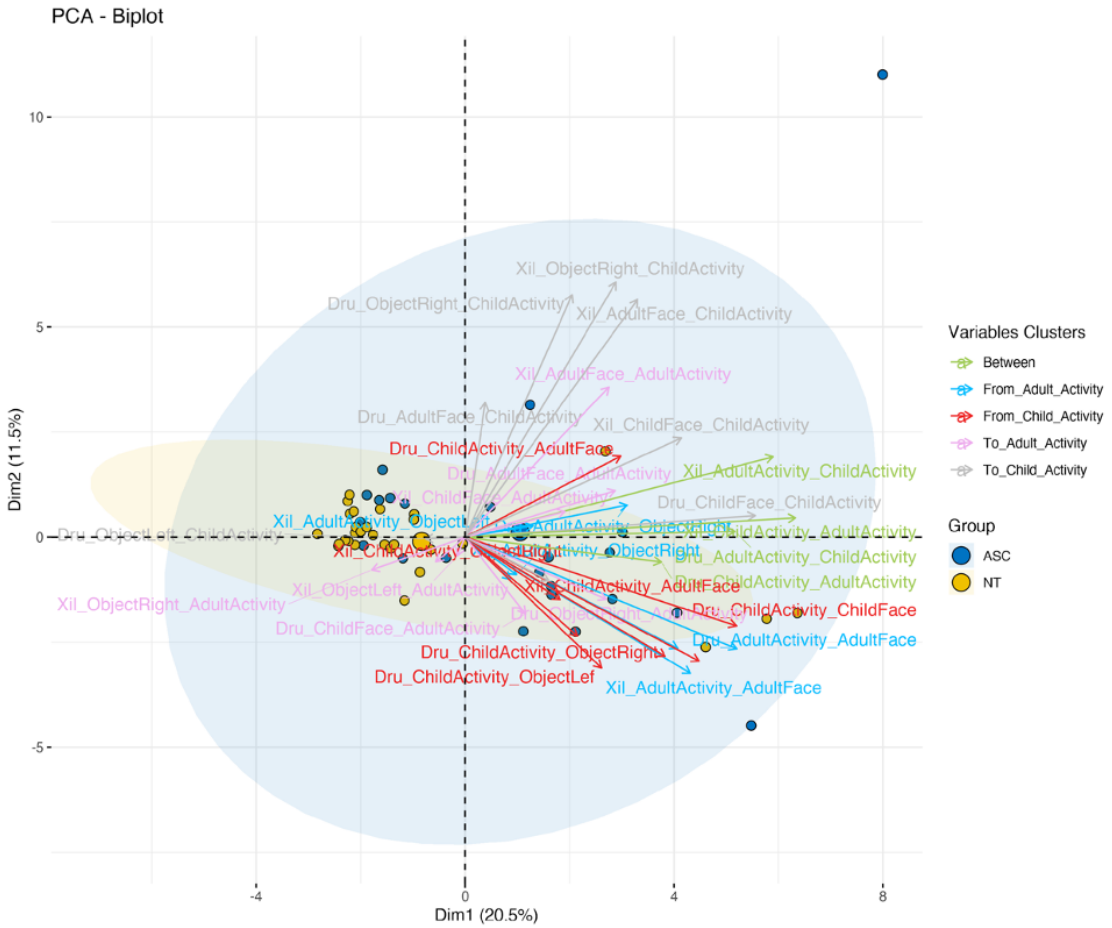
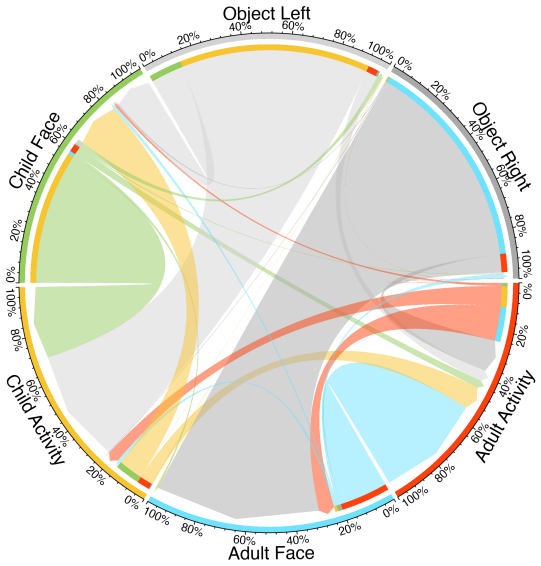
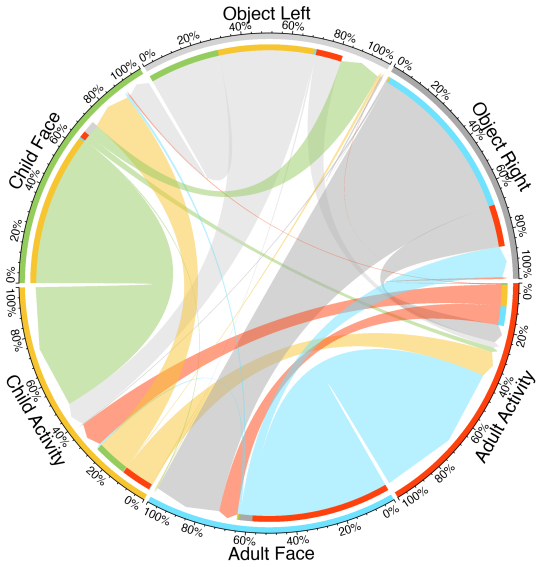


Figure 5.5: Scatter plot showing individuals based on the first and second principal components, with concentration ellipses around each group (ASC and NT). Overlaid, a correlation plot displaying the correlation between the first and second principal components, with variables (transition propensities) color-coded based on the categorization. Stimulus videos: *Xylophone* and *Drums*.

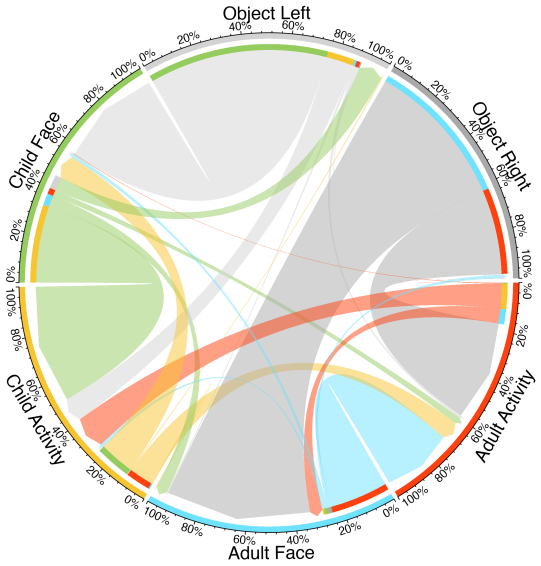


(a) Drums - NT Group

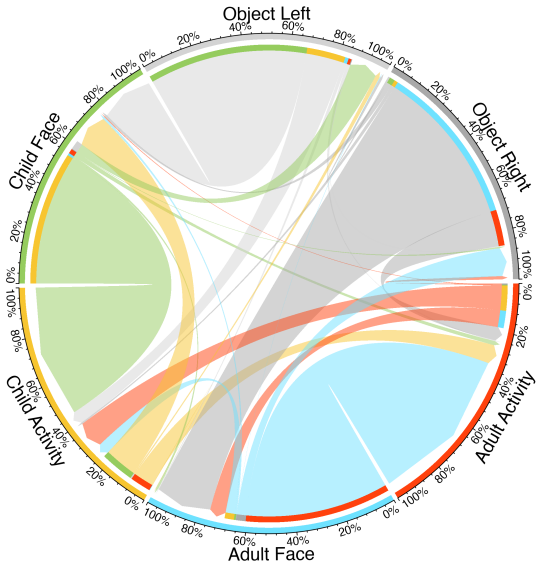


(b) Drums - ASC Group

Figure 5.6: Chord diagrams representing gaze transition propensities between areas of interest (AOIs) during the "Drums" trial. The diagram on the top illustrates transition patterns for the NT group, while the diagram on the bottom shows those for the ASC group. AOIs are color-coded as follows: Child Face (green), Child Activity (yellow), Adult Face (blue), Adult Activity (red), Object Right (dark gray), and Object Left (light gray). Arrow thickness indicates the strength of the transitions between AOIs.



(a) Xylophone - NT Group



(b) Xylophone - ASC Group

Figure 5.7: Chord diagrams representing gaze transition propensities between areas of interest (AOIs) during the "Xylophone" trial. The diagram on the top illustrates transition patterns for the NT group, while the diagram on the bottom shows those for the ASC group. AOIs are color-coded as follows: Child Face (green), Child Activity (yellow), Adult Face (blue), Adult Activity (red), Object Right (dark gray), and Object Left (light gray). Arrow thickness indicates the strength of the transitions between AOIs.

Table 5.2: Descriptive Statistics of Transition Propensities Across Defined AOIs in ASC and NT Children and Mann-Whitney Test Results Comparing the Two Groups in Sensory Social Routine Trials.

Trial	Transition	Group	N	Mean	Median	SD	Minimum	Maximum	Mann-Whitney U	Mann-Whitney p
Sheriff	AdultActivity - AdultFace	ASC	23	0.00738	0.00551	0.00553	0.00	0.01706	306	0.490
		NT	30	0.01021	0.00784	0.00964	0.00	0.03547		
Sheriff	AdultActivity - ChildFace	ASC	23	3.47e-4	0.00000	4.89e-4	0.00	0.00125	172	0.001
		NT	30	0.00263	0.00114	0.00446	0.00	0.02000		
Sheriff	AdultFace - AdultActivity	ASC	23	0.01262	0.01262	0.01185	0.00	0.04343	130	< 0.001
		NT	30	0.00499	7.12e-4	0.01549	0.00	0.08408		
Sheriff	AdultFace - ChildActivity	ASC	23	1.45e-4	0.00000	2.82e-4	0.00	8.55e-4	292	0.279
		NT	30	1.69e-4	0.00000	2.95e-4	0.00	0.00145		
Sheriff	AdultFace - ChildFace	ASC	23	1.60e-4	0.00000	2.73e-4	0.00	9.03e-4	150	< 0.001
		NT	30	4.83e-4	3.69e-4	3.61e-4	0.00	0.00116		
Sheriff	AdultFace - ObjectLeft	ASC	23	3.29e-5	0.00000	1.50e-4	0.00	7.20e-4	341	0.887
		NT	30	2.86e-5	0.00000	8.85e-5	0.00	3.50e-4		
Sheriff	AdultFace - ObjectRight	ASC	23	0.00178	0.00000	0.00393	0.00	0.01637	256	0.051
		NT	30	3.30e-4	0.00000	8.20e-4	0.00	0.00279		
Sheriff	ChildActivity - AdultFace	ASC	23	1.24e-4	0.00000	2.25e-4	0.00	8.21e-4	301	0.363
		NT	30	3.06e-4	0.00000	5.99e-4	0.00	0.00247		
Sheriff	ChildActivity - ChildFace	ASC	23	0.01428	0.00525	0.01688	0.00	0.06592	332	0.816
		NT	30	0.01388	0.00563	0.01331	0.00	0.04103		
Sheriff	ChildFace - AdultActivity	ASC	23	2.74e-4	2.27e-4	2.70e-4	0.00	0.00104	227	0.032
		NT	30	1.29e-4	9.44e-5	1.42e-4	0.00	5.11e-4		
Sheriff	ChildFace - AdultFace	ASC	23	7.71e-5	0.00000	1.44e-4	0.00	4.73e-4	165	< 0.001
		NT	30	1.82e-4	1.70e-4	1.20e-4	0.00	4.99e-4		
Sheriff	ChildFace - ChildActivity	ASC	23	0.00964	0.00800	0.01027	6.99e-4	0.04773	127	< 0.001
		NT	30	0.00311	3.76e-4	0.00583	0.00	0.02607		
Sheriff	ChildFace - ObjectLeft	ASC	23	0.00148	1.48e-4	0.00493	0.00	0.02387	223	0.024
		NT	30	2.44e-4	3.77e-5	6.63e-4	0.00	0.00357		
Sheriff	ChildFace - ObjectRight	ASC	23	5.23e-6	0.00000	2.39e-5	0.00	1.15e-4	328	0.446
		NT	30	5.94e-5	0.00000	3.26e-4	0.00	0.00178		
Sheriff	ObjectLeft - AdultFace	ASC	23	1.30e-4	0.00000	3.08e-4	0.00	0.00146	342	0.961
		NT	30	4.01e-5	4.01e-5	7.58e-5	0.00	3.70e-4		
Sheriff	ObjectLeft - ChildFace	ASC	23	0.00293	6.12e-4	0.00820	0.00	0.04000	292	0.338
		NT	30	0.01295	0.00282	0.04000	0.00	0.22222		
Sheriff	ObjectRight - AdultFace	ASC	23	0.00603	0.00386	0.00966	0.00	0.04429	211	0.014
		NT	30	0.01012	0.01010	0.01712	5.52e-4	0.09823		
Sheriff	ObjectRight - ChildFace	ASC	23	1.11e-4	0.00000	2.84e-4	0.00	0.00137	210	< 0.001
		NT	30	0.00000	0.00000	0.00000	0.00	0.00000		
Witches	AdultActivity - AdultFace	ASC	23	0.00467	0.00254	0.00508	1.02e-4	0.01733	287	0.302

Continued on next page...

Table 5.2: (Continued)

Trial	Transition	Group	N	Mean	Median	SD	Minimum	Maximum	Mann-Whitney U	Mann-Whitney p
		NT	30	0.00562	0.00305	0.00631	4.94e-4	0.02915		
Witches	AdultActivity - ChildFace	ASC	23	5.07e-4	3.31e-4	6.71e-4	0.00	0.00294	266	0.156
		NT	30	3.92e-4	2.43e-4	6.06e-4	0.00	0.00288		
Witches	AdultFace - AdultActivity	ASC	23	0.01054	0.00389	0.01529	0.00	0.06339	247	0.080
		NT	30	0.00438	0.00135	0.00858	3.01e-4	0.04134		
Witches	AdultFace - ChildActivity	ASC	23	6.44e-5	0.00000	1.09e-4	0.00	4.30e-4	295	0.331
		NT	30	1.21e-4	6.53e-5	2.24e-4	0.00	9.91e-4		
Witches	AdultFace - ChildFace	ASC	23	2.08e-4	1.05e-4	2.79e-4	0.00	9.78e-4	332	0.812
		NT	30	1.88e-4	1.31e-4	2.09e-4	0.00	8.99e-4		
Witches	AdultFace - ObjectLeft	ASC	23	2.89e-5	0.00000	1.10e-4	0.00	5.13e-4	344	0.972
		NT	30	8.38e-6	0.00000	2.60e-5	0.00	1.03e-4		
Witches	AdultFace - ObjectRight	ASC	23	0.00256	3.29e-4	0.00828	0.00	0.04000	156	< 0.001
		NT	30	1.36e-4	0.00000	3.55e-4	0.00	0.00178		
Witches	ChildActivity - AdultFace	ASC	23	1.75e-4	0.00000	3.87e-4	0.00	0.00132	290	0.229
		NT	30	3.82e-4	0.00000	6.92e-4	0.00	0.00319		
Witches	ChildActivity - ChildFace	ASC	23	0.00775	0.00537	0.00850	2.25e-4	0.03283	328	0.767
		NT	30	0.00617	0.00389	0.00565	0.00	0.02137		
Witches	ChildFace - AdultActivity	ASC	23	0.00103	8.33e-4	8.85e-4	1.14e-4	0.00365	237	0.054
		NT	30	7.52e-4	5.85e-4	8.07e-4	0.00	0.00352		
Witches	ChildFace - AdultFace	ASC	23	1.37e-4	0.00000	1.99e-4	0.00	6.13e-4	206	0.009
		NT	30	5.11e-4	2.85e-4	5.91e-4	0.00	0.00206		
Witches	ChildFace - ChildActivity	ASC	23	0.00645	0.00366	0.00645	4.66e-4	0.02190	260	0.129
		NT	30	0.00513	0.00224	0.00657	0.00	0.02709		
Witches	ChildFace - ObjectLeft	ASC	23	4.32e-4	3.12e-4	4.62e-4	0.00	0.00195	211	0.014
		NT	30	1.92e-4	8.28e-5	2.58e-4	0.00	0.00106		
Witches	ObjectLeft - AdultFace	ASC	23	1.16e-4	0.00000	1.94e-4	0.00	7.29e-4	205	0.008
		NT	30	4.23e-4	4.69e-4	4.47e-4	0.00	0.00168		
Witches	ObjectLeft - ChildFace	ASC	23	0.00293	6.12e-4	0.00820	0.00	0.04000	292	0.338
		NT	30	0.01295	0.00282	0.04000	0.00	0.22222		
Witches	ObjectRight - AdultFace	ASC	23	0.00603	0.00386	0.00966	0.00	0.04429	211	0.014
		NT	30	0.01012	0.01010	0.01712	5.52e-4	0.09823		
Witches	ObjectRight - ChildFace	ASC	23	1.11e-4	0.00000	2.84e-4	0.00	0.00137	210	< 0.001
		NT	30	0.00000	0.00000	0.00000	0.00	0.00000		

Table 5.3: Descriptive Statistics of Transition Propensities Across Defined AOIs in ASC and NT Children and Mann-Whitney Test Results Comparing the two groups in Musical Activities Trials.

Trial	Transition	Group	N	Mean	Median	SD	Minimum	Maximum	Mann-Whitney U	Mann-Whitney p
Drums	AdultActivity - AdultFace	ASC	23	0.00212	0.00193	0.00197	0.00000	0.00790	293	0.461
		NT	29	0.00179	9.86e-4	0.00209	2.88e-4	0.01027		
Drums	AdultActivity - ChildActivity	ASC	23	0.00234	0.00135	0.00239	2.56e-4	0.00864	181	0.004
		NT	29	0.00109	6.38e-4	0.00121	1.12e-4	0.00524		
Drums	AdultActivity - ChildFace	ASC	23	9.22e-5	0.00000	1.25e-4	0.00000	3.64e-4	333	0.992
		NT	29	1.24e-4	0.00000	2.05e-4	0.00000	7.23e-4		
Drums	AdultActivity - ObjectLeft	ASC	23	1.09e-5	0.00000	3.60e-5	0.00000	1.27e-4	329	0.885
		NT	29	1.25e-5	0.00000	3.84e-5	0.00000	1.55e-4		
Drums	AdultActivity - ObjectRight	ASC	23	8.58e-5	0.00000	1.58e-4	0.00000	6.87e-4	238	0.024
		NT	29	3.47e-5	0.00000	1.08e-4	0.00000	5.05e-4		
Drums	AdultFace - AdultActivity	ASC	23	0.01598	0.00942	0.01806	0.00000	0.07020	133	< 0.001
		NT	29	0.00566	0.00127	0.01537	3.03e-4	0.08353		
Drums	AdultFace - ChildActivity	ASC	23	1.69e-4	0.00000	3.08e-4	0.00000	0.00131	229	0.045
		NT	29	3.07e-4	2.25e-4	3.25e-4	0.00000	0.00110		
Drums	ChildActivity - AdultActivity	ASC	23	0.00275	0.00143	0.00364	2.84e-4	0.01735	236	0.074
		NT	29	0.00121	7.11e-4	0.00130	1.95e-4	0.00575		
Drums	ChildActivity - AdultFace	ASC	23	8.73e-5	0.00000	1.36e-4	0.00000	5.43e-4	324	0.850
		NT	29	8.97e-5	0.00000	1.54e-4	0.00000	5.55e-4		
Drums	ChildActivity - ChildFace	ASC	23	0.00307	0.00121	0.00365	0.00000	0.01271	314	0.729
		NT	29	0.00231	0.00107	0.00315	1.43e-4	0.01455		
Drums	ChildActivity - ObjectLeft	ASC	23	6.44e-5	0.00000	9.17e-5	0.00000	2.63e-4	262	0.094
		NT	29	2.77e-5	0.00000	6.37e-5	0.00000	2.30e-4		
Drums	ChildActivity - ObjectRight	ASC	23	5.70e-6	0.00000	2.73e-5	0.00000	1.31e-4	319	0.278
		NT	29	0.00000	0.00000	0.00000	0.00000	0.00000		
Drums	ChildFace - AdultActivity	ASC	23	5.00e-4	0.00000	9.38e-4	0.00000	0.00415	294	0.455
		NT	29	3.94e-4	3.09e-4	4.19e-4	0.00000	0.00169		
Drums	ChildFace - ChildActivity	ASC	23	0.01125	0.00831	0.00910	0.00118	0.03012	198	0.012
		NT	29	0.00674	0.00246	0.00873	3.00e-4	0.03366		
Drums	ObjectLeft - AdultActivity	ASC	23	6.00e-4	2.80e-4	0.00120	0.00000	0.00571	319	0.780
		NT	29	5.70e-4	5.70e-4	7.83e-4	0.00000	0.00386		
Drums	ObjectLeft - ChildActivity	ASC	23	0.00219	0.00174	0.00286	0.00000	0.01268	197	0.010
		NT	29	0.01060	0.01060	0.02043	0.00000	0.11364		
Drums	ObjectRight - AdultActivity	ASC	23	0.00194	0.00194	0.00213	0.00000	0.00706	282	0.327
		NT	29	0.00165	0.00165	0.00228	0.00000	0.01260		
Drums	ObjectRight - ChildActivity	ASC	23	7.86e-5	0.00000	2.05e-4	0.00000	9.69e-4	183	0.003
		NT	29	8.27e-5	8.27e-5	1.41e-4	0.00000	7.85e-4		
Xylophone	AdultActivity - AdultFace	ASC	23	0.00183	8.26e-4	0.00200	0.00000	0.00567	295	0.484

Continued on next page...

Table 5.3: (Continued)

Trial	Transition	Group	N	Mean	Median	SD	Minimum	Maximum	Mann-Whitney U	Mann-Whitney p
		NT	29	9.75e-4	7.25e-4	9.42e-4	1.13e-4	0.00356		
Xylophone	AdultActivity - ChildActivity	ASC	23	0.00267	0.00195	0.00247	0.00000	0.00988	209	0.021
		NT	29	0.00165	0.00102	0.00194	4.21e-4	0.00846		
Xylophone	AdultActivity - ChildFace	ASC	23	8.62e-5	0.00000	1.61e-4	0.00000	7.08e-4	299	0.461
		NT	29	5.07e-5	0.00000	9.46e-5	0.00000	4.21e-4		
Xylophone	AdultActivity - ObjectLeft	ASC	23	0.00000	0.00000	0.00000	0.00000	0.00000		
		NT	29	0.00000	0.00000	0.00000	0.00000	0.00000		
Xylophone	AdultActivity - ObjectRight	ASC	23	1.31e-4	0.00000	4.13e-4	0.00000	0.00197	260	0.023
		NT	29	1.28e-5	0.00000	6.89e-5	0.00000	3.71e-4		
Xylophone	AdultFace - AdultActivity	ASC	23	0.01690	0.00653	0.02623	0.00000	0.11905	149	< 0.001
		NT	29	0.00461	0.00175	0.00836	4.96e-4	0.04150		
Xylophone	AdultFace - ChildActivity	ASC	23	0.00106	7.48e-5	0.00317	0.00000	0.01534	258	0.122
		NT	29	1.60e-4	0.00000	2.59e-4	0.00000	0.00103		
Xylophone	ChildActivity - AdultActivity	ASC	23	0.00202	0.00153	0.00159	2.69e-4	0.00572	212	0.025
		NT	29	0.00122	6.83e-4	0.00142	0.00000	0.00559		
Xylophone	ChildActivity - AdultFace	ASC	23	4.42e-5	0.00000	7.49e-5	0.00000	2.69e-4	303	0.515
		NT	29	7.45e-5	0.00000	1.26e-4	0.00000	4.97e-4		
Xylophone	ChildActivity - ChildFace	ASC	23	0.00324	0.00186	0.00331	0.00000	0.01139	253	0.141
		NT	29	0.00177	0.00144	0.00140	1.05e-4	0.00551		
Xylophone	ChildActivity - ObjectLeft	ASC	23	1.34e-4	0.00000	2.79e-4	0.00000	9.99e-4	298	0.420
		NT	29	4.92e-5	0.00000	9.69e-5	0.00000	3.50e-4		
Xylophone	ChildActivity - ObjectRight	ASC	23	0.00000	0.00000	0.00000	0.00000	0.00000		
		NT	29	0.00000	0.00000	0.00000	0.00000	0.00000		
Xylophone	ChildFace - AdultActivity	ASC	23	4.50e-4	3.79e-4	4.62e-4	0.00000	0.00128	332	0.985
		NT	29	4.60e-4	3.07e-4	5.21e-4	0.00000	0.00168		
Xylophone	ChildFace - ChildActivity	ASC	23	0.01298	0.00991	0.01355	5.19e-4	0.06497	169	0.002
		NT	29	0.00592	0.00263	0.00735	0.00000	0.02696		
Xylophone	ObjectLeft - AdultActivity	ASC	23	8.65e-5	0.00000	1.83e-4	0.00000	6.88e-4	254	0.107
		NT	29	1.58e-4	1.58e-4	2.78e-4	0.00000	0.00124		
Xylophone	ObjectLeft - ChildActivity	ASC	23	0.00122	5.80e-4	0.00240	0.00000	0.01112	328	0.923
		NT	29	0.00120	0.00120	0.00257	0.00000	0.01384		
Xylophone	ObjectRight - AdultActivity	ASC	23	0.00147	0.00144	0.00272	0.00000	0.01355	136	< 0.001
		NT	29	0.00652	0.00652	0.00849	0.00000	0.04843		
Xylophone	ObjectRight - ChildActivity	ASC	23	1.46e-4	7.46e-5	3.66e-4	0.00000	0.00179	160	< 0.001
		NT	29	0.00000	0.00000	0.00000	0.00000	0.00000		

5.1.4 Discussion

This part of the research project was dedicated to deepening our understanding of social attention development by exploring gaze behavior in children with ASC compared to NT peers. Focusing on dynamic, interactive tasks that simulate real-world social engagement, we employed videos of child–adult interactions involving singing and instrument playing. These naturalistic stimuli, observed by children wearing eye trackers, enabled us to capture detailed gaze patterns and move beyond traditional paradigms that rely on static images or brief social stimuli. Overall, this approach offers a more comprehensive perspective on how children with ASC process and respond to social information in everyday contexts.

Central to this investigation is the concept of turn-taking imitation, which enhances the ecological validity of the study by replicating the dynamics of everyday social exchanges. The tasks encompassed both dyadic interactions, such as SSRs involving motor imitation during songs, and triadic interactions that included shared engagement and turn-taking with musical instruments. This design allowed for an in-depth exploration of gaze behaviors across varying social contexts.

The study employs CTMCs to model gaze transitions and PCA to distill key patterns from the data. These methodological innovations have revealed valuable insights into the temporal dynamics of social attention and highlighted notable differences in gaze behaviors between children with ASC and their NT peers.

5.1.4.1 Interpretation of Results

Our study identified context-dependent gaze patterns during the observation of both dyadic and triadic interactions, revealing significant differences in attentional dynamics between NT and ASC children. In SSRs, which involved dyadic interactions and motor imitation during songs, NT children demonstrated a clear preference for social engagement, frequently transitioning their gaze among face-related AOIs. They also showed a strong tendency to reorient their gaze from distractor objects or activity areas back to faces, highlighting the prioritization of social elements during visual exploration. This behavior aligns with previous findings indicating that NT children naturally focus on socially salient stimuli, such as faces, which are critical for reciprocal communication and social learning [229, 230]. Fre-

quent gaze shifts between faces may facilitate their ability to synchronize and share attention, a foundational skill in social development [231, 232, 233, 234].

In contrast, ASC children exhibited distinct gaze patterns during SSRs, characterized by a higher likelihood of gaze aversion from faces and increased transitions toward nonsocial elements, such as distractor objects or activity areas. They were less likely to reorient their gaze back to faces after being distracted, suggesting difficulties in integrating visual information across multiple social cues. This pattern supports the notion that ASC children may prefer predictable, nonsocial stimuli over dynamic, socially demanding ones [235, 236, 225, 226]. These findings are consistent with previous research indicating that reduced social preference [237, 238, 239] or adaptive responses to sensory overload or social anxiety [142, 240, 241] may contribute to gaze aversion in ASC children.

During the musical instrument trials, which involved triadic interactions characterized by turn-taking, shared engagement with parallel toys, and congruent musical sounds, NT children predominantly focused their visual attention on the activity area. When distracted by peripheral objects, they consistently redirected their gaze back to the activity, demonstrating their ability to maintain shared focus. Furthermore, NT children displayed gaze triangulation, frequently shifting attention among the adult's face, the child's activity area, and other socially salient elements. In contrast, ASC children were more likely to transition their gaze between the activity areas of the two partners and exhibited a stronger tendency to shift gaze away from faces toward a single activity area. These behaviors suggest reduced attention to social engagement cues and difficulties in coordinating attention across multiple social partners. Additionally, ASC children often shifted their gaze from the activity area to peripheral distractor objects, underscoring differences in attentional control and motivation compared to NT children. These results align with previous findings that ASC toddlers are more easily distracted by background objects [242, 243].

Moreover, PCA visualization revealed that gaze shifts in ASC children were more broadly distributed across principal components, reflecting a higher degree of variability in attentional strategies. This heterogeneity is consistent with studies reporting idiosyncratic gaze behaviors in ASC populations [244, 245, 246], and has been associated with reduced comprehension of dynamic scenes at the neural level

[247] as well as a stronger presence of autistic traits [248].

5.1.4.2 Significance and Value

This study enhances autism research by combining dynamic, ecologically valid stimuli with advanced analytical techniques to uncover nuanced gaze behavior patterns in both NT and ASC children. By incorporating both SSRs and object-based shared activities with musical instruments, we were able to explore distinct gaze behaviors in dyadic versus triadic interactions, shedding light on how varying levels of social complexity influence attentional dynamics. The findings underscore the diagnostic importance of gaze aversion as a critical marker for understanding social engagement in ASC children. By employing CTMCs and PCA, the study was able to capture dynamic gaze patterns and reduce data complexity, offering insights that traditional static AOIs methods might overlook.

5.1.4.3 Limitations & Future Improvements

In this study, the sample size, while adequate for initial analyses, limits the generalizability of the findings and restricts the exploration of subgroups within the ASC population. Furthermore, the gender imbalance between the NT and ASC groups may have influenced the results, and future research should aim for more balanced samples. Finally, incorporating longitudinal data would allow for a deeper exploration of how gaze aversion and social attention patterns develop over time, providing further insights into their long-term impact on social and communication outcomes.

Chapter 6

Neurodivergent Trajectories in Middle Childhood: Effects in Expressive Kinematics and Speech

In this chapter, we investigate neurodivergent trajectories in middle childhood by examining two critical domains of expressive behavior: motor kinematics and speech production. The research focuses on how expressive motor patterns, specifically, the dynamic expression of Vitality Forms (VFs), and speech disturbances such as dysarthria can serve as indicators of underlying neurodevelopmental differences. By integrating advanced AI-driven analysis into both kinematic and speech domains, the studies presented here aim to provide objective, scalable frameworks for assessing complex communication challenges in children with NDD. The chapter is divided into two sections: the first explores divergent kinematic profiles during social and non-social tasks, and the second presents the development of a hierarchical model for automated speech assessment.

6.1 Exploring Divergent Kinematics in Children with NDD Across Social and Non-Social Vitality Forms

This section continues the exploration of communicative skills by focusing on how children with ASC express different VFs. Introduced by Daniel Stern, VFs are dynamic qualities of actions that convey affective tones, such as gentleness or rudeness, adding emotional depth to social interactions [249, 250]. Unlike basic emotions, which are brief, involuntary responses to stimuli involving regions like the amygdala, VFs are continuous, intentional adjustments in behavior supported by brain areas such as the dorso-central insula and middle cingulate cortex [251, 252, 253]. For instance, handing an object gently or rudely communicates warmth or irritation, transcending the action's functional goal. In children with ASC, however, these nuanced cues are often challenging to express and interpret, reflecting impairments in kinematic subtleties and intentional communication [254, 255].

Theories such as the Enactive Mind approach further highlight reduced social attention and altered salience processing in ASC, contributing to differences in embodied social cognition [256, 257, 237]. These challenges may also manifest as distinct expressions of VFs in both social and non-social contexts, although current research offers limited insights into this phenomenon.

To address this gap, the study examines (1) the spatiotemporal characteristics of gentle and rude VFs in both ASC and NT children, and (2) the impact of social presence on VFs expression in children with ASC. The experimental design involved ASC and NT children performing two tasks: moving a bottle to a designated spot (non-social context) and handing it to another person (social context), while expressing neutral, gentle, or rude VFs. It is hypothesized that children with ASC will exhibit distinct kinematic profiles for each VF, with social presence significantly influencing their expressions. By exploring VFs in autism through AI-driven kinematic analysis, this section advances our understanding of how motor behavior reflects social communication challenges.

6.1.1 Participants

The study was carried out on a group of children with ASC ($n = 25$) and a group of NT peers ($n = 23$) (range age: 7–13 years). Both ASC and NT children were male, right-handed and with an Intellectual Quotient (IQ) > 70. The two groups were comparable in terms of age and IQ ($p > 0.05$). Exclusion criteria for autistic children included the presence of a neurometabolic or genetic syndrome, epileptic encephalopathies and/or epilepsy, structural malformations of the central nervous system and major movement disorders. Exclusion criteria for NT children included a family history of autism, a personal history of language delay, intellectual disability, or any neurodivergent conditions such as autism, ADHD, motor dyspraxia, dyslexia, anxiety, and so forth. Two ASC children showed outliers kinematic values during the data analysis process and were therefore excluded from the retrospective research.

Autistic children were recruited and tested at the National Research Council of Italy, Institute for Biomedical Research and Innovation (CNR-IRIB) in Messina and Catania (Italy), by referring the study to family associations, and via the research center's website. NT children were recruited via mainstream schools in the local territory. The study was approved by the ethics committee of the AOUP of Palermo (protocol number 09/2021), and all the families that voluntarily participated in the study signed a written informed consent. All methods were performed in accordance with the relevant guidelines and regulations.

All autistic children had received a clinical diagnosis of an ASC by a multidisciplinary team including experienced developmental psychologists and child neuropsychiatrists, supported by gold standard assessments such as the Autism Diagnostic Observation Schedule - 2nd Edition (ADOS-2) [258, 179]. The Test of Emotion Comprehension (TEC) [259] was used to assess the understanding of nine different domains of emotional comprehension in both the ASC and the NT children. Additionally, the Italian translation [260] of the Wechsler Intelligence Scale for Children, 4th Edition (WISC-IV) [261] was employed to measure intellectual ability of children in the autism spectrum. The WISC-IV allowed us to assess the Verbal Comprehension Index (VCI), Visual Spatial Index (VSI), Fluid Reasoning Index (FRI), Working Memory Index (WMI), and the Processing Speed Index

(PSI). In the NT group, the Raven's Standard Progressive Matrices (RPM) [262] were used to estimate non-verbal intelligence and logical thinking. None of the children had a clinical diagnosis of Developmental Coordination Disorder (DCD) or motor dyspraxia. Demographic and clinical characteristics of the sample are reported in Table 6.1.

Characteristic	**ASC children (n = 25)**	**NT children (n = 23)**
Age (years)	9.87 ± 2.2	9.63 ± 2.14
IQ total score	98.4 ± 14.4	91.2 ± 8.5
ADOS-2 SA	9.2 ± 4.1	n.a
ADOS-2 RRB	3 ± 2.2	n.a
ADOS-2 total score	12.3 ± 4	n.a
TEC total score	6 ± 1.7	8.3 ± 0.8

Table 6.1: Demographic and clinical characteristics of the sample reported as Mean ± SD. IQ: intellectual quotient, SA: social affect, RRB: restricted repetitive behaviors, TEC: test of emotion comprehension, n.a: not applicable.

To assess the understanding of the meanings associated with "gentle" and "rude," a specially designed semi-structured survey was administered to each child before commencing the experimental phase. The survey took about 15 minutes to complete. The experimenter presented three types of questions. Initially, the child was prompted to provide definitions for the terms "gentle" and "rude" ("What does the word gentle/rude mean?"). Subsequently, the child was asked to offer examples of both gentle and rude behaviors ("Can you please give me an example of being gentle/rude?"). Following this, the comprehension of the consequences of exhibiting gentle or rude behavior towards others was explored ("What happens if you are gentle/rude to another person?").

After this initial survey, all participants received explicit definitions of "gentle" and "rude," along with their synonyms, aimed at establishing a clear and unambiguous conceptual understanding of the two terms. Subsequently, each child was presented with 12 brief video clips, each lasting a few seconds, depicting actions classified as either gentle or rude (e.g., "kicking a ball in a rude way" or "stirring a soup in a gentle way"). To prevent a possible imitation effect of the VFs observed

in these videos on the experimental task, the presented actions differed from those performed during the experiment (passing an object). Additionally, some actions were performed with a different effector (foot).

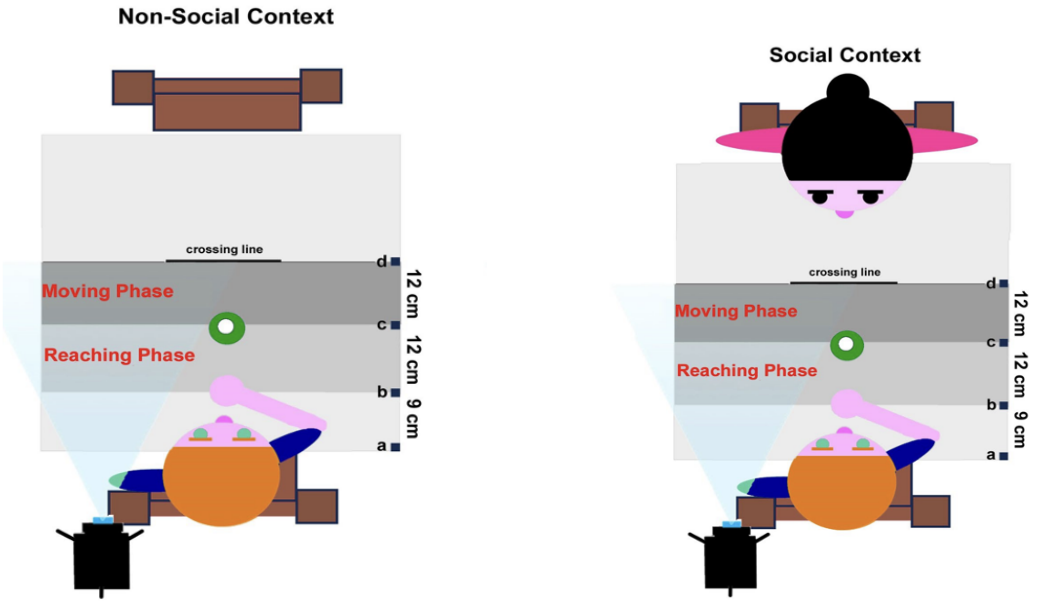
After each scene, participants were asked to verbally identify whether the observed action was gentle or rude, and the accuracy of their responses was recorded by the experimenter. Only participants who achieved a minimum score of 75% correct answers (9 out of 12) were considered eligible to proceed to the subsequent experimental phase. Four children were excluded because they fell below the accuracy ratio and their understanding of the concepts of rude and gentle had not been acquired yet.

6.1.2 Methods

6.1.2.1 Experimental Paradigm

The setup featured a square table with a small bottle positioned on it, accompanied by two chairs, one for the child and the other opposite the child. A high-resolution HC-X920 digital video camera, angled at 30 degrees on a tripod, recorded the table and the child's hands. Another high-resolution camera on a tripod was placed in front of the child and the table, capturing the scene from the frontal perspective. The experimental design comprised two conditions: social and non-social. In the social scenario, a female experimenter (receiver) occupied the chair opposite the child, while an empty chair was placed in front of the child in the non-social scenario. The child was instructed to sit in the chair facing the table, placing the right hand on the table in a pinching position, aligned with the participant's mid-sagittal plane and 9 cm away from the table edge (Figure 6.1, distance a–b). A bottle was positioned 12 cm from the starting position (Figure 6.1, distance b–c). Before starting the task, the experimenter explained the instructions. Children were tasked with grasping the bottle and moving it forward with varying VFs (rude, neutral, or gentle), as specified by the experimenter.

In the non-social condition, the child moved the bottle forward beyond a marked line on the table 12 cm far from the bottle (Figure 6.1, distance c–d); in the social condition, the child moved the bottle toward a second experimenter seated on the opposite chair (Figure 6.1).



(a) Non-social (b) Social

Figure 6.1: Experimental Setting adopted during the non-social (a) and social (b) contexts.

To establish a baseline for each participant, children initially performed the bottle movement in both social and non-social conditions for seven repetitions each, using their natural approach without specific instructions on VFs. Subsequently, children executed the same action with either gentle or rude VFs, according to the instruction provided by the experimenter in both social and non-social conditions for seven repetitions each, resulting in a total of 42 actions per child [7 actions $\times$ 2 VFs $\times$ 2 Contexts + 7 neutral actions (baseline) $\times$ 2 Contexts]. To mitigate potential performance bias due to factors like fatigue, boredom, or attention issues, all trials were counterbalanced concerning VFs and conditions.

6.1.2.2 Video Editing

Video footage of all the children was carefully reviewed by a trained researcher. A total of 201 and 143 trials were excluded for ASC and NT respectively, because they did not align with task requirements, such as instances where the child moved the bottle outside the designated area or the camera failed to properly capture the

action. Following the cleaning procedures, the remaining number of trials included in the analysis was 849 for ASC and 823 for NT.

Before starting the analysis, each action was divided into two distinct phases using Virtual Dub software:

- **Phase 1 - Reaching phase:** Spanning from the initiation of the reaching movement to the grasping of the bottle (Figure 6.2 - A1, A2 and B1, B2);
- **Phase 2 - Moving phase:** Covering the period from grasping the bottle to completing the action (Figure 6.2 - A3, A4 and B3, B4).

In the reaching phase, our primary aim was to explore the way children grasped the bottle. In the moving phase, our attention shifted to examining how and to what extent the child manipulated the bottle to move it forward, taking into account the specific VFs and the social and non-social contexts surrounding the action (Figure 6.2).

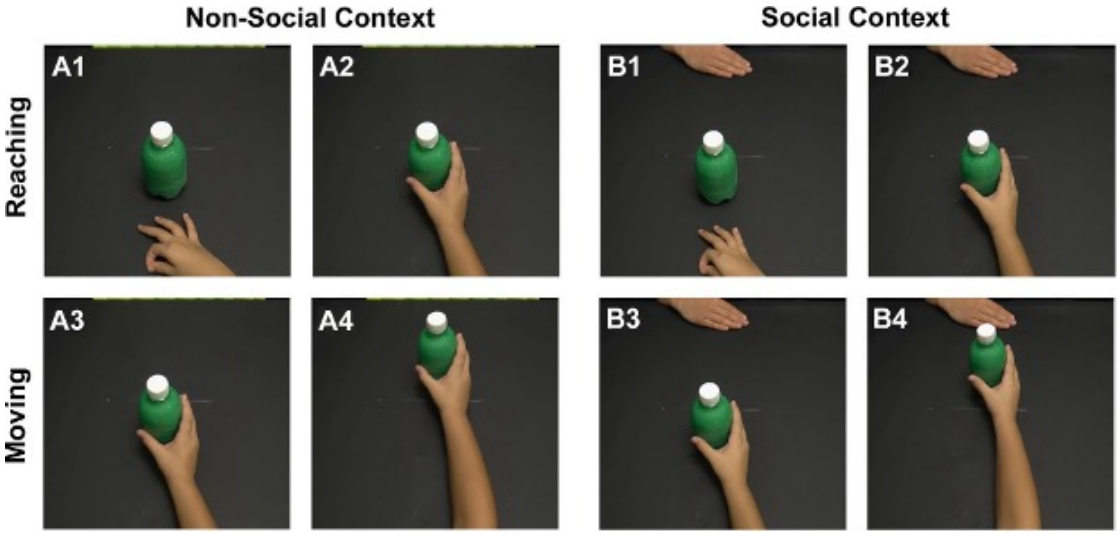


Figure 6.2: Phase 1: reaching phase, from the beginning of the movement (A1,B1) until grasping the bottle (A2,B2); Phase 2: moving phase, from grasping the bottle (A3,B3) until the end of the action (A4,B4) during both social and non-social contexts.

6.1.2.3 Automatic Hands Tracking

Following the pre-processing stage, each video was analyzed using DL models for the automatic frame-by-frame tracking of the child's hand movements during the task. The automatic extraction of trajectories was carried out by applying and customizing the MediaPipe Hand Landmarker solution [165]. As Pose Landmark Detection, MediaPipe Hand Landmarker is a powerful computer vision framework developed by Google's MediaPipe team. It utilizes advanced DL techniques, specifically CNNs, to enable real-time hand movement tracking and landmark estimation from video or webcam streams. With MediaPipe Hand Landmarker, developers can leverage pre-trained models and algorithms to identify and locate key reference points on the hand, such as fingertips, knuckles, and the palm. More precisely, this model was able to detect the localization of 21 hand-knuckle coordinates within the detected hand regions. The framework incorporates a multi-stage regression approach, thanks to which the model predicts the landmarks in a sequential manner, progressively refining its estimates at each stage. This process enables it to handle challenges like occlusions and improve the accuracy of landmark estimation.

In our task, we modified the network architecture to accurately distinguish between the child's right hand and the experimenter's hand in the social context. Specifically, we customized the network to provide as output the rotation of the rectangular bounding box for each hand, in order to differentiate between them. This modification allowed for automatic and reliable identification of the child's hand, even during the social context, and facilitated accurate tracking of their movements. The recorded videos have a frame rate of 25 fps, resulting in estimated trajectories that are sampled every 40 ms.

6.1.2.4 Feature Extraction

For the characterization of our specific task, we focused on three reference points (markers) on the hand. Specifically, the wrist was considered the most stable point for analyzing the kinematic parameters of the whole action (Figure 6.3a - Figure 6.3c), while the tips of the thumb and index finger were used to analyze the grasping of the bottle (Figure 6.3b).

The features were extracted for each marker using the trajectories in two di-

mensions, specifically the x and y coordinates of the Cartesian plane, with the origin located at the top left corner of the frame (Figure 6.3d). Specifically, we extracted the following features:

- **Velocity and acceleration (both phases):** The velocity of the wrist was computed as the Euclidean distance between the location of the reference point in every two subsequent frames divided by their temporal distance. The acceleration was calculated as the difference between two consecutive velocity samples. For both signals, the x and y components and the total module were calculated. Additionally, the signals were resampled with a frame interval of 5 ms, and a moving average filter with a window of three samples was applied to reduce noisy fluctuations (Figure 6.3e). Mean and maximum values, along with their corresponding time, were computed for both velocity (px/s) and acceleration (px/s^2). Furthermore, the maximum deceleration and its corresponding time point were computed.

- **Maximum opening (Phase 1):** The maximum opening in pixels during the bottle grasping phase was calculated as the maximum Euclidean distance between the thumb and index fingertips. The corresponding time was also determined. The feature was normalized by the distance between the wrist and the tip of the middle finger to account for children's dimensions.

- **Grasping time (Phase 1):** The duration of grasping time in seconds was measured as the time interval between the beginning of the action and the instant of bottle grasping.

- **Maximum displacement along x and y during action execution (Phase 2):** The maximum extent of movement in pixels along the two axes was calculated by measuring the difference between the maximum and minimum ordinate and the maximum and minimum abscissa covered by the hand.

- **Coordinates of the hand position at the end of the action (Phase 2):** We checked whether the model correctly identified and tracked the child's hand in the final frames of the video. If the hand was not successfully

detected in the last part of the video, resulting in missing data, we excluded the final hand coordinates from the features.

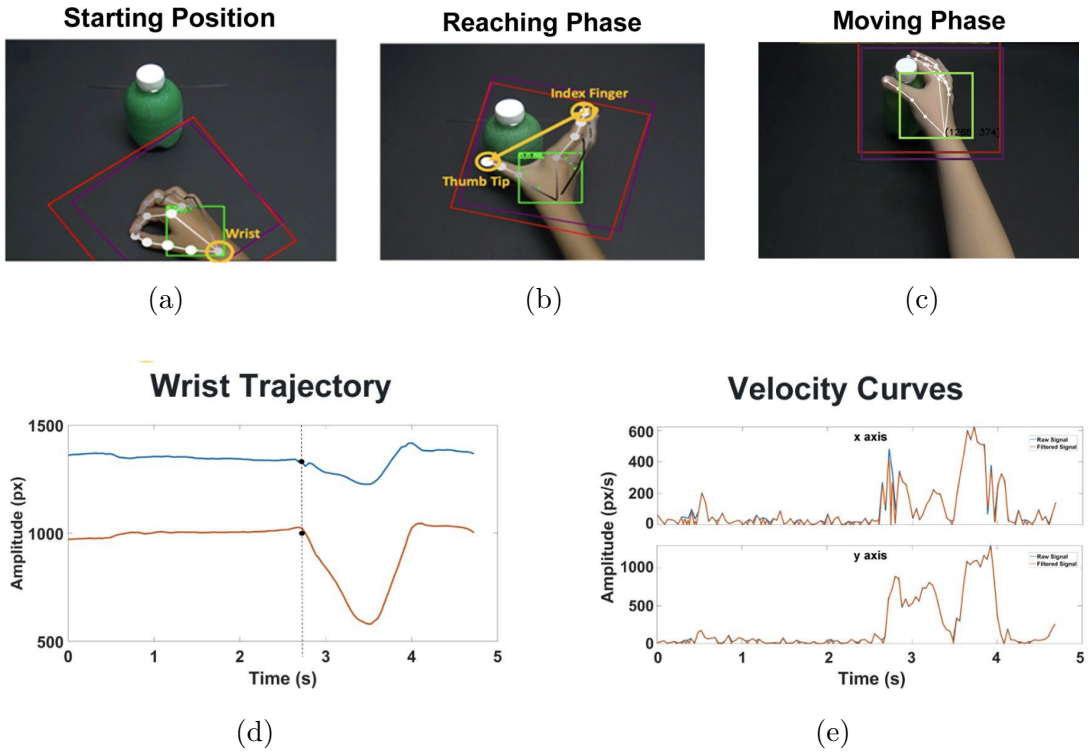


Figure 6.3: Hand markers employed for feature extraction: the wrist point was used as a reference to compute kinematic parameters for the entire action, from the starting position (a) to the moving phase (c). The thumb and index fingertips were used to characterize the grasping phase of the action (b). An example of a wrist trajectory (x-axis component in blue and y-axis component in red) is automatically extracted from one subject, with the starting point of the action, where the child begins to move, indicated by black dots (d). Additionally, an example of a velocity curve derived from the x and y components of the wrist trajectory of one subject is provided. The original signal is shown in blue, and the interpolated signal is indicated in red (e).

6.1.2.5 Statistical Analysis

The statistical analysis was conducted using RStudio (Version 1.4.1106) [263], an integrated development environment for R. Data normality was assessed with the

Shapiro-Wilk test. For variables that followed a normal distribution, a two-way analysis of variance (ANOVA) was performed with Context (social vs. non-social) and Vitality Form (gentle vs. rude) as within-subject factors. Post hoc comparisons were carried out using Tukey's HSD test. For non-normally distributed variables, non-parametric tests, including the Wilcoxon signed-rank test and the Kruskal-Wallis test, were employed. Statistical significance was set at $p < 0.05$. Data visualization was performed using `ggplot2`, an R package for creating data visualizations, to illustrate means and standard errors across conditions.

6.1.3 Results

The results described below pertain to the analysis of the following kinematic parameters recorded in both ASC and NT children: mean velocity (v_m), maximum velocity ($v_{\max}$), time to maximum velocity ($v_{t\max}$), mean acceleration (a_m), maximum acceleration ($a_{\max}$), maximum deceleration ($D_{\max}$), time to maximum deceleration ($D_{t\max}$), maximum opening ($O_{\max}$), time to maximum opening ($O_{t\max}$), maximum displacement in the X-axis ($D_{X\max}$), and maximum displacement in the Y-axis ($D_{Y\max}$). The main significant results are summarized in Table 6.2 and Figure 6.4 (for more details, refer to the *Supplementary Information*).

Main effects Analysis A significant main effect of **GROUP** was observed during the reaching phase for the kinematic parameters a_m ($p < 0.01$), $a_{\max}$ ($p < 0.01$), $O_{\max}$ ($p < 0.05$) and $O_{t_{\max}}$ ($p < 0.01$). In the moving phase, the main effect of **GROUP** was significant for a_m ($p < 0.05$), $a_{\max}$ ($p < 0.05$), $D_{\max}$ ($p < 0.06$), $D_{t_{\max}}$ ($p < 0.06$), $D_{X_{\max}}$ ($p < 0.01$) and $D_{Y_{\max}}$ ($p < 0.01$) as shown in Figure 6.4 and Table 6.2. Regarding the effect of **VITALITY**, significant differences were observed in both phases. In the reaching phase, the parameters v_m ($p < 0.0001$), $v_{\max}$ ($p < 0.0001$), a_m ($p < 0.0001$), $a_{\max}$ ($p < 0.01$), $D_{\max}$ ($p < 0.01$), and $D_{t_{\max}}$ ($p < 0.0001$) were significant (Table 6.2). Planned contrasts (*a priori*, Bonferroni correction, $p < 0.008$) revealed differences in v_m ($p < 0.0001$), $v_{\max}$ ($p < 0.0001$), a_m ($p < 0.0001$), $a_{\max}$ ($p < 0.001$), and $D_{\max}$ ($p < 0.001$), while $D_{t_{\max}}$ was not significant. In the moving phase, significant effects of **VITALITY** were observed for v_m ($p < 0.0001$), $v_{\max}$ ($p < 0.001$), a_m ($p < 0.01$), $a_{\max}$ ($p < 0.001$), and $D_{\max}$

($p < 0.0001$). Planned contrasts showed differences for v_m ($p < 0.0001$), $v_{\max}$ ($p = 0.003$), $a_{\max}$ ($p < 0.001$), and $D_{\max}$ ($p < 0.001$), while a_m and $D_{t_{\max}}$ were not significant.

Finally, the effect of **CONTEXT** was significant only during the moving phase, where $a_{\max}$ ($p < 0.05$) and $v_{\max}$ ($p < 0.05$) showed differences (Figure 6.4 and Table 6.2).

Interaction effects Analysis The interaction **VITALITY * GROUP** was significant exclusively in the moving phase for the following kinematic parameters: $D_{\max}$ ($p < 0.03$), $a_{\max}$ ($p = 0.06$), and a_m ($p < 0.03$) (Figure 6.4C).

The interaction **CONTEXT * VITALITY** also showed significance only in the moving phase, involving $D_{Y_{\max}}$ ($p < 0.05$), $D_{t_{\max}}$ ($p = 0.06$), $a_{\max}$ ($p < 0.05$), a_m ($p < 0.05$), and $v_{\max}$ ($p < 0.05$), as detailed in Table 6.2.

The interaction **CONTEXT * GROUP** was significant during the reaching phase for $D_{\max}$ ($p < 0.05$) and $v_{t_{\max}}$ ($p < 0.05$). In the moving phase, the same interaction revealed a significant difference for v_m ($p < 0.01$) (Figure 6.4D).

Finally, the interaction **GROUP * VITALITY * CONTEXT** yielded significant results in the reaching phase for $D_{\max}$ ($p < 0.05$) and v_m ($p = 0.06$), as well as in the moving phase for $D_{X_{\max}}$ ($p < 0.05$) and v_m ($p < 0.01$) (Figure 6.4E).

		Reaching phase		**Moving phase**	
		F(1,41)	**p**	**F(1,40)**	**p**
$v_{\mathbf{m}}$	Vitality	80.92	< 0.001	59.76	< 0.001
	Group*Vitality*Context	3.81	0.06	6.59	0.01
	Context*Group			7.61	0.01
$v_{\mathbf{max}}$	Group	3.35	0.07		
	Vitality	66.84	< 0.001	24.69	< 0.001
	Context			4.64	0.03
$v_{\mathbf{tmax}}$	Context*Group	6.13	0.02		
$a_{\mathbf{m}}$	Group	7.94	0.01	6.21	0.02
	Vitality	86.09	< 0.001	21.89	< 0.001
	Vitality*Group			4.66	0.03
$a_{\mathbf{max}}$	Group	8.19	0.01	4.66	0.003
	Vitality	20.25	0.01	13.23	< 0.001
	Context			5.76	0.02
	Vitality*Group			3.74	0.06
	Context*Vitality			6.09	0.02
$D_{\mathbf{max}}$	Vitality	21.53	0.01	37.17	< 0.001
	Context*Group	4.46	0.04		
	Group*Vitality*Context	5.81	0.02	4.96	0.03
$D_{\mathbf{tmax}}$	Vitality	5.08	0.01		
	Context*Vitality			3.58	0.06
$D_{X_{\mathbf{max}}}$	Vitality			14.69	< 0.001
	Group*Vitality*Context			5.85	0.02
$D_{Y_{\mathbf{max}}}$	Vitality			27.67	< 0.001
	Context*Vitality			5.32	0.03
$O_{\mathbf{max}}$	Group	7.11	0.01		
	Vitality	12.24	0.01		

Table 6.2: Main and interaction effects: significant F and p-values for reaching and moving phases.

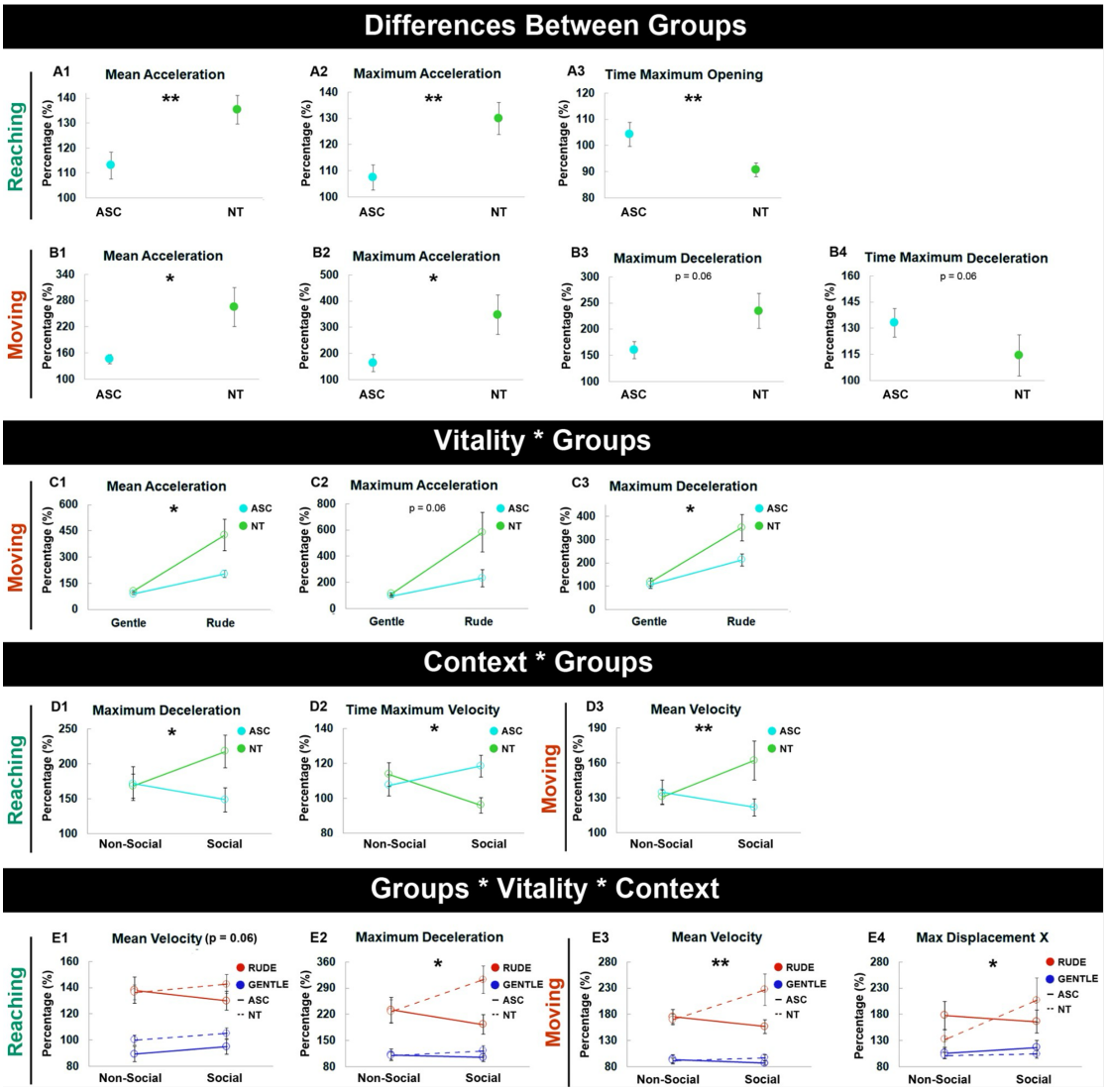


Figure 6.4: Overview of the main differences between ASC and NT groups during the reaching and moving phases in the expression of vitality forms. All values are normalized to the baseline condition, where 100% represents the baseline. Vertical bars indicate standard errors. Statistical significance is denoted as follows: $^{*}p \leq 0.05$, $^{**}p \leq 0.01$.

Correlation analyses Table 6.3 shows results regarding the correlation analyses carried out between neuropsychological measures and VFs kinematic parameters.

ASC children									
		Gentle	Rude	Social	Non-social	Gentle social	Gentle non soc.	Rude social	Rude non soc.
Mean acceleration	ADOS	0.099	-2.4	-0.26	-0.03	0.07	0.10	-0.28	-0.09
	TEC	-0.201	0.28	0.20	0.13	-0.18	-0.18	0.23	0.25
	WISC-IV	0.241	-0.04	-0.11	0.17	0.06	0.32	-0.46	0.19
	AS	**0.59****	-0.23	-0.33	0.30	**0.60****	**0.46***	-0.46*	0.19
Max acceleration	ADOS	0.07	-0.24	-0.25	0.04	-0.07	0.17	-0.24	-0.02
	TEC	-0.98	0.08	0.07	0.02	0.03	-0.18	0.07	0.11
	WISC-IV	**0.58****	-0.13	-0.17	**0.52***	**0.54***	**0.54***	-0.18	**0.46***
	AS	0.21	**-0.46***	**-0.47***	0.08	0.20	0.19	**-0.48***	0.03
Time max acceleration	ADOS	0.09	0.14	0.12	0.14	-0.38	0.15	0.14	0.13
	TEC	-0.21	-0.17	-0.19	-0.16	-0.18	-0.19	-0.18	-0.15
	WISC-IV	0.17	0.20	0.20	0.19	-0.04	0.18	0.20	0.19
	AS	-0.21	-0.17	-0.17	-0.18	-0.30	-0.17	-0.15	-0.18
Max deceleration	ADOS	0.06	-0.14	0.05	-0.24	0.15	0.06	-0.02	-0.27
	TEC	-0.33	0.08	-0.03	-0.16	-0.31	0.24	0.14	-0.03
	WISC-IV	-0.05	0.07	-0.15	0.29	-0.31	0.26	-0.02	0.18
	AS	**0.63****	0.18	0.34	**0.45***	**0.60****	**0.46***	0.09	0.24
Time max deceleration	ADOS	-0.08	0.13	0.08	0.10	-0.18	0.14	0.16	0.01
	TEC	-0.09	-0.13	-0.19	0.01	-0.18	0.10	-0.15	-0.04
	WISC-IV	0.33	0.19	0.31	0.18	0.34	0.16	0.22	0.09
	AS	**-0.61****	-0.27	-0.36	**-0.65****	**-0.48***	**-0.54***	-0.22	-0.37
NT children									
		Gentle	Rude	Social	Non-social	Gentle social	Gentle non soc.	Rude social	Rude non soc.
Mean acceleration	RAVEN	-0.13	0.009	0.01	-0.07	0.04	-0.22	0.01	-0.01
	TEC	-0.35	-0.067	-0.04	-0.27	-0.01	**-0.49***	-0.04	-0.16
Max acceleration	RAVEN	-0.29	0.12	0.10	0.02	-0.19	-0.23	0.11	0.10
	TEC	**-0.55****	-0.11	-0.10	0.25	0.20	**-0.53****	-0.10	-0.12
Time max acceleration	RAVEN	0.16	0.13	0.12	0.33	0.15	0.27	0.09	0.30
	TEC	0.20	0.15	0.18	0.004	0.19	0.17	0.17	-0.11
Max deceleration	RAVEN	-0.20	-0.22	-0.16	-0.31	-0.28	-0.001	-0.11	-0.35
	TEC	-0.15	-0.14	-0.03	-0.37	0.31	0.11	0.04	-0.48
Time max deceleration	RAVEN	0.26	0.18	0.19	0.27	0.18	0.34	0.20	0.14
	TEC	-0.10	-0.05	-0.06	-0.09	-0.12	-0.06	-0.01	-0.12

Table 6.3: Results of correlations analysis carried out in ASC and NT children between neuropsychological tests scores and VFs kinematic parameters ($*p < 0.05$, $**p < 0.01$). The numbers in bold are those marked with an asterisk ($*/**$) and signify statistical significance.

6.1.4 Discussion

This section investigated how children with ASC and NT peers communicate positive and negative VFs through actions, in both social and non-social contexts. The study further explored kinematic differences in the performance of action VFs between the two groups. To address these objectives, we designed tasks in

which both groups moved a small bottle either to a designated point on a table (non-social context) or toward a receiver (social context) while expressing neutral, gentle, or rude VFs. DL models were employed to extract kinematic parameters from videos, thereby characterizing these VFs during two distinct action phases: reaching and moving.

6.1.4.1 Interpretation of Results

The findings revealed that ASC children are capable of modulating their motor profiles to express both gentle and rude VFs, as evidenced by adjustments in kinematic parameters such as mean velocity, maximum velocity, mean acceleration, and maximum acceleration. These results suggest that once ASC children comprehend the meanings of VFs, with only participants scoring above 75% accuracy in a preliminary interview being included, they can adapt their actions accordingly. This observation aligns with the findings of Csartelli et al. [264], who reported modifications in peak velocity and acceleration when ASC children performed actions with distinct VFs.

Despite this adaptive capacity, significant divergences in motor profiles were observed when comparing ASC to NT children. For example, during the moving phase, ASC children exhibited reduced mean acceleration, maximum acceleration, and maximum deceleration when moving the bottle forward. In the reaching phase, differences emerged in the timing of maximum hand opening and maximum deceleration. These kinematic differences were particularly pronounced during the expression of rude VFs, indicating greater challenges in motor control for ASC children. Such findings support prior research suggesting that motor planning difficulties in ASC may arise from challenges in processing multiple pieces of information simultaneously, leading to inefficiencies in action execution.

A critical focus of the study was on the modulation of VFs across social and non-social contexts. The results indicated that social contexts negatively impacted VFs expression in ASC children, as evidenced by reduced mean velocity during the moving phase. In contrast, NT children exhibited an increase in kinematic parameters, such as maximum acceleration and mean velocity, when shifting from non-social to social contexts. The ability of NT children to adapt and maintain

clear distinctions between gentle and rude actions in social settings underscores their superior modulation skills compared to ASC peers. Conversely, ASC children demonstrated diminished differentiation between gentle and rude VFs during social interactions, suggesting that social communication challenges may influence motor behavior.

Moreover, our analysis revealed that higher social communication difficulties, as indicated by elevated SA scores on the ADOS-2, were associated with reduced VFs modulation, particularly in social contexts and for the expression of rude VFs. This finding highlights the potential influence of social anxiety or action inhibition on the motor expressions of ASC children, further emphasizing the interplay between social communication challenges and motor behavior in autism.

6.1.4.2 Significance and Value

This study makes a significant contribution to understanding the modulation of VFs in ASC children within authentic social contexts. By employing an innovative, non-invasive method, automatic DL video tracking, we captured continuous motion data without imposing physical constraints. This approach is particularly advantageous when working with children who may resist wearing markers or sensors, a common challenge in autism due to sensory sensitivities. The AI-based kinematic motion tracking method provides an authentic representation of motor behavior, offering distinct advantages over traditional motion capture systems.

Furthermore, from a clinical perspective, the study underscores the interconnectedness of social communication and motor domains in ASC, revealing how social communication difficulties extend into motor behavior and affect the nuanced expression of vitality in actions. This knowledge supports the development of interventions that target both motor and social domains, thereby offering a more holistic therapeutic approach.

6.1.4.3 Limitations & Future Improvements

The small sample size poses a limitation, as it may not adequately capture the diversity within ASC and NT populations, thereby reducing the generalizability of the findings. Increasing the sample size and incorporating a broader range of

participants would strengthen both the robustness and applicability of the results. Additionally, the study's focus on male participants restricts the ability to investigate sex-based differences in VFs expression, an essential avenue for future research.

Moreover, the tasks were designed specifically around moving a bottle in either social or non-social contexts. While these controlled scenarios offered valuable insights into VFs expression, their situational specificity might limit the relevance of the findings to everyday interactions. Future studies should examine VFs expression across a wider variety of naturalistic contexts to yield a deeper understanding of its role in social communication.

6.2 Artificial Intelligence for Speech Assessment in Children: A Hierarchical Approach

This section is dedicated to the development of an AI-based model to support speech assessment in children with dysarthria. The proposed system is a Hierarchical Machine Learning Model (HMLM) that combines conventional Machine Learning (ML) with Deep Learning (DL) techniques to address both the diagnosis and severity assessment of dysarthria. Employing the standardized "PA-TA" speech disturbance test from the Scale for the Assessment and Rating of Ataxia (SARA) [265], the model first differentiates between healthy children and those with dysarthria, subsequently evaluating the severity of the condition. This innovative approach marks a significant step forward in leveraging automated tools to identify and quantify dysfunctions in speech production, thereby facilitating the evaluation of language-related impairments.

The model enhances current therapeutic practices by offering precise and objective insights into linguistic abilities, contributing to a deeper understanding and improved management of dysarthria within neurodevelopmental contexts. Developed using speech data from children with ataxia [266], a neurological condition affecting motor coordination, the model demonstrates strong generalizability. Its effectiveness in assessing speech disturbances underscores its value as a diagnostic and therapeutic tool for Neurodevelopmental Disorders involving communication impairments.

Dysarthria, defined as a motor speech disorder caused by neurological impairments that affect the muscles responsible for speech [267], presents challenges in articulation, phonation, respiration, and prosody. Its manifestation in different NDD varies depending on the underlying neurological deficits. For example, in CP, dysarthria often arises from motor control issues due to central nervous system damage [268], while children with Rett syndrome may experience speech challenges linked to motor and communicative regression [269]. Congenital neuromuscular disorders and other conditions that impair motor coordination can also result in dysarthria as a secondary effect of weakened or poorly coordinated speech muscles [270]. Recognizing these patterns is essential for developing effec-

tive, multidisciplinary therapeutic interventions that address both the neurological and communicative aspects of these conditions.

6.2.1 Participants

The study population was recruited in 2018 at the Movement Analysis and Robotics laboratory (MARlab) of the Intensive Neurorehabilitation and Robotics Departments of IRCCS Bambino Gesù Children's Hospital (Rome, Italy). Overall, it is composed of 55 subjects: 18 Healthy (H), 21 with Progressive Ataxia (PA) and 16 with Congenital Non-Progressive Ataxia (CA). H group included sex/age matched healthy volunteers without personal/familiar history of neurological diseases and no signs at clinical examination (age $12[7.6]; 12F/6M$). All patients had genetically confirmed diagnosis and a routine diagnostic workup, including general and neurological examination, brain Magnetic Resonance Imaging (MRI), sensory evoked potentials, nerve conduction study and visual acuity evaluation; moreover, they were in follow-up at the MARlab for at least 2 years, to ensure a correct group classification. None of the enrolled subjects had relevant cognitive impairment or were taking psychoactive drugs (other usual medications, such as vitamin or antioxidant were allowed). Patients with severe disability, moderate severe cognitive impairment affecting tests execution were excluded. Demographic data were collected for the three groups. The research conformed to the ethical standards laid down in the 1964 Declaration of Helsinki. All subjects participated on a voluntary basis, after that they or their legal responsible signed the informed consent (the study was approved by local ethical committee Protocol NET-201302356160 WP3, nr. 1619-2018, received 03 July 2018).

6.2.2 Methods

6.2.2.1 Experimental Setup

After receiving a clinical evaluation, all the 55 subjects were asked to perform the "PATA" test in a quiet room. Each vocal task was recorded with SaraHome, a novel technology for the assessment at home of patients with ataxia symptoms [271], using the microphone array mounted on the Microsoft Kinect V2 for 10

seconds at a sampling frequency (F_s) of 16 KHz. Each subject was asked to repeat the word "PATA" as many times as possible in 10 seconds, as reported in [272, 273]. At the end of each task, speech disturbance was scored by expert personnel using a standardized clinical scale: SARA [274]. For each patient with CA and PA, the same test was repeated after 12 months (time t_1) to monitor the possible evolution of disturbances. It was possible to repeat the test only for 21/34 patients (12 PA and 9 CA). For this reason, 76 audio recordings were totally considered. All the data were analyzed using Matlab version 2020 [164].

6.2.2.2 Overall Architecture

Figure 6.5 illustrates the overall architecture of the proposed hierarchical model for speech assessment. The approach combines classical ML techniques with DL architectures to address both classification and severity stratification tasks. Initially, raw speech signals are pre-processed to extract relevant acoustic features, which serve as input to the hierarchical model. The first layer performs a binary classification to distinguish between healthy individuals and those with dysarthria. The second layer refines the assessment by stratifying dysarthria into different severity levels. This two-tiered framework ensures precise and granular insights into speech impairments, facilitating a deeper understanding of the underlying neurological conditions.

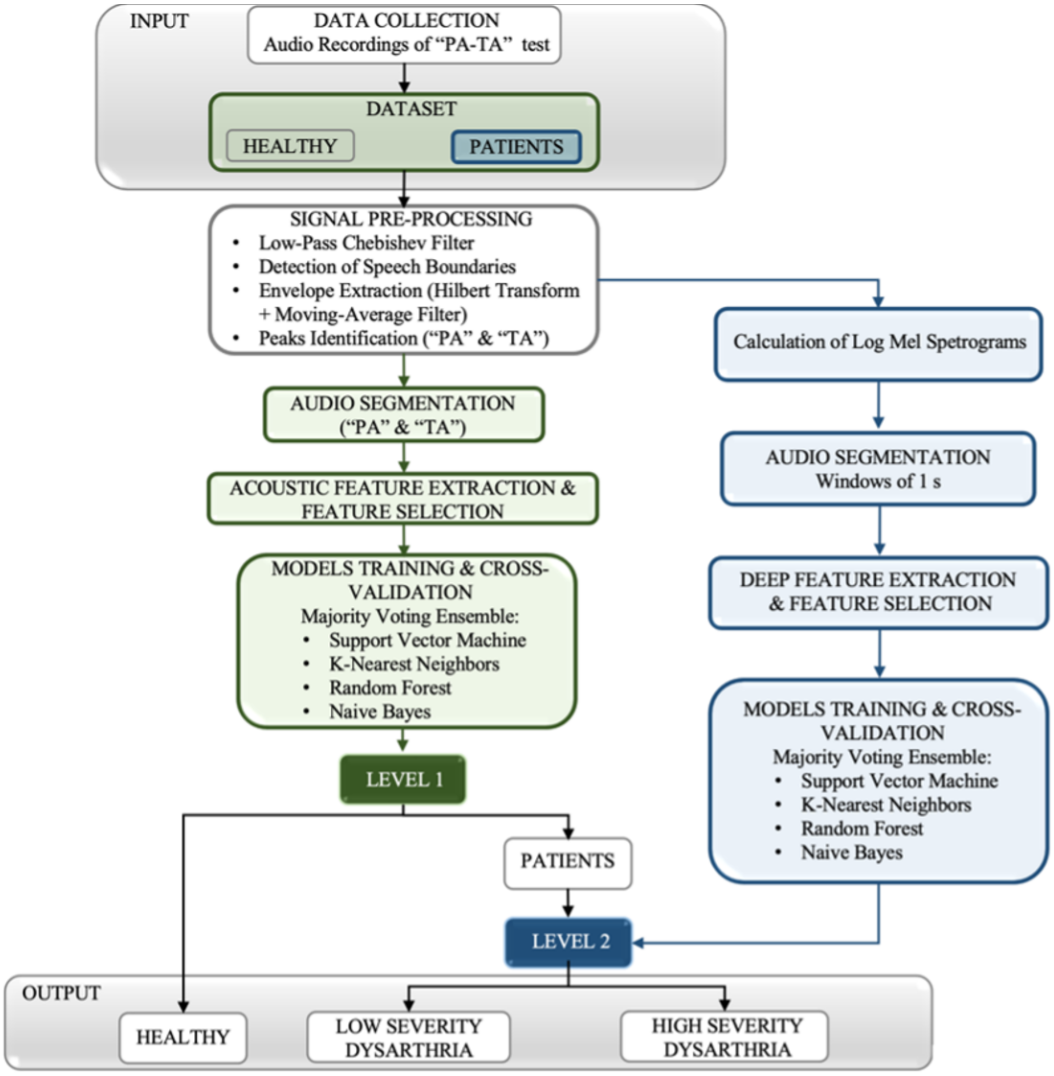


Figure 6.5: Flowchart of the overall architecture.

6.2.2.3 Signal Pre-Processing and "PA-TA" Segmentation

Sometimes the collected data were affected by background noise such as external voices, door slamming sounds, or environmental noises; therefore, a step of pre-processing and clean-up was necessary. Initially, we evaluated the average signal spectrum (Short Time Fourier periodogram) to detect the frequency range of interest (Figure 6.6). Since patients' voices repeating "PA-TA" were mostly under the frequency of 1 kHz, in order to reduce all the noise above this frequency, we

applied an eleven-order low-pass Chebyshev filter with a cut-off frequency of 1 kHz and a Hanning window with length equal to the 0.5% of F_s. Afterward, we applied the method based on fine-tuning of threshold short-term energy and spectral spread [275] to detect speech boundaries and remove the remaining noise.

The envelope of each signal was extracted by applying firstly the module of the Hilbert Transform and later a zero-phase moving-average filter whose parameters were tuned according to the signal approximate entropy. If signal approximate entropy was lower than the empirical threshold of 0.8, a single moving-average filter was applied to the Hilbert Transform; otherwise, two cascade filters were employed, as reported in Table 6.4.

The main steps of signal pre-processing are shown in Figure 6.7. After these steps, "PA" and "TA" peaks were detected from the envelope, selecting only maxima with a minimum prominence equal to 10% of the absolute value of the Hilbert Transform mean. Instead, signal minima were recognized by computing the energy and by choosing only the minimum prominence of 0.01 and at least 10% of the sampling frequency apart.

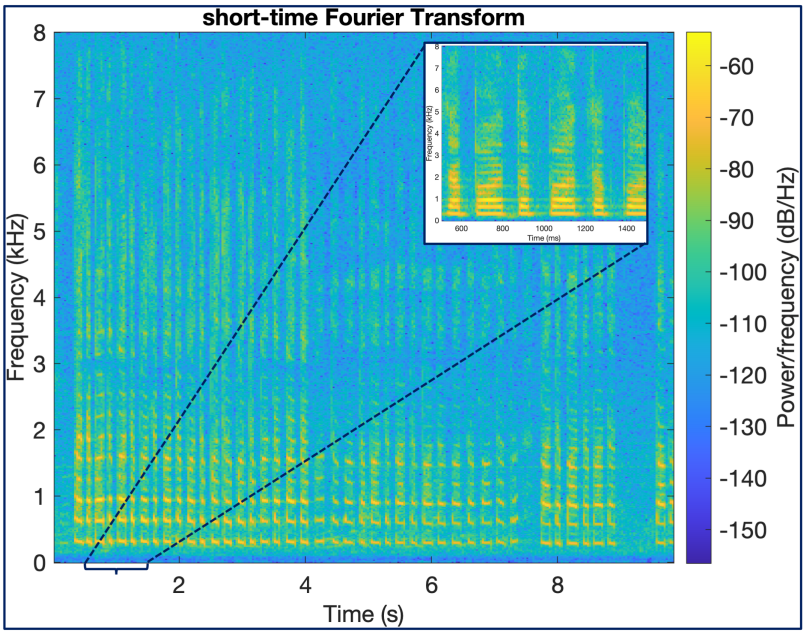


Figure 6.6: Signal short-time Fourier Transform with a detailed view between 500 ms and 1500 ms.

Approximate Entropy Threshold	**Numerator Coefficients of the Rational Transfer Function (b)**	**Denominator Coefficients of the Rational Transfer Function (a)**
Entropy < 0.8	c_n, where $c = 1/1500$, $n = 1500$	0.7
Entropy ≥ 0.8	Filter 1: c_n, where $c = 1/1000$, $n = 1000$	0.7
	Filter 2: c_n, where $c = 1/500$, $n = 500$	1.0

Table 6.4: Moving average filter parameters.

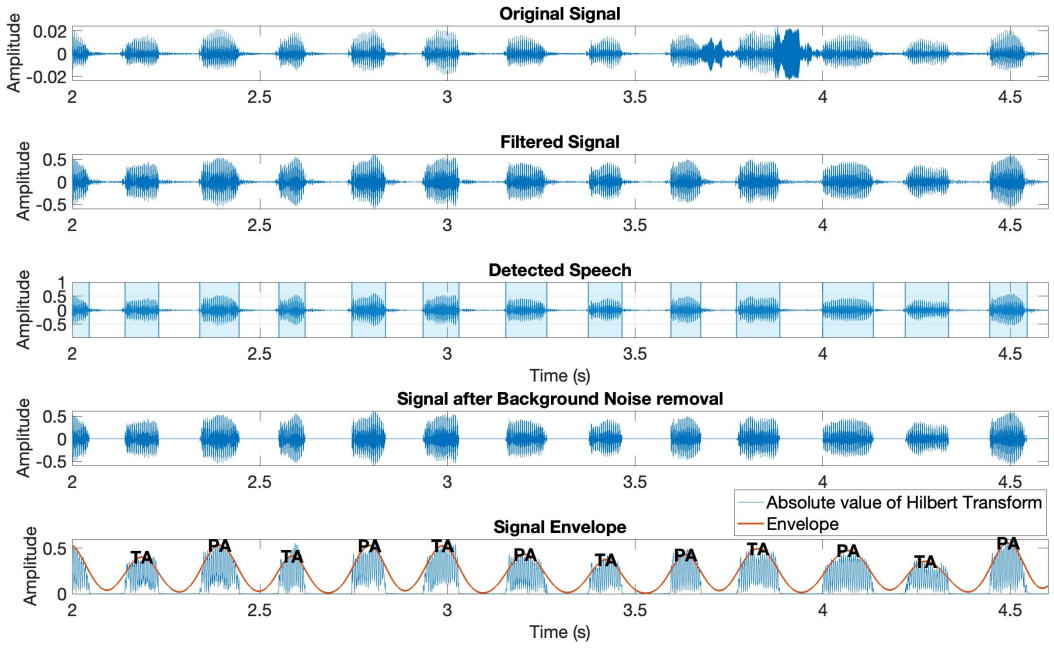


Figure 6.7: Main steps of signal pre-processing from noise removal to "PA" and "TA" peaks identification. The first row shows an example of background noise (squeak) that is removed through the low-pass filter. Later, the detect speech method based on threshold short-term energy and spectral spread is employed to remove the remaining background noise. The cleaned-up signal is used to obtain the envelope and consequently to identify peaks.

6.2.2.4 Feature Extraction and Selection for Machine Learning

Audio signals were segmented in order to increase the statistical significance of the dataset in terms of inter-subject and inter-class variability [276]. Because of windowing the signals, it was possible to assume their quasi-stationary within each frame, easing the subsequent analysis [277]. Since the performance of the system depends largely on noise reduction among peaks and the selection of useful acoustic events only (Figure 6.8), it was necessary to carry out the segmentation using the "PA" & "TA" peaks as reference points, and the samples between the closest preceding and consecutive minima considering each PA-TA cycle.

After performing audio segmentation, we investigated the most relevant conventional features of our targeted application. In literature, the issue of feature extraction in the field of audio processing is quite challenging because of several factors such as the simultaneous presence of different sound sources and the background noises that may affect machine performance [277]. These characteristics are considered to identify the most reliable parameters. In Figure 6.5, all the features extracted and grouped by time domain (*PATAfreq*, Approximate Entropy), frequency domain (spectral values, MFCCs and GTCCs coefficients), chaotic domain (Lyapunov Exponent), and age of children, are listed. All the features were extracted from each PA-TA cycle and then the average value for each subject was calculated.

PATA frequency ($PATA_{freq}$) is a simple time domain physical feature, whose calculation is directly performed from the temporal envelope of signals in order to assess a fundamental parameter of our specific task, according to the following equation:

$$PATA_{freq} = \frac{n_{peaks}}{2 \cdot l} \tag{6.1}$$

where n_{peaks} is the total number of recognized peaks and l is the length of the signal.

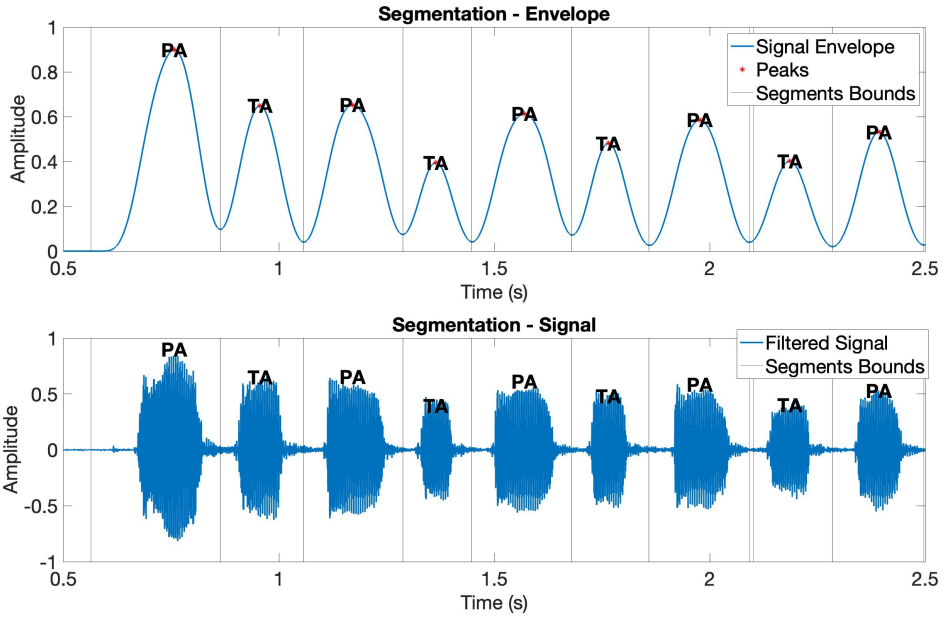


Figure 6.8: Audio segmentation. Each signal was segmented using "PA" & "TA" peaks identified on the envelope as reference points and considering, for each PA-TA cycle, only the samples between the closest preceding and consecutive minima. Each peak corresponds to a syllable.

Approximate Entropy is calculated to measure the complexity and possible fluctuations of the signals [278] and for its strength to discriminate human voice components from corrupted speech [279]. Lyapunov Exponent is calculated to consider the non-linearity of speech [280, 281, 282, 283, 284]. Frequency domain features, conventionally used for lots of applications [285, 286], are the most described in literature. These variables are intended to describe the physical properties of the signal frequency content, and they cover a large number of different categories. Among this wide range of possibilities, we computed the following features:

- **Mel-Frequency Cepstral Coefficients (MFCCs):** They are one of the most popular features employed in speech processing. They constitute the mel-frequency cepstrum (MFC), a compact representation of the short-term power spectrum of an audio signal, obtained through a linear cosine transform from the log power spectrum to the nonlinear mel scale frequency [286].

- **Gammatone Cepstral Coefficients (GTCCs):** They are a modification of MFCCs inspired by biology and are obtained by applying Gammatone filters with equivalent rectangular bandwidth bands [287].

- **Spectral Centroid:** It can be considered the barycenter of the spectrum and indicates where most of the signal energy is contained:

$$\text{Spectral Centroid} = \frac{\sum_{k=b_1}^{b_2} f_k s_k}{\sum_{k=b_1}^{b_2} s_k}, \tag{6.2}$$

  where f_k is the frequency in Hz and s_k is the spectral value that corresponds to bin k, while b_1 and b_2 are the band edges, in bins, over which to calculate the spectral centroid [286, 288].

- **Spectral Spread:** It is a measure of the spread of the spectrum around its mean value:

$$\text{Spectral Spread} = \sqrt{\frac{\sum_{k=b_1}^{b_2} (f_k - \mu_1)^2 s_k}{\sum_{k=b_1}^{b_2} s_k}}, \tag{6.3}$$

  where f_k is the frequency in Hz and s_k is the spectral value that corresponds to bin k, while b_1 and b_2 are the band edges, in bins, over which to calculate the spectral spread and μ_1 is the spectral centroid [287, 289].

- **Spectral Skewness:** It is a measure of the asymmetry of the spectrum around its mean value and is computed from the 3rd order moment:

$$\text{Spectral Skewness} = \frac{\sum_{k=b_1}^{b_2} (f_k - \mu_1)^3 s_k}{(\mu_2)^3 \sum_{k=b_1}^{b_2} s_k}, \tag{6.4}$$

  where f_k is the frequency in Hz and s_k is the spectral value that corresponds to bin k, while b_1 and b_2 are the band edges, in bins, over which to calculate the spectral skewness, μ_1 is the spectral centroid and μ_2 is the spectral spread. Skewness $= 0$ indicates a symmetric distribution, Skewness < 0 more energy on the right, Skewness > 0 more energy on the left [290].

- **Spectral Kurtosis:** It gives a measure of the flatness of the spectrum around its mean value and indicates a possible non-stationary or non-Gaussian

behavior in the frequency domain. It is the 4th order moment and is computed starting from the short-time Fourier Transform of the signal $S(t, f)$:

$$\text{Spectral Kurtosis} = \frac{\langle |S(t,f)|^4 \rangle}{\langle |S(t,f)|^2 \rangle^2} - 2, \quad f \neq 0, \tag{6.5}$$

where $\langle \cdot \rangle$ is the time-average operator. Kurtosis = 3 indicates a normal distribution, Kurtosis < 3 a flatter distribution, and Kurtosis > 3 a peaked distribution [291, 292, 293].

- **Spectral Slope:** It represents the amount of decrease of the spectral amplitude and is computed as the linear regression of the spectral amplitude:

$$\text{Spectral Slope} = \frac{\sum_{k=b_1}^{b_2} (f_k - \mu_f)(s_k - \mu_s)}{\sum_{k=b_1}^{b_2} (f_k - \mu_f)^2}, \tag{6.6}$$

where f_k is the frequency in Hz and s_k is the spectral value that corresponds to bin k, while b_1 and b_2 are the band edges, in bins, over which to calculate the spectral slope, μ_f is the mean frequency and μ_s is the mean spectral value [294].

- **Spectral Decrease:** It represents the amount of decrease of spectral amplitude, defined from perceptual studies to be more correlated to human perception:

$$\text{Spectral Decrease} = \frac{\sum_{k=b_1+1}^{b_2} (s_k - s_{b_1})}{k-1} \Bigg/ \sum_{k=b_1+1}^{b_2} s_k, \tag{6.7}$$

where s_k is the spectral value that corresponds to bin k, while b_1 and b_2 are the band edges, in bins, over which to calculate the spectral decrease.

- **Spectral Rolloff Point:** It is the frequency below which 95% of the signal energy resides:

$$\text{Spectral Rolloff Point} = i \text{ such that } \sum_{k=b_1}^{i} s_k = 0.95 \sum_{k=b_1}^{b_2} s_k \tag{6.8}$$

where s_k is the spectral value corresponding to bin k, and b_1 and b_2 are the band edges, in bins, over which to calculate the spectral rolloff point [295].

- **Spectral Flatness:** It is a measure of the noisiness/sinusoidality of a spectrum and is computed as the ratio between the geometric mean and the arithmetic mean of the energy spectrum:

$$\text{Spectral Flatness} = \frac{\left(\prod_{k=b_1}^{b_2} s_k\right)^{1/(b_2-b_1)}}{\frac{1}{b_2-b_1}\sum_{k=b_1}^{b_2} s_k} \tag{6.9}$$

where s_k is the spectral value corresponding to bin k, and b_1 and b_2 are the band edges, in bins, over which to calculate the spectral flatness. For tonal signals, it is close to 0, and for noisy signals, it is close to 1 [296].

- **Spectral Crest:** It is a measure of the noisiness/sinusoidality of a spectrum, calculated as the ratio between the maximum value within the band and the arithmetic mean of the energy spectrum:

$$\text{Spectral Crest} = \frac{\max(s_{k\in[b_1,b_2]})}{\frac{1}{b_2-b_1}\sum_{k=b_1}^{b_2} s_k} \tag{6.10}$$

where s_k is the spectral value corresponding to bin k, and b_1 and b_2 are the band edges, in bins, over which to calculate the spectral crest.

- **Spectral Entropy:** It describes the complexity of the spectral distribution:

$$\text{Spectral Entropy} = \frac{-\sum_{k=b_1}^{b_2} s_k \log(s_k)}{\log(b_2 - b_1)} \tag{6.11}$$

where s_k is the spectral value corresponding to bin k, and b_1 and b_2 are the band edges, in bins, over which to calculate the spectral entropy [285].

- **Pitch:** It is the fundamental frequency of the audio signal, such that its integer multiples best explain the content of the signal spectrum [297, 298].

- **Harmonic Ratio (H-Ratio):** It is the ratio between the power of the

fundamental frequency and the total power in an audio frame:

$$H - Ratio = \frac{\sum_{n=1}^{N} s(n)s(n-m)}{\sqrt{\sum_{n=1}^{N} s(n)^2 \sum_{n=0}^{N} s(n-m)^2}}, \quad 1 \leq m \leq M \tag{6.12}$$

where $s(n)$ is a single frame of audio data with N elements, and M is the maximum lag in the calculation [299, 300].

Once features have been extracted, the next step was to eliminate redundant variables preserving the amount of information and increasing computational speed and performances [301], [302]. Among the highly correlated variables (Spearman correlation $\geq$ 75% [301], [302]), the least correlated variables with the output of classification were removed as shown in Figure 6.9, and features min-max normalization was implemented. After this step, the ranking of the univariate features according to the predictor importance score, was performed using chi-square tests [303], [304], [305], [306]. Then the optimal subset of features was defined selecting the highest difference between consecutive scores as the break-point. Finally, the best combination of features was achieved by selecting 6 main features (mfcc3, Age, PATAfreq, Spectral Centroid, Spectral Kurtosis, mfcc8) as shown in Figure 6.10. We tested different techniques for feature selection obtaining comparable results so that we chose the best feature selection technique in terms of the computational cost.

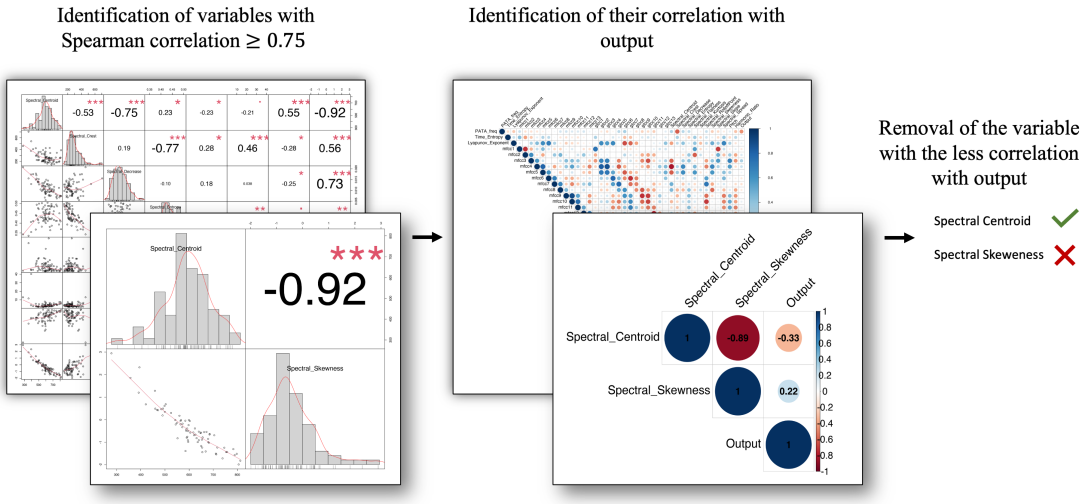


Figure 6.9: First Feature Selection step based on removal of variables with Spearman correlation $\geq$ 75% and lowest correlation with the output. As example, we report the removal of Spectral Skewness with correlation of 89% with Spectral Centroid and 22% with output classes.

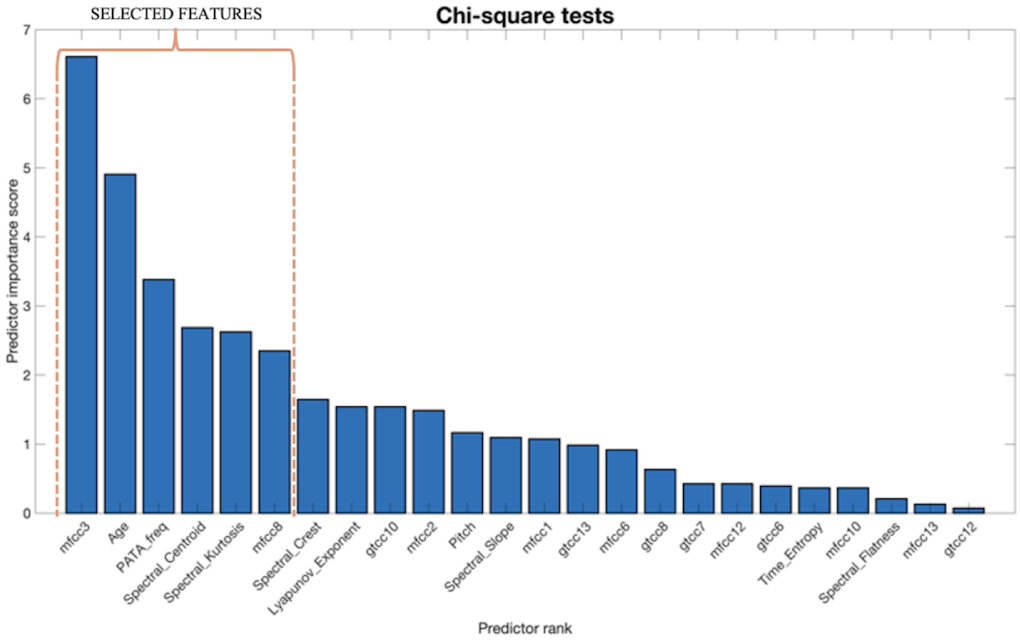


Figure 6.10: Second Feature Selection step based on Chi-square test. Features are ranked, and the break-point is chosen as the highest difference between consecutive scores with the constraint of at least 3 features.

6.2.2.5 Feature Extraction and Selection with Deep Learning

DL Networks are complex architectures used to detect specific features directly from data. They can have hundreds of layers and a huge number of parameters such as weights and biases to be learned. Training from scratch a deep architecture in order to extract specific features avoiding overfitting, requires a large amount of data (hundreds, thousands or even millions, it depends on the application), resulting in high computational and timing costs. Generally, the time of training is related to lots of different factors like the number of epochs, dataset size, computational power, etc., but to reach a certain accuracy even months could be necessary. Usually, GPUs are employed to speed up the process. In many real applications, it is difficult and expensive to obtain training data that match the feature space and predict the distribution characteristics of the test data. Therefore, in practice, there is a need to create a high-performance learner for a target domain trained

from a related source domain. This is the motivation for the transfer learning [307]. Leveraging a pretrained network that has already learned many features on a big dataset to exploit it for a new task, and to specialize the model on a new similar task [276], [308], [309].

There are two main techniques for Transfer Learning:

- **Fine Tuning:** The approach of "fine-tuning" the deeper layers of the pre-trained network on the new dataset is typically much faster and easier than training the model from scratch. Although it requires the least amount of data and computational resources [310], the new dataset must be large enough and similar to the pre-trained one.

- **Feature Extraction:** A more specialized method in which data of the new dataset are passed only once through the pre-trained network and then features are extracted from one of the pools of the network. These features are then used to train a ML model such as SVM, etc. This technique is the most suitable for small datasets.

In this work, the Transfer Learning approach with Feature Extraction was used because of the limited dimensions of the dataset. In particular, we chose the pre-trained Visual Geometry Group CNN (VGGish) CNN [311], [312], developed by Google and inspired by the famous VGG networks used for image classification. Its structure consists of a series of convolution and activation layers, optionally followed by a max pooling layer. The VGGish CNN contains 17 layers in total and it is designed for audio classification tasks. Originally, it was employed to classify the soundtracks of a dataset of 70M training videos (5.24 million hours) with 30,871 video-level labels. In our method, signals were first segmented starting from the first "PA-TA" peak and considering windows with a length of 1 second and 50% overlapped. Then, they were preprocessed to obtain the format required for the network. In particular, they were resampled to 16 kHz, then a one-sided Short-Time Fourier Transform (STFT) was computed, only the magnitude of the complex spectral values was considered, discarding the phase. Finally, the Mel spectrogram was calculated, and it was converted to a log scale. Overlapped segments of 96 spectra were given as input to the network. Activations of the pooling layer "pool 4" were extracted as features to train ML models. We selected "pool 4" since it

was the most discriminative pool layer of the pre-trained model of VGGish CNN [311], [312]. The choice of the pool depends on the similarity between the dataset of the pre-trained model and the dataset of the new application. Since the deeper layers extract higher level features while earlier levels extract lower level ones, the correct depth is as deeper as more similar the datasets are [276], [308], [309]. The structure of VGGish CNN is reported in detail in Table 6.5. The flowchart of features extracted by the VGGish from each data frame of one subject is reported in Figure 6.5. We extracted 12,288 features from layer "pool 4". After the feature extraction step, we selected the best combination of 1,444 deep features using the same approach described in the previous paragraph for ML. Two variables PATAfreq and Age were added also for their high predictive power.

Layer Depth	**Layer Name**	**Layer Type**	**Layer Details**
1	'InputBatch'	Image Input	96x64x1 images
2	'conv1'	Convolution	64 3x3x1 convolutions with stride [1 1] and padding "same"
3	'relu'	ReLU	ReLU
4	'pool1'	Max Pooling	2x2 max pooling with stride [2 2] and padding "same"
5	'conv2'	Convolution	128 3x3x64 convolutions with stride [1 1] and padding "same"
6	'relu2'	ReLU	ReLU
7	'pool2'	Max Pooling	2x2 max pooling with stride [2 2] and padding "same"
8	'conv3_1'	Convolution	256 3x3x128 convolutions with stride [1 1] and padding "same"
9	'relu3_1'	ReLU	ReLU
10	'conv3_2'	Convolution	256 3x3x256 convolutions with stride [1 1] and padding "same"
11	'relu3_2'	ReLU	ReLU
12	'pool3'	Max Pooling	2x2 max pooling with stride [2 2] and padding "same"
13	'conv4_1'	Convolution	512 3x3x256 convolutions with stride [1 1] and padding "same"
14	'relu4_1'	ReLU	ReLU
15	'conv4_2'	Convolution	512 3x3x512 convolutions with stride [1 1] and padding "same"
16	'relu4_2'	ReLU	ReLU
17	'pool4'	Max Pooling	2x2 max pooling with stride [2 2] and padding "same"
18	'fc1_1'	Fully Connected	4096 fully connected layer
19	'relu5_1'	ReLU	ReLU
20	'fc1_2'	Fully Connected	4096 fully connected layer
21	'relu5_2'	ReLU	ReLU
22	'fc2'	Fully Connected	128 fully connected layer
23	'EmbeddingBatch'	ReLU	ReLU
24	'regressionoutput'	Regression Output	Mean-squared-error

Table 6.5: VGGish Network Structure

6.2.2.6 Classification

Classification was conducted by processing audio signals as input to a hierarchical model, which discerns healthy subjects, low severity patients, and high severity patients using the Speech Disturbance score of the SARA Scale as clinical output. As shown in Figure 6.5, we defined binary labels for each level: the first layer discriminates subjects with dysarthria vs healthy, and the second layer, trained only on patients, recognizes speech disturbance severity (Low [0-1] vs High [2-3]). The Speech Disturbance item is one of the eight items that compose the SARA scale. It has a score between 0 (normal) and 6 (anarthria), assigned according to hearing words intelligibility [274].

In our dataset, since the enrolled subjects do not cover the full range of the score, we decided to label it considering the maximum observed value of the 3 to obtain a balanced dataset. Detailed information about the dataset is summarized in Table 6.6. Classification was performed by testing four conventional classifiers: SVM, k-Nearest Neighbors (k-NN), Naïve Bayes (NB), and Decision Tree, and combaining their outputs using the majority voting ensemble technique [313]. In our approach, we used two binary levels of classifiers. The first level discriminates healthy vs patients, and the second level assesses the speech disturbance severity (Low vs High).

Subjects (t0)	**Subjects (t0 + t1)**	**Clinical score range**
Level 1:	**Level 1:**	
18 Healthy	18 Healthy	/
37 Patient	58 Patient	
Level 2:	**Level 2:**	
20 Low	33 Low	Low=[0-1]
17 High	25 High	High=[2-3]

Table 6.6: Dataset Information.

We tested the best combination of features extracted with ML and DL for HMLM and then we compared the results with a flat classification approach (a parallel multi-classifier with three classes: Healthy vs Low severity vs High sever-

ity). Cross-validation techniques such as 5-fold, 10-fold, and leave-one-out were applied to check for overfitting and to avoid data selection bias. Finally, the majority voting ensemble technique was used to aggregate the outputs of the single audio frames into related subjects.

6.2.2.7 Performance Metrics

Classification performances were assessed using Accuracy, Precision, Recall, and F1-Score [314]. These metrics are summarized in Table 6.7. For HMLM, the employed definitions of Precision, Recall, and F1-Score discriminate and weight differently each type of misclassification error, taking into account the output of each level instead of just the final one [314], [315]. As it regards accuracy, we reported the result for each level and the overall one. Given that cross-validation was carried out, we have computed correctly and incorrectly predictions of each class for each fold and we have summed them up at the end of all iterations before calculating the performance measures, as shown in Figure 6.11.

Measure	**Binary Classification**	**Multi-class Classification**	**Hierarchical Classification**
Accuracy	$\frac{tp+tn}{tp+tn+fp+fn}$	$\frac{\sum_{i=1}^{l}(tp_i+tn_i)}{\sum_{i=1}^{l}(tp_i+tn_i+fp_i+fn_i)}/l$	$\frac{\sum_{i=1}^{l}(tp_i+tn_i)}{\sum_{i=1}^{l}(tp_i+tn_i+fp_i+fn_i)}/l$
Precision	$\frac{tp}{tp+fp}$	$P_\mu = \frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l}(tp_i+fp_i)}$	$P_\downarrow = \frac{\lvert C_\downarrow^c \cap C_\downarrow^d \rvert}{\lvert C_\downarrow^c \rvert}$ $P_M = \frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i+fp_i}}{l}$
Recall	$\frac{tp}{tp+fn}$	$R_\mu = \frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l}(tp_i+fn_i)}$	$R_\downarrow = \frac{\lvert C_\downarrow^c \cap C_\downarrow^d \rvert}{\lvert C_\downarrow^d \rvert}$ $R_M = \frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i+fn_i}}{l}$
F1-Score	$\frac{2\cdot(Precision\cdot Recall)}{Precision+Recall}$	$F1S_\mu = \frac{2\cdot(P_\mu\cdot R_\mu)}{P_\mu+R_\mu}$	$F1S_\downarrow = \frac{2\cdot(P_\downarrow\cdot R_\downarrow)}{P_\downarrow+R_\downarrow}$ $F1S_M = \frac{2\cdot(P_M\cdot R_M)}{P_M+R_M}$

Table 6.7: Performance Metrics: In the first column, measures for binary classification are reported: tp represent the true positive, tn the true negative, fp the false positive and fn the false negative. In the second column, the same measures are generalized for a multi-class problem considering many classes C_i. μ and M are referred to micro- and macro-averaging. Finally, in the third column, there are the measures for hierarchical classification: $C_\downarrow^c$ are the subclasses of C assigned by the classifier while $C_\downarrow^d$ are the labels.

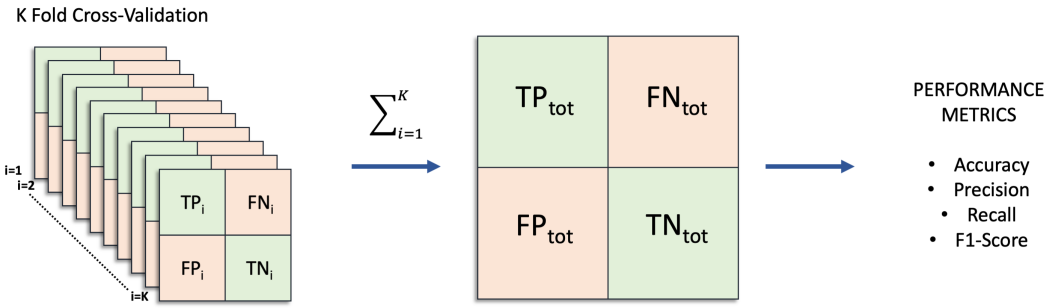


Figure 6.11: K Folds results aggregation. For each typology of cross-validation, we created a unique confusion matrix and then we computed the performance measures.

6.2.3 Results

Results obtained with the HMLM are shown in Table 6.8. We reported the performance of ML with canonical features, ML with DL features, and their combination, respectively for level 1 (healthy vs patients) and level 2 (low vs high severity). All the models were tested with the ensemble majority voting and three different cross-validation techniques (5-fold, 10-fold, and leave-one-out). No significant differences were found between the performance metrics of the three cross-validation methods. The combination of ML (level 1) and DL (level 2) approaches achieved optimal results in discriminating patients from healthy individuals with a mean accuracy of approximately 90%, and in identifying speech disorders severity with an accuracy of about 80%. For level 1 and for level 2 with 5-fold cross-validation (see other details of 10-fold & Leave-one-out in Table 6.8) we achieved a precision of 93.44% and 78.55%, a recall of 98.28% and 79.50%, and an F1-score of 95.80% and 79.02%, respectively. While the overall precision, recall, and F1-score achieved by the model is 84.67%, the overall accuracy obtained is 76.32%.

Measure	Machine Learning			Transfer Learning			Combination: Machine Learning (Level 1) + Transfer Learning (Level 2)		
	5-FOLD	10-FOLD	LEAVE-ONE-OUT	5-FOLD	10-FOLD	LEAVE-ONE-OUT	5-FOLD	10-FOLD	LEAVE-ONE-OUT
Accuracy	1: 88.08% 2: 61.02% Total: 57.89%	1: 86.84% 2: 51.72% Total: 51.32%	1: 88.16% 2: 60.66% Total: 56.58%	1: 78.95% 2: 75% Total: 60.53%	1: 84.11% 2: 77.42% Total: 65.79%	1: 84.21% 2: 78.33% Total: 67.11%	1: 93.42% 2: 78.69% Total: 76.32%	1: 88.16% 2: 75.44% Total: 69.74%	1: 88.16% 2: 78.69% Total: 71.05%
Precision	1: 91.53% 2: 58.52% Total: 73%	1: 91.38% 2: 48.95% Total: 68.48%	1: 90.16% 2: 56.37% Total: 72.37%	1: 87.50% 2: 79.18% Total: 69.96%	1: 87.10% 2: 77.81% Total: 74.73%	1: 88.33% 2: 78.70% Total: 75.66%	1: 93.44% 2: 78.55% Total: 84.67%	1: 92.98% 2: 75.62% Total: 78.93%	1: 90.16% 2: 78.87% Total: 78.61%
Recall	1: 93.10% 2: 55.36% Total: 73%	1: 91.38% 2: 49.02% Total: 68.48%	1: 94.83% 2: 53.66% Total: 72.37%	1: 84.48% 2: 75.52% Total: 69.96%	1: 93.10% 2: 77.81% Total: 79.28%	1: 91.33% 2: 79.86% Total: 75.66%	1: 98.28% 2: 79.50% Total: 84.67%	1: 91.38% 2: 76.60% Total: 78.93%	1: 94.83% 2: 80.24% Total: 78.61%
F1-Score	1: 92.31% 2: 56.90% Total: 73%	1: 91.38% 2: 48.98% Total: 48.95%	1: 92.44% 2: 54.98% Total: 72.37%	1: 85.96% 2: 75.26% Total: 69.96%	1: 90% 2: 78.54% Total: 74.73%	1: 89.33% 2: 79.28% Total: 75.66%	1: 95.80% 2: 79.02% Total: 84.67%	1: 92.98% 2: 75.62% Total: 78.93%	1: 92.44% 2: 79.55% Total: 78.61%

Table 6.8: Hierarchical approach performance metrics [Level 1: Healthy vs Patients – Level 2: Low severity vs High severity]: Detailed performance metrics of HMLM for each level (1-2) in cascade combining ML, transfer learning and ML + transfer learning respectively. Each parameter has been extracted using the ensemble majority voting technique with four classifiers (SVM, k-Nearest Neighbors, Naïve Bayes and Decision Tree).

Detailed information about the collected dataset and the classification output is reported in Figure 6.12, showing the confusion matrix of leave-one-out and in the Supplementary Table A.11.

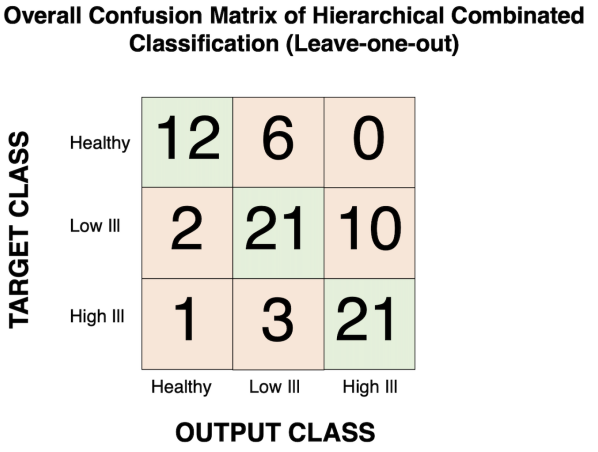


Figure 6.12: Final confusion matrix of the Hierarchical model using leave-one-out validation.

We observed that healthy subjects were never classified as patients with high severity, although they were sometimes confused with the low severity patients. Since many of these patients had a speech disturbance score of 0 as healthy sub-

jects, the two classes were partially overlapped. For the same reason, the network made few mistakes distinguishing patients with low and high severity. Table 6.9 reports results achieved with a flat multi-class approach. In this case, the use of a unique level with three classes reached a maximum overall accuracy of about 65% with the use of the DL approach.

Measure	**Machine Learning**			**Transfer Learning**		
	5-fold	10-fold	Leave-one-out	5-fold	10-fold	Leave-one-out
Accuracy	57.14%	54.23%	57.89%	65.58%	65.89%	65.79%
Precision Macro (M)	63.91%	62.43%	63.97%	66.58%	66.37%	66.05%
Precision Micro (μ)	57.14%	54.23%	57.89%	65.58%	65.89%	65.79%
Recall Macro (M)	41.24%	40.10%	41.59%	48.35%	48.71%	48.66%
Recall Micro (μ)	40%	37.20%	40.74%	48.81%	49.19%	49.02%
F1-Score Macro (M)	50.13%	48.84%	50.41%	55.83%	56.26%	56.15%
F1-Score Micro (μ)	47.06%	44.13%	47.83%	55.97%	56.33%	56.18%

Table 6.9: Flat multi-class approach performance metrics (single model with three-classes [Healthy, Patients with low severity and Patients with high severity]): Detailed performance metrics of flat multi-class approach testing ML and transfer learning. Each parameter has been extracted using the ensemble majority voting technique with four classifiers (SVM, k-Nearest Neighbors, k-Nearest Neighbors, and Decision Tree).

6.2.4 Discussion

For the final step of our research project, we turned our attention to another essential domain of child development: language. As the primary means through which communication occurs, language plays a pivotal role in shaping children's ability to interact, express themselves, and engage socially, making it a cornerstone of their communicative development. Specifically, we leveraged AI methodologies, including ML and DL techniques, to explore innovative strategies in speech analysis for the assessment and stratification of dysarthria. The primary goal was to develop a reliable and accurate tool to support clinical practice by enabling the automatic identification of dysarthria and its severity through audio recordings. Central to this effort was the implementation of a novel HMLM, designed to clas-

sify individuals at two distinct levels: first, distinguishing healthy individuals from those with dysarthria, and second, categorizing the severity of the condition. This structured, two-stage approach optimizes feature selection and enhances classification precision, outperforming traditional flat multi-class models. Importantly, while this model was trained on a dataset including children with ataxia, it is generalizable and equally applicable to the assessment of dysarthria in children with NDD. By providing precise speech evaluation, the model has the potential to enhance the accuracy of assessments and support the development of personalized therapeutic interventions.

6.2.4.1 Interpretation of Results

This study represents the first application of an HMLM for dysarthria assessment, particularly through the standardized speech-based "PA-TA" test. Preprocessing and segmenting the "PA-TA" signals enabled the implementation of two binary classifiers in cascade, enhanced by ensemble majority voting techniques. Conventional and DL models were integrated into three configurations, ML, DL, and a hybrid approach, tested across three cross-validation methods: 5-fold, 10-fold, and leave-one-out.

The results demonstrated that conventional features performed more effectively in differentiating healthy individuals from those with dysarthria at the first classification level, while transfer learning features exhibited superior performance in assessing severity at the second level. Across all cross-validation methods, the HMLM consistently outperformed the flat classification approach, achieving significantly higher overall accuracies: 76.32% versus 65.58% (5-fold), 69.74% versus 65.89% (10-fold), and 71.05% versus 65.79% (leave-one-out). These consistent outcomes underscore the robustness of the HMLM framework in dysarthria assessment. Furthermore, the findings highlight the model's capacity to capture subtle variations in speech patterns that are often challenging to detect, suggesting that the hierarchical structure, by tailoring feature sets to each classification level, is particularly effective in improving the detection of early signs and severity of dysarthria.

6.2.4.2 Significance and Value

This research introduces significant advancements in AI-based dysarthria assessment by integrating conventional and DL approaches within a hierarchical model. By optimizing feature extraction at each classification level, the HMLM provides a more nuanced and precise analysis of speech patterns compared to traditional methods. The study also introduces a structured dataset tailored for dysarthria research, addressing a critical gap in available resources. This dataset, supports the development of AI models capable of capturing a wide range of speech characteristics, including subtle distinctions in acoustic parameters.

Beyond its technical contributions, this study holds practical implications for clinical applications, particularly in telemedicine. By enabling remote monitoring and diagnosis through simple voice recordings, the proposed framework offers a scalable solution that reduces the need for in-person assessments. This is especially beneficial for individuals in underserved areas or those with limited access to specialized care. Additionally, the system's ability to evaluate dysarthria in children with Neurodevelopmental Disorders opens new avenues for integration into therapeutic workflows. Providing objective, precise metrics for assessing speech impairments could significantly enhance early detection, facilitate personalized therapy, and improve overall care outcomes. These advancements represent a significant step toward more accessible, efficient, and patient-centered care for individuals with speech impairments.

6.2.4.3 Limitations & Future Improvements

In this case, too, the primary limitation of the study is related to the dataset. The relatively small sample size, compounded by the rarity of the condition and the narrow range of severity scores (0–6), constrained the statistical variability and generalizability of the findings. Nevertheless, the dataset represents a foundational resource for future dysarthria research, offering a starting point for more extensive investigations.

Analysis of the cross-validation results revealed misclassifications primarily due to overlapping clinical scores across different classes, such as between healthy individuals and those with mild dysarthria or between mild and severe cases. This

overlap highlights the inherent challenges in distinguishing subtle variations in speech patterns, which are not always consistently captured by clinical scoring systems. These findings emphasize the need for larger and more diverse datasets to improve classification accuracy and help identify more homogeneous phenotypes. Future work should focus on expanding the dataset to include a broader spectrum of severity and more diverse speech profiles. Additionally, the development of longitudinal datasets would enable exploration of dysarthria progression over time, providing deeper insights into its developmental trajectories. Such efforts could lead to the refinement of AI-based tools and their integration into broader clinical workflows, enhancing the accuracy and accessibility of dysarthria assessment and treatment strategies. The system's applicability to NDD holds promise for supporting clinicians in assessing speech impairments, improving therapeutic decision-making, and optimizing outcomes for diverse patient populations.

Chapter 7

Conclusions

The results presented in the preceding chapters provide critical insights into the multifaceted nature of NDD, shedding light on their manifestations and progression across various developmental domains. Leveraging a range of advanced AI techniques, these findings unveil novel patterns within both typical and divergent neurodevelopmental processes, thereby contributing to a deeper understanding of these complex phenomena. From the early stages of this research, particular emphasis was placed on exploring the interplay between developmental domains to comprehend the emergence and evolution of Neurodevelopmental Disorders. Central to this investigation was the search for novel early behavioral biomarkers capable of predicting atypical trajectories from the first weeks of life and examining how these deviations unfold throughout growth and across functional areas. Such insights are essential for guiding timely and targeted interventions, particularly in light of the limitations of traditional approaches, which often overlook the complexities of early development within the multimodal framework of the neurodevelopmental cascade.

AI has emerged as a transformative tool in this context, offering unprecedented sensitivity in detecting subtle behavioral and kinematic patterns. Its application enables a more nuanced understanding of the dynamic connections between motor, cognitive, and social processes, while simultaneously facilitating the development of scalable, automated systems that enhance the accuracy and accessibility of early diagnosis and developmental evaluation. By investigating how these domains inter-

act and influence one another over time, this dissertation underscores the potential of an integrated, AI-driven approach to uncover both early indicators and the cascading effects characteristic of NDD. The findings presented advance theoretical knowledge and offer practical implications for clinical practice, highlighting how early detection through sophisticated computational models can guide intervention strategies and improve long-term developmental outcomes.

This chapter synthesizes the main conclusions drawn from the research, emphasizing the significance of the observed patterns and their implications for understanding neurodevelopmental trajectories. Additionally, it outlines prospective avenues for future work, aiming to build upon the progress achieved and further explore the complex interactions underlying neurodevelopment, with a continued focus on leveraging technological innovations to refine early identification and intervention methodologies.

7.1 Summary of Contributions

Within the scope of this research project, we employed advanced AI methodologies to enhance the early detection and understanding of neurodevelopmental delays by identifying key behavioral biomarkers and distinguishing unique patterns in neurodivergent children across multiple developmental domains. Grounded in the "neurodevelopmental cascade" theory, our work aimed to optimize the evaluation of early indicators, such as movements, gestures, speech, social interaction, visual attention, and emotional expression, to facilitate personalized diagnostics and timely interventions. This integrative approach ensured that each domain, motor, social, communicative, and cognitive, was analyzed within a unified framework, emphasizing their interdependence and the collective impact on developmental trajectories.

Spanning critical milestones from infancy to middle childhood, the project employed a multimodal approach that is organized into the following chapters:

7.1.1 Motor Development in Early Infancy: From Spontaneous Movements to Reaching Behaviors

Marker-less Analysis of Spontaneous Movements in Newborns

This part of the research introduces a novel, AI-based, marker-less approach for automatically tracking newborns' spontaneous movements from recorded videos, spanning five distinct time points between 10 days and 24 weeks of age. Kinematic parameters extracted from these trajectories enabled the characterization of GMs and the identification of motor delays in the lower limbs, in infants later diagnosed with NDD. Notably, the most pronounced differences appeared at the earliest assessment, shortly after birth, gradually diminishing over time but later manifesting as delays in other developmental domains. These findings highlight the need for methods that can detect early signs of abnormal behavior in infants, allowing for timely and personalized interventions.

AI Analysis of Infants' Hands Movements in Social Engagement and Object-Reaching Tasks:

By the age of six months, our focus expanded to include the analysis of hands movements during interactive tasks. We compared typically developing infants with those who later exhibited language delays at 36 months. Our investigation revealed significant differences in hands movement patterns during both social engagement and object-reaching tasks. These differences were consistently identified using both AI-based video tracking and kinematic sensor data. In particular, results demonstrated that variations in hands velocity, especially in rotational movements, and differences in reaching times were crucial in distinguishing between the two groups. These results illustrate how early motor abilities are intricately connected to later communicative and cognitive outcomes, forming a cascade in which early anomalies influence subsequent developmental challenges.

DL improved these analyses, creating a fully automated tool that works with basic video devices like smartphones. This design has the potential to make early screening easier in both clinical and home settings, allowing families and healthcare providers to monitor developmental progress. By providing a scalable, mobile-friendly solution, this approach enables proactive care and early detection of developmental challenges. Using AI for marker-less movement analysis, this work offers

practical tools to help caregivers identify developmental issues early and improve outcomes for at-risk infants.

7.1.2 Emerging Communicative Skills in Toddlers: Gestures, Gaze and Vocalizations in Naturalistic Interactions

Early Multimodal Behavioral Cues in Autism: A Microanalytic Exploration of Actions, Gestures and Speech during Naturalistic Parent-Child Interactions:
In this chapter, we moved beyond single-domain analyses by adopting a multimodal approach to early autism characterization, examining a range of motor and socio-communicative behaviors and their integration. Specifically, we employed a microanalytic approach to examine naturalistic parent-child interactions, focusing on the production and coordination of communicative gestures, with language and gaze. Our findings indicate that autistic children exhibit fewer and less diverse communicative behaviors compared to NT peers. In particular, ASC children show limited use of pointing, reduced alternation between looking at objects and social partners and more vocalizations than words. They also showed a preference for object manipulation over functional play. In contrast, NT children demonstrated greater use of declarative gestures, frequent gaze alternation, and behaviors indicative of social interest and a desire to share experiences.

A Deep Learning Approach for Automatic Video Coding of Deictic Gestures in Children with Autism:
To overcome the labor-intensive nature of manual coding, we developed an AI-based system utilizing a transformer-based DL model. This system processes entire video sequences holistically, thereby capturing the continuity of social interactions and accurately distinguishing subtle differences in gesture production. The approach provides a scalable tool for the automatic characterization of socially relevant behaviors by recognizing four key deictic gestures—showing, requesting, giving, and pointing—that are essential for understanding early social behaviors.

Built on a flexible Python-based system, it can easily analyze new video data and train models for various applications in gesture recognition and neurodevelop-

mental research. Moreover, its adaptability makes it suitable for both clinical and home settings. Integrating this technology into a broader multimodal framework provides scalable tools for identifying early markers of autism and related conditions. Automating complex behavioral analyses further enhances early screening efforts and helps develop personalized treatment protocols for each child.

7.1.3 Advancing Visual Attention to Social Cues in Preschoolers: Exploring Gaze Patterns Development

Decoding Social Attention in Preschoolers: A New Eye-Tracking Paradigm with Markov Chain Analysis:

In this study, we applied CTMCs and PCA to examine gaze transitions between social and nonsocial elements during interactive play tasks, such as dyadic social interactions (SSRs) and object-based musical activities. Our analysis revealed that neurotypical children display a clear preference for social stimuli, frequently alternating their gaze between faces and social partners—especially during dyadic SSRs, where motor imitation during songs emphasizes joint attention and shared engagement. In contrast, children with ASC exhibit less gaze triangulation, more frequent transitions among nonsocial elements (e.g., distractor objects and activity areas), and a reduced tendency to reorient their gaze to faces after distractions. They also engage in repetitive shifts between nonsocial AOIs, indicating different environmental scanning strategies. These patterns suggest that gaze disengagement is a critical marker of social attention challenges in autism, potentially reflecting reduced social motivation, alternative attentional strategies, or coping mechanisms for sensory overload. Moreover, the variability in gaze patterns within the ASC group—as highlighted by PCA—underscores the need for flexible, individualized interventions tailored to diverse sensory and attentional profiles.

7.1.4 Neurodivergent Trajectories in Middle Childhood: Effects in Expressive Kinematics and Speech

Exploring Divergent Kinematics in Children with NDD across Social and Non-Social Vitality Forms:

This study examined how children with ASC and NT peers express different vitality forms (VFs), the dynamic, affective qualities of actions, during motor tasks. Participants were required to move a small bottle under both social (handing the bottle to a receiver) and non-social (moving it to a designated point) conditions while expressing neutral, gentle, or rude VFs. Our detailed kinematic analysis revealed that although children with ASC can modulate their motor profiles to express both gentle and rude VFs, key kinematic parameters such as mean acceleration and maximum deceleration differ significantly from those observed in neurotypical children. Moreover, the social context appeared to attenuate the expression of negative VFs in children with ASC, with the extent of modulation correlating with autism severity. These findings highlight the intricate relationship between motor behavior and social communication skills.

Artificial Intelligence for Speech Assessment in Children: A Hierarchical Approach:
In our final study, we developed an AI-based hierarchical model (HMLM) for the assessment of dysarthria using the standardized "PA-TA" speech test. The model employs a two-stage classification process: first, it distinguishes healthy individuals from those with dysarthria, and then it stratifies the severity of the condition. By integrating conventional ML techniques with DL methods, leveraging frequency-domain parameters and features extracted via CNNs, the HMLM consistently outperformed flat multi-class models across various cross-validation methods. This hierarchical approach is capable of capturing subtle variations in speech patterns, providing an objective and scalable tool for dysarthria assessment in children with NDD. Its ability to enable remote, non-invasive monitoring further highlights its clinical value.

By integrating sensory, cognitive, and behavioral dimensions within a unified framework, this research underscores the transformative role of AI in early neurodevelopmental assessment. The findings presented not only enhance our understanding of NDD trajectories but also provide a robust foundation for data-driven, personalized diagnostic and intervention strategies. Across the studies presented, the cascading nature of neurodevelopmental processes becomes increasingly evident. Early deviations in motor skills and gestures are linked to later socio-

communicative delays, while gaze behavior patterns offer crucial insights into attentional allocation in social contexts. These interconnections highlight the need for an integrated approach that captures the dynamic interplay between motor, communicative, and attentional domains. By adopting this perspective, this dissertation advances both theoretical knowledge and practical applications, paving the way for more precise, multimodal early detection and intervention methodologies.

7.2 Future Work

Building upon the insights gained from this research and addressing its current limitations, future studies should aim to develop a comprehensive multimodal framework for evaluating child development. This framework should integrate data across multiple domains—including motor skills, gesture production, gaze behavior, and speech—to capture their interdependencies and cascading effects. By doing so, it will provide a holistic and dynamic understanding of neurodevelopment, moving beyond isolated domain analyses toward an integrated perspective on early risk markers and their longitudinal impact.

A critical step forward involves expanding and diversifying datasets to improve the generalizability of AI-driven models across different populations, particularly those with diverse genetic and environmental risk factors for NDD. Large-scale, longitudinal studies will be essential to track developmental trajectories over time, revealing how early deviations in motor, communicative, and attentional behaviors evolve and interact throughout childhood. Such studies, although challenging due to the need for extensive follow-up data, are crucial for refining predictive models and enhancing early intervention strategies.

In motor tracking, the next challenge is to develop even more precise and accessible tools for detecting early motor anomalies. Future efforts should focus on refining AI-driven movement analysis to improve sensitivity to subtle kinematic variations and better characterize the interplay between early motor skills and later cognitive and communicative development. Additionally, integrating multimodal data—such as linking motor assessments with gaze behavior and social interaction metrics—may provide a richer understanding of how early motor function

contributes to broader developmental trajectories.

For gesture analysis, automating the recognition of key communicative gestures in naturalistic settings remains a priority. While deep learning-based tools have demonstrated promise, improving their accuracy and robustness requires expanding training datasets, optimizing feature extraction pipelines, and developing methods to handle class imbalances in rare but clinically significant behaviors. A future research direction could involve integrating gesture recognition with real-time eye-tracking and speech processing, allowing for a more complete characterization of multimodal communication in young children.

In the domain of gaze behavior, analyzing visual attention patterns across diverse tasks and social contexts will yield richer insights into how attentional strategies adapt across developmental stages. Future work should explore the integration of gaze tracking with neural and physiological markers (e.g., EEG or pupillometry) to better understand the underlying mechanisms driving social attention differences in ASC. Additionally, real-time AI-based gaze analysis could be leveraged for interactive, adaptive screening tools in both clinical and home settings, enhancing early detection accessibility.

For speech analysis, advancing models capable of detecting early communicative disturbances from unstructured audio and video data is a crucial next step. Future research should focus on multi-source data fusion, combining speech processing with motor and gesture analysis to refine diagnostic precision. Moreover, developing lightweight, mobile-compatible AI tools will be essential for real-world applications, ensuring that early screening can be seamlessly integrated into naturalistic environments without requiring structured assessment protocols. Collaborations with clinicians and therapists will play a key role in aligning these technological advancements with practical intervention needs.

By addressing these research priorities, future studies can bridge the gap between AI-driven developmental assessment and clinical applications, creating scalable, precise, and accessible tools for the early detection of neurodevelopmental challenges. The ultimate goal is to develop an integrated, multimodal framework that combines motor, social, and communicative data into a unified model, offering a transformative approach to early screening and intervention. This paradigm shift holds the potential to revolutionize the field of neurodevelopmental research, pro-

viding data-driven, personalized, and adaptive solutions for improving outcomes in neurodivergent children.

References

[1] Deborah J. Morris-Rosendahl and Marc-André Crocq. Neurodevelopmental disorders—the history and future of a diagnostic concept. *Dialogues in Clinical Neuroscience*, 22(1):65–72, March 2020.

[2] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)*. American Psychiatric Publishing, Washington, DC, 5th edition, 2013.

[3] Istituto di Neuroscienze. Istituto di neuroscienze del cnr website, 2024.

[4] World Health Organization. World health organization website, 2024.

[5] Yao Yang, Shuo Zhao, Min Zhang, Mengmeng Xiang, Jia Zhao, Shun Chen, Hong Wang, Li Han, and Jing Ran. Prevalence of neurodevelopmental disorders among us children and adolescents in 2019 and 2020. *Frontiers in Psychology*, 13:997648, November 2022.

[6] Environmental Protection Agency. Environmental protection agency website, 2024.

[7] Istituto Superiore di Sanità. Istituto superiore di sanità website, 2024.

[8] E. Mark Mahone and Martha B. Denckla. Attention-deficit/hyperactivity disorder: A historical neuropsychological perspective. *Journal of the International Neuropsychological Society*, 23(9-10):916–929, October 2017.

[9] Elliot H. Sherr. Chapter 36 - neurodevelopmental disorders, causes, and consequences. In Thomas Lehner, Bruce L. Miller, and Matthew W. State, ed-

itors, *Genomics, Circuits, and Pathways in Clinical Neuropsychiatry*, pages 587–599. Academic Press, San Diego, 2016.

[10] Johanna Löytömäki, Marja-Leena Laakso, and Kaisa Huttunen. Social-emotional and behavioural difficulties in children with neurodevelopmental disorders: Emotion perception in daily life and in a formal assessment context. *Journal of Autism and Developmental Disorders*, 53(12):4744–4758, December 2023.

[11] Alessandro De Felice, Laura Ricceri, Alessandra Venerosi, Fiorenzo Chiarotti, and Gemma Calamandrei. Multifactorial origin of neurodevelopmental disorders: Approaches to understanding complex etiologies. *Toxics*, 3(1):89–129, March 2015.

[12] Paul A. Eubig, Andreza Aguiar, and Susan L. Schantz. Lead and pcbs as risk factors for attention deficit/hyperactivity disorder. *Environmental Health Perspectives*, 118(12):1654–1667, December 2010.

[13] Omar Y. Muthaffar, Abdullah Y. Abbar, and Maha T. Fitaih. Prevalence of seizures in children diagnosed with neurodevelopmental disorders. *Cureus*, 16(7):e63765, July 2024.

[14] Masahiro Doi, Naomi Usui, and Shuichi Shimada. Prenatal environment and neurodevelopmental disorders. *Frontiers in Endocrinology (Lausanne)*, 13:860110, March 2022.

[15] Eleni Bonti, Ioanna K. Zerva, Christina Koundourou, and Maria Sofologi. The high rates of comorbidity among neurodevelopmental disorders: Reconsidering the clinical utility of distinct diagnostic categories. *Journal of Personalized Medicine*, 14(3):300, March 2024.

[16] B. V. Prasad, V. Patil, and K. K. Sony. Neurodevelopmental disorders: Role of non-invasive neuromodulation therapies. *Annals of Neurosciences*, 31(2):77–79, April 2024.

[17] Daniel Z. Wetmore and Craig C. Garner. Emerging pharmacotherapies for neurodevelopmental disorders. *Journal of Developmental and Behavioral Pediatrics*, 31(7):564–581, September 2010.

[18] Michael J. Wesley and Joshua A. Lile. Combining noninvasive brain stimulation with behavioral pharmacology methods to study mechanisms of substance use disorder. *Frontiers in Neuroscience*, 17:1150109, July 2023. This article is part of the Research Topic: New Discoveries in the Field of Brain Stimulation and Addiction Disorders.

[19] Alex Tseng, Bruno Biagianti, Sarah M. Francis, Christine A. Conelea, and Stacey Jacob. Social cognitive interventions for adolescents with autism spectrum disorders: A systematic review. *Journal of Affective Disorders*, 274:199–204, September 2020.

[20] Rebecca Landa. Efficacy of early interventions for infants and young children with, and at risk for, autism spectrum disorders. *International Review of Psychiatry*, 30:1–15, March 2018.

[21] Yvette Hus and Osnat Segal. Challenges surrounding the diagnosis of autism in children. *Neuropsychiatric Disease and Treatment*, 17:3509–3525, 2021.

[22] S. Maleki Varnosfaderani and M. Forouzanfar. The role of ai in hospitals and clinics: Transforming healthcare in the 21st century. *Bioengineering (Basel)*, 11(4):337, March 2024.

[23] Esther Thelen. *Motor development: A new synthesis*. Harvard University Press, 1992.

[24] Esther Thelen and Linda B. Smith. *A dynamic systems approach to the development of cognition and action*. MIT Press, 1994.

[25] K. E. Adolph and C. S. Tamis-LeMonda. The costs and benefits of development: The transition from crawling to walking. *Child Development Perspectives*, 8(4):187–192, 2014.

[26] I. Babik, J. C. Galloway, and M. A. Lobo. Early exploration of one's own body, exploration of objects, and motor, language, and cognitive development related dynamically across the first two years of life. *Developmental Psychology*, 58(2):222–235, 2022.

[27] Jana M. Iverson. Early motor development and its relation to social and communicative skills in infancy. *Developmental Psychobiology*, 63(5):637–645, 2021.

[28] Ann S. Masten and Dante Cicchetti. Developmental cascades. *Development and Psychopathology*, 22(3):491–495, 2010.

[29] Gilbert Gottlieb. *Developmental-behavioral genetics*. Elsevier, 2003.

[30] Esther Thelen and Linda B. Smith. *Dynamic systems theories of development*. MIT Press, 2003.

[31] Arnold J. Sameroff. The transactional model of development: How children and contexts shape each other. *American Psychological Association*, 2009.

[32] Annette Karmiloff-Smith. *Beyond modularity: A developmental perspective on cognitive science*. MIT Press, 1998.

[33] L. X. Guo, A. Pace, L. R. Masek, R. M. Golinkoff, and K. Hirsh-Pasek. Cascades in language acquisition: Re-thinking the linear model of development. *Advances in Child Development and Behavior*, 64:69–95, 2023.

[34] Amanda E. Guyer, Koraly Pérez-Edgar, and Eveline A. Crone. Opportunities for neurodevelopmental plasticity from infancy through early adulthood. *Child Development*, 89(3):687–697, May 2018.

[35] L. M. Oakes and D. Rakison. Developmental cascades in infancy. *Child Development Perspectives*, 13(4):240–245, 2019.

[36] Ann S. Masten and Dante Cicchetti. Developmental cascades in developmental psychopathology. *Development and Psychopathology*, 17:491–495, 2005.

[37] L. M. Oakes. The development of visual attention in infancy: A cascade approach. *Advances in Child Development and Behavior*, 64:1–32, 2023.

[38] Klaus Libertus. Editorial: Motor skills and their foundational role for perceptual, social, and cognitive development. *Frontiers in Psychology*, 8:301, March 2017.

[39] B. I. Bertenthal and C. Von Hofsten. The relation between visual and motor development in infancy. *Current Directions in Psychological Science*, 7(2):61–65, 1998.

[40] Stephanie A. Soska and Karen E. Adolph. Opportunities for learning and social interaction in infant sitting experiences. *Infancy*, 19(3):255–269, 2014.

[41] Jean Yingling. *Temporal Features of Infant Speech: A Description of Babbling Patterns Circumscribed by Postural Achievement*. PhD thesis, University of Denver, 1981. Unpublished doctoral dissertation.

[42] Erin N. Jarvis, Kelsey L. West, and Jana M. Iverson. Object exploration during the transition to sitting: A study of infants at heightened risk for autism spectrum disorder. *Infancy*, 25(5):640–657, 2020.

[43] Miranda M. Mlincek, Emily J. Roemer, Christen Kraemer, and Jana M. Iverson. Posture matters: Object manipulation during the transition to arms-free sitting in infants at elevated vs. typical likelihood for autism spectrum disorder. *Physical & Occupational Therapy in Pediatrics*, 42(4):351–365, 2022.

[44] Sally Ozonoff, Gregory S. Young, Alice Carter, Daniel Messinger, Nurit Yirmiya, Lonnie Zwaigenbaum, et al. Recurrence risk for autism spectrum disorders: A baby siblings research consortium study. *Pediatrics*, 128(3):e488–e495, 2011.

[45] Jana M. Iverson and Susan Goldin-Meadow. Gesture paves the way for language development. *Psychological Science*, 16(5):367–371, 2005.

[46] Yu Ye, Alex Heckman, Leslie K. MacNeil, Mary G. Baxter, Daniel S. Messinger, and Jana M. Iverson. The gestures in 2–4-year-old children with autism spectrum disorder. *Frontiers in Psychology*, 12:604542, 2021.

[47] Eve Sauer LeBarton and Jana M. Iverson. Associations between gross motor and communicative development in at-risk infants. *Infant Behavior and Development*, 44:59–67, 2016.

[48] Amy I. Mendez, Heather Tokish, Emily McQueen, Suchitra Chawla, Ami Klin, Nathalie L. Maitre, and Cheryl Klaiman. A comparison of the clinical presentation of preterm birth and autism spectrum disorder: Commonalities and distinctions in children under 3. *Clinics in Perinatology*, 50(1):81–101, March 2023.

[49] Linda R. Watson, Elizabeth R. Crais, Grace T. Baranek, Jennifer R. Dykstra, and Kathleen P. Wilson. Communicative gesture use in infants with and without autism: a retrospective home video study. *American Journal of Speech-Language Pathology*, 22(1):25–39, 2013.

[50] Christos Ioannou, Frouke Hermens, Paul Hodgkins, Charmaine Plisson, Connor Kelly, and Elizabeth Milne. Comorbidity matters: Social visual attention in a comparative study of autism spectrum disorder, attention-deficit/hyperactivity disorder and their comorbidity. *Journal of Abnormal Child Psychology*, 48(4):579–593, 2020.

[51] Francesca Canu, Elisabetta Livi, Laura Piccardi, Cecilia Guariglia, Stefano Vicari, and Deny Menghini. Evidence towards a continuum of impairment across neurodevelopmental disorders from basic ocular-motor tasks. *Frontiers in Psychology*, 13:825049, 2022.

[52] Ami Klin, Sarah Shultz, and Warren Jones. Affording autism an early brain development re-definition. *Development and Psychopathology*, 32(4):1175–1189, 2020.

[53] E. Braithwaite, V. Kyriakopoulou, L. Mason, et al. Objective assessment of visual attention in toddlerhood. *eLife*, 12, 2023.

[54] Amy Pace, Rebecca Alper, Margaret Burchinal, Kathy Hirsh-Pasek, and Roberta M. Golinkoff. Identifying pathways between socioeconomic status and language development. *Annual Review of Linguistics*, 3:285–308, 2017.

[55] Kathy Hirsh-Pasek, Lauren B. Adamson, Roger Bakeman, Margaret T. Owen, Roberta M. Golinkoff, Amy Pace, Pamella K. S. Yust, and Catherine S. Tamis-LeMonda. The contribution of early communication quality to low-income children's language success. *Psychological Science*, 26(7):1071–1083, 2015.

[56] J. M. Iverson, K. L. West, J. L. Schneider, S. N. Plate, J. B. Northrup, and E. Roemer Britsch. Early development in autism: How developmental cascades help us understand the emergence of developmental differences. *Advances in Child Development and Behavior*, 64:109–134, 2023.

[57] M. G. Carelli et al. Eye-tracking insights into gaze behavior in neurodevelopmental disorders. *Advances in Child Development and Behavior*, 64:33–38, 2023.

[58] Mehak Mengi and Deepti Malhotra. Artificial intelligence based techniques for the detection of socio-behavioral disorders: A systematic review. *Archives of Computational Methods in Engineering*, 29, 2021.

[59] C. Lord, M. Rutter, P. C. DiLavore, and S. Risi. Autism diagnostic observation schedule: A standardized observation of communicative and social behavior. *Journal of Autism and Developmental Disorders*, 19(2):185–212, 1989.

[60] Catherine Lord, Michael Rutter, and Ann Le Couteur. Autism diagnostic interview—revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders*, 24(5):659–685, 1994.

[61] Christa Einspieler and Heinz F. R. Prechtl. Prechtl's assessment of general movements: a diagnostic tool for the functional assessment of the young nervous system. *Mental Retardation and Developmental Disabilities Research Reviews*, 11(1):61–67, 2005.

[62] Małgorzata Eliks and Ewa Gajewska. The alberta infant motor scale: A tool for the assessment of motor aspects of neurodevelopment in infancy and early childhood. *Frontiers in Neurology*, 13:927502, 2022.

[63] Nadia C. Valentini, Larissa Wagner Zanella, and Fernando Copetti. Peabody developmental motor scales - second edition. In *Fisioterapia Neuropediátrica: Abordagem Biopsicossocial*, pages 111–116. Manole, 2022.

[64] Christa Einspieler, Arend F. Bos, Megan E. Libertus, and Peter B. Marschik. The general movement assessment helps us to identify preterm infants at risk for cognitive dysfunction. *Frontiers in Psychology*, 7:406, 2016.

[65] Michael G. O'Grady and Stacey C. Dusing. Reliability and validity of play-based assessments of motor and cognitive skills for infants and young children: A systematic review. *Physical Therapy*, 95(1):25–38, 2015.

[66] Andrea Zunino. Video gesture analysis for autism spectrum disorder detection. In *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018.

[67] A. Hill, Á. MacNamara, and D. Collins. Psycho-behaviourally based features of effective talent development in rugby union: a coach's perspective. *Sport Psychologist*, 29:201–212, 2015.

[68] Meredith L. Rowe. A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child Development*, 83(5):1762–1774, 2012.

[69] Jana M. Iverson and Barbara A. Braddock. Gesture and motor skill in relation to language in children with language impairment. *Journal of Speech, Language, and Hearing Research*, 54(1):72–86, 2011.

[70] Olga Chorna, Andrea Guzzetta, and Nathalie L. Maitre. Correlation between early visual functions and cognitive outcome in infants at risk for cerebral palsy or other neurodevelopmental disorders: A systematic review. *Neuroscience & Biobehavioral Reviews*, 156:105196, 2024.

[71] Anna Rebreikina, Dmitry Zakharchenko, Antonina Shaposhnikova, et al. Voluntary attention assessing tests in children with neurodevelopmental disorders using eye tracking. *Children*, 11(11):1333, 2024.

[72] Elisabeth H. Wiig, Eleanor Semel, and Wayne A. Secord. *Clinical Evaluation of Language Fundamentals: Fifth Edition (CELF-5).* Pearson, San Antonio, TX, 2013.

[73] Douglas M. Dunn. *Peabody Picture Vocabulary Test: Fifth Edition (PPVT-5).* Pearson, Bloomington, MN, 2019.

[74] Kari Fulcher-Rood, Anny Castilla-Earls, and Julie Higginbotham. Diagnostic decisions in child language assessment: Findings from a case review assessment task. *Language, Speech, and Hearing Services in Schools*, 50(3):385–398, 2019.

[75] Rhea Paul and Courtenay Frazier Norbury. *Language Disorders from Infancy Through Adolescence: Listening, Speaking, Reading, Writing, and Communicating.* Mosby, 4th edition, 2012.

[76] Eleanor Semel, Elisabeth H. Wiig, and Wayne A. Secord. *Clinical Evaluation of Language Fundamentals - Fourth Edition (CELF-4).* Psychological Corporation, 2003.

[77] L. M. Dunn and D. M. Dunn. *Peabody Picture Vocabulary Test–Fourth Edition (PPVT-4).* APA PsycTests, 2007.

[78] Jon Miller. Using language sample analysis to assess spoken language production in adolescents. *Language, Speech, and Hearing Services in Schools*, 47(2):1–13, 2016.

[79] Jadyn Trayvick, Sarah B. Barkley, Alessia McGowan, et al. Speech and language patterns in autism: Towards natural language processing as a research and clinical tool. *Frontiers in Psychology*, 15, 2024.

[80] Samuel V. Wass, Celia G. Smith, Louise Stubbs, Kaili Clackson, and Farhan U. Mirza. Physiological stress, sustained attention, emotion regulation, and cognitive engagement in 12-month-old infants from urban environments. *Developmental Psychology*, 57(8):1179–1194, 2021.

[81] Kevin B. Johnson, Wei-Qi Wei, Dinith Weeraratne, Mark E. Frisse, Katherine Misulis, Kyoungmi Rhee, Jie Zhao, and Jennifer L. Snowdon. Precision medicine, ai, and the future of personalized health care. *Clinical and Translational Science*, 14(1):86–93, January 2021.

[82] Steffi L. Colyer, Murray Evans, Darren P. Cosker, and Aki I. T. Salo. A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system. *Sports Medicine - Open*, 4(1):24, 2018.

[83] Marco Leo, Giuseppe Massimo Bernava, Pierluigi Carcagnì, and Cosimo Distante. Video-based automatic baby motion analysis for early neurological disorder diagnosis: State of the art and future directions. *Institute of Applied Sciences and Intelligent Systems, National Research Council of Italy*, 2024.

[84] Nelson Silva, Dajie Zhang, Tomas Kulvicius, Alexander Gail, Carla Barreiros, Stefanie Lindstaedt, Marc Kraft, Sven Bölte, Luise Poustka, Karin Nielsen-Saines, Florentin Wörgötter, Christa Einspieler, and Peter B. Marschik. The future of general movement assessment: The role of computer vision and machine learning - a scoping review. *Research in Developmental Disabilities*, 114:103854, 2021.

[85] Muhammad Tausif Irshad, Muhammad Adeel Nisar, Philip Gouverneur, Marion Rapp, and Marcin Grzegorzek. Ai approaches towards prechtl's assessment of general movements: A systematic literature review. *Sensors*, 20(18):5321, 2020.

[86] Lars Adde, Jorunn L. Helbostad, Alexander R. Jensenius, Gunnar Taraldsen, Kristine H. Grunewaldt, and Ragnhild Støen. Early prediction of cerebral palsy by computer-based video analysis of general movements: a feasibility study. *Developmental Medicine & Child Neurology*, 52(3):268–275, 2010.

[87] Chiara Tacchino, Martina Impagliazzo, Erika Maggi, Marta Bertamino, Isa Blanchi, Francesca Campone, Paola Durand, Marco Fato, Psiche Giannoni, Riccardo Iandolo, Massimiliano Izzo, Pietro Morasso, Paolo Moretti, Luca Ramenghi, Keisuke Shima, Koji Shimatani, Toshio Tsuji, Sara Uccella,

Nicolò Zanardi, and Maura Casadio. Spontaneous movements in the newborns: a tool of quantitative video analysis of preterm babies. *Computer Methods and Programs in Biomedicine*, 199:105838, 2021.

[88] Annette Stahl, Christian Schellewald, Øyvind Stavdahl, Ole Morten Aamo, Lars Adde, and Harald Kirkerød. An optical flow-based method to predict infantile cerebral palsy. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 20(4):605–612, 2012.

[89] Espen A.F. Ihlen, Ragnhild Støen, Lynn Boswell, Raye-Ann de Regnier, Toril Fjørtoft, Deborah Gaebler-Spira, Cathrine Labori, Marianne C. Loennecken, Michael E. Msall, Unn I. Möinichen, Colleen Peyton, Michael D. Schreiber, Inger E. Silberg, Nils T. Songstad, Randi T. Vågen, Gunn K. Øberg, and Lars Adde. Machine learning of infant spontaneous movements for the early prediction of cerebral palsy: A multi-site cohort study. *Journal of Clinical Medicine*, 9(1):5, 2020.

[90] Hodjat Rahmati, Ole Morten Aamo, Øyvind Stavdahl, Ralf Dragon, and Lars Adde. Video-based early cerebral palsy prediction using motion segmentation. *Proceedings of the 2014 IEEE Engineering in Medicine and Biology Society Conference (EMBC)*, pages 30–42, 2015.

[91] A. Caruso, L. Gila, F. Fulceri, T. Salvitti, M. Micai, W. Baccinelli, and M. L. Scattoni. Early motor development predicts clinical outcomes of siblings at high-risk for autism: Insight from an innovative motion-tracking technology. *Brain Sciences*, 10(6):379, 2020.

[92] W. Baccinelli, M. Bulgheroni, V. Simonetti, F. Fulceri, A. Caruso, L. Gila, and M. L. Scattoni. Movidea: A software package for automatic video analysis of movements in infants at risk for neurodevelopmental disorders. *Brain Sciences*, 10(4):203, 2020.

[93] Devleena Das, Katelyn Fry, and Ayanna M. Howard. Vision-based detection of simultaneous kicking for identifying movement characteristics of infants at-risk for neuro-disorders. *IEEE Transactions on Biomedical Engineering*, 2023.

[94] Hyun Iee Shin, Hyung-Ik Shin, Moon Suk Bang, Don-Kyu Kim, Seung Han Shin, Ee-Kyung Kim, Yoo-Jin Kim, Eun Sun Lee, Seul Gi Park, Hye Min Ji, and Woo Hyung Lee. Deep learning-based quantitative analyses of spontaneous movements and their association with early neurological development in preterm infants. *Scientific Reports*, 12:3138, 2022.

[95] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6880–6897, 2022.

[96] Simon Reich, Dajie Zhang, Tomas Kulvicius, Sven Bölte, Karin Nielsen-Saines, Florian B. Pokorny, Robert Peharz, Luise Poustka, Florentin Wörgötter, Christa Einspieler, and Peter B. Marschik. Novel ai driven approach to classify infant motor functions. *Scientific Reports*, 11:9888, 2021.

[97] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *CVPR*, pages 7291–7299, 2017.

[98] Claire Chambers, Nidhi Seethapathi, Rachit Saluja, Helen Loeb, Samuel R. Pierce, Daniel K. Bogen, Laura Prosser, Michelle J. Johnson, and Konrad P. Kording. Computer vision to automatically assess infant neuromotor risk. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(11):2431–2442, 2020.

[99] Daniel Groos, Lars Adde, Ragnhild Støen, Heri Ramampiaro, and Espen A.F. Ihlen. Towards human-level performance on automatic pose estimation of infant spontaneous movements. *Computerized Medical Imaging and Graphics*, 95:102012, 2022.

[100] Mingliang Zhai, Yulin Li, Xiameng Qin, and Chen Yi. Fast-structext: An efficient hourglass transformer with modality-guided dynamic token merge for document understanding. *Pattern Recognition Letters*, 167:123–134, 2022.

[101] Yannick Bukschat and Marcus Vetter. Efficientpose: An efficient, accurate and scalable end-to-end 6d multi-object pose estimation approach. *ArXiv Preprint*, 2020.

[102] M. Moro, V. P. Pastore, C. Tacchino, P. Durand, I. Blanchi, P. Moretti, and M. Casadio. A markerless pipeline to analyze spontaneous movements of preterm infants. *Computer Methods and Programs in Biomedicine*, 226:107119, 2022.

[103] Alexander Mathis, Pranav Mamidanna, Kevin M. Cury, Taiga Abe, Venkatesh N. Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. Deeplabcut: Markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 21:1281–1289, 2018.

[104] E. Passmore, A.L. Kwong, S. Greenstein, J.E. Olsen, A.L. Eeles, J.L.Y. Cheong, A.J. Spittle, and G. Ball. Automated identification of abnormal infant movements from smartphone videos. *PLOS Digital Health*, 3:e0000432, 2024.

[105] Lisa Letzkus, J. Vince Pulido, Abiodun Adeyemo, Stephen Baek, and Santina Zanelli. Machine learning approaches to evaluate infants' general movements in the writhing stage—a pilot study. *Scientific Reports*, 14:4522, 2024.

[106] Hirokazu Doi, Naoya Iijima, Akira Furui, Zu Soh, Rikuya Yonei, Kazuyuki Shinohara, Mayuko Iriguchi, Koji Shimatani, and Toshio Tsuji. Prediction of autistic tendencies at 18 months of age via markerless video analysis of spontaneous body movements in 4-month-old infants. *Scientific Reports*, 12:18045, 2022.

[107] Hirokazu Doi, Akira Furui, Rena Ueda, Koji Shimatani, Midori Yamamoto, Kenichi Sakurai, Chisato Mori, and Toshio Tsuji. Spatiotemporal patterns of spontaneous movement in neonates are significantly linked to risk of autism spectrum disorders at 18 months old. *Scientific Reports*, 13:13869, 2023.

[108] Gurpreet Singh, Abhishek Raj Shekhar, Xinrui Yu, and Jafar Saniie. Smart Infant Monitoring System Using Computer Vision and AI. In *2023 IEEE*

International Conference on Electro Information Technology (eIT), pages 1–6, 2023.

[109] Davis E. King. Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.*, 10:1755–1758, December 2009.

[110] Toshio Tsuji, Shota Nakashima, Hideaki Hayashi, Zu Soh, Akira Furui, Taro Shibanoki, Keisuke Shima, and Koji Shimatani. Markerless Measurement and Evaluation of General Movements in Infants. *Scientific Reports*, 10:1422, 2020.

[111] H. Abbasi, S.R. Mollet, S.A. Williams, L. Lim, M.R. Battin, T.F. Besier, and A.J.C. McMorland. Deep-learning for automated markerless tracking of infants' general movements. *International Journal of Information Technology*, 15:4073–4083, 2023.

[112] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, 2016.

[113] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 97, pages 6105–6114. PMLR, 2019.

[114] K. D. McCay, E. S. L. Ho, H. P. H. Shum, G. Fehringer, C. Marcroft, and N. D. Embleton. Abnormal Infant Movements Classification With Deep Learning on Pose-Based Features. *IEEE Access*, 8:51582–51592, 2020.

[115] D. Sakkos, K. D. McCay, C. Marcroft, N. D. Embleton, S. Chattopadhyay, and E. S. L. Ho. Identification of Abnormal Movements in Infants: A Deep Neural Network for Body Part-Based Prediction of Cerebral Palsy. *IEEE Access*, 9:94281–94292, 2021.

[116] Iwona Doroniewicz, Daniel J. Ledwoń, Alicja Affanasowicz, Katarzyna Kieszczyńska, Dominika Latos, Małgorzata Matyja, Andrzej W. Mitas, and

Andrzej Myśliwiec. Writhing Movement Detection in Newborns on the Second and Third Day of Life Using Pose-Based Feature Machine Learning Classification. *Sensors*, 20(21):5986, 2020.

[117] Sara Moccia, Lucia Migliorelli, Virgilio Carnielli, and Emanuele Frontoni. Preterm Infants' Pose Estimation With Spatio-Temporal Features. *IEEE Transactions on Biomedical Engineering*, 67(8):2370–2380, August 2020. Epub 2019 Dec 23.

[118] H. Alkahtani, Z.A.T. Ahmed, T.H.H. Aldhyani, M.E. Jadhav, and A.A. Alqarni. Deep learning algorithms for behavioral analysis in diagnosing neurodevelopmental disorders. *Mathematics*, 11(4208), 2023.

[119] Vaibhavi Lokegaonkar, Vijay Jaisankar, Pon Deepika, Madhav Rao, T K Srikanth, Sarbani Mallick, and Manjit Sodhi. Introducing ssbd+ dataset with a convolutional pipeline for detecting self-stimulatory behaviours in children using raw videos. *arXiv preprint*, 2311.15072, 2023.

[120] Amit Kumar Singh, Trapti Shrivastava, and Vrijendra Singh. Advanced gesture recognition in autism: Integrating yolov7, video augmentation and videomae for video analysis. *arXiv preprint arXiv:2410.09339*, 2024.

[121] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint*, 2203.12602, 2022.

[122] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[123] Alessandro Floris, Simone Porcu, and Luigi Atzori. Controlling media player with hands: A transformer approach and a quality of experience assessment. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(5), 2024.

[124] Chunyi Song, Shigang Wang, Meimei Chen, Honghua Li, Feiyong Jia, and Yunxiu Zhao. A multimodal discrimination method for the response to name

behavior of autistic children based on human pose tracking and head pose estimation. *Displays*, 76:102360, 2023.

[125] D.Q. McDonald, E. DeJardin, E. Sariyanidi, J.D. Herrington, B. Tunç, C.J. Zampella, and R.T. Schultz. Predicting autism from head movement patterns during naturalistic social interactions. *Proc 7th Int Conf Med Health Inform ICMHI*, pages 55–60, 2023.

[126] Soumitra Samanta, Colin Bannard, Julian Pine, and The Language05 Team. Can automated gesture recognition support the study of child language development? In *Department of Psychological Sciences, University of Liverpool*, 2020.

[127] Varun P. Gopi, Bibin Francis, and Anju Thomas. Chapter 18 - early-stage identification of autism in children using gesture monitoring based on artificial intelligence. In Kunal Pal, Bala Chakravarthy Neelapu, and J. Sivaraman, editors, *Advances in Artificial Intelligence*, pages 491–522. Academic Press, 2024.

[128] J. Hashemi, G. Dawson, and K.L.H. et al. Carpenter. Computer vision analysis for quantification of autism risk behaviors. *IEEE Transactions on Affective Computing*, 12(1):215–226, 2021.

[129] Alessia Silvia Ivani, Alice Giubergia, Laura Santos, Alice Geminiani, Silvia Annunziata, Arianna Caglio, Ivana Olivieri, and Alessandra Pedrocchi. A gesture recognition algorithm in a robot therapy for asd children. *Biomedical Signal Processing and Control*, 74:103512, 2022.

[130] U.A. Siddiqui, F. Ullah, A. Iqbal, A. Khan, R. Ullah, S. Paracha, H. Shahzad, and K.-S. Kwak. Wearable-sensors-based platform for gesture recognition of autism spectrum disorder children using machine learning algorithms. *Sensors*, 21(3319), 2021.

[131] Mayada Elsabbagh, Janice Fernandes, Sara Jane Webb, Geraldine Dawson, Tony Charman, and Mark H. Johnson. Disengagement of visual attention in infancy is associated with emerging autism in toddlerhood. *Journal of Child Psychology and Psychiatry*, 54(11):1215–1224, 2013.

[132] Sania Zahan, Zulqarnain Gilani, Ghulam Mubashar Hassan, and Ajmal Mian. Human gesture and gait analysis for autism detection. *arXiv preprint*, 2304, 2023.

[133] Jing Li, Yihao Zhong, Junxia Han, Gaoxiang Ouyang, Xiaoli Li, and Honghai Liu. Classifying asd children with lstm based on raw videos. *Neurocomputing*, 390:226–238, 2020.

[134] A. Vabalas, E. Gowen, E. Poliakoff, and A.J. Casson. Applying machine learning to kinematic and eye movement features of a movement imitation task to predict autism diagnosis. *Scientific Reports*, 10:8346, 2020.

[135] R. Asmetha Jeyarani and Radha Senthilkumar. Eye tracking biomarkers for autism spectrum disorder detection using machine learning and deep learning techniques: Review. *Research in Autism Spectrum Disorders*, 108:102228, 2023.

[136] Tania Akter, Satu Md, Imran Khan Md, Mohammad Ali, Shahadat Uddin, Pietro Lio, Julian Quinn, and Mohammad Ali Moni. Machine learning model to predict autism investigating eye-tracking dataset. *IEEE Access*, 9:383–387, 2021.

[137] Ruohan Zhang, Akanksha Saran, Bo Liu, Yifeng Zhu, Sihang Guo, Scott Niekum, Dana Ballard, and Mary Hayhoe. Human gaze assisted artificial intelligence: A review. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 4951–4958, 2020.

[138] Z.A.T. Ahmed, E. Albalawi, T.H.H. Aldhyani, M.E. Jadhav, P. Janrao, and M.R.M. Obeidat. Applying eye tracking with deep learning techniques for early-stage detection of autism spectrum disorders. *Data*, 8:168, 2023.

[139] Noah J. Sasson and Jed T. Elison. Eye tracking young children with autism. *J. Vis. Exp.*, 61:e3675, 2012.

[140] J. Kang, X. Han, J. Song, Z. Niu, and X. Li. The identification of children with autism spectrum disorder by svm approach on eeg and eye-tracking data. *Comput Biol Med*, 120:103722, 2020.

[141] Ranjeet Vasant Bidwe, Sashikala Mishra, Simi Kamini Bajaj, and Ketan Kotecha. Attention-focused eye gaze analysis to predict autistic traits using transfer learning. *International Journal of Computational Intelligence Systems*, 2024.

[142] N. Stuart, A. Whitehouse, R. Palermo, E. Bothe, and N. Badcock. Eye gaze in autism spectrum disorder: A review of neural evidence for the eye avoidance hypothesis. *J Autism Dev Disord*, 53(5):1884–1905, 2023.

[143] K. W. Cho et al. Gaze-wasserstein: a quantitative screening approach to autism spectrum disorders. In *2016 IEEE Wireless Health (WH)*, pages 1–8, 2016.

[144] Sahar Moradizeyveh, Mehnaz Tabassum, Sidong Liu, Robert Ahadizad Newport, Amin Beheshti, and Antonio Di Ieva. When eye-tracking meets machine learning: A systematic review on applications in medical image analysis. *arXiv*, 2403.07834, 2024.

[145] A. Bhardwaj, M. Sharma, S. Kumar, S. Sharma, and P. C. Sharma. Transforming pediatric speech and language disorder diagnosis and therapy: The evolving role of artificial intelligence. *Health Sciences Review*, 12:100188, 2024.

[146] H. Gonzalez Villasanti, L. M. Justice, L. J. Chaparro-Moreno, T. J. Lin, and K. Purtell. Automatized analysis of children's exposure to child-directed speech in preschool settings: Validation and application. *PLOS One*, 15(11):e0242511, 2020.

[147] Y. Sharma and B. K. Singh. Prediction of specific language impairment in children using speech linear predictive coding coefficients. In *2020 First International Conference on Power, Control and Computing Technologies (ICPC2T)*, pages 305–310, 2020.

[148] R. Gale, J. Dolata, E. Prud'hommeaux, J. van Santen, and M. Asgari. Automatic assessment of language ability in children with and without typical development. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6111–6114, 2020.

[149] A. Pahwa. A machine learning approach for identification & diagnosing features of neurodevelopmental disorders using speech and spoken sentences. In *2016 International Conference on Computing, Communication and Automation (ICCCA)*, 2016.

[150] K. Radha, D. V. Rao, K. V. K. Sai, R. T. Krishna, and A. Muneera. Detecting autism spectrum disorder from raw speech in children using stft layered cnn model. In *2024 International Conference on Green Energy, Computing and Sustainable Technology (GECOST)*, pages 437–441, 2024.

[151] Z. Wang, J. Liu, K. He, Q. Xu, X. Xu, and H. Liu. Screening early children with autism spectrum disorder via response-to-name protocol. *IEEE Transactions on Industrial Informatics*, 17(1):587–595, 2021.

[152] Mohana Sree Venkata Sai Krishna Narala. A comparative study for autism spectrum disorder using efficientnet, resnet50. In *2023 International Conference on Computational Intelligence, Networks and Security (ICCINS)*, December 2023.

[153] F. Costanzo, E. Fucà, C. Caciolo, D. Ruà, S. Smolley, D. Weissberg, and S. Vicari. Talkitt: toward a new instrument based on artificial intelligence for augmentative and alternative communication in children with down syndrome. *Frontiers in Psychology*, 14:1176683, 2023.

[154] A. Utepbayeva, N. Zhiyenbayeva, L. Assylbekova, and O. Tapalova. Artificial intelligence applications (fluency sis, articulation station pro, and apraxia farm) in the psycholinguistic development of preschool children with speech disorders. *International Journal of Information and Education Technology (IJIET)*, 2024.

[155] Ashwini B., Deeptanshu, S. Gulati, and J. Shukla. Artificial intelligence driven predictive analysis of acoustic and linguistic behaviors for asd identification. *IEEE Transactions on Artificial Intelligence*, 5(11):5709–5719, 2024.

[156] J. Trayvick, S.B. Barkley, A. McGowan, A. Srivastava, A.W. Peters, G.A. Cecchi, J.H. Foss-Feig, and C.M. Corcoran. Speech and language patterns

in autism: Towards natural language processing as a research and clinical tool. *Psychiatry Research*, 340:116109, 2024.

[157] E.I. Toki, I.G. Tsoulos, V. Santamato, and J. Pange. Machine learning for predicting neurodevelopmental disorders in children. *Applied Sciences*, 14:837, 2024.

[158] Mohammed Almutairi. Application of artificial intelligence in assessing speech, language, and voice disorders: A scoping review. *Paper Publications*, 11:109–119, 2024.

[159] E. Donolato, E. Toffalini, K. Rogde, A. Nordahl-Hansen, A. Lervåg, C. Norbury, and M. Melby-Lervåg. Oral language interventions can improve language outcomes in children with neurodevelopmental disorders: A systematic review and meta-analysis. *Campbell Systematic Reviews*, 19(4):e1368, 2023.

[160] I. Sindhu and A. Sainin. Automatic speech and voice disorder detection using deep learning—a systematic literature review. *IEEE Access*, 2024.

[161] Chuanbo Hu, Xin Li, Mindi Ruan, Xiangxu Yu, Lynn Paul, and Shuo Wang. Exploiting chatgpt for diagnosing autism-associated language disorders and identifying distinct features. *Article*, 2024.

[162] A. Caruso, M. Micai, L. Gila, F. Fulceri, and M. L. Scattoni. The italian network for early detection of autism spectrum disorder: Research activities and national policies. *Psychiatria Danubina*, 33(Suppl 11):65–68, 2021.

[163] Python Software Foundation. Welcome to python.org. 2023. Accessed November 23, 2023.

[164] Mathworks—makers of matlab and simulink. 2022. Accessed November 18, 2024.

[165] Google Developers. Pose landmark detection guide. 2023. Accessed November 23, 2023.

[166] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann. Blazepose: On-device real-time body pose tracking. 2020. arXiv.

[167] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. 2019. arXiv.

[168] L. Meinecke, N. Breitbach-Faller, C. Bartz, R. Damen, G. Rau, and C. Disselhorst-Klug. Movement analysis in the early detection of newborns at risk for developing spasticity due to infantile cerebral palsy. *Human Movement Science*, 25(2):125–144, 2006.

[169] Bryan Manly and Jorge Navarro Alberto. *Multivariate Statistical Methods: A Primer*. CRC Press, fourth edition edition, 2016.

[170] N. Elssied, A. Ibrahim, and A. H. Osman. A novel feature selection based on one-way anova f-test for e-mail spam classification. *Research Journal of Applied Sciences, Engineering and Technology*, 7:625–638, 2014.

[171] S. Sawilowsky. Fermat, schubert, einstein, and behrens-fisher: The probable difference between two means when $\sigma_1^2 \neq \sigma_2^2$. *Journal of Modern Applied Statistical Methods*, 1(2), 2002.

[172] H. Liu and H. Motoda. *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers, USA, 1998.

[173] Two-stage analysis versus linear mixed-effects models for longitudinal data [the metafor package], 2024. Accessed: June 18, 2024.

[174] Daniel Berrar. Cross-validation. In S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach, editors, *Encyclopedia of Bioinformatics and Computational Biology*, pages 542–545. Academic Press, 2019.

[175] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2009.

[176] Mohammad Hossin and M. N. Sulaiman. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5:01–11, 2015.

[177] J. Bradshaw, A. J. Schwichtenberg, and J. M. Iverson. Capturing the complexity of autism: Applying a developmental cascades framework. *Child Development Perspectives*, 16(1):18–26, 2022.

[178] E. Thelen. The central role of action in typical and atypical development. 2004.

[179] Catherine Lord, Susan Risi, Lisa Lambrecht, Edwin H. Cook, Bennett L. Leventhal, Pamela C. DiLavore, Andrew Pickles, and Michael Rutter. The autism diagnostic observation schedule—generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders*, 30(3):205–223, 2000.

[180] M. Rutter, A. Bailey, and C. Lord. *The Social Communication Questionnaire: Manual*. Western Psychological Services, Los Angeles, 2003.

[181] American Psychiatric Association. *DSM IV: Manuale diagnostico e statistico dei disturbi mentali*. Masson, Milano, 1994. Traduzione italiana, 1995.

[182] L. Fenson, P. S. Dale, J. S. Reznick, D. J. Thal, E. Bates, and J. Hartung. *MacArthur Communicative Development Inventories: User's Guide and Technical Manual*. Brookes Publishing Co., Baltimore, MD, 1993.

[183] E. M. Mullen. *Mullen Scales of Early Learning*. American Guidance Service Inc., Circle Pines, MN, ags ed. edition, 1995.

[184] A. N. Bhat, J. C. Galloway, and R. J. Landa. Object exploration and reaching in infants at risk for autism. *Poster Presented at the International Meeting for Autism Research*, 2005.

[185] M. A. Lobo and J. C. Galloway. Postural and object-oriented experiences advance early reaching, object exploration, and means-end behavior. *Child Development*, 79(6):1869–1890, 2008.

[186] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2019.

[187] J. W. Graham. Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60:549–576, 2009.

[188] E. Niechwiej-Szwedo, T. A. Brin, B. Thompson, and L. W. T. Christian. Kinematic assessment of fine motor skills in children: Comparison of a kinematic approach and a standardized test. *TBD*, 2023. Add the journal information if known.

[189] J. C. Pinheiro and D. M. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer, 2000.

[190] B. T. West, K. B. Welch, and A. T. Galecki. *Linear Mixed Models: A Practical Guide Using Statistical Software*. CRC Press, 2014.

[191] C. E. McCulloch, S. R. Searle, and J. M. Neuhaus. *Generalized, Linear, and Mixed Models*. Wiley, 2008.

[192] W. W. Stroup. *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. CRC Press, 2012.

[193] M. Hadders-Algra. The role of the environment in the development of the infant's motor system. *Neuroscience & Biobehavioral Reviews*, 24(5):619–629, 2000.

[194] K. E. Adolph and S. R. Robinson. *Scaling motor development*. Wiley, 2015.

[195] MathWorks. Matlab 2023b, 2023.

[196] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. Version 4.2.1.

[197] Jana M. Iverson et al. Early motor abilities in infants at heightened versus low risk for asd: A baby siblings research consortium (bsrc) study. *Journal of Abnormal Psychology*, 128(1):69–80, 2019.

[198] Karen E. Adolph and John M. Franchak. The development of motor behavior. *Wiley Interdisciplinary Reviews: Cognitive Science*, 7(6):1430–1442, 2016.

[199] Mijna Hadders-Algra. Early human motor development: From variation to the ability to vary and adapt. *Neuroscience and Biobehavioral Reviews*, 90:411–427, 2018.

[200] A. M. Wetherby. Ontogeny of communicative functions in autism. *Journal of Autism and Developmental Disorders*, 16:295–316, 1986.

[201] L.-A. R. Sacrey, L. Zwaigenbaum, S. Bryson, J. Brian, I. M. Smith, W. Roberts, P. Szatmari, T. Vaillancourt, C. Roncadin, and N. Garon. Screening for behavioral signs of autism spectrum disorder in 9-month-old infant siblings. *Journal of Autism and Developmental Disorders*, 2020.

[202] J. M. Iverson, O. Capirci, and M. C. Caselli. From communication to language in two modalities. *Cognitive Development*, 9(1):23–43, 1994.

[203] D. J. Thal and S. Tobias. Communicative gestures in children with delayed onset of oral expressive vocabulary. *Journal of Speech and Hearing Research*, 35(6):1281–1289, 1992.

[204] B. Choi, P. Shah, M. L. Rowe, C. A. Nelson, and H. Tager-Flusberg. Gesture development, caregiver responsiveness, and language and diagnostic outcomes in infants at high and low risk for autism. *Journal of Autism and Developmental Disorders*, 50:2556–2572, 2020.

[205] S. B. Campbell, N. B. Leezenbaum, A. S. Mahoney, T. N. Day, and E. N. Schmidt. Social engagement with parents in 11-month-old siblings at high and low genetic risk for autism spectrum disorder. *Autism*, 19(8):915–924, 2015.

[206] E. S. LeBarton and J. M. Iverson. Gesture development in toddlers with an older sibling with autism. *International Journal of Language and Communication Disorders*, 51(1):18–30, 2016.

[207] A. D. Delehanty and A. M. Wetherby. Rate of communicative gestures and developmental outcomes in toddlers with and without autism spectrum disorder during a home observation. *American Journal of Speech-Language Pathology*, 30(2):649–662, 2021.

[208] A. Morawska, A. Basha, M. Adamson, and L. Winter. Microanalytic coding versus global rating of maternal parenting behavior. *Early Child Development and Care*, 185(3):448–463, 2015.

[209] H. Sloetjes and P. Wittenburg. Annotation by category - ELAN and ISO DCR. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, 2008.

[210] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[211] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[212] A. Kraskov, H. Stogbauer, and P. Grassberger. Estimating mutual information. *Physical Review E*, 69, 2004.

[213] B. C. Ross. Mutual information between discrete and continuous data sets. *PLOS ONE*, 9(2), 2014.

[214] S. M. Lundberg and S. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774, 2017.

[215] Adam McCrimmon and Kristen Rostad. Test review: Autism diagnostic observation schedule, second edition (ados-2) manual (part ii): Toddler module. *Journal of Psychoeducational Assessment*, 32:88–92, 2014.

[216] Marilina Mastrogiuseppe, Olga Capirci, Simone Cuva, and Paola Venuti. I gesti deittici nei bambini con disturbi dello spettro autistico: un'analisi quantitativa e qualitativa all'interno di interazioni spontanee madre-bambino. *Sistemi intelligenti*, pages 11–32, 2016.

[217] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, Honolulu, HI, 2017. IEEE.

[218] Dima Amso, Sara Haas, and Julie Markant. An eye tracking investigation of developmental change in bottom-up attention orienting to faces in cluttered natural scenes. *PLOS ONE*, 9(1):e85701, 2014.

[219] Teodora Gliga, Mayada Elsabbagh, Athina Andravizou, and Mark Johnson. Faces attract infants' attention in complex displays. *Developmental Science*, 13(4):600–610, 2010.

[220] Teresa Farroni, Gergely Csibra, Francesca Simion, and Mark H. Johnson. Eye contact detection in humans from birth. *Proceedings of the National Academy of Sciences*, 99(14):9602–9605, 2002.

[221] Teresa Farroni, Mark H. Johnson, Enrica Menon, Luisa Zulian, Dino Faraguna, and Gergely Csibra. Newborns' preference for face-relevant stimuli: effects of contrast polarity. *Proceedings of the National Academy of Sciences*, 102(47):17245–17250, 2005.

[222] Athena Vouloumanos and Janet F. Werker. Listening to language at birth: evidence for a bias for speech in neonates. *Developmental Science*, 10(2):159–164, 2007.

[223] Mayada Elsabbagh, Pat Walsh, Patrick Bolton, and Ilina Singh. In search of biomarkers for autism: scientific, social and ethical challenges. *Nature Reviews Neuroscience*, 12(10):603–612, 2011.

[224] Coralie Chevallier, Julia Parish-Morris, Alana McVey, Keiran M. Rump, Noah J. Sasson, John D. Herrington, and Robert T. Schultz. Measuring social attention and motivation in autism spectrum disorder using eye-tracking: Stimulus type matters. *Autism Research*, 8(5):620–628, 2015.

[225] Nicholas Hedger, Indu Dubey, and Bhismadev Chakrabarti. Social orienting and social seeking behaviors in asd. a meta-analytic investigation. *Neuroscience and Biobehavioral Reviews*, 119:292–306, 2020.

[226] Wenwen Hou, Yingying Jiang, Yunmei Yang, Liqi Zhu, and Jing Li. Evaluating the validity of eye-tracking tasks and stimuli in detecting high-risk infants later diagnosed with autism: A meta-analysis. *Journal of Autism and Developmental Disorders*, 2024.

[227] Ann N. Esler, Vanessa H. Bal, Whitney Guthrie, Amy Wetherby, Susan Ellis Weismer, and Catherine Lord. The autism diagnostic observation schedule, toddler module: Standardized severity scores. *Journal of Autism and Developmental Disorders*, 45(9):2704–2720, 2015.

[228] Katarzyna Chawarska, Suzanne Macari, and Frederick Shic. Context modulates attention to social scenes in toddlers with autism. *Child Development*, 87(3):825–833, 2016.

[229] Francesca Simion, Viola Macchi Cassia, Chiara Turati, and Eloisa Valenza. The origins of face perception: specific versus non-specific mechanisms. *Infant and Child Development*, 10(1-2):59–65, 2001.

[230] Michael C. Frank, Edward Vul, and Scott P. Johnson. Development of infants' attention to faces during the first year. *Cognition*, 110:160–170, 2009.

[231] Giacomo Vivanti and Sally J. Rogers. Autism and the mirror neuron system: insights from learning and teaching. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1644):20130184, 2014.

[232] Michael H. Johnson and Annette Karmiloff-Smith. Perspectives on infant development. In *Theoretical Perspectives on Infant Development*, pages 121–141. 2004.

[233] Bennett I. Bertenthal and Ty W. Boyer. The development of social attention in human infants. In A. Puce and B. I. Bertenthal, editors, *The Many Faces of Social Attention*, pages 21–65. Springer, 2015.

[234] Ami Klin, Sarah Shultz, and Warren Jones. Social visual engagement in infants and toddlers with autism: Early developmental transitions and a model of pathogenesis. *Developmental Psychopathology*, 27(4):853–869, 2015.

[235] Meia Chita-Tegmark. Social attention in asd: A review and meta-analysis of eye-tracking studies. *Research in Developmental Disabilities*, 48:79–93, 2016.

[236] Terje Falck-Ytter, Johan Lundin Kleberg, Ana Maria Portugal, and Emilia Thorup. Social attention: Developmental foundations and relevance for autism spectrum disorder. *Biological Psychiatry*, 91:245–257, 2022.

[237] Geraldine Dawson, Andrew N. Meltzoff, and John Osterling. Children with autism fail to orient to naturally occurring social stimuli. *Journal of Autism and Developmental Disorders*, 28:479–485, 1998.

[238] Coralie Chevallier, Gregory Kohls, Victoria Troiani, Edward S. Brodkin, and Robert T. Schultz. The social motivation theory of autism. *Trends in Cognitive Sciences*, 16(4):231–239, 2012.

[239] Jennifer M. Moriuchi, Ami Klin, and Warren Jones. Mechanisms of diminished attention to eyes in autism. *American Journal of Psychiatry*, 174:26–35, 2017.

[240] Johan Lundin Kleberg, Jens Högström, Martina Nord, Sven Bölte, Eva Serlachius, and Terje Falck-Ytter. Autistic traits and symptoms of social anxiety are differentially related to attention to others' eyes in social anxiety disorder. *Journal of Autism and Developmental Disorders*, 47:3814–3821, 2017. S.I.: Anxiety in Autism Spectrum Disorders.

[241] Wei Ni, Haoyang Lu, Qiandong Wang, Ci Song, and Li Yi. Vigilance or avoidance: How do autistic traits and social anxiety modulate attention to the eyes? *Frontiers in Neuroscience*, 16, 2023. This article is part of the Research Topic "Eye Movement Tracking in Ocular, Neurological, and Mental Diseases".

[242] Frederick Shic et al. Limited activity monitoring in toddlers with autism spectrum disorder. *Developmental Psychopathology*, 23(3):775–790, 2011.

[243] Clare Harrop, Desiree Jones, Shuting Zheng, Sallie Nowell, Robert Schultz, and Julia Parish-Morris. Visual attention to faces in children with autism spectrum disorder: are there sex differences? *Journal of Autism and Developmental Disorders*, 49(1):44–56, 2019.

[244] Tamami Nakano, Kyoko Tanaka, Yuuki Endo, Yui Yamane, et al. Atypical gaze patterns in children and adults with autism spectrum disorders dissociated from developmental changes in gaze behaviour. *Proceedings of the Royal Society B: Biological Sciences*, 277(1696):2935–2943, 2010.

[245] Quan Wang, Daniel J. Campbell, Suzanne L. Macari, et al. Operationalizing atypical gaze in toddlers with autism spectrum disorders: a cohesion-based approach. *Autism Research*, 11:1690–1700, 2018.

[246] Inbar Avni, Gal Meiri, Asif Bar-Sinai, et al. Children with autism observe social interactions in an idiosyncratic manner. *Autism*, 24(7):1803–1813, 2020.

[247] Lisa Byrge, Julien Dubois, J. Michael Tyszka, et al. Idiosyncratic brain activation patterns are associated with poor social comprehension in autism. *Journal of Neuroscience*, 35(14):5837–5850, 2015.

[248] Thomas A.W. Bolton, Lorena G.A. Freitas, et al. Neural responses in autism during movie watching: Inter-individual response variability co-varies with symptomatology. *Journal of Neuroscience*, 40:6121–6132, 2020.

[249] Daniel Stern. *The Interpersonal World of the Infant: A View from Psychoanalysis and Developmental Psychology.* Basic Books, New York, 1985.

[250] Daniel N. Stern. *Forms of Vitality: Exploring Dynamic Experience in Psychology, the Arts, Psychotherapy, and Development.* Oxford University Press, 2010.

[251] William James. What is an emotion? *Mind*, 34:188–205, 1884.

[252] M. Malezieux, A. S. Klein, and N. Gogolla. Neural circuits for emotion. *Annual Review of Neuroscience*, 46:211–231, 2023.

[253] G. Di Cesare et al. The middle cingulate cortex and dorso-central insula: A mirror circuit encoding observation and execution of vitality forms. *Proceedings of the National Academy of Sciences*, 118(44):e2111358118, 2021.

[254] G. Di Cesare, M. Gentilucci, and et al. Differences in action style recognition in children with autism spectrum disorders. *Frontiers in Psychology*, 8:1456, 2017.

[255] M. J. Rochat et al. Impaired vitality form recognition in autism. *Neuropsychologia*, 51(10):1918–1924, 2013.

[256] K. Ami, J. Warren, S. Robert, and V. Fred. The enactive mind, or from actions to cognition: Lessons from autism. *Philosophical Transactions of the Royal Society of London. B*, 358:345–360, 2003.

[257] A. Klin, D. J. Lin, P. Gorrindo, G. Ramsay, and W. Jones. Two-year-olds with autism orient to non-social contingencies rather than biological motion. *Nature*, 459:257–261, 2009.

[258] K. Gotham, A. Pickles, and C. Lord. Standardizing ados scores for a measure of severity in autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 39(5):693–705, 2009.

[259] F. Pons and P. L. Harris. Longitudinal change and longitudinal stability of individual differences in children's emotion understanding. *Cognition and Emotion*, 19(8):1158–1174, 2005.

[260] A. Orsini, L. Pezzuti, and L. Picone. *Wechsler Intelligence Scale for Children IV Edizione Italiana*. Giunti, 2012.

[261] D. Wechsler. *Wechsler Intelligence Scale for Children; Manual*. The Psychological Corporation, 1949.

[262] J. Raven. Raven progressive matrices. In *Handbook of Nonverbal Assessment*, pages 223–237. Springer, 2003.

[263] RStudio Team. RStudio: Integrated Development for R. http://www.rstudio.com/, 2020. RStudio, PBC.

[264] L. Casartelli et al. Vitality form expression in autism. *Scientific Reports*, 10(1):17182, 2020.

[265] T. Schmitz-Hübsch, S. T. Du Montcel, and L. Baliko. Scale for the assessment and rating of ataxia: Development of a new clinical scale. *Neurology*, 66(11):1717–1720, June 2006.

[266] R. Brandsma, T. F. Lawerman, M. J. Kuiper, R. J. Lunsing, H. Burger, and D. A. Sival. Reliability and discriminant validity of ataxia rating scales in early onset ataxia. *Develop. Med. Child Neurol.*, 59(4):427–432, April 2017.

[267] American Speech-Language-Hearing Association (ASHA). Dysarthria in adults.

[268] Małgorzata Sadowska, Beata Sarecka-Hujar, and Ilona Kopyta. Cerebral palsy: Current opinions on definition, epidemiology, risk factors, classification and treatment options. *Developmental Medicine and Child Neurology*, 2017.

[269] S. Budden, M. Meek, and C. Henighan. Communication and oral-motor function in rett syndrome. *Developmental Medicine and Child Neurology*, 1990.

[270] Katherine C. Hustad, Kristin Gorton, and Jimin Lee. Classification of speech and language profiles in 4-year-old children with cerebral palsy: A prospective preliminary study. *Journal of Speech, Language, and Hearing Research*, 2010.

[271] S. Summa, T. Schirinzi, G. M. Bernava, A. Romano, M. Favetta, E. M. Valente, E. Bertini, E. Castelli, M. Petrarca, G. Pioggia, and G. Vasco. Development of sarahome: A novel, well-accepted, technology-based assessment tool for patients with ataxia. *Comput Methods Programs Biomed*, 188:105257, May 2020.

[272] D. R. Lynch, J. M. Farmer, A. Y. Tsou, S. Perlman, S. H. Subramony, C. M. Gomez, T. Ashizawa, G. R. Wilmot, R. B. Wilson, and L. J. Balcer.

Measuring friedreich ataxia: Complementary features of examination and performance measures. *Neurology*, 66(11):1711–1716, Jun. 2006.

[273] S. H. Subramony, W. May, D. Lynch, C. Gomez, K. Fischbeck, M. Hallett, P. Taylor, R. Wilson, T. Ashizawa, and Cooperative Ataxia Group. Measuring friedreich ataxia: Interrater reliability of a neurologic rating scale. *Neurology*, 64(7):1261–1262, Apr. 2005.

[274] T. Schmitz-Hübsch, S. T. du Montcel, L. Baliko, J. Berciano, S. Boesch, C. Depondt, P. Giunti, C. Globas, J. Infante, J.-S. Kang, B. Kremer, C. Mariotti, B. Melegh, M. Pandolfo, M. Rakowicz, P. Ribai, R. Rola, L. Schöls, S. Szymanski, B. P. van de Warrenburg, A. Dürr, T. Klockgether, and R. Fancellu. Scale for the assessment and rating of ataxia: development of a new clinical scale. *Neurology*, 66(11):1717–1720, Jun. 2006.

[275] M. Asad Ullah and S. Nisar. A silence removal and endpoint detection approach for speech processing. *Sarhad University International Journal of Basic and Applied Science*, 4(1):1, September 2016.

[276] F. Ponzio, E. Macii, E. Ficarra, and S. Di Cataldo. Colorectal cancer classification using deep convolutional networks - an experimental study. In *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies*, pages 58–66, 2018.

[277] F. Alias, J.C. Socoro, and X. Sevillano. A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. *Appl. Sci.*, 6(5), 2016.

[278] S. Pincus. Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences of the United States of America*, 88:2297–2301, 1991.

[279] R. Metzger, J. Doherty, and D. Jenkins. Using approximate entropy as a speech quality measure for a speaker recognition system. In *2016 Annual Conference on Information Science and Systems (CISS)*, pages 292–297, 2016.

[280] M. Banbrook and S. McLaughlin. Is speech chaotic? invariant geometrical measures for speech data. In *IEE Colloquium on Exploiting Chaos in Signal Processing*, pages 8/1–8/10, 1994.

[281] H.M. Teager and S.M. Teager. Evidence for nonlinear sound production mechanisms in the vocal tract. In W.J. Hardcastle and A. Marchal, editors, *Speech Production and Speech Modelling*, pages 241–261. Springer Netherlands, 1990.

[282] F. Gonzalez, A. Guillamon, J.C. Alcaraz, and M.C. Alcaraz. Detection of chaotic behaviour in speech signals using the largest lyapunov exponent. In *14th International Conference on Digital Signal Processing Proceedings*, volume 1, pages 317–320, 2002.

[283] V. Pitsikalis, I. Kokkinos, and P. Maragos. Nonlinear analysis of speech signals: Generalized dimensions and lyapunov exponents. In *8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, pages 817–820, 2003.

[284] M. Banbrook, S. McLaughlin, and I. Mann. Speech characterization and synthesis by nonlinear methods. *IEEE Transactions on Speech and Audio Processing*, 7(1):1–17, 1999.

[285] H. Misra, S. Ikbal, H. Bourlard, and H. Hermansky. Spectral entropy based feature for robust asr. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–193, 2004.

[286] G. Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. *Accessed: Mar.2021.*

[287] M. Darji. Audio signal processing: A review of audio signal classification features. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 2:227–230, 2017.

[288] MATLAB. Spectral centroid for audio signals and auditory spectrograms. 2020.

[289] MATLAB. Spectral spread for audio signals and auditory spectrograms. 2020.

[290] MATLAB. Spectral skewness for audio signals and auditory spectrograms. 2020.

[291] J. Antoni. The spectral kurtosis: a useful tool for characterising non-stationary signals. *Mechanical Systems and Signal Processing*, 20(2):282–307, 2006.

[292] J. Antoni and R. B. Randall. The spectral kurtosis: application to the vibratory surveillance and diagnostics of rotating machines. *Mechanical Systems and Signal Processing*, 20(2):308–331, 2006.

[293] MATLAB. Spectral kurtosis from signal or spectrogram. https://it.mathworks.com/help/signal/ref/pkurtosis.html, 2020. Accessed: Nov. 2024.

[294] X. Valero and F. Alias. *Gammatone Cepstral Coefficients: Biologically Inspired Features for Non-Speech Audio Classification*, volume 14. 2012.

[295] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator, 1997.

[296] J. Johnston. Transform coding of audio signals using perceptual noise criteria. *IEEE J. Sel. Areas Commun.*, 6:314–323, 1988.

[297] B. S. Atal. Automatic speaker recognition based on pitch contours. *The Journal of the Acoustical Society of America*, 52(6B):1687–1697, 1972.

[298] D. Hermes. Measurement of pitch by subharmonic summation. *The Journal of the Acoustical Society of America*, 83:257–264, 1988.

[299] MPEG-7. *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval*. Wiley, 2003. ISBN: 9780470093344.

[300] Stanford CCRMA. Quadratic interpolation of spectral peaks. https://ccrma.stanford.edu/~jos/sasp/Quadratic_Interpolation_Spectral_Peaks.html, 2003.

[301] M. Reddy and L. Reddy. Dimensionality reduction: An empirical study on the usability of ife-cf (independent feature elimination-by c-correlation and f-correlation) measures. *International Journal of Computer Science Issues*, 7, Feb. 2010.

[302] I. Guyon and A. Elisseeff. An introduction of variable and feature selection. *J. Machine Learning Research Special Issue on Variable and Feature Selection*, 3:1157–1182, Jan. 2003.

[303] O. Al-Harbi. A comparative study of feature selection methods for dialectal arabic sentiment classification using support vector machine. *arXiv:1902.06242 [cs]*, Feb. 2019. Online, Available: http://arxiv.org/abs/1902.06242.

[304] S. An and X. Fan. Study on method of feature selection in speech content classification. *IJACSA*, 5(4), 2014.

[305] Y. Zhai, W. Song, X. Liu, L. Liu, and X. Zhao. A chi-square statistics based feature selection method in text classification. pages 160–163, Nov. 2018.

[306] J. R. Delgado-Contreras, J. P. Garćıa-Vázquez, R. F. Brena, C. E. Galván-Tejada, and J. I. Galván-Tejada. Feature selection for place classification through environmental sounds. *Procedia Computer Science*, 37:40–47, Jan. 2014.

[307] K. Weiss, T. Khoshgoftaar, and D. Wang. A survey of transfer learning. *Journal of Big Data*, 3, May 2016.

[308] Extract image features using pretrained network, 2020. Available: https://it.mathworks.com/help/deeplearning/ug/extract-image-features-using-pretrained-network.html.

[309] Pretrained deep neural networks, 2020. Available: https://it.mathworks.com/help/deeplearning/ug/pretrained-convolutional-neural-networks.html.

[310] Rete neurale convoluzionale, 2020. Available: https://it.mathworks.com/discovery/convolutional-neural-network-matlab.html.

[311] S. Hershey et al. Cnn architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135, Mar. 2017.

[312] J. F. Gemmeke et al. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, New Orleans, LA, Mar. 2017.

[313] M. Re and G. Valentini. Ensemble methods: A review. In *Advances in Machine Learning and Data Mining for Astronomy*, pages 563–594. 2012.

[314] M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing Management*, 45(4):427–437, Jul. 2009.

[315] S. Kiritchenko, S. Matwin, R. Nock, and A. F. Famili. Learning and evaluation in the presence of class hierarchies: Application to text categorization. In A. Y. Tawfik and S. D. Goodwin, editors, *Advances in Artificial Intelligence*, volume 3060, pages 395–406. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[316] S. S. Sparrow, D. V. Cicchetti, and D. A. Balla. *Vineland Adaptive Behavior Scales, (Vineland-II).* American Guidance Services, Circle Pines, MN, 2005.

[317] M. C. Caselli, A. Bello, P. Rinaldi, S. Stefanini, and P. Pasqualetti. *Il Primo Vocabolario del Bambino: Gesti, Parole e Frasi. Forme lunghe e forme brevi del questionario e valori di riferimento per la fascia 8–36 mesi.* Franco Angeli, 2015.

[318] L. Fenson, V. A. Marchman, D. J. Thal, P. S. Dale, J. S. Reznick, and E. Bates. *MacArthur-Bates Communicative Development Inventories, Second Edition (CDIs).* 2006.

[319] C. Lord, M. Rutter, P. C. DiLavore, S. Risi, K. Gotham, and S. L. Bishop. *Autism Diagnostic Observation Schedule (ADOS-2), Part 1: Modules 1–4 (2nd ed.)*. Western Psychological Services, Los Angeles, CA, 2012.

[320] E. Thelen, D. Corbetta, K. Kamm, J. P. Spencer, K. Schneider, and R. F. Zernicke. The transition to reaching: mapping intention and intrinsic dynamics. *Child Development*, 64(4):1058–1098, 1993.

Bibliography

Publications Related to the Thesis

Journal Articles

R. Bruschetta, A. Caruso, M. Micai, S. Campisi, G. Tartarisco, G. Pioggia, and M. L. Scattoni. Marker-Less Video Analysis of Infant Movements for Early Identification of Neurodevelopmental Disorders. *Diagnostics*, vol. 15, no. 136, 2025. DOI: 10.3390/diagnostics15020136.

G. Di Cesare, **R. Bruschetta**, A. Vitale, A. Pelosi, E. Leonardi, F. I. Famà, M. Mastrogiuseppe, C. Carrozza, S. Aiello, A. Campisi, R. Minutoli, P. Chilà, S. Campisi, F. Marino, G. Pioggia, G. Tartarisco, V. Cuccio, and L. Ruta. Exploring Divergent Kinematics in Autism Across Social and Non-Social Vitality Forms. *Scientific Reports*, vol. 14, no. 24164, 2024. DOI: 10.1038/s41598-024-74232-8. (**Co-first Author**)

G. Tartarisco, **R. Bruschetta**, S. Summa, L. Ruta, M. Favetta, M. Busà, A. Romano, E. Castelli, F. Marino, A. Cerasa, T. Schirinzi, M. Petrarca, E. Bertini, G. Vasco, and G. Pioggia. Artificial Intelligence for Dysarthria Assessment in Children With Ataxia: A Hierarchical Approach. *IEEE Access*, vol. 9, 2021. DOI: 10.1109/ACCESS.2021.3135078. (**Co-first Author**)

Conference Papers

R. Bruschetta, S. Campisi, M. Mastrogiuseppe, E. Leonardi, S. Aiello, C. Salvatore, A. Venturi, E. Schiavon, A. Campisi, F. I. Famà, C. Carrozza, C. Blandino, F. Marino, A. Cerasa, O. Capirci, G. Pioggia, L. Ruta, and G. Tartarisco. A Deep Learning Approach for Automatic Video Coding of Deictic Gestures in Children With Autism. in *Proc. of*

the International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME 2023), 19-21 July 2023, Tenerife, Canary Islands, Spain. DOI: 10.1109/ICECCME57830.2023.10253245.

Appendices

Appendix A

Supplementary Information

A.1 Supplementary Information for Section 3.1

Table A.1: Comprehensive Dataset Overview providing details on the videos available for each participant, along with their corresponding labels.

Subject	10 days	6 weeks	12 weeks	18 weeks	24 weeks	Label
LO0002		X			X	NDD
LO0015			X		X	NDD
RM0001	X	X	X	X	X	NDD
RM0002	X	X	X		X	NDD
RM0003		X	X	X		NDD
RM0006	X	X	X	X	X	NDD
RM0007	X	X	X	X	X	NT
RM0008	X	X		X	X	NDD
RM0011		X	X	X		NT
RM0012		X				NT
RM0013	X		X	X	X	NT
RM0014		X	X	X	X	NT
RM0018	X	X	X	X	X	Drop-out
RM0019	X	X				Drop-out
RM0020	X	X	X	X	X	NT
RM0021					X	No Label
RM0022		X	X	X	X	NT
RM0023		X	X			NDD
RM0024	X		X		X	NDD
RM0025	X	X	X	X		NT
RM0028			X	X	X	NT
RM0029		X	X		X	NDD

Continued on next page...

Table A.1: (Continued)

Subject	10 days	6 weeks	12 weeks	18 weeks	24 weeks	Label
RM0030	X	X	X	X		NT
RM0031					X	NDD
RM0032		X	X	X	X	NT
RM0033			X			NDD
RM0034	X		X			NDD
RM0036	X		X	X		NT
RM0037		X	X	X	X	NT
RM0038	X	X	X	X	X	NT
RM0040	X		X	X	X	NDD
RM0041			X	X	X	NT
RM0042		X	X	X		NDD
RM0043					X	NDD
RM0044	X	X	X			NT
RM0048	X	X				Drop-out
RM0050				X		NT
RM0054			X	X		NT
RM0055			X			Drop-out
RM0057		X				NT
RM0059		X	X		X	NDD
RM0066	X					NDD
RM0069	X	X			X	NT
RM0073	X	X		X	X	NT
RM0074		X	X		X	NT
RM0075	X	X	X	X	X	NT
RM0077				X		NT
RM0078			X	X	X	NT
RM0079	X		X	X		NT
RM0080		X	X	X		NT
RM0083	X	X		X		Drop-out
RM0086		X	X	X		NT
RM0087	X	X	X			NT
RM0088	X	X	X			No Label
RM0091	X	X	X	X	X	NDD
RM0092	X	X	X	X		NT
RM0094				X		NT
RM0095		X				NT
RM0096	X	X	X	X	X	No Label
RM0097			X	X		NDD
RM0101		X	X	X	X	No Label
RM0102	X		X	X	X	NT
RM0103			X	X	X	NDD
RM0104		X	X	X	X	NDD
RM0105		X	X	X	X	No Label
RM0107		X	X	X	X	NDD
RM0108		X	X	X		No Label

Continued on next page...

Table A.1: (Continued)

Subject	**10 days**	**6 weeks**	**12 weeks**	**18 weeks**	**24 weeks**	**Label**
RM0113				X	X	No Label
RM0114	X	X	X			NDD
RM0115		X	X		X	No Label
RM0119		X				NDD
RM0121		X	X	X		NDD
RM0122	X	X	X	X	X	NDD
RM0125	X	X	X	X	X	NDD

A.2 Supplementary Information for Section 3.2

Timepoint (months)	Group	N	Mean	SE	Median	SD	Minimum	Maximum	Statistics	p	Median Difference	Effect size
2.5	Language delay	6	0.404	0.0876	0.333	0.2145	0.2269	0.746	37.0	0.718	-0.0341	0.119
	No symptoms	14	0.430	0.0441	0.462	0.1649	0.1861	0.714				
3	Language delay	6	0.429	0.0625	0.422	0.1532	0.1832	0.611	33.0	0.639	-0.0569	0.154
	No symptoms	13	0.475	0.0495	0.495	0.1783	0.2079	0.867				
3.5	Language delay	7	0.449	0.0711	0.356	0.1882	0.2051	0.735	32.0	0.596	-0.0844	0.169
	No symptoms	11	0.522	0.0602	0.543	0.1997	0.2486	0.832				
4	Language delay	8	0.431	0.0690	0.395	0.1952	0.2146	0.735	70.0	0.935	0.0142	0.0278
	No symptoms	18	0.413	0.0421	0.414	0.1788	0.1645	0.782				
4.5	Language delay	8	0.466	0.0657	0.501	0.1858	0.0875	0.720	61.0	0.567	0.0877	0.153
	No symptoms	18	0.441	0.0601	0.361	0.2551	0.1504	0.943				
5	Language delay	7	0.398	0.0316	0.425	0.0836	0.2864	0.521	30.0	0.123	0.0738	0.429
	No symptoms	15	0.329	0.0221	0.319	0.0858	0.1971	0.502				
5.5	Language delay	7	0.440	0.0484	0.435	0.1280	0.2882	0.645	37.0	0.930	0.00542	0.0390
	No symptoms	11	0.440	0.0686	0.456	0.2276	0.1759	0.858				
6	Language delay	5	0.268	0.0190	0.257	0.0426	0.2234	0.339	2.00	0.006	-0.180	0.900
	No symptoms	8	0.462	0.0523	0.455	0.1479	0.3022	0.748				

Table A.2: Descriptive statistics of *Median Velocity of Hands* (1/s) for each timepoint and for both groups, extracted using AI during SM trials and results of the Mann-Whitney U test comparing the two groups.

Timepoint (months)	Group	N	Mean	SE	Median	SD	Minimum	Maximum	Statistics	p	Median Difference	Effect size
2.5	Language delay	4	1.132	0.782	0.362	1.564	0.3255	3.48	16.0	0.635	-0.0496	0.200
	No symptoms	10	1.078	0.268	0.842	0.846	0.0935	2.56				
3	Language delay	6	0.808	0.214	0.811	0.525	0.2574	1.58	22.0	0.607	-0.0125	0.185
	No symptoms	9	0.928	0.166	0.852	0.497	0.3152	1.65				
3.5	Language delay	6	0.825	0.125	0.900	0.305	0.3812	1.19	20.0	0.456	0.0330	0.259
	No symptoms	9	0.660	0.113	0.663	0.339	0.0377	1.15				
4	Language delay	8	0.934	0.315	0.669	0.892	0.0517	2.67	38.0	0.657	-0.0170	0.136
	No symptoms	11	1.053	0.229	0.951	0.760	0.1175	2.22				
4.5	Language delay	4	0.886	0.476	0.691	0.952	0.1037	2.06	23.0	0.953	-0.0077	0.0417
	No symptoms	12	0.736	0.202	0.304	0.700	0.0891	1.89				
5	Language delay	8	0.556	0.154	0.544	0.435	0.1173	1.21	39.0	0.717	-0.00758	0.114
	No symptoms	11	0.717	0.194	0.514	0.643	0.0764	1.97				
5.5	Language delay	7	0.722	0.215	0.467	0.568	0.0816	1.53	33.0	0.887	-0.00804	0.0571
	No symptoms	10	0.701	0.126	0.641	0.398	0.2434	1.55				
6	Language delay	6	0.391	0.162	0.293	0.398	0.0858	1.15	5.00	0.041	-0.0791	0.722
	No symptoms	6	0.852	0.136	0.733	0.334	0.5772	1.48				

Table A.3: Descriptive statistics of *Median Angular Velocity of Hands* (rad/s) for each timepoint and for both groups, extracted using APDM during SM trials and results of the Mann-Whitney U test comparing the two groups.

A.3 Supplementary Information for Section 4.1

	ASC (N = 17)	NT (N = 15)	U	p-value
Demographics				
Sex (M/F)	10/7	7/8	-	0.491
Age (months)	28.6 (4.8)	22.1 (5.8)	42.50	0.0008**
GMDS-ER Performance AE	24 (8.5)	-	-	-
PVB[a]				
Exp-LQ	62.5 (18.9)	95.6 (20.1)	11.50	0.0207*
Rec-LQ	73.3 (15.2)	102.5 (15.5)	10.50	0.0146*
AGQ	62.3 (7.10)	95.8 (14.9)	0	0.0001**
VABS-II standard score[b]				
ABC	68 (14.9)	103 (9.59)	5	<0.0001***
Communication	72.23 (11.39)	106 (15.6)	3	<0.0001***
Daily living skills	72.54 (10.68)	100 (10.27)	1.500	<0.0001***
Socialization	69.45 (7)	94.4 (7.34)	2.500	<0.0001***
Motor	83.27 (11.37)	104.3 (10.17)	15.50	0.0002**
ADOS-2[c]				
SA	14.59 (3.91)	-	-	-
RRB	2.23 (1.95)	-	-	-
CSS	6 (1.54)	-	-	-
GMDS-ER DQ score[d]				
General	71.46 (12.16)	-	-	-
Hearing and Language	47.41 (19.09)	-	-	-
Personal Social	70.41 (14.92)	-	-	-
Performance	84 (20.04)	-	-	-
Eye-hand Coordination	71.94 (10.88)	-	-	-
Locomotor	83.52 (18.67)	-	-	-

Table A.4: Demographic and clinical information of the ASC and NT groups. Measures report mean (SD) values. Abbreviations: PVB, Primo Vocabolario del Bambino; Exp-LQ, expressive vocabulary quotient; Rec-LQ, receptive vocabulary quotient; AGQ, Actions and Gestures Quotient; VABS-II, Vineland Adaptive Behavior Scales II Ed.; ABC, Adaptive Behavior Composite; ADOS-2, Autism Diagnostic Observation Schedule-2; SA, Social Affect; RRB, Restricted Repetitive Behaviors; CSS, Calibrated Symptom Severity score; GMDS-ER, Griffiths Mental Development Scales-Extended Revised; DQ, Developmental Quotient; AE, Age Equivalent. [a]Number of subjects ADOS-2 = 17; [b]Number of subjects GMDS = 17; [c]Number of subjects PVB (ASC=10; NT=15); [d]Number of subjects VABS-II (ASC=11; NT=15). Signif. codes: 0 '***' 0.001 '**' 0.01 '*'.

Assessment Measures The following assessment measures were administered to both ASC and NT children by experienced and trained clinical psychologists of the CNR-IRIB research team.

The Vineland Adaptive Behavior Scales, Second Edition (VABS-II). The VABS-II [316] is a psychometrically validated instrument administered via semi-structured parent interview to measure a child's adaptive behavior in daily life in four domains: communication, daily living skills, locomotor, and socialization. Additionally, VABS-II also allows for a final standardized adaptive behavior composite score (ABC).

Primo Vocabolario del Bambino (PVB). The PVB [317] is the Italian adaptation of the MacArthur-Bates Communicative Development Inventory, Second Edition (MB-CDI) [318]. The questionnaire is aimed at parents to assess verbal and non-verbal communication in children aged 8–36 months. Specifically, it gathers information on early communicative and linguistic development, starting from the first non-verbal cues, through expansion of vocabulary, the emergence of grammar, and the first combinations of words. Two sections, 'gestures and words' and 'words and phrases', are administered for children aged 8–24 months and 18–36 months respectively. Appropriate sections were administered based on the age range in the NT group and the verbal developmental age in the ASC group.

Inclusion and exclusion criteria Inclusion criteria in the ASC group were the following: (a) autism diagnosis according to the DSM-5 [2], within a multidisciplinary team and supported by the ADOS-2 [319], (b) Italian as the main language spoken at home. Exclusion criteria were the following: (a) any other identifiable genetic condition associated with autism (e.g., Fragile X syndrome, Cornelia de Lange syndrome, Tuberous Sclerosis Complex, Rett syndrome, Angelman syndrome, Prader-Willi syndrome), (b) epileptic encephalopathy with onset in infancy, and (c) significant sensory or motor impairment (e.g., vision or hearing impairment, CP).

Inclusion criteria for NT children included: (a) typical neurodevelopmental milestones for their age, as determined by parental reports, (b) typical language and social development appropriate for their age, confirmed by average scores on the communication and social domains of the Vineland Adaptive Behavior Scales, Second Edition (VABS-II) [316], (c) no significant medical conditions that could impact NT development, (d) Italian as the main language spoken at home. Exclusion criteria: (a) family history of autism, (b) psychomotor and language delay, (c) prematurity and low birth weight, (d) early infantile epileptic encephalopathy, (e) significant sensory or motor im-

pairment (e.g., hearing and vision deficits, CP).

Behavioral coding scheme: variables description All behaviors exhibited by the children were categorized into one of the following categories:

1. **Actions** were classified into one of the following categories:

   (a) **Manipulation**: Generic manipulations and explorations of the object encompass activities involving a non-specific motor scheme, such as mouthing, rotating, shaking, banging, or turning an object. This category also includes tactile exploration, like touching or scraping to examine surface characteristics, particularly when objects have grooves or diverse textures. Visual-only exploration is considered only in explicit cases where the child picks up an object and deliberately moves his head to view it from different angles.

   (b) **Motor actions:** include the following subcategories:

      i. **Index-touch action**: This category involves motor patterns requiring the use of the index finger to interact with an object. Examples include pressing a pop-up with the index finger, touching a doll's foot with the index finger for exploration, or popping soap bubbles with the index finger.

      ii. **Reaching**: This behavior is characterized by an attempt to stretch or extend their arm toward an object of interest, held by the partner, with associated trunk, arm, and hand tension directed forward. Following Thelen et al. [320], an arm movement will be coded as a reach if: (a) an object is located in the infant's reachable space; (b) the infant looks at the object before reaching; and (c) one or two hands contact the object (i.e., successful grasp is not required).

      iii. **Receiving action**: The child accepts an object that the adult is explicitly offering.

   (c) **Transitive actions**: This category encompasses all actions with objects, including using a concrete object appropriately (e.g., bringing a miniature cup to the mouth, stacking circles) and inventive use, where the child substitutes one object for another (e.g., bringing a wooden cube to the mouth, using a fork as a comb).

2. **Gestures** were classified into one of the following categories:

(a) **Showing**: A gesture is defined as showing when one holds up an object in the partner's line of sight.

(b) **Giving**: A gesture is defined as giving when one gives an object to their partner.

(c) **Pointing**: A gesture is defined as pointing when there is a clear extension of the arm, hand, and index finger directed toward a specific object, location, or event.

(d) **Conventional-interactive**: Are arbitrarily related to their meanings, culturally defined, and used for the purpose of regulating interaction (e.g., nodding one's head for "yes," shaking one's head for "no," shaking one's extended hand for "more" or "less," or moving both hands back and forth with one's palms toward a partner for "wait").

(e) **Instrumental**: These gestures are used to involve the partner in the action, utilizing the partner's body as a tool to achieve an immediate goal, aiming to prompt the partner to take immediate action (e.g., the child takes his or her mother's hand and brings it closer to the desired object, or the child takes his mother's hand and places it on the door to indicate the desire to go out).

3. **Alternate gaze**: Child's gaze behavior is specifically analyzed when it is accompanied by a gesture. The classification of gaze behavior as "alternate" occurs in two scenarios: firstly, when the child's gaze alternates between the focus object (i.e., the object referred to by the gesture), the partner's face, and the focus object itself, or vice versa; secondly, when the child's gaze shifts from the focus object to the partner's face, or vice versa.

4. **Pragmatic functions**: Each gesture produced by the child has been systematically coded to align with its corresponding pragmatic function.

   (a) **Requesting function**: Is assigned when the child employs a gesture to make a specific request, such as pointing to an object to indicate a desire for his mother to give it to him.

   (b) **Declarative function**: Is assigned when the purpose of the gesture is to express shared interest in an object or event with someone, such as the child pointing to an object to convey his enthusiasm to an adult.

5. **Speech acts**: Vowel utterances were classified into the following three main categories: vocalizations, words, and phrases:

   (a) **Vocalizations**: Vocalizations encompassed vowel strings (e.g., "eeaa"), reduplicated babble (e.g., "gaga"), and variegated babble (e.g., "bama"), while non-vowel sounds like laughter, crying, and vegetative noises (e.g., sneezing, coughing, breathing) are not coded.

   (b) **Words**: Include verbal productions referring to a specific referent on multiple occasions or in different contexts, phonetically similar to the adult model. This category encompasses actual Italian words (e.g., "ball," "cow," "yes," "no"), consistent sound patterns used by a child for a specific object or event (e.g., using "pa" to refer to a ball), articles, onomatopoeic sounds systematically used for specific referents (e.g., "woof," "meow," "choo choo"), and evaluative sounds (e.g., "wow!", "hey!", "uh-oh!").

   (c) **Phrases**: Phrases consist of the sequential combination of two or more individual intelligible words in verbal productions.

A.4 Supplementary Information for Section 5.1

Video Stimulus	NT (out of 30)	ASC (out of 23)
Xylophone	29	22
Drums	29	23
Witches	30	20
Sheriff	29	21

Table A.5: Number of no omitted trials for each video.

Median Equilibrium Probability in NT group	Adult Activity	Adult Face	Child Activity	Child Face	Object Left	Object Right
Sheriff	0.02582	0.17288	0.04068	0.54664	4.69e-4	0.00000
Witches	0.12615	0.41577	0.06360	0.11043	0.00561	0.00000
Drums	0.25877	0.17135	0.23481	0.08380	0.00000	0.00000
Xylophone	0.29490	0.11327	0.32166	0.09395	0.00811	0.00000

Table A.6: Median equilibrium probabilities across AOIs for NT gaze patterns in the four trials.

Table A.7: Distribution of equilibrium probabilities across AOIs for NT gaze patterns in the four trials.

Trial	Kruskal-Wallis				Dwass-Steel-Critchlow-Fligner		
	χ^2	gdl	p	ε^2	Comparison	W	p
Sheriff	150	7	<.001	0.648	Child Face – Adult Activity	8.192	<.001
					Child Face – Adult Face	6.257	<.001
					Child Face – Child Activity	7.555	<.001
					Child Face – Object Left	-8.915	<.001

Continued on next page...

Table A.7: (Continued)

Trial	Kruskal-Wallis				Dwass-Steel-Critchlow-Fligner		
	χ^2	gdl	p	ε^2	Comparison	W	p
					Child Face – Object Right	-9.362	<.001
Witches	165	7	<.001	0.689	Adult Face – Adult Activity	6.921	<.001
					Adult Face – Child Activity	-8.656	<.001
					Adult Face – Child Face	-8.196	<.001
					Adult Face – Object Left	-9.425	<.001
					Adult Face – Object Right	-9.517	<.001
Drums	171	7	<.001	0.739	Adult Activity – Adult Face	-5.773	0.001
					Adult Activity – Child Activity	-1.792	0.911
					Adult Activity – Child Face	-8.478	<.001
					Adult Activity – Object Left	-9.389	<.001
					Adult Activity – Object Right	-9.389	<.001
					Child Activity – Adult Face	4.190	0.061
					Child Activity – Child Face	-7.445	<.001
					Child Activity – Object Left	-9.389	<.001
					Child Activity – Object Right	-9.344	<.001
Xylophone	179	7	<.001	0.773	Adult Activity – Adult Face	-8.368	<.001
					Adult Activity – Child Activity	1.001	0.997
					Adult Activity – Child Face	-8.830	<.001
					Adult Activity – Object Left	-9.279	<.001
					Adult Activity – Object Right	-9.443	<.001
					Child Activity – Adult Face	8.478	<.001
					Child Activity – Child Face	-8.742	<.001
					Child Activity – Object Left	-9.279	<.001
					Child Activity – Object Right	-9.443	<.001

Table A.8: Categorization of transition features for both trial groups and corresponding color-coding in the PCA plot.

Sensory Social Routine Trials (Sheriff - Witches)	Transition Propensity	Musical Activities Trials (Drums - Xylophone)	Transition Propensity
Between	Adult Face - Child Face Child Face - Adult Face	Between	Adult Activity - Child Activity Child Activity - Adult Activity
From Adult Face	Adult Face - Adult Activity	From Adult Activity	Adult Activity - Adult Face
	Adult Face - Child Activity		Adult Activity - Child Face
	Adult Face - Object Left		Adult Activity - Object Left
	Adult Face - Object Right		Adult Activity - Object Right

Continued on next page...

Table A.8: (Continued)

Sensory Social Routine Trials (Sheriff - Witches)	**Transition Propensity**	**Musical Activities Trials (Drums - Xylophone)**	**Transition Propensity**
From Child Face	Child Face - Adult Activity	**From Child Activity**	Child Activity - Adult Face
	Child Face - Child Activity		Child Activity - Child Face
	Child Face - Object Left		Child Activity - Object Left
	Child Face - Object Right		Child Activity - Object Right
To Adult Face	Adult Activity - Adult Face	**To Adult Activity**	Adult Face - Adult Activity
	Child Activity - Adult Face		Child Face - Adult Activity
	Object Left - Adult Face		Object Left - Adult Activity
	Object Right - Adult Face		Object Right - Adult Activity
To Child Face	Adult Activity - Child Face	**To Child Activity**	Adult Face - Child Activity
	Child Activity - Child Face		Child Face - Child Activity
	Object Left - Child Face		Object Left - Child Activity
	Object Right - Child Face		Object Right - Child Activity

A.5 Supplementary Information for Section 6.1

Table A.9: Mean and standard deviation of 13 kinematic features for NT and ASC children during Phase 1 and Phase 2, under both social and non-social conditions with gentle and rude VFs.

		NT children		**ASC children**	
		Gentle	Rude	Gentle	Rude
			PHASE 1		
Mean velocity	Non social	100.4 (17.9)	136.8 (28.1)	89.2 (27.1)	138.1 (45.9)
	Social	105.5 (19.9)	143.1 (34.5)	94.9 (25.7)	129.9 (32.7)
Max velocity	Non social	106.8 (20.7)	134.1 (31.3)	96.3 (26.6)	123.5 (33.2)
	Social	104.2 (17.9)	133.9 (28.4)	93.6 (25.6)	125.4 (26.7)
Time max velocity	Non social	114.6 (35.4)	112.6 (34.9)	105.2 (36.6)	112.6 (34.9)
	Social	97.6 (33.1)	94.2 (19.9)	113.2 (33.1)	94.2 (38.7)
Mean acceleration	Non social	109.5 (25.5)	162.2 (64.5)	88.9 (25.7)	138.4 (50.2)
	Social	106.3 (26.7)	163.1 (45.4)	91.5 (26.7)	132.9 (36.2)
Max acceleration	Non social	114.8 (51.7)	153.4 (95.7)	96.8 (23.9)	120.3 (35.1)
	Social	112.1 (33.2)	139.7 (37.4)	93.4 (28.8)	119.3 (30.4)
Time max acceleration	Non social	155.7 (91.6)	163.5 (116.9)	136.5 (84.4)	125.8 (73.8)
	Social	136.8 (126.4)	171.5 (165.2)	110.4 (53.1)	122.3 (50.7)
Max deceleration	Non social	108.9 (54.9)	227.4 (149.9)	111.9 (69.9)	231.4 (154.5)
	Social	122.4 (65.2)	312.9 (177.6)	104.3 (48.7)	192.8 (121.1)
Time max deceleration	Non social	106.1 (36.1)	76.3 (29.8)	112.4 (49.9)	75.9 (34.9)
	Social	100.9 (32.2)	68.4 (25.3)	120.4 (42.7)	92.3 (63.8)
Max opening	Non social	102.9 (8.7)	109.9 (7.9)	97.5 (15.1)	102.1 (18.9)
	Social	101.1 (9.3)	108.4 (9.8)	99.3 (12.4)	104.2 (13.9)
Time max opening	Non social	100.6 (17.7)	83.3 (14.9)	110.5 (35.3)	96.4 (36.5)
	Social	97.7 (19.2)	80.6 (15.8)	97.7 (24.8)	97.3 (23.6)
Movement time	Non social	103.6 (14.5)	83.3 (12.7)	112.7 (26.8)	86.4 (33.9)
	Social	97.3 (14.4)	79.4 (12.5)	105.4 (17.5)	88.9 (28.9)
			PHASE 2		
Mean velocity	Non social	91.7 (20.7)	169.9 (50.4)	93.2 (40.2)	172.6 (63.2)
	Social	95.9 (29.4)	227.0 (146.1)	86.5 (18.7)	155.6 (58.4)
Max velocity	Non social	100.5 (24.0)	176.6 (59.9)	95.8 (31.2)	173.7 (60.7)
	Social	97.6 (21.8)	327.1 (349.1)	92.6 (21.9)	229.0 (275.3)

Continued on next page...

Table A.9: (Continued)

		NT children		ASC children	
		Gentle	Rude	Gentle	Rude
Time max velocity	Non social	102.7 (52.4)	116.3 (99.5)	102.1 (33.2)	145.4 (151.4)
	Social	86.5 (35.1)	98.3 (53.9)	95.4 (33.5)	154.7 (196.7)
Mean acceleration	Non social	106.8 (38.9)	286.2 (128.7)	89.6 (31.9)	195.3 (80.3)
	Social	102.6 (30.4)	566.4 (755.9)	95.8 (22.9)	207.5 (130.2)
Max acceleration	Non social	115.6 (53.9)	273.1 (168.9)	94.3 (28.4)	163.5 (61.1)
	Social	109.1 (33.4)	892.9 (1364.9)	93.1 (24.8)	293.9 (561.7)
Time max acceleration	Non social	126.8 (61.9)	144.5 (92.1)	244.9 (666.3)	675.9 (2058.8)
	Social	288.2 (809.7)	262.9 (608.6)	124.7 (97.1)	508.7 (1742.9)
Max deceleration	Non social	103.2 (83.1)	293.2 (192.9)	95.7 (79.3)	221.4 (107.7)
	Social	133.1 (107.3)	410.1 (426.1)	115.8 (2.1)	195.3 (154.9)
Max displacement along X	Non social	100.3 (15.9)	110.4 (25.8)	97.2 (19.3)	115.7 (22.4)
	Social	103.6 (18.7)	120.8 (27.1)	95.4 (17.8)	113.5 (25.6)
Max displacement along Y	Non social	98.7 (13.7)	105.2 (21.3)	93.4 (20.9)	102.7 (18.5)
	Social	96.5 (12.8)	108.9 (23.2)	90.2 (19.5)	105.8 (20.7)

Table A.10: Main and interaction effect: mean(sd), significant F and p value for NT and ASC children during the Reaching and Moving Phases.

Parameter	**Condition**	**NT**	**ASC**	**F and p values**
		Reaching Phase		
Mean velocity	Gentle	102.9 (18.9)	92.1 (26.4)	Vitality: $F(1,41) = 80.92$, $p < 0.001$
	Rude	139.9 (31.3)	134.1 (39.3)	Group*Vitality*Context: $F(1,41) = 3.81$, $p = 0.06$
	Non social	118.6 (23.1)	113.6 (36.5)	
	Social	124.3 (27.2)	112.4 (29.2)	
Max velocity	Gentle	105.2 (19.3)	94.9 (26.1)	Vitality: $F(1,41) = 66.84$, $p < 0.001$
	Rude	134.1 (29.9)	124.5 (29.9)	Group: $F(1,41) = 3.35$, $p = 0.07$

Continued on next page...

Table A.10: (Continued)

Parameter	Condition	NT	ASC	F and p values
	Non social	120.2 (26.1)	109.9 (29.9)	
	Social	119.1 (23.2)	109.6 (26.2)	
Time max velocity	Gentle	106.1 (34.3)	109.2 (34.5)	Context*Group: $F(1,41) = 6.13$, $p < 0.02$
	Rude	103.4 (27.4)	116.6 (37.2)	
	Non social	113.6 (35.1)	107.4 (36.1)	
	Social	95.9 (26.5)	118.4 (35.5)	
Mean acceleration	Gentle	101.7 (26.1)	90.6 (26.2)	Group: $F(1,41) = 7.94$, $p < 0.01$
	Rude	162.6 (54.9)	135.7 (43.2)	Vitality: $F(1,41) = 86.09$, $p < 0.001$
	Non social	135.9 (45.1)	113.7 (37.9)	
	Social	134.7 (36.1)	112.2 (31.5)	
Max acceleration	Gentle	113.4 (42.4)	95.1 (26.3)	Group: $F(1,41) = 8.19$, $p < 0.01$
	Rude	146.5 (66.5)	119.8 (32.8)	Vitality: $F(1,41) = 20.25$, $p < 0.01$
	Non social	134.1 (73.7)	108.5 (29.5)	
	Social	125.9 (35.3)	106.3 (29.6)	
Time max acceleration	Gentle	146.3 (108.9)	123.4 (68.8)	
	Rude	167.5 (141.1)	124.1 (62.2)	
	Non social	159.6 (104.3)	131.2 (79.1)	
	Social	154.2 (145.8)	116.3 (51.9)	
Max deceleration	Gentle	115.7 (60.1)	108.1 (59.3)	Vitality: $F(1,41) = 21.53$, $p < 0.01$
	Rude	270.1 (163.8)	212.1 (137.8)	Context*Group: $F(1,41) = 4.46$, $p < 0.04$
	Non social	168.1 (102.4)	171.6 (112.2)	Group*Vitality*Context: $F(1,41) = 5.81$ $p < 0.02$
	Social	217.6 (121.4)	148.6 (84.9)	
Time max deceleration	Gentle	103.5 (34.4)	116.4 (46.3)	Vitality: $F(1,41) = 5.08$, $p < 0.01$
	Rude	72.4 (27.5)	84.1 (49.4)	
	Non social	91.2 (32.9)	94.2 (42.4)	
	Social	84.7 (28.7)	106.4 (53.3)	
Max opening	Gentle	102.1 (8.9)	98.4 (13.7)	Group: $F(1,41) = 7.11$, $p < 0.01$

Continued on next page...

Table A.10: (Continued)

Parameter	Condition	NT	ASC	F and p values
	Rude	109.2 (8.8)	103.1 (16.4)	Vitality: F(1,41) = 12.24, p < 0.01
	Non social	106.5 (8.3)	99.8 (16.9)	
	Social	104.7 (9.5)	101.7 (13.2)	
Time max opening	Gentle	99.2 (18.4)	116.6 (30.1)	
	Rude	82.1 (15.4)	96.8 (30.1)	
	Non social	92.1 (16.3)	103.5 (35.9)	
	Social	89.2 (17.5)	105.1 (24.2)	
Movement time	Gentle	100.5 (14.5)	109.1 (22.1)	
	Rude	81.4 (12.6)	87.7 (31.4)	
	Non social	93.4 (13.6)	99.6 (30.3)	
	Social	88.4 (13.5)	97.2 (23.2)	
		Moving Phase		
Mean velocity	Gentle	94.4 (25.1)	89.9 (29.5)	Vitality: F(1,40) = 59.76, p < 0.001
	Rude	198.4 (98.2)	164.1 (60.8)	Context*Group: F(1,40) = 7.61, p < 0.01
	Non social	130.8 (35.5)	132.9 (51.7)	Group*Vitality*Context: F(1,40) = 6.59, p < 0.01
	Social	161.9 (87.7)	121.1 (38.5)	
Max velocity	Gentle	99.1 (22.9)	94.2 (26.5)	Vitality: F(1,40) = 24.69, p < 0.001
	Rude	251.8 (204.4)	201.4 (168.1)	Context: F(1,40) = 4.64, p < 0.03
	Non social	138.6 (41.9)	134.7 (45.9)	
	Social	212.2 (185.4)	160.8 (148.6)	
Time max velocity	Gentle	94.6 (43.7)	98.7 (33.4)	Group: F(1,40) = 6.21, p < 0.02
	Rude	107.3 (76.7)	150.1 (174.0)	Vitality: F(1,40) = 21.89, p < 0.001
	Non social	109.5 (76.0)	123.7 (92.3)	
	Social	92.4 (44.4)	125.1 (115.1)	
Mean acceleration	Gentle	104.7 (34.7)	87.7 (27.4)	Group: F(1,40) = 6.21, p < 0.02
	Rude	426.3 (442.3)	201.4 (105.3)	Vitality: F(1,40) = 21.89, p < 0.001
	Non social	196.5 (83.8)	142.5 (56.1)	Vitality*Group: F(1,40) = 4.66, p < 0.03

Continued on next page...

Table A.10: (Continued)

Parameter	Condition	NT	ASC	F and p values
	Social	334.5 (393.3)	146.7 (76.6)	
Max acceleration	Gentle	112.4 (43.6)	93.7 (26.6)	Vitality: $F(1,40) = 13.23$, $p < 0.001$
	Rude	582.9 (766.9)	228.7 (311.4)	Context: $F(1,40) = 5.76$, $p < 0.02$
	Non social	194.3 (111.4)	128.9 (44.7)	Group: $F(1,40) = 4.66$, $p < 0.003$
	Social	500.9 (699.2)	193.5 (293.2)	Vitality*Group: $F(1,40) = 3.74$, $p = 0.06$ Context*Vitality: $F(1,40) = 6.09$, $p < 0.02$
Max deceleration	Gentle	118.1 (95.2)	105.8 (85.7)	Group: $F(1,40) = 3.76$, $p = 0.06$
	Rude	351.6 (309.5)	208.3 (131.4)	Vitality: $F(1,40) = 37.17$, $p < 0.001$
	Non social	198.2 (138.0)	158.5 (93.5)	Vitality*Group: $F(1,40) = 4.96$, $p < 0.03$
	Social	271.5 (266.7)	155.6 (123.6)	
Time max deceleration	Gentle	110.1 (65.7)	134.7 (47.4)	Vitality: $F(1,40) = 5.08$, $p < 0.01$
	Rude	80.9 (64.6)	126.1 (107.4)	Context*Vitality: $F(1,40) = 3.58$, $p = 0.06$
	Non social	95.6 (52.4)	120.7 (50.6)	
	Social	95.4 (77.9)	139.9 (104.3)	
Max displacement along X	Gentle	102.7 (32.1)	123.3 (77.8)	Vitality: $F(1,40) = 14.69$, $p < 0.001$
	Rude	169.7 (142.9)	183.2 (122.8)	Group*Vitality*Context: $F(1,40) = 5.85$, $p < 0.02$
	Non social	116.7 (55.6)	156.8 (111.5)	
	Social	155.7 (119.5)	149.7 (89.1)	
Max displacement along Y	Gentle	96.3 (16.7)	102.9 (38.7)	Vitality: $F(1,40) = 27.67$, $p < 0.001$
	Rude	121.5 (34.4)	134.7 (53.3)	Context*Vitality: $F(1,40) = 5.32$, $p < 0.03$
	Non social	106.8 (19.2)	122.5 (57.9)	

Continued on next page...

Table A.10: (Continued)

Parameter	Condition	NT	ASC	F and p values
	Social	111.1 (31.9)	115.1 (34.1)	

A.6 Supplementary Information for Section 6.2

Table A.11: Comparison between clinical assessment (target) and hierarchical model (predicted) for each subject.

CLINICAL ASSESSMENT (TARGET)			**HIERARCHICAL MODEL (PREDICTED)**	
Presence of ATAXIA	Speech Disturbance Severity	Speech Disturbance Score	Presence of ATAXIA	Speech Disturbance Severity
Patient	High	2	Patient	High
Patient	Low	1	Patient	High
Patient	High	2	Patient	High
Patient	High	2	Patient	High
Patient	Low	1	Patient	Low
Patient	Low	1	Patient	Low
Patient	Low	1	Patient	High
Patient	Low	1	Patient	Low
Healthy	Low	0	Healthy	Low
Patient	Low	1	Patient	Low
Patient	Low	1	Patient	High
Patient	Low	0	Patient	Low
Patient	Low	1	Patient	Low
Healthy	Low	0	Healthy	Low
Patient	Low	1	Patient	High
Patient	Low	1	Patient	Low
Patient	High	2	Patient	High
Patient	High	2	Patient	High
Patient	High	2	Patient	High
Healthy	Low	0	Patient	Low
Healthy	Low	0	Healthy	Low
Healthy	Low	0	Healthy	Low
Patient	Low	0	Patient	Low
Patient	Low	0	Patient	Low
Patient	Low	1	Healthy	Low
Patient	Low	1	Patient	Low
Patient	High	2	Patient	High
Patient	High	2	Patient	High
Patient	High	3	Patient	High
Patient	High	3	Patient	High
Healthy	Low	0	Healthy	Low
Patient	Low	0	Healthy	Low
Patient	Low	1	Patient	Low
Patient	High	2	Patient	High
Patient	High	2	Patient	Low
Patient	High	2	Patient	High

Continued on next page...

Table A.11: (Continued)

CLINICAL ASSESSMENT (TARGET)			HIERARCHICAL MODEL (PREDICTED)	
Presence of ATAXIA	Speech Disturbance Severity	Speech Disturbance Score	Presence of ATAXIA	Speech Disturbance Severity
Patient	High	2	Patient	High
Healthy	Low	0	Healthy	Low
Patient	Low	1	Patient	Low
Patient	Low	1	Patient	Low
Patient	Low	1	Patient	Low
Patient	Low	1	Patient	Low
Patient	Low	1	Patient	High
Patient	High	2	Patient	High
Patient	High	3	Patient	Low
Patient	Low	1	Patient	High
Healthy	Low	0	Patient	Low
Healthy	Low	0	Healthy	Low
Patient	High	2	Patient	High
Patient	High	2	Patient	High
Patient	High	2	Patient	Low
Patient	High	2	Patient	High
Healthy	Low	0	Healthy	Low
Patient	High	2	Patient	High
Healthy	Low	0	Healthy	Low
Patient	High	2	Patient	High
Patient	High	2	Patient	High
Patient	Low	1	Patient	Low
Patient	Low	1	Patient	Low
Patient	High	3	Patient	High
Healthy	Low	0	Healthy	Low
Patient	High	2	Healthy	Low
Patient	Low	1	Patient	Low
Healthy	Low	0	Patient	Low
Healthy	Low	0	Patient	Low
Patient	Low	1	Patient	High
Patient	Low	1	Patient	High
Healthy	Low	0	Patient	Low
Patient	Low	1	Patient	Low
Patient	Low	1	Patient	Low
Healthy	Low	0	Patient	Low
Patient	Low	1	Patient	High
Healthy	Low	0	Healthy	Low
Patient	Low	1	Patient	High
Healthy	Low	0	Healthy	Low
Patient	Low	1	Patient	Low