Use of Machine Learning and Computer Vision Methods for Building Behavioral and Electrophysiological Biomarkers for Brain Disorders

by

Dmitry Yu. Isaev

Department of Biomedical Engineering
Duke University

Date:_______________________

Approved:

___________________________
Guillermo Sapiro, Supervisor

___________________________
Geraldine Dawson

___________________________
Sina Farsiu

___________________________
Marc Sommer

___________________________
David Carlson

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biomedical Engineering in the Graduate School of Duke University

2023
ABSTRACT

Use of Machine Learning and Computer Vision Methods for Building Behavioral and Electrophysiological Biomarkers for Brain Disorders

by

Dmitry Yu. Isaev

Department of Biomedical Engineering
Duke University

Date:_______________________

Approved:

___________________________
Guillermo Sapiro, Supervisor

___________________________
Geraldine Dawson

___________________________
Sina Farsiu

___________________________
Marc Sommer

___________________________
David Carlson

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biomedical Engineering in the Graduate School of Duke University

2023
Abstract

Research on biomarkers of brain disorders is an actively developing area. Biomarkers may allow for the early detection of diseases, which is essential for early intervention and improved outcomes. Biomarkers for monitoring the changes in the patient’s state can potentially increase the efficiency of clinical trials. Digital biomarkers, which emerged in recent years, rely on applications of machine learning methods to the data gathered by low-cost sensors, often embedded in consumer devices. Digital biomarkers have the potential to provide low-cost and more objective, granular, and sensitive to change metrics than traditional clinical ratings used in assessments of neurological and neurodevelopmental disorders. On the other hand, in traditional electrophysiological methods measuring brain activity, such as electroencephalography (EEG), biomarkers historically were based on visual analysis by clinicians, classical signal processing measures, or event-related potential (ERP) technique. Search for machine learning-based EEG biomarkers is an active area of research. This dissertation aims to build novel digital behavioral and EEG-based biomarkers and outcome measures by applying machine learning to behavioral, EEG, and concurrently recorded behavioral and EEG data.

Machine learning models for the detection of gaze, human face and body landmarks, and automatic speech recognition achieve good performance on publicly
available datasets. However, applying these models to a new clinical dataset immediately incurs a dataset shift problem, since the conditions under which real clinical video and audio data are recorded differ from the training dataset (e.g. different video camera angles, or audio noise). Furthermore, clinical datasets are in general much smaller than those used for training such models, and there are not enough human resources in the clinical setting to perform data labeling, making re-training not feasible.

Yet, the question remains – whether the predictions from pre-trained models can provide valuable insight into human behavior and neurophysiology in the clinical setting, and whether they can be a source of clinically relevant findings.

In this dissertation, we first explore this question in two use cases: (1) building digital measures of caregiver-child interaction in neurodevelopmental disorders using pre-trained pose detection deep learning models; and (2) creating a digital biomarker of ataxic dysarthria using pre-trained automatic speech recognition deep learning models. We show that in the first case, our method enables to distinguish different clusters of caregiver responsiveness which are associated with a child’s caregiver- and clinician-reported socialization, communication, and language abilities, thus demonstrating the feasibility of using digital measures of caregiver-child interaction in clinical trials. In the second case, we demonstrate the convergent validity of our novel biomarker with
clinician-reported scores and the greater sensitivity to change than clinician-reported scores on a longitudinal dataset.

Second, we propose a novel deep learning model for detecting seizures in neonates from EEG data. We demonstrate the model’s high generalizability by evaluating it on an independent dataset from another hospital and show that model by design can be applied in different facilities with different EEG hardware. This approach has the potential to be clinically validated and will allow to scale up studies of neonatal seizures by increasing the sample sizes (including data from multiple clinical centers).

Finally, we turn to the problem of combining EEG and behavioral biomarkers, which can improve biomarker sensitivity, but also provide new insights into brain-behavior relationships. In the study of autism, we propose a new metric of attentional preference to social/non-social stimuli and show that not only it distinguishes between autistic and neurotypical children, but also is differently associated with brain activity as measured by EEG. Then we turn to the question of scaling up EEG and behavior studies and provide the tool that allows measuring participants’ attention to the screen during EEG recording. This tool will allow to reduce human effort and make measurements of participants’ visual attention more objective, thus scaling up data preprocessing and allowing for multi-center studies of concurrent EEG and behavior.
Dedication

To Roza
Contents

Abstract .............................................................................................................................................. iv

List of Tables ....................................................................................................................................... xiv

List of Figures ..................................................................................................................................... xvi

Acknowledgements ............................................................................................................................. xviii

Chapter 1. Introduction ....................................................................................................................... 1
  Overview ........................................................................................................................................... 1
  Dissertation organization and key contributions ............................................................................. 2

Chapter 2. Background ....................................................................................................................... 6
  Historical perspective on biomarkers in brain disorders ................................................................. 6
  Electrophysiological biomarkers ...................................................................................................... 10
  Digital behavioral biomarkers ......................................................................................................... 11
  Combination of behavioral and EEG biomarkers .......................................................................... 14
  Machine learning challenges in biomarkers development ............................................................. 16

Chapter 3. Computer vision – based behavioral measurements of caregiver-child interaction in neurodevelopmental disorders ........................................................................... 21
  Introduction ....................................................................................................................................... 23
  Methods ............................................................................................................................................ 26
  Participants ........................................................................................................................................ 26
  Social, communication and language assessments ......................................................................... 30
  Caregiver-child interaction assessment ......................................................................................... 30
Automated coding via computer vision analysis .................................................. 31
Association of cluster membership and clinical measures.............................. 37
Results .................................................................................................................. 38
Discussion .......................................................................................................... 43
Caregiver responsiveness and language abilities .............................................. 44
Limitations .......................................................................................................... 46
Conclusions ......................................................................................................... 47
Chapter 4. Use of automatic speech recognition models to create biomarkers for
cerebellar ataxia ..................................................................................................... 49
  Introduction ......................................................................................................... 51
  Methods ........................................................................................................... 56
  Participants ....................................................................................................... 56
  Clinical data collection ...................................................................................... 59
  Audio data collection and extraction .............................................................. 60
  Neural network training .................................................................................... 61
  Entropy extraction ............................................................................................. 61
  Acoustic measures of vowels in vowel segments extraction ....................... 62
  Statistical analysis ............................................................................................ 63
  Results .............................................................................................................. 65
  Transformation of metrics and effect of age .................................................... 65
  Comparison of ataxia and control groups. ....................................................... 65
  Associations of AVE and acoustic measures with BARS ......................... 68
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Longitudinal analysis</td>
<td>69</td>
</tr>
<tr>
<td>Associations of AVE and acoustic measures</td>
<td>71</td>
</tr>
<tr>
<td>Test-retest reliability of AVE and acoustic measures</td>
<td>71</td>
</tr>
<tr>
<td>Discussion</td>
<td>72</td>
</tr>
<tr>
<td>Conclusions</td>
<td>77</td>
</tr>
<tr>
<td>Chapter 5. Deep learning models for EEG-based detection of neonatal seizures</td>
<td>78</td>
</tr>
<tr>
<td>Introduction</td>
<td>79</td>
</tr>
<tr>
<td>Technical significance</td>
<td>82</td>
</tr>
<tr>
<td>Clinical relevance</td>
<td>82</td>
</tr>
<tr>
<td>Generalizable insights about machine learning in the context of healthcare</td>
<td>83</td>
</tr>
<tr>
<td>Cohort</td>
<td>84</td>
</tr>
<tr>
<td>Data collection and annotation</td>
<td>84</td>
</tr>
<tr>
<td>Data extraction</td>
<td>87</td>
</tr>
<tr>
<td>Feature choices</td>
<td>88</td>
</tr>
<tr>
<td>Methods</td>
<td>89</td>
</tr>
<tr>
<td>Machine learning models</td>
<td>89</td>
</tr>
<tr>
<td>Data balancing approaches</td>
<td>94</td>
</tr>
<tr>
<td>Post-processing</td>
<td>94</td>
</tr>
<tr>
<td>Results</td>
<td>95</td>
</tr>
<tr>
<td>Evaluation approach/Study design</td>
<td>95</td>
</tr>
<tr>
<td>Results on machine learning approaches on different balancing techniques</td>
<td>97</td>
</tr>
</tbody>
</table>
Chapter 7. Computer vision analysis for labeling inattention during EEG recordings with visual stimuli

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>141</td>
</tr>
<tr>
<td>Method</td>
<td>145</td>
</tr>
<tr>
<td>Participants</td>
<td>145</td>
</tr>
<tr>
<td>Extracting CVA features</td>
<td>147</td>
</tr>
<tr>
<td>Data attrition</td>
<td>149</td>
</tr>
<tr>
<td>Data preprocessing</td>
<td>149</td>
</tr>
<tr>
<td>Data labeling</td>
<td>150</td>
</tr>
<tr>
<td>Training and evaluating machine learning model</td>
<td>151</td>
</tr>
<tr>
<td>Transfer learning; adjusting ML model to a new subject</td>
<td>151</td>
</tr>
<tr>
<td>Agreement measurements between model and human and between two humans</td>
<td>153</td>
</tr>
<tr>
<td>Graphical User Interface for visualizing and retraining the model</td>
<td>153</td>
</tr>
<tr>
<td>Results</td>
<td>154</td>
</tr>
<tr>
<td>Dataset statistics</td>
<td>154</td>
</tr>
<tr>
<td>Transfer learning results</td>
<td>154</td>
</tr>
<tr>
<td>Cohen’s κ analysis</td>
<td>157</td>
</tr>
<tr>
<td>Agreement between model and human coder and between two human coders</td>
<td>158</td>
</tr>
<tr>
<td>GUI for visualizing and preprocessing pipeline</td>
<td>159</td>
</tr>
<tr>
<td>Discussion</td>
<td>161</td>
</tr>
<tr>
<td>Conclusion</td>
<td>164</td>
</tr>
</tbody>
</table>
List of Tables

Table 1: Sex, racial and ethnic breakdown and maternal education level per diagnostic group................................................................. 29

Table 2: Means and Standard Deviations for Vineland Adaptive Behavior Scale (VABS-3) Communication and Socialization Domain Scores, and Verbal IQ scores. ...................... 30

Table 3: Cluster 2 - Cluster 1 differences in Transition Matrices (%).................................. 40

Table 4: Demographic information of the participants included in the analysis.............. 58

Table 5: Details of age and time interval between first and second timepoints (TP) for eight subjects included in longitudinal analysis.......................................................... 59

Table 6: Summary of BARS speech scores. ............................................................................ 66

Table 7: Medians and confidence intervals (CIs) for raw Average Vowel Entropy (AVE) and Mean Intensity Standard Deviation (MISD). ................................................................. 66

Table 8: Summary of seizure amount, duration, and total recording duration in the Duke dataset. .................................................................................................................. 87

Table 9: Summary of seizure amount, duration, and total recording duration in a subset of 39 patients from the Helsinki dataset who had seizures by consensus. .............. 87

Table 10: Results of different balancing approaches and their influence on the performance on Duke dataset (average AUC on leave one patient out cross-validation). ................. 97

Table 11: Results of cross-dataset validation as measured by average AUC on per-patient evaluation of models.............................................................. 98

Table 12: Attention network performance of DL2 (Class balance) model on the Duke dataset and the Helsinki dataset. .................................................................................. 102

Table 13: Associations of log-ratio of EEG Relative Power in Social and Toys videos and \( LR_{TBR, Social-Toys} \) and RALD .................................................................................. 132

Table 14: List of input features per frame for the machine learning model. ............... 150
Table 15: Total frames per participant and percentage of frames labeled as inattention. ........................................................................................................................................................................ 156

Table 16: Thresholds and Cohen’s κ levels at highest median value of κ in the two sampling approaches at epochs 5,10,20. ........................................................................................................................................................................ 157

Table 17: Agreement level (Cohen’s κ) between human annotators and between the model and consensus annotation at epochs 5, 10 and 20. ........................................................................................................................................................................ 159

Table 18: Average precision, AUC, and Maximal Cohen’s κ percentiles at different epochs with two sampling/adaptation alternatives ........................................................................................................................................................................ 160

Table 19: Pairwise associations between all study variables. ............................................................. 174

Table 20: Model parameters (degrees of freedom, required effect size) of Models M1-M10. ........................................................................................................................................................................ 175

Table 21: Statistics of ANOVA models M1-M5............................................................................................ 176

Table 22: Statistics of ANOVA models M6-M10............................................................................................ 177

Table 23: P-values of Wilcoxon signed-rank tests between leave-one-patient out AUCs on each level of balancing of each deep learning model........................................................................................................................................................................ 182
List of Figures

Figure 1: Video recording screenshots with extracted bending angles (BA). .................. 31

Figure 2: A schema of two subjects (a caregiver and a child) skeleton landmarks when playing on the floor................................................................. 35

Figure 3: Differences on VABS Communication and Socialization, and DAS Verbal IQ scores for Cluster 1 versus Cluster 2 revealed by Mixed Markov Model. ............... 41

Figure 4: Association of VABS scores and Age in Cluster 1 (C1) and Cluster 2 (C2) .... 43

Figure 5: Comparison of control group and subgroups of ataxia.................................. 67

Figure 6: Associations between BARS speech and BARS total with the proposed Average Vowel Entropy (AVE) and Mean Intensity Standard Deviation (MISD) .................. 68

Figure 7: Longitudinal dynamics of clinical measures and proposed metrics................. 70

Figure 8: Associations between transformed Mean Pitch and Intensity Standard Deviations (MPSD and MISD, respectively) and transformed Average Vowel Entropy. 71

Figure 9: Histogram of seizure rate per patient in the Duke dataset (left) and the Helsinki dataset (right) on a log-scale................................................................. 86

Figure 10: Graphical schema of the two deep learning architectures for seizure predictions and a schema for the feature extractor ................................................. 92

Figure 11: Scatterplot of average Cohen’s κ of inter-rater agreement vs cross- dataset AUC on patients with consensus seizures/non-seizures (on the Helsinki dataset). ........ 99

Figure 12: Ranges of Cohen’s κ, sensitivity, and specificity as decision threshold changes for DL2 Class balanced model.............................................................. 101

Figure 13: Average attention scores across all samples for one of the patients from Duke dataset. ........................................................................................................ 103

Figure 14: Examples of seizure predictions................................................................. 104

Figure 15: Screenshots of three stimuli used in the study............................................. 121
Figure 16: Visual attention measurements. ................................................................. 129

Figure 17: Relations between $RALD_{Social,Toys}$ and $LR_{Social,Toys,Theta}$, $LR_{Social,Toys,Beta1}$, and $LR_{TBR,Social-Toys}$ in Posterior Region for TD and ASD groups. ........................................................................ 131

Figure 18: EEG Recording setup. ..................................................................................... 147

Figure 19: A: Visualization of CVA features together with the video of the participant (here blurred to protect privacy). B: Interface for labeling the frames. ......................... 154

Figure 20: Performance metrics on different sampling/adaptation methods...................... 157

Figure 21: Median (thick line) and Interquartile Range (shaded area) of Cohen’s $\kappa$ at different threshold levels at epochs 5, 10, and 20. ................................................................. 158
Acknowledgements

If I had to choose only one lesson I learned along the Ph.D. journey to take with me to my future endeavors, it would not be about how to do science. It would be how to treat people with kindness. First and foremost, I am grateful for this lesson to my advisor, Guillermo Sapiro, for the living example of such an attitude. He taught me how to do research, how to write papers, and how to present my work, but it is his kindness and support, a constant context of all those lessons, that I will never forget.

I would also like to thank Geraldine Dawson, David Carlson, and Matias Di Martino, who in a big way were my co-advisors along this journey. I greatly appreciate them constantly navigating me in the world of interdisciplinary research of machine learning, brain, and behavior, and giving me valuable feedback both about the details of my work and its broader perspective.

I want to thank my committee members, Sina Farsiu and Marc Sommer, for their helpful criticism and advice when presenting this work in the preliminary stages.

While working on this dissertation I witnessed how hard it is to acquire and prepare clinical data. I would like to thank Samantha Major, Jordan Grapel, Todd Calnan, Michael Murias, Maura Sabatos-DeVito, and many other people at Duke Center for Autism and Brain Development, for their colossal effort in organizing the collection, collecting, and annotating the clinical data which I used in the research. I am also
grateful to Dmitry Tchapyjnikov, who did a ton of work annotating vast amounts of neonatal EEG.

I want to thank Anoopum Gupta for advising me on the clinical aspects of ataxia, and his entire team for the wonderfully collected data that was generously shared with me.

Since my return to academia in 2012 and before joining Duke I had incredible mentors, who defined the direction of my work and helped me grow tremendously. I would like to thank Anastasia Bonch-Osmolovskaya, Maria Ivanova, Olga Dragoy, Nina Dronkers, Boris Gutman, Neda Jahanshad, and Paul Thompson for their support and all the lessons they taught me.

I am also grateful to my labmates Martin Bertran, Natalia Martinez, Jordan Hashemi, Anish Simhal, Qiang Qiu, Ze Wang, Steven Espinosa, Sam Perochon, Pradeep Raj, Oded Schlesinger, Lun Huang, Young Kyung Kim, and Vikram Aikat. Now I wish I spent more time playing ping-pong or hanging out together, and not just meeting over coffee, lunches, and occasional dinners.

Special thanks to Maria Bagonis, Matias Di Martino and Farren Hilliard, Zhuoqing Chang and Judith Leng, Anjani Ragothaman and Vishnu Vijayakumar, for having a great time together and randomly chatting about everything when it was most needed.
Thanks to Melanie Trost and the entire Duke Student Health team for their work in keeping me and all other Duke students physically and mentally healthy. This was a solid support for me over these years, and I do not take it for granted.

I would like to thank my beloved old friends whom I have seen mostly on the smartphone screen, but whose support I always felt, Oleg Petukhov, Andrei Loktev, Sasha Neverova, Zina Polonskaya, Alexei Savin, Anya Zhelamkova, Lyuda Litvinova, and Fyodor Schklyaruk. Valera Piontkovsky, your help was life-changing at a time when it was hard to imagine that I will ever pursue a Ph.D., and I will always remember that.

This work would not ever be possible without the love and support of my parents, Olga Polyakova and Yuriy Isaev, and my sister, Anya Isaeva. Mom, dad, Anya, thank you for everything.

It is a stellar list of people above, and I am greatly indebted to each and every one of them. Yet there’s one more. Roza Vlasova, my wife. If one person can be a whole world for another one, that is what Roza is for me. A loved one, a friend, a colleague, a mentor, and beyond. Her love and faith in me were the cornerstone of all my efforts to get a Ph.D. Truly it is her love that helped me to live through these years and make this work happen.
Chapter 1. Introduction

Overview

Biomarkers of brain disorders are an actively researched area, which can help early detection of diseases, providing an opportunity to intervene at early stages, when it can be most effective (Klin, 2018; Pino et al., 2021; Pratt & Hall, 2018; Robb et al., 2016; Shen & Piven, 2017). Biomarkers that enable tracking the changes in the patient’s state can potentially increase the efficiency of clinical trials (Engel, 2011; Samtani et al., 2013; Shic et al., 2022). Digital behavioral biomarkers and outcome measures of brain disorders, backed by progress in hardware and machine learning algorithms, are cost-effective and at the same time provide better scalability, granularity, objectivity, and repeatability than traditional patient-reported, observer-reported, or clinician-reported measures (Insel, 2017; Sapiro et al., 2019). At the same time, the development of new machine learning algorithms drives the progress in quantitative analysis of instrumental measures of brain activity such as electroencephalography (EEG) (Li et al., 2017; O'Shea et al., 2020; Temko et al., 2013), and neuroimaging (Drysdale et al., 2017; Kushki et al., 2019; Thompson et al., 2020).

This dissertation covers a range of applications, proposing novel electrophysiological and digital behavioral biomarkers for brain disorders. Our first aim is to develop biomarkers of social behavior in neurodevelopmental disorders such as autism and attention-deficit/hyperactivity disorder (ADHD). Our second aim is to
develop a new low-cost digital behavioral biomarker of cerebellar ataxia through the assessment of ataxic dysarthria using the deep learning models of automatic speech recognition. Our third aim is to build a deep-learning algorithm for the detection of neonatal seizures in EEG recordings. Finally, the fourth aim is to build the basis for the convergence of behavioral and electrophysiological biomarkers in the lab setting. As an initial step, we build a metric of social/nonsocial visual attentional preference during the EEG experiment in the autism population and correlate it with the EEG activity, and then build a tool for visual attention detection based on computer vision and machine learning. The developed tool should reduce the costs and time burden for EEG data preprocessing and lay down a path for computer vision analysis of behavior recorded concurrently with EEG.

**Dissertation organization and key contributions**

The organization of the dissertation is as follows.

*Chapter 2. Background.*

In this chapter, we review historical developments that lead to the modern concept of biomarkers in brain disorders. We also review common electrophysiological biomarkers and the recent developments in digital behavioral biomarkers in the contexts of neurology and neurodevelopment. We close the chapter with a review of specific challenges in machine learning pertinent to the development of biomarkers and their use in clinical care.

In this chapter, we present a method for analyzing social behavior in a lab-based setting of free play between a caregiver and a child in a cohort of neurotypical children, and children with autism, ADHD, and combined autism and ADHD. We propose the usage of the ‘reaching to the toy’ metrics as the derivatives of pose detection algorithms output and analyze the time series of these metrics using dyadic data analysis methods. We show that this technique allows for revealing patterns of leading-following behavior which are correlated with children’s social, communication, and verbal abilities. The work presented corresponds to the paper “Computer vision analysis of caregiver-child interactions in children with neurodevelopmental disorders – a preliminary report”, currently under review.

Chapter 4. Use of automatic speech recognition models to create biomarkers for cerebellar ataxia.

In this chapter, we present a method for building a digital behavioral biomarker from the speech of patients with ataxic dysarthria, recorded on a smartphone in a clinical setting. We propose to use the entropy of predictions of vowel tokens by a pre-trained automatic speech recognition model as a biomarker. We show that it correlates with disease severity and allows to differentiate between the groups of mildly impaired patients and controls. Moreover, it correlates with the assessments of the disease
progression. The work presented corresponds to the paper “Uncertainty of vowel predictions as a digital biomarker for ataxic dysarthria” (Isaev et al., in press).


In this chapter, we present a method for detecting seizures in neonates treated in Neonatal Intensive Care Unit (NICU). We present a deep convolutional neural network with a multi-instance learning layer, which allows to highlight the EEG channels exhibiting seizure activity. We demonstrate the method’s generalizability by assessing it on an independent dataset and provide an assessment of agreement between the algorithm and the human rater. The work presented corresponds to the paper “Attention-Based Network for Weak Labels in Neonatal Seizure Detection” (Isaev, Tchapyjnikov, et al., 2020).

Chapter 6. Combining EEG and behavioral measures in autism studies.

In this chapter, we present a novel metric for assessing the social/nonsocial preference in visual attention during the EEG experiment, Relative Average Look Duration. We show that it allows to distinguish between autistic and neurotypical children participating in the EEG experiment. Additionally, it is correlated with brain activity in autistic and neurotypical children in different ways, suggesting different mechanisms of social information processing. The work presented corresponds to the paper “Relative Average Look Duration and its Association with Neurophysiological Activity in Young children with Autism Spectrum Disorder” (Isaev, Major, et al., 2020).
Chapter 7. Computer vision analysis for labeling inattention during EEG recordings with visual stimuli.

In this chapter, we present a computer vision and machine learning-based method to detect inattention periods during EEG recordings with visual stimuli. It will provide a faster, more objective, and repeatable way to code looks away from the screen, which is an inevitable part of preprocessing the EEG recordings with visual stimuli. Additionally, it will scale up experiments like the one described in chapter 6 by substantially reducing the time and effort for labeling looks at the screen. We show that the method has a good agreement with the human annotator and provide a publicly available tool that includes the visualization of model predictions, an instrument for adjusting the model for a particular participant, and post-processing the detection results. The work presented corresponds to the paper “Computer vision analysis for labeling inattention during EEG recordings with visual stimuli” (in preparation).

Chapter 8. Conclusion.

In this chapter, we summarize the dissertation and discuss the limitations of the work, the challenges of the machine learning approaches to biomarker development, and potential future research directions.
Chapter 2. Background

Historical perspective on biomarkers in brain disorders

The term ‘biomarker’ was used in literature since 1980s (Aronson, 2005) and received a formal definition from the United States National Institute of Health (NIH) in 2001 as “a characteristic that is measured as an indicator of normal biological processes, pathogenic processes or responses to an exposure or intervention” ("Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework," 2001).

Nowadays, biomarkers in neurology and psychiatry can be derived from electrophysiological, radiological, quantitative behavioral measures, and laboratory-based tests. Still, the oldest method of assessment, surviving until today, was the observation of behavioral alterations in various disease conditions (Barberis & Wright, 2022).

The oldest surviving text that bridges the observed behavioral deficits and traumatic injury dates to 1600 BCE and comes from ancient Egypt (Breasted, 1930). It includes a description of 33 cases of head and neck traumas and connects them to the deficits including hemiplegia, loss of speech, and seizures. Classification of different symptoms which we would now call neurological is found in ancient Babylonian texts, providing evidence that the link between behavior and brain damage was known to ancient physicians, even though the explanation of the mechanisms was far from being biologically true (Scurlock & Andersen, 2005). From that time till the end of the XIX
century, the connection between observed alterations and organic lesions was the main source of knowledge about the functional organization of the nervous system.

The search for a biological basis for behavior went on in ancient Greece. Greeks were the first to introduce the connection of sensory organs and the brain, as can be inferred from the works of Alcmaeon of Croton (about 490 BCE). The authors of Corpus Hippocraticum were strong proponents of natural causes of diseases and advocated for the diagnosis, prognosis, and treatment to be based on careful observation (Breitenfeld et al., 2014). For example, the work “On the Sacred Disease” suggested the natural cause of epilepsy, though incorrectly ascribing it to a build up of phlegm (one of the ‘four humours’) in the brain (Magiorkinis et al., 2010; Von Storch, 1930). In Ancient Rome, systematic observation of specific symptoms caused by particular injuries and dissections allowed Galenus of Pergamon (129-216 CE) to make functional differentiation of cranial nerves (Marketos & Skiadas, 1999; Porras-Gallo et al., 2019).

While studies of the human body and nature were under severe pressure from the clergy in medieval Europe, significant contributions to the investigation of the nervous system were done in Persia during the Golden Age. Works of Joveini (? - 983 AD) contributed to the systematic symptoms classification and detailed descriptions of psychiatric conditions, such as mania, dementia, and psychosis (Yarmohammadi et al., 2014a, 2014b).
Centuries later in Europe, despite Vesalius’s (1514 – 1564) advancements in human nervous system anatomy (Cambiaghi, 2017), systematization of nervous system disorders was still based on symptoms classification rather than biological basis, as can be found in *De Cerebri Morbis*, a work by the Vesalius’ contemporary Pratensis (Peastonk, 1988) published in 1549.

At the end of the XIX century, William Gowers’ *Manual of Diseases of the Nervous System* (1886), was a work on the etiology, diagnosis, prognosis, and treatment which was ultimately based on neuroanatomy and biological organization of the nervous system (Gowers, 1970; Mulholland, 1996). His work, along with works of Joseph Babinski and William Erb at the end of the 19th century established neurological examination as we know it now (McHenry & Garrison, 1969). The use of the reflex hammer and ophthalmoscopy was promoted as a means of instrumental diagnostics. It is important to note that chapters on psychiatric conditions in Gowers’ *Manual*, such as “Hysteria”, contained symptom descriptions and still lacked biological basis.

Thus, the mechanistic connections between sensory and motor symptoms and the nervous system were mostly established by the end of the XIX century. At the same time, the idea of localization of complex mental functions in the brain was just emerging. Pierre Paul Broca reported patients with a lesion in the lower portion of the left frontal lobe and severed abilities to produce speech and intact intellectual abilities at the same time (Broca, 1861). This was the first influential argument for the localization of
language production in the brain. Later, Wernicke suggested the ‘psychic reflex’ as a mechanism for language function mapping. He assumed that there is a representation of sensory stimulus in the sensory part of the brain, and motor representation in the motor part of the brain connected by associative fiber tracts, which provide the link between language production and comprehension (Rutten, 2017). Even though such an understanding of mental functions in the brain was oversimplified, it initiated the search for the brain basis of mental phenomena and complex behavior.

Work of Kraepelin (Compendium der Psychiatrie, 1883) and his student, Alzheimer, who described behavioral and pathologo-anatomical markers of Alzheimer’s disease in 1906 (Maurer et al., 1997), corroborated the notion that the brain is the potential site of mental disturbances, and set the foundation for neuropsychiatry. In the next 60 years, however, there remained a great divide between neurology and psychiatry. Brain correlates of major illnesses like schizophrenia, manic-depressive psychosis, and depression were not found, and psychiatry was under the heavy influence of psychoanalysis. Only in the second part of the XX century, the progress in molecular biology, genetics, neurophysiology, and neuroimaging brought about the search for biologically-based biomarkers of psychiatric disorders in addition to behavioral observations (Kandel, 1998; Price et al., 2000).

The XXI century brought up progress in sensing and communication technologies, together with the progress of methods in computer vision and machine
learning. The new concept of digital biomarkers became ubiquitous and refers to “a characteristic or set of characteristics, collected from digital health technologies, that is measured as an indicator of normal biological processes, pathogenic processes, or responses to an exposure or intervention, including therapeutic interventions” (Vasudevan et al., 2022). Across neurology and psychiatry, the search is going on for digital biomarkers that will allow for more objective, low-cost measurements, facilitating the diagnostics and monitoring of the disease and treatment progression, ultimately making healthcare more affordable for the broader population.

**Electrophysiological biomarkers**

Instrumental methods in neurophysiology, such as EEG, invented in 1924 (Haas, 2003; Petsche et al., 1984), allowed for objective and quantitative measurement of brain activity. EEG became a primary measure to evaluate epileptic activity (Feyissa & Tatum, 2019), even in the absence of clinical symptoms of epilepsy (Wietstock et al., 2016). The EEG signal is inspected visually by an epileptologist to identify abnormal patterns of brain activity, such as interictal epileptiform discharges, or seizures. Quantitative analysis of EEG for epilepsy includes spectral analysis, coherence measures, automatic spikes, and seizure detection (Andrzejak et al., 2015; Bartolomei et al., 2008; Höller & Nardone, 2021). Such EEG biomarkers improve outcomes by more precise diagnosis (locating foci for the surgical resection) and allowing for early intervention (by administering anticonvulsant drugs earlier) (Lai et al., 2013; Ronen et al., 2007; van der
Heide et al., 2012). In chapter 5 of this work, we propose a novel algorithm for the detection of seizures in neonates, who often do not present any visual signs of the seizure due to immaturity of the brain.

Besides EEG, electromyography (EMG) is a widely used diagnostic tool for diseases of muscles, peripheral nerves, and neuromuscular connections. Research is being performed on its use as a biomarker in cerebral palsy and chronic neck pain (in walking assessment), and Parkinson’s disease (in tremor assessment) (Jiménez-Grande et al., 2021; Vinti et al., 2021; Zhang et al., 2017).

In neurodevelopmental disorders, spectral powers in different EEG bands are extensively researched as potential biomarkers of attention-deficit/hyperactivity disorder (ADHD) (Arns et al., 2012), and the EEG-based Event-Related-Potential technique (Luck, 2014) is studied as a potential source of biomarkers for autism (Shic et al., 2022). It is research in progress, and still, the gold standard for diagnostics of these disorders, as stated in the latest edition of the Diagnostic and Statistical Manual for Mental Disorders (DSM-5), is the behavioral assessment. The challenges for these studies are the low number of participants and the wide variety of participants inside the samples, hindering the statistical power.

**Digital behavioral biomarkers**

Unlike EEG, the use of digital behavioral biomarkers has the prospects of going outside the lab environment and gathering objective data at scale, producing larger
samples covering more broad populations (Insel, 2017). Neurological assessment, which relies on the measurement of motor function, can benefit a lot from objective measurements provided by digital behavioral biomarkers (Masanneck et al., 2023; Youn et al., 2021). Accelerometers are used to measure gross motor movements, such as gait and steps symmetry, and fine motor movements, such as tremor (Kourtis et al., 2019). On a smartphone, analysis of touchscreen swipes and tapping patterns can be used to assess fine motor skills (Goñi et al., 2022).

Digital cameras in smartphone devices, combined with computer vision-based gaze analysis, allow the assessment of oculomotor dysmetria and abnormalities in smooth pursuit eye movements, which is a common symptom of cerebellar damage (Azami et al., 2022; Chang et al., 2020). Nystagmus was also analyzed by computer vision and machine learning methods (Friedrich et al., 2022). Changes in saccades, smooth pursuit, and more complex behavior like visual search and scene exploration are found in Alzheimer’s disease, offering an opportunity to explore them with digital cameras (Kourtis et al., 2019).

Digital microphones make speech biomarkers a low-cost solution for the analysis of voice features in neurological disorders. For example, a prominent sign of Parkinson’s disease (PD) is hypokinetic dysarthria, manifesting as reduced loudness, prosody impairments, and difficulties in articulation (Logemann et al., 1978; Tjaden, 2008). Digital biomarkers to detect those changes are being proposed in the literature
(Jeancolas et al., 2021; Novotný et al., 2014; Orozco-Arroyave et al., 2015). Ataxic dysarthria is another specific speech impairment that is caused by cerebellar damage. Speech characteristics in ataxia arise from inaccuracy of force, range, timing and directions of speech movements, which creates a distinct speech pattern, different from other dysarthrias (Kent et al., 2000). Digital biomarkers for ataxic dysarthria from speech analysis are being made (Kashyap et al., 2020; Noffs et al., 2020; Vogel et al., 2020; Vogel et al., 2017). In chapter 4 of this work, we explore the potential of combining acoustic analysis of speech and automatic speech recognition to create a new digital biomarker for ataxic dysarthria.

In autism, alterations in attentional preferences often precede social communication impairments and restricted and repetitive behavior (Elsabbagh et al., 2013; Elsabbagh et al., 2011; Werner et al., 2000). To probe these alterations, strategically designed stimuli are being used, and child’s behavior is recorded while viewing such stimuli on a monitor or smartphone screen (Sapiro et al., 2019). Application of computer vision and machine learning to videos of children recorded by smartphones and 3D cameras allowed to find differences in social attention (Chang et al., 2021; Martin et al., 2018), head turns in response to name (Perochon et al., 2021), facial dynamics (Krishnappababu et al., 2021), and imitative behaviors (Lidstone et al., 2021; Tunçgenç et al., 2021). Head pose, facial landmarks, gaze, and body pose were among the features extracted by computer vision methods.
In addition, a growing body of evidence demonstrates the usefulness of smartphone or tablet touchscreen data to find visual-motor impairments in children with autism (Anzulewicz et al., 2016; Perochon et al., 2023; Simeoli et al., 2021).

Only few studies so far were focused on digital assessment of interaction between a caregiver and a child in autism (Chen et al., 2021; Kojovic et al., 2021). Our work, presented in chapter 3, is a step to fill in this gap, and develop an objective digital outcome measure from analyzing caregiver-child interaction in a free play setting. Automatic analysis of caregiver-child interaction may pave the way to enhancing caregiver-mediated interventions by providing on-line behavioral feedback to the caregiver.

**Combination of behavioral and EEG biomarkers**

Recent works on EEG biomarkers show that applying machine learning to EEG data improves biomarker sensitivity (Bosl et al., 2011; Bosl et al., 2017; Bosl et al., 2018; McPartland, 2016). In neurodevelopmental disorders, recordings of brain activity are used as phenotypic biomarkers and to reflect the mechanism of change in clinical trials, but behavioral measurements remain the primary metric for assessing outcomes. Eye-tracking is a widely used tool to study eye movement behavior as a proxy of visual attentional preferences in neurodevelopmental disorders. Combined analysis of neural correlates and gaze patterns during the same cognitive task may lead to novel insights on neural underpinnings of visual attentional preference in autism, and to the
development of novel biomarkers. Combined eye-tracking and EEG studies in autism only started to emerge, and the results are yet inconclusive (see Tan et al. (2022) for a review).

In chapter 6 we present one of the first works that combine behavioral (gaze at the screen) and EEG (power in frequency bands) signals, showing that correlative analysis of simultaneously captured EEG and gaze can provide insight into differences in autistic and neurotypical perception of social and nonsocial stimuli.

Digital approaches to behavioral analysis may also help in EEG data preprocessing, becoming a low-cost solution to facilitate objectivity, saving researchers’ time and thus scaling up multi-center studies. For example, collecting EEG data using visual tasks in children is significantly more challenging than in adults due to their reduced ability to continuously pay attention to visual stimuli (DeBoer et al., 2007; Thierry, 2005). Due to that reason, the first step of EEG data preprocessing is removing periods where the child was not attentive to the screen. This is done either by coding the child’s attention on-line (researcher pressing the button when the kid is not attentive, see, for example, Dawson et al. (2012); Ellis and Nelson (1999); Orekhova et al. (2006)), or by reviewing the video synchronized with EEG post-hoc (Murias, Major, Compton, et al., 2018). This is a burdensome and subjective process requiring researchers’ time. In chapter 7 we propose the use of computer vision algorithms and machine learning for
detecting periods of inattention during EEG, which may significantly reduce the time and effort for EEG preprocessing and make the process more objective.

**Machine learning challenges in biomarkers development**

Along with the increased availability of technologies for acquiring and storing data, machine learning is an important driver of biomarkers development nowadays. Most of the algorithms extracting knowledge from a wide variety of medical-related data are currently based on classical machine learning or deep learning techniques (see Rajpurkar et al. (2022) for a recent review). Since the breakthrough made by convolutional neural networks in 2012 (Krizhevsky & Hinton, 2012), deep learning-based algorithms for image and video processing proved very promising in the fields of medical imaging, histopathology, and ophthalmology (Esteva et al., 2021). Deep learning computer vision algorithms for human activity analysis are also gaining prominence both in the clinical setting (Yeung et al., 2019), for healthcare monitoring outside the clinic (Wang et al., 2016), and in the field of developmental biomarkers research (Sapiro et al., 2019). However, the progress of deep learning in health is hindered by multiple factors, which we discuss below.

The main problem for the progress of supervised learning algorithms in medicine is the scarcity of labeled data. Labeling data is a costly and time-consuming process, due to the large amounts of data that need to be labeled (e.g. drawing a mask of the structure on a brain MRI 3D scan, or labeling video frame-by-frame). Annotating
human interaction is an example of the prohibitive costs of labeling. Labeling of about 5 minutes of the video of caregiver-child free play for ‘reaching to the toy’ events takes about 5 hours.' On the other hand, privacy concerns occur in the case of outsourcing the annotation routines.

Another problem of supervised models is label noise, specifically in the case of multiple human raters. Human annotations are subjective and model performance may vary greatly when trained on annotations from different raters (Armato et al., 2009). Agreement between the human raters is often suboptimal, for example, humans tend to disagree on the beginning and end of the events in time series data (seizures, emotional periods, looks at the screen, see e.g. Erel et al. (2022); Hashemi et al. (2015); Stevenson et al. (2015)). Additional time often needs to be taken to train and set up a clear labeling protocol for human raters to reach high inter-rater reliability, which further increases costs and time for acquiring labeled data (Sabatos-DeVito et al., 2019). Still, evidence exists that manual annotations of behavioral data have low test-retest reliability, indicating labeling issues (Webb et al., 2020). Despite these concerns, often the data is labeled by only one rater, due to practical reasons of time and money, potentially introducing bias to the model being trained.

* From personal correspondence with Dr. Sabatos-DeVito, a co-author of the paper described in chapter 3.
Due to the costs of data labeling, applying the ready-to-use models trained on public datasets to build biomarkers may become a feasible option. However, this incurs a problem of dataset shift, a condition where a machine learning algorithm is applied to the data whose distribution differs from the training data distribution (Finlayson et al., 2021). There may be many reasons for dataset shift in clinical data, including but not limited to a change of data acquisition hardware, changes in population and setting where the model is applied, and change in demographics. Approaches like domain adaptation (HassanPour Zonoozi & Seydi, 2022; Wang & Deng, 2018; Wilson & Cook, 2020), few-shot learning (Song et al., 2023), and multi-task learning (Zhang & Yang, 2022) may help with the distribution shift, yet they often require training the original model on multiple distributions in a specialized way, which is often not feasible for the already available trained models. As such, monitoring the model performance and model fine-tuning on newly available data are still the most common practical ways to address the dataset shift (Finlayson et al., 2021; Schulam & Saria, 2019). Understanding the model limitations due to the dataset shift and extracting the metrics most robust to it, in case of the post-processing of model outputs, are critical for the interpretation of the findings.

The assessment of the clinical utility of machine learning-based biomarkers requires a discussion of the metrics upon which the biomarker is evaluated (Pletcher & Pignone, 2011). In case the biomarker is intended for diagnostic testing, high values of
area under the receiver-operating characteristics curve (AUC), which are commonly used as an indicator of model performance, often do not guarantee the usability of the algorithm in the clinical practice (Lobo et al., 2008; Saito & Rehmsmeier, 2015). Deep learning models are often poorly calibrated (Guo et al., 2017), and the selection of the best operating threshold for a binary yes/no decision may require additional steps to calibrate the model outputs (Cook, 2007). The area under the precision-recall curve (also known as average precision), can serve as a more informative measurement (Lobo et al., 2008) in this case.

Additionally, when evaluating the clinical utility of the biomarker, its usability relates to the problem of error costs (Pletcher & Pignone, 2011). From the machine learning perspective, the data in health applications is highly imbalanced, with a much lower presence of positive (disease/health condition) than negative (healthy controls) labels. In addition, the cost of false negatives (missing a diagnosis) and the cost of false positives may be vastly different in many applications. To mitigate that, methods exist for incorporating the costs of false negatives and false positives into the learning process with imbalanced data (Elkan, 2001; Khan et al., 2018; Liu & Zhou, 2006; Thai-Nghe et al., 2010).

In conclusion, it is also important to notice that when machine learning-based methods are used in clinical practice, clinicians tend to perceive the technology as a partner, rather than a replacement for their clinical judgment (Henry et al., 2022). Based
on the nature of this partnership (providing a second opinion for a specialist after or at the same time when they perform interpretation, prioritizing the cases for the specialist, or ruling out the negative cases), different protocols of algorithm evaluation need to be performed (see Hadjiiski et al. (2023) for a discussion). As a part of the evaluation process, it is essential to provide the specialist with the reasons for the decision made by an algorithm, a requirement known as model interpretability. Multiple approaches to establishing interpretability exist (see Salahuddin et al. (2022) for a review). For example, visual saliency maps, highlighting the regions which contribute most to the final prediction, are often used in imaging, but may also be used in multi-channel time series such as EEG.

Many of the challenges described above were relevant to the work described in the following chapters, prompting to mitigate data imbalance, assess model generalization and interpretability (in the case new models were developed from scratch), and perform thorough quality checks if pre-trained models were applied to the data. In some cases, these assessments imposed limitations on the interpretation of the results, while in others served as a promise of the work to have broader implications and be generalizable beyond the data at our disposal.
Chapter 3. Computer vision – based behavioral measurements of caregiver-child interaction in neurodevelopmental disorders

Caregiver-child interactions are critical for child’s development and serve as a target for intervention in behavioral therapies for neurodevelopmental disorders. Quantitative evaluation of caregiver-child interaction is usually done manually by human raters in a labor-intensive and time-consuming process. Moreover, manual labeling requires the additional effort of annotators’ training to achieve high inter-rater reliability. Thus, automated and objective evaluation of caregiver-child interactions is essential for assessing proximal clinical outcomes in clinical trials, especially those involving caregiver-mediated interventions. In this chapter, we describe our study aimed at analyzing patterns of initiation and response to a play bout by the partner in a free play setting. We hypothesized that patterns of leading/following (namely the probability of following a partner’s lead) may be associated with diagnostic groups, and the socialization, communication, and language abilities of the children. We proposed a pipeline based on a pre-trained deep learning pose estimation algorithm to extract time series of ‘reaching to the toy’ movements by the caregiver and a child and then analyzed the resulting time series using Mixed Markov Models. Our findings provided evidence of the association between the patterns of dyadic behavior and socialization, communication, and language abilities, but not the diagnosis of the children. The results
hold promise for use of computer vision algorithms to build digital behavioral biomarkers and outcome measures of dyadic behavior in clinical trials.
Introduction

Dyadic interactions between a caregiver and child are foundational to children’s development in multiple domains including joint attention, language, and self-regulation. For example, longitudinal studies of autistic children have found that a caregiver’s responsiveness to their child’s attention and activity during play predicted the rate of children’s language growth (Adamson et al., 2019; Siller & Sigman, 2002, 2008). Measurement of play-based caregiver-child interactions is important for assessing proximal clinical outcomes in clinical trials, especially those that involve caregiver-mediated interventions (Conrad et al., 2021; Nevill et al., 2018; Rogers et al., 2012). A traditional method of measuring behaviors observed during caregiver-child interactions involves human coding, a time-consuming and labor-intensive process that requires training and assessment of inter-rater reliability, making it less scalable for large clinical trials. Thus, there is a need for automated and objective methods for assessing dyadic interactions.

Both autism and attention-deficit/hyperactivity disorder (ADHD) have been found to be associated with differences in early caregiver-child interaction. Reduced frequency of joint engagement and attention has been found to characterize interactions between young autistic children and their caregivers (Adamson et al., 2012; Adamson et al., 2019). Caregivers interacting with their children with ADHD have been found to have a more directive style of interaction (Tallmadge & Barkley, 1983). Caregiver
responsiveness has been identified as a mechanism for facilitating autistic children’s language acquisition and social communication in natural play and in the context of caregiver-mediated interventions (Davis et al., 2022; Siller & Sigman, 2002, 2008; Warlaumont et al., 2014).

*Computer vision approaches for behavior analysis.* Computer vision-based approaches have been used to objectively characterize behavioral patterns in children, including those associated with autism (Campbell et al., 2019; Carpenter et al., 2021; Chang et al., 2021; Dawson et al., 2018; Lidstone et al., 2021; Tunçgenç et al., 2021). In the study of interaction behavior Chen et al. (2021) automatically tracked head movements in mothers and infants during a face-to-face still face procedure and found that infant-mother head movement dynamics varied based on the infant’s attachment security.

To the best of our knowledge, only one study so far (Kojovic et al., 2021) has used computer vision analysis to measure dyadic interactions in autistic and neurotypical children. Kojovic and colleagues (2021) successfully distinguished autistic from neurotypical children based on automated analysis of clinician-child interactions recorded during the Autism Diagnostic Observation Schedule (Lord et al., 2000), including the Toddler Module and Modules 1 and 2, which are sensitive to non-verbal communication (Lord et al., 2012).

*Analysis of the interactive behaviors.* In the present study, we examined caregiver-child interactions with children diagnosed with autism alone, autism and co-occurring
ADHD, ADHD alone, and neurotypical development in order to better characterize variations in caregiver-child interaction using a novel computer vision-based method. We analyzed behaviors of caregivers and children recorded during a 6-minute period of free play with a set of standardized, developmentally appropriate toys. Interactive behaviors in a dyadic context are complex and multimodal, often involving a combination of facial expressions, gestures, vocalizations, gaze, and movements of hands and body in a sequence which can be contingent or non-contingent on the previous behavioral patterns of a partner. As a first step in applying computer vision analysis (CVA) techniques to an unconstrained dyadic free play setting, we focused our measurement on a commonly observed behavior that was exhibited by both the child and their caregiver, namely, reaching to one of the toys. Given that caregivers and children were sitting on the floor with a set of toys located between them, reaching for a toy was observed frequently for both the caregiver and child, which we considered as a proxy to initiating or responding to a toy play bout by the partner.

We were particularly interested in whether the caregiver’s reaching behavior was contingent on the child’s behavior, as this could be interpreted as responding to the child’s initiation. Using CVA, we automatically extracted timepoints of ‘reaching to a toy’ behaviors from 2D video recordings of dyadic interaction and analyzed initiating-responding patterns in this behavior using dyadic data analysis approaches (Kenny et al., 2006). We then used time-series analysis of the ‘reaching to a toy’ behavior to identify
distinct clusters of dyadic interaction styles based on the contingent or non-contingent (person following or not following other person’s lead, respectively) relationship between caregiver and child reaching. Finally, we examined whether these distinct clusters were correlated with the child’s communication and socialization abilities (as reported by a caregiver on a standardized measure), and language skills (verbal IQ). We hypothesized that cluster membership would be associated with the diagnostic group and level of social and communicative abilities across diagnostic groups. We viewed this as a first step towards exploring the validity of interpretable automatic annotation of dyadic behavior second by second (i.e. micro-analytic coding, Bakeman and Gottman (1997)) to assess outcomes in clinical trials.

**Methods**

**Participants**

Participants were seventy-eight children, ranging from 41-100 months of age, and their caregivers who were part of a study funded by the National Institute of Health. In this study, all caregivers were familiar and included 6 fathers, 71 mothers, and 1 grandmother. Children were recruited through brochures posted on the university website and given out at community events attended by families with children with developmental disabilities (e.g., walks), email, and social media. The ethnic and racial composition of the sample was as follows: White, 70.51%; Black, 10.26%; Asian, 5.13%; Other and mixed race, 14.10%; Hispanic, 10.26%. The mean level of maternal education
was having received a bachelor’s degree. The sample included 29 children diagnosed with autism spectrum disorder (ASD; Mean age = 66 months, SD = 14.5, 23 males), 20 children diagnosed with both autism and attention-deficit/hyperactivity disorder (ADHD; Mean age = 78.7, SD =13.5, 13 males), 22 children diagnosed with ADHD (Mean age = 76.1, SD = 13.7, 18 males), and 7 children with neurotypical development (NT; Mean age = 79.6, SD = 14.9, 4 males). Information regarding demographic characteristics by group is provided in Table 1. A chi-square test was performed to check for associations between diagnostic group and race, ethnicity and maternal education level. Inclusion criteria for each subgroup were as follows: (1) **autism alone**: DSM-5 diagnosis of ASD, based on ADOS-2 and Autism Diagnostic Interview-Revised (Le Couteur et al., 2003) and a score on the ADHD rating scale (DuPaul et al., 1998; McGoey et al., 2007) < 80%ile; (2) **ADHD alone**: Score on ADHD-RS ≥ 93%ile and an expert consensus DSM-5 diagnosis of ADHD and score on the Social Responsiveness Scale – 2nd Edition (SRS-2) of < 60 (Constantino & Gruber, 2012); (3) **autism+ADHD** met DSM-5 criteria for ASD based on ADOS-2 and ADI-R, score ≥93%ile on ADHD-RS, and meet consensus expert DSM-5 diagnosis of ADHD; (4) **NT**: score < 80%ile on ADHD-RS and < 60 on SRS-2, and Full Scale IQ > 80 as assessed by the Differential Ability Scales – II (Elliott, 2007). These diagnostic measures were administered as part of this study by clinical psychologists who were research-reliable on the ADOS-2 and ADI-R. Exclusion criteria included known genetic or neurological syndrome or condition, history of epilepsy or current
seizure disorders, significant vision, hearing, and/or serious motor impairment, and for clinical groups, a clinically elevated score (t score ≥ 65) on the Child Behavior Checklist (Achenbach & Rescorla, 2000) in domains other than those related to autism or ADHD. We excluded children with other mental health conditions to better understand the specific effects of autism and/or ADHD in caregiver-child interactions. NT participants were excluded if they had a known or suspected developmental, neurological, or psychiatric disorder and/or clinically elevated scores on the SRS-2, ADHD-RS, and/or Child Behavior Checklist, and/or sibling/first degree relative with autism or ADHD. The study was approved by the Institutional Review Board at Duke University.
Table 1: Sex, racial and ethnic breakdown and maternal education level per diagnostic group.

<table>
<thead>
<tr>
<th>Group</th>
<th># Participants</th>
<th>Age range (months)</th>
<th># Male</th>
<th># Hispanic / Latino</th>
<th># W</th>
<th># B</th>
<th># A</th>
<th>#MTO</th>
<th># Other</th>
<th>Mean level of maternal education</th>
</tr>
</thead>
<tbody>
<tr>
<td>NT</td>
<td>7</td>
<td>55-95</td>
<td>4</td>
<td>2</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Some graduate school</td>
</tr>
<tr>
<td>Autism</td>
<td>28</td>
<td>41-91</td>
<td>22</td>
<td>2</td>
<td>15</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>Bachelor’s degree</td>
</tr>
<tr>
<td>ADHD</td>
<td>22</td>
<td>48-100</td>
<td>18</td>
<td>1</td>
<td>13</td>
<td>3</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>Bachelor’s degree</td>
</tr>
<tr>
<td>Autism+ADHD</td>
<td>20</td>
<td>56-98</td>
<td>13</td>
<td>3</td>
<td>19</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Bachelor’s degree</td>
</tr>
</tbody>
</table>

Race breakdown: W – White, B – Black, A – Asian, MTO – more than one race
Social, communication and language assessments

Social and communication skills were assessed with a caregiver-report measure, the Vineland Adaptive Behavior Scales - 3rd edition (VABS-3, Sparrow et al. (2016)).

Verbal IQ was assessed with the Differential Abilities Scales-II (DAS-II, Elliott, 2007).

Means and standard deviations for the VABS Socialization (VSoc), VABS Communication (VCom), and DAS Verbal IQ (VIQ) scores for the autism, ADHD, autism+ADHD, and NT groups are shown in Table 2.

Table 2: Means and Standard Deviations for Vineland Adaptive Behavior Scale (VABS-3) Communication and Socialization Domain Scores, and Verbal IQ scores.

<table>
<thead>
<tr>
<th>Group</th>
<th>VABS-3 Communication</th>
<th>VABS-3 Socialization</th>
<th>Verbal IQ*</th>
</tr>
</thead>
<tbody>
<tr>
<td>NT</td>
<td>107.29 (6.95)</td>
<td>104.71 (6.02)</td>
<td>114.43 (7.00)</td>
</tr>
<tr>
<td>Autism</td>
<td>75.76 (17.02)</td>
<td>73.28 (15.27)</td>
<td>76.97 (31.86)</td>
</tr>
<tr>
<td>ADHD</td>
<td>93.09 (11.84)</td>
<td>87.68 (10.49)</td>
<td>113.77 (21.35)</td>
</tr>
<tr>
<td>Autism+ADHD</td>
<td>83.45 (10.81)</td>
<td>76.95 (14.74)</td>
<td>97.60 (20.08)</td>
</tr>
</tbody>
</table>

* Differential Abilities Scale (DAS-II) Verbal Ability Standard Score

Caregiver-child interaction assessment

As part of a series of caregiver-child play-based interactions, participants were invited to play together on the floor “as they naturally would” for 6 minutes. Caregivers and children wore purple and red T-shirts, respectively, to be more easily detected by CVA. A standardized set of toys representing 7 toy categories (building, pretend, puzzle, game, transportation, books, drawing/writing) was placed in a bin in the center of the room. Toys (puzzle, sensory) were also available at a table in the corner of the room.
room, but we chose to analyze the behavior only in the center of the room because reaching events could not be reliably detected when at the table. The interaction was recorded synchronously at 30 frames per second by 2 RGB cameras located in the corners of the room, resulting in the video recording from 2 viewpoints (see Figure 1).

![Figure 1](image)

**Figure 1**: Video recording screenshots with extracted bending angles (BA). Cosine of bending angle signals from the left and right video streams of the child (top left) and caregiver (bottom left). Detected Contiguous Reaching Events (CREs) are blue-shaded vertical bars on each graph. A, B, C vertical lines correspond to video frames on the right. A - child sits straight, caregiver touches the toy; B - child touches the toy and caregiver sits straight; C - caregiver touches the toy. Child CRE happens between A and B, caregiver CRE happens between B and C.

**Automated coding via computer vision analysis**

**Preprocessing and feature extraction**

Body pose landmarks were extracted first. After initial extraction of landmarks time series, we transformed them into Bending Angle (BA) time series, a composite
metric to describe the process of torso rotation towards the floor, which accompanies reaching movement. Continuous Reaching Events (CREs) were a result of segmentation of BA time-series based on constraints on the cosine of angular velocity. Finally, Reaching Events (REs) were extracted as a binary time series indicating whether the participant exhibited a reaching movement (i.e. CRE) with 0.5 seconds precision.

**Landmark detection.** Videos were split into left and right streams, corresponding to the 2 cameras. Participants were then detected on each video using the DensePose algorithm (Güler et al., 2018) and identified as child and caregiver using in-house code for classifying colors of T-shirts. A 3D multi-person pose estimation algorithm (3DMMPPE, Moon et al. (2019)) was then run separately on the left and right video streams. 3DMMPPE provides estimated 3D coordinates of body landmarks of both participants in the same 3D space relative to the camera axis. The OpenPose (Cao et al., 2021) 2D pose detection algorithm was run in parallel, providing confidence scores per landmark, a proxy for data quality. Parts of signals with a low confidence score (average of upper body confidence scores < 0.45) were removed from further analysis (Upper body confidence scores mean and standard deviation were 0.62±0.14 for the entire dataset).

**Bending Angle time series construction.** Torso directions, a vertical vector, and bending angles were computed (from the above mentioned landmarks) as preliminary steps for the detection of reaching events time series, which was our primary variable of
interest for the analysis. A torso direction (ToD) vector was computed for each participant as a normalized cross-product of the left-right shoulder and neck-pelvis landmark vectors. We observed that when caregiver and child sit in front of each other (angle between ToDs of caregiver and child is about 180 degrees), their spines are vertical. This observation was used to compute the vertical vector (VV) as the average of child’s and caregiver’s pelvis-neck vectors taken together across all frames where the scalar product $\text{ToD}_{\text{child}} \cdot \text{ToD}_{\text{caregiver}} < -0.95$. After VV for each video stream was computed, bending angle (BA) was defined as $BA(t) = \arccos(\text{ToD}(t) \cdot \text{VV})$. See Figure 2 for visualization of these vectors and angles. When the child or caregiver is reaching for the toy, they bend their torso towards the floor, which is reflected in a temporary BA increase. For each dyad, the BA is the time-series signal per participant (caregiver/child) and camera view (left/right).

**Continuous reaching events extraction.** To remove noise, the signal was low-pass filtered at 5 Hz, then a sliding window of 15 frames (0.5s) and step of 1 frame was applied, a first-order linear approximation to the signal was fitted in each window, and slope of the linear approximation was computed. In parallel, the confidence signal was computed as the median confidence of landmarks in each window. The final BA signal per participant was combined from the left and right BA, selecting the signal with higher confidence at each timepoint.
Contiguous reaching events (CREs) were then defined as contiguous periods where the slope of the \( \cos(BA(t)) \) signal was bounded between -1.0 and -0.1 (see Fig.1). These boundaries were selected empirically based on reviewing the time-series for reaching events together with participants’ videos, and the specificity of CRE detections was assessed independently.

**Reaching events time series construction.** We then discretized the entire time of the experiment into 1-second windows with 0.5 second overlap. For the caregiver and the child, we separately defined reaching event time series (‘RE time series’) as a binary value per each window. The window was assigned a value of “1” if CRE start timepoint fell into it, and “0” otherwise. RE time series effectively define the binary signal of ‘beginning of the reaching to the toy movement’ with the frequency of 2 Hz. In each dyad, there are two RE time series (for the caregiver and for the child). They are the basis of the time series dyadic analysis which we performed to reveal patterns of interaction.
Figure 2: A schema of two subjects (a caregiver and a child) skeleton landmarks when playing on the floor. Torso Direction (ToD) vector is computed as a cross-product of Pelvis-Neck and Left-Right shoulder vectors. Vertical Vector (VV) is computed as an average of Pelvis-Neck vectors for all frames where the cosine of angle between ToD_{caregiver} and ToD_{child} less than -0.95 (torsos are in front of each other). Bending angles (BA) are angles between each persons’ ToD and VV.

Specificity Assessment

For each dyad video, ten CREs per caregiver and child were randomly selected from all CREs (1560 segments total). A random subsample of 200 segments per caregiver and per child was then selected from this sample and labeled by two independent manual raters. Reaching events extraction was done via a fully automatic pipeline, and human raters were involved only in the specificity assessment. CRE was manually labeled as true detection if a person bent or touched/moved a toy with their hand; otherwise CRE was labeled as false detection. Inter-rater reliability (Cohen’s $\kappa$) and
specificity, measured as a percentage of consensus true positive labels was computed. Labeling was performed with ‘pigeon’ widget for Jupyter Notebook (https://github.com/agermanidis/pigeon).

**Time-Series Dyadic Analysis**

For each dyad, a pair of binary RE time-series was transformed into a single time-series with four states (‘No RE,’ ‘RE Child,’ ‘RE Caregiver,’ ‘RE Both’). Then dyadic data analysis (Kenny et al., 2006) methods (Actor-Partner Interdependence Model, APIM), implemented as Markov and Mixture Markov models (Fuchs et al., 2017; Helske & Helske, 2019; van de Pol & Langeheine, 1990) were applied to the time series. Transition probabilities between the states of the Markov model (MM) allow us to characterize how one participant’s RE state at the current timepoint influenced both themselves and their partner at the next timepoint, capturing interdependence between participants. Applying a Mixture Markov Model (MixMM) to the same data reveals clusters with different transition probabilities matrices, a proxy for different interaction patterns in dyads. Bayesian Information Criteria (BIC, Schwarz (1978)) was used to measure goodness-of-fit of the models. Cluster stability was additionally assessed via model-based distance methods and bootstrapping (see Appendix A for details).

**Potential influence of demographic factors**

To assess whether the CVA methods are biased based on the race of the participant, we tested whether the amount of dropped frames during reaching events
detection process was associated with race. We also tested whether the amount of detected Continuous Reaching Events (CREs) can be associated with demographic factors, including age, sex, maternal education level, race, and ethnicity. Results of this analysis are reported in the Appendix A.

**Association of cluster membership and clinical measures**

After clustering the dyadic time series, we examined the association between cluster membership and clinical measures, including VCom and VSoc and VIQ. Our aim was to examine whether cluster variable helps to explain the variance in the clinical scores. For that we ran a sequential analysis of variance on linear model fit (models M1-M10). Measure in the model formulation is either VCom, VSoc, or VIQ; and Group stands for the diagnostic group.

Models M1-M4 serve as a baseline, which allowed us to estimate how well the variance in Measure is explained by Group and Age variables alone.

- **M1**: \( \text{Measure} \sim \text{Age} \)
- **M2**: \( \text{Measure} \sim \text{Group} \)
- **M3**: \( \text{Measure} \sim \text{Age} + \text{Group} \)
- **M4**: \( \text{Measure} \sim \text{Age} + \text{Group} + \text{Group} \times \text{Age} \)

Models M5-M10 are designed to test whether adding Cluster variable helps to explain the variance in Measure besides what is explained by the Group variable.
Additionally, models M9 and M10 explore potential interaction between Cluster and Age or Group.

\[ M5: \text{Measure} \sim \text{Cluster} \]
\[ M6: \text{Measure} \sim \text{Age} + \text{Cluster} \]
\[ M7: \text{Measure} \sim \text{Group} + \text{Cluster} \]
\[ M8: \text{Measure} \sim \text{Age} + \text{Group} + \text{Cluster} \]
\[ M9: \text{Measure} \sim \text{Age} + \text{Group} + \text{Cluster} + \text{Age} \times \text{Cluster} \]
\[ M10: \text{Measure} \sim \text{Age} + \text{Group} + \text{Cluster} + \text{Group} \times \text{Cluster} \]

Model parameters, such as between- and within-group degrees of freedom are provided in Table 20 in Appendix A. Additionally, we performed a power sensitivity analysis using G*Power software package (Faul et al., 2009) to assess the effect size that was needed to reject the null hypothesis with a sample size of 78 subjects at the power level of 0.8. Finally, we also checked for an association between diagnostic group, race, or ethnicity and cluster with a Chi-squared test.

**Results**

Diagnostic group was associated with race ($X^2 (12,N=78) = 23.8, p = 0.022$), but was not associated with ethnicity or maternal education level, based on Chi-square test statistics. See Table 1 and Table 19 in Appendix A for details.

Cohen’s kappas for CRE detections for children and caregivers were 0.85 and 0.86, respectively. Specificity of detections for children was 0.76 (95%CI: 0.67 – 0.82) and for caregivers was 0.70 (95%CI: 0.67 – 0.82). Kruskal-Wallis test (Kruskal & Wallis, 1952)
was not significant when comparing specificity of CRE detections between diagnostic
groups both for children \((H(3)=0.58, p=0.90)\) and caregivers \((H(3)=0.78, p=0.86)\).

MixMM with 2 clusters showed a better fit than MM when compared by
Bayesian Information Criteria (BIC=86302.33 for MixMM; BIC=87166.67 for MM). Behavioral differences between Cluster 1 and Cluster 2 are presented in Table 3.

Clusters were highly stable, based on bootstrapping assessment (see Appendix A). Cluster 2 was characterized by more frequent caregiver responsiveness to the child’s behavior and more overall reaching movement. Compared to Cluster 1, Cluster 2 had more transitions from ‘RE Child’ to ‘RE Caregiver’ state (by 130.2%), from ‘RE Caregiver’ to ‘RE Child,’ and from ‘No RE’ to ‘RE Child’ states (by 15.5% and 29.2% respectively), and overall more transitions to ‘RE Both’ and fewer transitions to ‘No RE’ states (more amount of reaching movements).
Table 3: Cluster 2 - Cluster 1 differences in Transition Matrices (%).\textsuperscript{a,b}

<table>
<thead>
<tr>
<th>From State</th>
<th>To State</th>
<th>No RE</th>
<th>RE Child</th>
<th>RE Caregiver</th>
<th>RE Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>No RE</td>
<td>RE Caregiver</td>
<td>-20.4</td>
<td>29.2</td>
<td>122.9</td>
<td>221.0</td>
</tr>
<tr>
<td>RE Child</td>
<td>RE Caregiver</td>
<td>-23.3</td>
<td>-10.1</td>
<td>130.2</td>
<td>142.2</td>
</tr>
<tr>
<td>RE Caregiver</td>
<td>RE Both</td>
<td>-23.0</td>
<td>15.5</td>
<td>-1.5</td>
<td>48.7</td>
</tr>
<tr>
<td>RE Both</td>
<td>RE Both</td>
<td>-31.7</td>
<td>-17.4</td>
<td>-2.3</td>
<td>18.4</td>
</tr>
</tbody>
</table>

\textsuperscript{a} Computed as \((TM_2 - TM_1)/TM_1 \times 100\%\).
\textsuperscript{b} There are four potential states at each point of behavioral time series: No Reaching Event (No RE), Child Reaching Event (RE Child), Caregiver Reaching Event (RE Caregiver), Caregiver and Child Reaching Events happening at the same timepoint (Both RE). Each potential state transition is defined by an entry in the matrix. The value corresponding to each transition is the relative percentage difference between transition probabilities of dyads in Cluster 2 and Cluster 1 referenced to dyads in Cluster 1.

Relationships between cluster classification and clinical characteristics

Clusters obtained in MixMM also differed significantly in terms of their VABS-3 Communication and Socialization Domain Scores and Verbal IQ, as measured by one-way ANOVA model \(F(1,76) = 9.522, f = 0.35, p=0.003\) for VABS Communication;

\(F(1,76)=5.863, f = 0.28, p= 0.018\) for VABS Socialization; \(F(1,76)=10.53, f = 0.37, p=0.002\) for Verbal IQ). Children in Cluster 2 were found to have lower communication, socialization, and language skills (See Figure 3).
Figure 3: Differences on VABS Communication and Socialization, and DAS Verbal IQ scores for Cluster 1 versus Cluster 2 revealed by Mixed Markov Model. Cluster 2 is characterized by more frequent caregiver responsiveness to the child’s behavior and more overall reaching movement.

We next examined whether autism, autism+ADHD, ADHD, and NT children were more likely to fall into Cluster 1 or 2. Results showed that clusters were not associated with diagnostic groups or with demographic characteristics based on a Chi-squared test (see Table 19 in Appendix A for details). Additionally, clusters were not associated with age (F(1,76)=0.008, p=0.927). Thus, levels of language and social abilities were related to cluster membership, but not to clinical diagnosis.

Next, using models M1-M10 listed above, we examined how adding cluster variable to group and age helps to explain the variance in clinical scores. F-statistics, p-
values and effect sizes (Eta-squared and Cohen’s f) for each variable of the sequentially applied linear models (M1-M10) and adjusted $R^2$ per each model are shown in Table 21 and Table 22 in Appendix A.

In a set of baseline models, though Age was significantly associated with the clinical scores (model M1, adjusted $R^2 = 0.10, 0.05$, and 0.09, for VCom, VSoc, and VIQ, respectively), the biggest single contribution to the variances of clinical scores was by diagnostic Group (model M2, adjusted $R^2 = 0.33, 0.32$, and 0.27, for VCom, VSoc, and VIQ, respectively). Adding Group variable to explain variance unexplained by Age helps to explain the data better (models M3 and M4).

In a set of models exploring potential of the Cluster variable to explain the variance in data, Cluster by itself (model M5) demonstrated adjusted $R^2 = 0.10$, 0.06 and 0.11 for VCom, VSoc and VIQ measures. However, both separately adding Age (model M6) and Group (model M7), and then further adding both Age and Group (model M8), interaction of Cluster and Age (model M9), and Cluster and Group (model M10) significantly increased the model’s explained variance. The model that explained the data best was M9, which included the interaction Age*Cluster term. It showed adjusted $R^2 = 0.44$, 0.40 and 0.36 for VCom, VSoc and VIQ respectively. Note that for VIQ, measure model M10 was on par with M9 on adjusted $R^2$; however, neither the Age*Cluster term in M9, nor the Group*Cluster term in M10 were significant. For VCom and VSoc, the Age*Cluster interaction term in model M9 was significant ($F(1,71) = 4.216$, $f = 0.24$, $p =$
0.044 and $F(1,71) = 4.798$, $f = 0.26$, $p = 0.032$ for $V_{Com}$ and $V_{Soc}$, respectively). It implies different patterns of association between $V_{Com}$ and $V_{Soc}$ and $Age$ for different clusters, which is shown in Figure 4. Specifically, increase of age in Cluster 2 was associated with increase of $V_{Com}$ and $V_{Soc}$ scores, while there was no significant association of age and VABS scores in Cluster 1.

![Figure 4: Association of VABS scores and Age in Cluster 1 (C1) and Cluster 2 (C2). In Cluster 2, which was characterized by more caregiver responsiveness and more overall reaching behavior, VABS Communication and Socialization scores are positively associated with age.](image)

**Discussion**

We applied computer vision analytics to a 6-minute, lab-based, caregiver-child free play interaction to measure initiating and responding to reaching for a toy in caregiver-child dyads and examined correlations of these interaction patterns with
standardized clinical measures of children’s communication, social, and verbal skills. These methods detected two clusters of dyads that differed in their overall amount of reaching movements and in their patterns of caregiver responsiveness. Specifically, as compared to Cluster 1, dyads in Cluster 2 exhibited reaching movements more frequently and the caregiver was more likely to respond to the child’s reaching for a toy by also reaching for a toy. Children in Cluster 2 were also found to have lower levels of communication, social, and verbal skills. Alternatively, children who had higher social and language skills were more likely to be in Cluster 1, in which caregivers were less likely to be contingently responding to their child’s reaching for a toy. Furthermore, we found that, within Cluster 2 which was characterized by caregiver responsiveness and more movement, children’s age was also correlated with level of social abilities; specifically, children who were older had higher social abilities. Interestingly, this relationship was not found for children in Cluster 1, which was characterized by non-contingent caregiver responding and less movement. We also found that cluster membership was not associated with a child’s diagnostic status or age. Cluster membership was predictive of a child’s language and social abilities, but not their clinical diagnosis of autism or ADHD.

**Caregiver responsiveness and language abilities**

While caregiver responsiveness has been identified as a mechanism for facilitating language and social growth in autistic children, longitudinal studies of
caregiver-child interaction have identified important changes in the nature of caregiver responsiveness as a child acquires more advanced social and language skills. Bornstein and colleagues defined caregiver responsiveness as the prompt, contingent responses to their child’s activities, which can include toy exploration and verbalizations (Bornstein et al., 2008). In a prospective longitudinal study, these authors found that as children acquired more advanced language skills, caregivers’ reduced their responsiveness to children’s exploratory toy behavior (i.e., manipulating objects) and increased their responsiveness to their child’s verbalizations by imitating and expanding on their verbalizations. These results may help to explain the finding in the present study that children with more advanced language were more likely to be in Cluster 1, which was characterized by lower levels of caregiver contingent responding to the child’s reaching for a toy. It would be of interest in future analyses to explore whether the lower level of caregiver’s contingent responding to the child’s reaching behavior was accompanied by an increased level of contingent responding to the child’s verbalizations.

It was of particular interest that associations between caregivers’ contingent responsiveness was associated with children’s language and communication abilities, but did not differ based on diagnostic group. This finding suggests that the relationship between key features of caregiver behavior in the context of caregiver-child interaction, namely, contingent responsiveness and language acquisition, are trans-diagnostic,
reflecting general principles regarding the importance of caregiver-child interaction for facilitating children’s acquisition of communication skills (Tomasello & Farrar, 1986).

**Limitations**

A limitation of the study is that only reaching behavior was considered in the analysis. As an initial attempt to apply computer vision to the measurement of dyadic interaction, we chose reaching events derived from bending angle because it has a strong signal-to-noise ratio. Furthermore, it is possible that more frequent reaching behaviors in Cluster 2 might be a sign of more movement activity of children in Cluster 2 in general, although the lack of an association between an ADHD diagnosis and Cluster 2 membership does not support this. As responsiveness is a multi-modal concept, physical behavior, like reaching movement, can elicit a reply in other modalities, such as verbalizations as mentioned above. Adding automated detection of vocalizations is a natural extension of this study.

An additional limitation of the study is the quality of reconstructing 3D from 2D and extracting ‘reaching’ signals, which is reflected in the specificity levels of reaching events detection of 0.76 and 0.70 for caregiver and child respectively. Using 3D depth cameras and estimating landmarks from the 3D data can be a potential solution in the future. Usage of purple and red T-shirts to identify the participants constrains the proposed method’s scalability, and can be addressed in the future studies by using more advanced machine learning technologies for human detection.
Analysis of the behavior only in the center of the room is another study limitation coming from the method constraints. When children are standing, which inevitably happens when they approach the table in the corner of the room, or walk away from the toys on the floor, it is not possible to detect reaching events both due to the relative position of child’s body to the camera and the absence of torso rotation needed to reach to the object on the table. Nevertheless, this happens rarely because toys for this part of the caregiver-child interaction are intentionally placed on the floor and participants are specifically instructed to move to the floor in the center of the room to play with the toys from the box.

Another limitation of the study is the sample, which was relatively small, limited diversity, and included small neurotypical comparison group. Seventy percent of the sample was White and the mean maternal education level was at least a bachelor’s degree. Children with other co-occurring conditions, such as anxiety, were not included which limits the generalizability of the findings. Future studies on CVA analysis of caregiver-child interaction should address this by including a more diverse sample. Specifically, the interaction of age and cluster findings, which have the least effect size, should be verified by future studies.

**Conclusions**

The long-term goal of this initial feasibility study is to use automated methods for micro-analytic coding to measure changes in caregiver-child interactions in the
context of clinical trials. This preliminary work is a first step in applying automated, tailored methods to naturalistic caregiver-child interactions to detect different behavioral patterns of dyads during free play. We presented the first attempt to micro-analytic coding of caregiver-child interactions with a focus on reaching behavior, which was used as a proxy for initiating and responding to a toy bout. Our findings offer promise that computer vision analysis of dyadic behavior may provide a useful way of automatically and objectively characterizing clinically meaningful aspects of caregiver-child interaction.
Chapter 4. Use of automatic speech recognition models to create biomarkers for cerebellar ataxia

Speech impairments are a common symptom of cerebellar ataxias, a family of disorders caused by damage of the cerebellum of various etiology. Common assessments of ataxia in the clinic, such as the Brief Ataxia Rating Scale (BARS), provide subjective and coarse measurements of ataxic speech. Thus, there’s a need for biomarkers that provide a more objective and granular assessment enabling both clinical trials and diagnostics. In this chapter, we focus on building such biomarkers from conversational speech using a pre-trained automatic speech recognition network. Given that distortion of vowels is a well-described clinical sign of ataxic speech, we propose Average Vowel Entropy (AVE), i.e. the entropy of vowel predictions by the speech recognition deep learning algorithm as a potential digital behavioral biomarker.

Automatic speech recognition additionally allowed to apply to conversational speech measures of variation of fundamental frequency and intensity which previously were used only in standardized tests (sustained phonation and oral diadochokinesis). We show that the entropy of vowels is associated with ataxia severity as measured by BARS and can distinguish mild to moderate ataxia from the control group, demonstrating its monitoring and diagnostic value. As a complementary measure, variation in intensity allows for discrimination between ataxia and controls at a very early stage of disease development. Additionally, we provide evidence that AVE can track longitudinal ataxia
progression more granularly than clinical scales (BARS), making it a promising biomarker for clinical trials.
Introduction

Cerebellar ataxias are a group of heterogeneous disorders that affect coordination, balance, and speech (Klockgether, 2007). Motor speech impairments in ataxia are known as ataxic dysarthria, a perceptually distinct motor speech disorder that emerges due to changes on respiratory, phonatory, resonatory, and articulatory levels, with the most prominent characteristics being articulation and prosody (Duffy, 2005; Ziegler, 2016). Ataxic speech characteristics arise from inaccuracy of force, range, and timing and directions of speech movements, and speech is commonly perceived as poorly timed and coordinated. Dysarthria is a common manifestation across cerebellar ataxias and leads to impairments in communication, reduced social connections, and decreased quality of life (Gibilisco & Vogel, 2013).

Among the most prominent and specific signs of ataxic dysarthria are prosodic abnormalities, such as pitch and loudness variations, along with speech rhythm and rate abnormalities. Distorted vowels are a hallmark of ataxic dysarthria perception, though consonant production can also be impaired (Darley et al., 1969; Duffy, 2005; Kent et al., 2000; Kent et al., 1997; Kent et al., 1979). Ataxia rating scales, including the International Cooperative Ataxia Rating Scale (Trouillas et al., 1997), the Brief Ataxia Rating Scale (BARS) (Schmahmann et al., 2009), and the Scale for Assessment and Rating of Ataxia (SARA) (Weyer et al., 2007), assess the intelligibility and fluency of speech in individuals with ataxia. These scales produce subjective and relatively coarse assessments of ataxic
dysarthria that limit their ability to detect small changes in speech over time as part of clinical care or clinical trials.

Acoustic analysis is an objective approach, serving as a quantitative way to represent perceived speech qualities. In acoustic studies, intelligibility as a perceptual characteristic was found to rely more on vowels than on consonants, both for young normal-hearing and elderly listeners with hearing impairment (Kewley-Port et al., 2007). Prosodic information is mostly carried by vowels, making acoustic analysis of vowels a potential way to objectively quantify differences in dysarthric and typical speech as well as between different diseases causing dysarthria (Lansford & Liss, 2014a, 2014b). Vowel acoustics is generally studied via the estimation of different statistics of intensity, fundamental frequency (F0), and first and second formants (F1 and F2). F1 and F2 frequency values are generally treated as proxies for tongue position in the superior-inferior and anterior-posterior axes respectively, generating a concept of Vowel Space Area (VSA). VSA is an area in the F1-F2 plane occupied by vowels of a particular person, and can be computed using various methodologies (see Kent and Rountrey (2020) for review). As a first approximation, it is considered to be a link between kinematics and acoustics of vowel production. Well established techniques exist for automated estimation of fundamental frequency and formants in the voiced intervals (see e.g., Praat package (Boersma & Weenink, 2021) for phonetic speech analysis). In previous studies of voice in ataxic dysarthria, sustained phonation and syllable repetition tests provided
evidence of difference in F0 and intensity variability between patients with ataxic
dysarthria and control participants (Kent et al., 2000). Prosodic abnormalities such as
abnormal stress patterns, influencing the duration of the vowels and speech rhythm,
were also found to be characteristic of the disease (Odell et al., 1991). Vowel Space Area
in ataxic patients can be reduced, as shown in a study of single-word pronunciation
(Delgado-Hernandez, 2017). All these changes are related to a perceived level of speech
intelligibility (Borrie et al., 2017; Delgado-Hernandez, 2017; Liss et al., 2000). However,
measures like Vowel Space Area and vowel durations are problematic to study in
conversational connected speech due to coarticulation effects which cannot be controlled
(Hertrich & Ackermann, 1999; Kent & Rountrey, 2020). Furthermore, vowel lengths can
also be influenced by the complexity of the articulatory gestures needed to pronounce
the words, which also is not controlled in experiments with conversational speech
(Ackermann et al., 1999). Additionally, vowel segments extraction in abnormal speech
often requires human involvement, which poses scalability and objectivity challenges.

Recent works on self-supervised speech representation learning for automatic
speech recognition via neural networks (ASR-NN) have shown great performance
(Baevski et al., 2020; Chi et al., 2020; Liu et al., 2020; Liu et al., 2019; Song et al., 2019).
Wav2vec 2.0 (wav2vec2, (Baevski et al., 2020)) achieved among the best up-to-date
performance, reaching a phoneme error rate metric of 8.3 percent when fine-tuned on
phonetic annotations of TIMIT acoustic-phonetic corpus of English language (Garofolo

53
et al., 1993), and 1.8 percent word error rate when fine-tuned on the LibriSpeech corpus (Panayotov et al., 2015). Though the speech representations in such models are not interpretable, their performance on speech-to-text and speech-to-phonemes tasks hints at the potential representation of the phonetic structure of speech in the latent layers of the neural network. The Average Vowel Entropy (AVE) measurement proposed in this paper is a novel biomarker built on this intuition. It assesses the entropy (uncertainty) of vowel phoneme prediction. Additionally, alignment of phonemes and audio signal (Zhu & Zhang, 2021) allowed to compute F0 and intensity variability in vowel segments.

Speech biomarkers for cerebellar dysfunction have been studied as a potential outcome measure for clinical trials. Dysphonia measures, including F0 Coefficient of Variance, as well as speech rate (syllables/s) measured in connected speech and syllable repetition tasks, and the fraction of pauses showed correlation with cerebellar dysfunction rating scales (in Friedreich’s ataxia and multiple sclerosis cohorts), as well as magnetic resonance imaging (MRI) and life quality measures (in multiple sclerosis cohort) (Blair et al., 2019; Noffs et al., 2020; Vogel et al., 2020; Vogel et al., 2017). In a cerebellar ataxia population, a study of features derived from phase-based and magnitude-based cepstral analysis of speech achieved 84% accuracy in cerebellar ataxia-control discrimination, as well as 74% accuracy in predicting SARA rating scale reduced to 3 levels (mild/moderate/severe) (Kashyap et al., 2020). The usage of automatic speech recognition has the potential to evaluate ataxic dysarthria both from acoustic (by
extracting vowel segments and analyzing their characteristics) and speech recognition perspective, the latter being regarded as a proxy for intelligibility.

A study of speech intelligibility using phonetic contrasts in Friedreich’s ataxia provided evidence that intelligibility was related to laryngeal timing and control (Blaney & Hewlett, 2007). Kent et al. (Kent et al., 2000) suggested that symptoms of ataxic dysarthria that are substantially related to laryngeal timing and control, such as irregular phonation, altered articulatory/voicing control, as well as poorly coordinated respiratory function, were represented by variability of fundamental frequency and intensity in sustained phonation and syllable repetition. In contrast to prior studies, we were interested in assessing vowel characteristics during more natural speech tasks with connected speech. Using speech recognition neural networks, we extracted vowel segments automatically and computed F0 and intensity variability of vowel segments (mean pitch standard deviation, MPSD, and mean intensity standard deviation, MISD) across all vowels in each subject’s speech data. Even though MPSD and MISD are not the same measures as variability metrics in sustained phonation and syllable repetition, we hypothesized that they may be a marker of distorted vowels with regard to respiration and phonation. In this work, we assess the potential of the proposed AVE, MPSD, and MISD, as digital biomarkers of ataxia severity. The convergent validity of these measures with the BARS speech and total score is evaluated, and the potential of
these measures to be more granular than clinical ratings is analyzed with longitudinal data of patients with cerebellar ataxia.

**Methods**

**Participants**

All experimental protocols were approved by the Partners Healthcare Institutional Review Board (Protocol# 2016P001048) and were in accordance with guidelines of the Declaration of Helsinki. All participants provided written informed consent and/or assent prior to participation in the study. Participants were recruited from the Massachusetts General Hospital (MGH) between September 2017 and March 2020 from the Ataxia and Movement Disorders Units. Additionally, individuals with ataxia-telangiectasia (A-T) were recruited through the Ataxia-Telangiectasia Children’s Project. Healthy control data were obtained from family members of patients (e.g., asymptomatic partners or gene negative family members) and MGH staff.

Participants underwent a BARS assessment and performed a speaking task which was recorded on an iPhone.

Ataxia and control participants older than 15 years were selected; the age limit was considered due to significant changes in voice characteristics during childhood and puberty (Kent & Vorperian, 2018; Ludlow et al., 2014). Furthermore, participants with ataxia were included only if they received a BARS speech assessment. This resulted in 86 participants, including 61 participants with ataxia (35 males, age range 16-82 y.o.), and
25 Controls (9 males, age range 18-86 y.o). Details of participants’ demographics as well as specific subtypes of ataxia are provided in Table 4.
Table 4: Demographic information of the participants included in the analysis.

<table>
<thead>
<tr>
<th>Specific Diagnosis</th>
<th>Age in Decades</th>
<th>N</th>
<th>Male (%)</th>
<th>BARS speech</th>
<th>BARS total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control group</td>
<td>2nd to 9th</td>
<td>25</td>
<td>36</td>
<td>0</td>
<td>1.5 1.5</td>
</tr>
<tr>
<td>1 Ataxia-Telangiectasia (A-T)</td>
<td>2nd to 3rd</td>
<td>5</td>
<td>60</td>
<td>1.5 3</td>
<td>N/A N/A</td>
</tr>
<tr>
<td>2 Adult onset cerebellar ataxia, other</td>
<td>3rd to 8th</td>
<td>19</td>
<td>59</td>
<td>0 4</td>
<td>0 22.5</td>
</tr>
<tr>
<td>3 Autosomal recessive cerebellar ataxia type 1 (ARCA1)</td>
<td>5th to 7th</td>
<td>2</td>
<td>100</td>
<td>2 2.5</td>
<td>12 23.5</td>
</tr>
<tr>
<td>4 Ataxia following intracerebral hemorrhage</td>
<td>4th</td>
<td>1</td>
<td>0</td>
<td>1.5 1.5</td>
<td>11.5 11.5</td>
</tr>
<tr>
<td>5 CANVAS</td>
<td>9th</td>
<td>1</td>
<td>1</td>
<td>0 0</td>
<td>4 4</td>
</tr>
<tr>
<td>6 Fragile X associated tremor ataxia syndrome (FXTAS)</td>
<td>8th</td>
<td>1</td>
<td>1</td>
<td>0 0</td>
<td>3 3</td>
</tr>
<tr>
<td>7 Friedreich’s Ataxia</td>
<td>4th to 7th</td>
<td>3</td>
<td>66</td>
<td>1.5 2</td>
<td>13.5 19</td>
</tr>
<tr>
<td>8 Gordon Holmes syndrome</td>
<td>4th</td>
<td>1</td>
<td>0</td>
<td>2 2</td>
<td>14.5 14.5</td>
</tr>
<tr>
<td>9 Multiple System Atrophy, cerebellar type (MSA-C)</td>
<td>6th to 8th</td>
<td>2</td>
<td>50</td>
<td>1 1</td>
<td>4 9</td>
</tr>
<tr>
<td>10 Progressive Supranuclear Palsy, with predominant cerebellar ataxia (PSP-C)</td>
<td>7th</td>
<td>1</td>
<td>0</td>
<td>1 1</td>
<td>7 7</td>
</tr>
<tr>
<td>11 SCA1</td>
<td>5th to 7th</td>
<td>3</td>
<td>0</td>
<td>1 1</td>
<td>5.5 11</td>
</tr>
<tr>
<td>12 SCA14</td>
<td>6th</td>
<td>1</td>
<td>100</td>
<td>0.5 0.5</td>
<td>7 7</td>
</tr>
<tr>
<td>13 SCA15</td>
<td>6th</td>
<td>1</td>
<td>0</td>
<td>1 1</td>
<td>11.5 11.5</td>
</tr>
<tr>
<td>14 SCA2</td>
<td>6th</td>
<td>1</td>
<td>100</td>
<td>1 1</td>
<td>13.5 13.5</td>
</tr>
<tr>
<td>15 SCA3</td>
<td>5th to 7th</td>
<td>6</td>
<td>33</td>
<td>0 1</td>
<td>4 13</td>
</tr>
<tr>
<td>16 SCA5</td>
<td>3rd</td>
<td>1</td>
<td>0</td>
<td>2 2</td>
<td>16.5 16.5</td>
</tr>
<tr>
<td>17 SCA6</td>
<td>7th to 8th</td>
<td>8</td>
<td>75</td>
<td>0.5 2.5</td>
<td>2 20</td>
</tr>
<tr>
<td>18 SCA8</td>
<td>5th</td>
<td>1</td>
<td>100</td>
<td>1.5 1.5</td>
<td>9 9</td>
</tr>
<tr>
<td>19 Sensory and cerebellar ataxia</td>
<td>7th</td>
<td>1</td>
<td>100</td>
<td>0.5 0.5</td>
<td>12 12</td>
</tr>
<tr>
<td>20 Spastic paraplegia type 7 (SPG7)</td>
<td>6th to 7th</td>
<td>2</td>
<td>100</td>
<td>0 0.5</td>
<td>8 10.5</td>
</tr>
<tr>
<td>Total</td>
<td>2nd to 9th</td>
<td>86</td>
<td>44</td>
<td>0 4</td>
<td>0 23.5</td>
</tr>
</tbody>
</table>

A-T patients did not have BARS total assessments.
Abbreviations: SCA – spinocerebellar ataxia; CANVAS – Cerebellar ataxia with neuropathy and vestibular areflexia syndrome;

Among the 61 participants with ataxia, all underwent at least one BARS Speech assessment and performed the speaking task, 55 had at least one full BARS assessment,
and 8 had data from multiple timepoints with at least two full BARS assessments and the speaking task. The first timepoint of assessment which included BARS Speech was taken for the cross-sectional analysis. For longitudinal assessments, values from the first and last timepoint with full BARS assessment were considered. The demographics and clinical details of the participants included in longitudinal assessment are provided in Table 5.

**Table 5: Details of age and time interval between first and second timepoints (TP) for eight subjects included in longitudinal analysis.**

<table>
<thead>
<tr>
<th>ID</th>
<th>Diagnostic group</th>
<th>Specific diagnosis</th>
<th>Decade of life</th>
<th>TP2 - TP1 (months)</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ataxia</td>
<td>Progressive Supranuclear Palsy, with predominant cerebellar ataxia (PSP-C)</td>
<td>7th</td>
<td>7</td>
<td>F</td>
</tr>
<tr>
<td>2</td>
<td>Ataxia</td>
<td>Multiple System Atrophy, cerebellar type (MSA-C)</td>
<td>8th</td>
<td>7</td>
<td>M</td>
</tr>
<tr>
<td>3</td>
<td>Ataxia</td>
<td>SCA1</td>
<td>6th</td>
<td>16</td>
<td>F</td>
</tr>
<tr>
<td>4</td>
<td>Ataxia</td>
<td>SCA6</td>
<td>7th</td>
<td>9</td>
<td>M</td>
</tr>
<tr>
<td>5</td>
<td>Ataxia</td>
<td>Multiple System Atrophy, cerebellar type (MSA-C)</td>
<td>6th</td>
<td>16</td>
<td>M</td>
</tr>
<tr>
<td>6</td>
<td>Ataxia</td>
<td>SCA6</td>
<td>8th</td>
<td>17</td>
<td>F</td>
</tr>
<tr>
<td>7</td>
<td>Ataxia</td>
<td>Multiple System Atrophy, cerebellar type (MSA-C)</td>
<td>6th</td>
<td>17</td>
<td>F</td>
</tr>
<tr>
<td>8</td>
<td>Ataxia</td>
<td>Polyneuropathy with mild cerebellar motor syndrome</td>
<td>7th</td>
<td>12</td>
<td>F</td>
</tr>
</tbody>
</table>

**Clinical data collection.**

Ataxia participants were scored by an ataxia specialist on the Brief Ataxia Rating Scale using the half-point version (Zhou et al., 2022). BARS is a clinical score with 5 subscales, including gait, knee-tibia test, finger-to-nose test, dysarthria, and oculomotor abnormalities. The dysarthria subscale provides a score between 0 and 4, with 0.5-point
increments, where 0 is normal speech and 4 is absent or unintelligible speech. Scores of 0.5 indicate abnormality of impairments of rate/rhythm/clarity only during syllable repetition test but not apparent during conversational speech, while scores ≥ 1 indicate the appearance of general conversational speech intelligibility impairments, ranging from mild to severe.

**Audio data collection and extraction.**

In the speaking task, participants were presented the ‘Cookie Theft’ picture from the Boston Diagnostic Aphasia Examination (Goodglass et al., 2001) on a 12.9-inch iPad Pro (2nd gen). Video instructions were presented via a custom iPad application which asked the participant to describe in as much detail as possible what they see in the picture after they hear the tone beep, and say ‘I’m done’ when they were finished. Video and audio of participants’ behavior during the task was recorded with an iPhone device placed immediately below the iPad in a portrait mode. Details of the setting can be found in (Chang et al., 2020). iPhone video and audio recording started at the beginning of the instruction, but the ‘Cookie Theft’ picture was not displayed on the iPad until the tone beep. After the instruction video was completed, the picture appeared on the iPad screen synchronously with the tone beep. When participants had finished describing the picture, the research staff member stopped the iPhone recording.

The audio stream was extracted from the iPhone recording for further analysis. Raw audio resolution was 44100 Hz. Tone beep was detected on the recording using in-
house custom python code, and the segment of audio recording after the tone beep was extracted.

**Neural network training.**

Before computing the entropy metrics, a wav2vec2 pretrained on 960 hours of the LibriSpeech corpus (Panayotov et al., 2015), denoted in (Baevski et al., 2020) as wav2vec2 Large LS-960, was fine-tuned to predict phonemes. Fine-tuning was needed since, by construction, wav2vec2 is a self-supervised model, meaning it first learns speech representations on a large set of unlabeled data and then can be adjusted to the task by providing a much smaller set of labeled data. In our case, the task was phoneme prediction, and the train subset of the TIMIT dataset (Garofolo et al., 1993) with phonetic annotations was used for fine-tuning. “Transformers” python package (Wolf et al., 2020) was used in the fine-tuning process. The fine-tuned neural network (FTNN) reached Phoneme Error Rate (PER) of 8.6 on a test subset of TIMIT.†

**Entropy extraction.**

The FTNN was run in inference mode for each participant. FTNN predicts a probability distribution of potential phoneme tokens (42 tokens total, following (Lee & Hon, 1989), including 14 vowel tokens and 4 special tokens) per 25 ms segment of the

---

† Fine-tuning was done following the same procedures as in the original speech-to-text training in Baevski et al. (2020). The original paper by Baevski et al. (2020) reported PER of 8.3 on TIMIT dataset, however the model was not released, and likely the discrepancy is caused by a slight difference in training parameters.
audio recording with 20 ms timesteps. For each segment, the most probable token was detected by computing the argument of the maximal FTNN prediction, and then the class of the most probable token was detected (vowel/consonant/special token). If the most probable token belonged to a vowel class, the entropy (a proxy for uncertainty) value was computed over the entire prediction distribution using the standard Shannon entropy $E = -\sum_{i=1}^{N} p_i \log p_i$ (Shannon & Weaver, 1949), where $N$ is the number of tokens in the prediction head. This was repeated for each 25 ms audio segment, and a time array of vowel entropies was formed per each participant’s recording. Then, the values of the array were averaged, forming the proposed Average Vowel Entropy (AVE) value per participant.¹

**Acoustic measures of vowels in vowel segments extraction.**

A vowel segment was defined as the period between the start timepoint of the vowel token and the start timepoint of the next token. The time of start of the token in the audio stream is equal to the index of the token in the inferred array of tokens multiplied by the FTNN input timestep (20ms). In each vowel segment, F0 (in Hz) and intensity (in dB) time series were extracted using the python package “parselmouth,” an interface to the Praat phonetic analysis software (Boersma & Weenink, 2021; Jadoul et al., 2018). Extraction of F0 via an independent tool (Praat) served as a quality control for

¹ The code for average vowel entropy computation is made publicly available at https://github.com/dyisaev/average-vowel-entropy.
vowel recognition for all subjects. As such, only vowel segments where the F0 measurement was present for the entire duration of the segment were included in the analysis. Standard deviation of F0 and intensity were extracted per each vowel segment and averaged, resulting in mean pitch standard deviation (MPSD), and mean intensity standard deviation (MISD).

**Statistical analysis.**

The aim of the statistical analysis was five-fold, (1) to evaluate the potential of AVE and acoustic measures to distinguish between ataxia and control groups; (2) to evaluate the convergent validity of AVE and the acoustic measures as a measure of ataxic dysarthria and overall ataxia severity; (3) to assess the sensitivity of AVE and acoustic measures to capture ataxia progression in comparison with the clinical rating scale (BARS); (4) to compute associations of AVE with acoustic measures as an interpretation of the information potentially captured by AVE; (5) to assess test-retest reliability of AVE, MPSD, and MISD.

Tukey’s Ladder of Powers method (Tukey, 1977)§ was applied to transform AVE, MPSD, and MISD to ensure normality of the metrics and to eliminate outliers. The Shapiro-Wilks test (Shapiro & Wilk, 1965) was then performed to assess normality of the distribution of transformed AVE and acoustic measures. Since the Ladder of Powers

---

§ Implemented in R as the `tukeyTransform` function in `rcompanion` package
method results in a monotonic transformation, it does not influence the Spearman’s correlations used in subsequent analysis.

Using a linear regression model, the analysis of associations between the extracted features and age in the control group was performed to rule out the potential effects of age and make a decision on whether to include age as a covariate in subsequent analysis. Then Student’s t-test was performed to compare the means of transformed AVE and acoustic measures between ataxia and control groups in three ways: all ataxia participants (N=61) vs. controls, individuals with ataxia and a BARS speech score less than or equal to 1.5 (N=41) versus controls, and individuals with ataxia with BARS speech score less than or equal to 0.5 (N=19) versus controls. This was done to additionally test the sensitivity of measures to mild speech impairment. Next, the associations between the extracted features and ataxia severity were analyzed via Spearman’s correlation with BARS speech and BARS total scales in the ataxia group. Additionally, to test for the significance of the association between BARS speech and the extracted features, an interval regression model with BARS speech as a dependent variable and speech features, age (if required), and sex as independent variables, was fitted (Long, 1997). The R intReg package was used for fitting the interval regression models.
To test the ability of extracted speech features to capture disease progression, a Wilcoxon signed rank paired test was performed between the two timepoints on BARS Speech, BARS Total, and each of the extracted features.

Linear regressions with AVE as a target variable and acoustic measures as predictors were performed to assess associations between AVE and acoustic measures.

To assess test-retest reliability of measures, for each subject, all extracted vowels were split into two partitions according to the time of vowel appearance in the speech recording, and values of AVE, MPSD and MISD were computed for each partition. Then Pearson’s correlation was computed between values of first and second partition.

**Results**

**Transformation of metrics and effect of age.**

The Shapiro-Wilks test of normality (Shapiro & Wilk, 1965) after the Tukey’s Ladder of Powers transformations was not significant for AVE and acoustic measures, indicating data conformity to normal distributions. None of the features were significantly associated with age in the control group. Age was then excluded from any subsequent analysis.

**Comparison of ataxia and control groups.**

In the ataxia group, out of 61 patients, 19 patients had a BARS speech score less than or equal to 0.5, and 41 patients had a BARS speech score less than or equal to 1.5. A summary of BARS speech scores in the ataxia group can be found in Table 6.
Table 6: Summary of BARS speech scores.

<table>
<thead>
<tr>
<th>BARS speech score</th>
<th># of patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>0.5</td>
<td>8</td>
</tr>
<tr>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>1.5</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>2.5</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

Means of transformed AVE and MISD were significantly higher in the ataxia group, compared with the control group, as measured by Student’s t-test ($t = 3.26, p = 0.002$ for AVE and $t = 4.36, p < 0.001$ for MISD). Medians and confidence intervals of raw AVE and MISD values in both groups are listed in Table 7.

Table 7: Medians and confidence intervals (CIs) for raw Average Vowel Entropy (AVE) and Mean Intensity Standard Deviation (MISD).

<table>
<thead>
<tr>
<th>Measure</th>
<th>Ataxia</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Vowel Entropy (AVE)</td>
<td>0.48 [0.41, 0.52]</td>
<td>0.33 [0.26, 0.41]</td>
</tr>
<tr>
<td>Mean Intensity Standard Deviation (MISD),dB</td>
<td>2.43 [2.27, 2.56]</td>
<td>2.07 [1.92, 2.20]</td>
</tr>
</tbody>
</table>

Based on t-test results, means of transformed AVE did not significantly differ between the control group and the subgroup of ataxia patients with a BARS speech score less than or equal to 0.5. Means of transformed MISD were significantly different between the two groups ($t = 2.11, p = 0.041$), with mean MISD of the control group being lower. For transformed AVE and MISD, control group mean was lower than mean of the subgroup of ataxia patients with BARS speech less than or equal to 1.5 ($t = 2.11, p = 0.041$).
for AVE, \( t = 2.96, p = 0.0046 \) for MISD). Differences between subgroups of ataxia with BARS speech score less than or equal to 0.5, 1.5, and the entire ataxia and control groups are shown in Figure 5.

 Means of transformed MPSD did not differ between ataxia group and control group in full ataxia group as well as in the subsets of ataxia group with low BARS speech scores (\( t = 0.88, p=0.38 \) for full group; \( t = 0.001, p=0.99 \) for BARS speech \( \leq 0.5 \); \( t = 0.60, p=0.55 \) for BARS speech \( \leq 1.5 \)).

 Figure 5: Comparison of control group and subgroups of ataxia with BARS speech less or equal than 0.5 and 1.5, and the entire ataxia group on transformed Average Vowel Entropy (top row), and Mean Intensity Standard Deviation (bottom row). P-values of Student’s t-test between the groups are displayed on the plots.
Associations of AVE and acoustic measures with BARS.

AVE and MISD were moderately correlated with BARS Speech (Spearman’s $\rho = 0.45$ and $0.51$ respectively, $p<0.001$), while MPSD did not significantly correlate with BARS Speech. AVE, but not MPSD or MISD, correlated with BARS Total ($\rho = 0.39$, $p=0.003$). In the interval regression models, associations of AVE and MISD with BARS Speech were significant ($t = 3.4$, $p=0.001$ in both cases) but not significant for MPSD. Associations of AVE and MISD with BARS Speech and Total are shown in Figure 6.

**Figure 6:** Associations between BARS speech and BARS total with the proposed Average Vowel Entropy (AVE) and Mean Intensity Standard Deviation (MISD) in the ataxia cohort. Spearman’s correlations ($\rho$) and $p$-values ($p$) for each pair of features are shown in the graph. Regression estimates visualizations are provided for associations with BARS speech and total.
Longitudinal analysis.

Wilcoxon signed rank paired tests demonstrated significant change over time (median 12, IQR 8.5 months) for BARS Total and AVE ($p=0.022$ and $p=0.039$ respectively for both metrics), and were not significant for BARS speech and acoustic measures. Figure 7 shows the dynamics of metrics across pairs of timepoints. Out of 8 individuals, 7 individuals showed an increase in AVE over time, representing increased uncertainty of vowels, and the same individuals had an increase in BARS total. Diagnoses of those subjects included spinocerebellar ataxias, multiple system atrophy of the cerebellar type (MSA-C), and progressive supranuclear palsy with predominant cerebellar ataxia (PSP-C) (see Table 5). The remaining individual had a diagnosis of polyneuropathy with mild cerebellar motor syndrome and demonstrated a slight decrease in AVE over time and no increase in BARS total and speech on the second assessment. In comparison, only two out of 8 individuals had an increase in BARS speech score during this time interval and the other 6 individuals demonstrated no change.
Figure 7: Longitudinal dynamics of clinical measures and proposed metrics. Top: Longitudinal dynamics of clinical measures (BARS total and BARS speech*) at two timepoints (TP). Raw scores are used, since only rank changes are assessed. Bottom: Longitudinal dynamics of Average Vowel Entropy (AVE), and Mean Intensity and Pitch standard deviation (MISD and MPSD, respectively). Higher AVE indicates more uncertainty of vowel recognition, and higher MISD and MPSD indicate more voice variability, all of which are characteristics of less intelligible speech. Only BARS total and Average Vowel Entropy have significant difference on Wilcoxon paired signed rank test between timepoint 2 and timepoint 1.

* Jitter (equal at both timepoints) was added to BARS speech values to make trajectories from all subjects visible.
Associations of AVE and acoustic measures.

AVE was significantly positively associated with MPSD (Student’s $t = 3.4$, $p=0.001$, adjusted $R^2 = 0.1$) and MISD ($t = 4.0$, $p<0.001$, adjusted $R^2=0.15$), see Figure 8.

![Figure 8: Associations between transformed Mean Pitch and Intensity Standard Deviations (MPSD and MISD, respectively) and transformed Average Vowel Entropy (AVE). These results provide evidence for a partial explanation of the proposed AVE via acoustic measures.](image)

Test-retest reliability of AVE and acoustic measures

For AVE, MISD and MPSD, Pearson’s correlation values between two partitions were 0.82, 0.79 and 0.65 respectively, indicating good, acceptable and questionable reliability, respectively.
Discussion

We tested the hypothesis that the level of uncertainty in vowel prediction by an automatic speech recognition system could represent speech impairments in individuals with ataxia. The entropy of vowel phoneme predictions by automatic speech recognition system captured vowel distortion resulting from respiratory, phonatory, and articulatory dysregulation in ataxia. Thus, average vowel entropy (AVE) may be considered an intermediate measure that bridges speech perception and acoustic characteristics of the speech such as F0, intensity, and formant dynamics. Furthermore, alignment of phonemes and audio signal (Zhu & Zhang, 2021) enabled us to evaluate average variability of F0 and intensity during vowel segments as potential biomarkers.

We found that AVE had good test-retest reliability and distinguished individuals with ataxia from controls, and was associated with ataxic dysarthria, as indicated by Spearman’s rank correlation between AVE and Brief Ataxia Rating Scale (BARS) speech clinical assessment. While mean intensity standard deviation (MISD) was also associated with BARS speech, only AVE was associated with BARS total. These results indicate the potential of AVE as a digital biomarker for cerebellar ataxia. Additionally, both AVE and MISD demonstrated the ability to distinguish control group from ataxia group of mild to moderate severity (BARS speech less or equal than 1.5). Mean pitch standard deviation (MPSD) could not discriminate ataxia from control group. Absence of findings on MPSD potentially can be a result of heterogeneity of less constrained connected speech data.
compared to sustained phonation or syllable repetition, but also may be a sign that variability of the F0 on short periods of time is not sensitive to ataxic speech disturbances.

AVE was also sensitive to capturing progression of the disease in the subset of 8 subjects who had longitudinal speech and clinical data. Seven of these individuals had diagnoses that would be expected to progress over time (see Table 5) and both AVE and BARS total (but not BARS speech) indicated disease progression in all 7 individuals. The 8th individual had idiopathic polyneuropathy with cerebellar ataxia and did not have a change in BARS total over a 1-year interval. In this individual, AVE demonstrated slight improvement over time. Additionally, AVE was the only derived speech feature that increased in the longitudinal cohort. This indicates that the uncertainty of vowel predictions, as here measured by AVE, may be more sensitive to capturing changes in ataxia severity than standard measures of F0 and intensity.

AVE and MISD capture different aspects of ataxic dysarthria, and these differences may underlie the clinical properties of these measures: AVE is better at capturing change over time, while MISD is better at distinguishing controls from individuals with mild ataxic dysarthria. It is possible that MISD is sensitive to respiratory and phonatory discoordination appearing early in disease (Folker et al., 2012; Kent et al., 2000), while AVE also captures articulation impairments, which contribute to continuing speech intelligibility decline. Using both measures together,
extracted in the same pipeline, can provide more granular picture of patient’s speech impairment.

As a first approximation, articulation can be captured by F1 and F2 formants, and vowel space area (VSA) derived from them. However, formant measures are highly dependent on the speech content, which defines prosodic contour (the vowel position in the word or sentence, and whether the vowel is stressed), and coarticulation influence on those metrics (Daniloff & Hammarberg, 1973). As such, experiments with word or sentence reading are more suitable for understanding of associations between AVE and articulation.

AVE, as a digital biomarker capturing the average uncertainty of automatic phoneme recognition, is related to the connection between predictability and intelligibility in speech. Work by Stilp and Kluender (2010) focused on the perceived intelligibility of speech, analyzing variability of auditory signal in the space of cochlear receptor sensitivity bands. They showed that intelligibility of speech is driven not as much by the signal in vowel or consonant segments, but rather by the segments of audio recording that are more dynamic and less predictable from the preceding audio signal (vowels, particularly diphthongs, and lateral/gliding consonants). To measure predictability they relied on the concept of “cochlea-scaled entropy”. While definitions of cochlea-scaled entropy and the average vowel entropy here used are different, the findings of Stilp and Kluender (Stilp & Kluender, 2010) suggest that potentially looking
into the classification entropy of each discrete token, rather than restricting the study to vowels or consonants, can also be a fruitful measure of intelligibility and a proxy to ataxic dysarthria severity.

An effort to develop ‘paralinguistic’ (dealing with the atypical speech of patients with speech impairments) models for automatic speech recognition (ASR) gained traction in recent years (Shor et al., 2021; Shor & Venugopalan, 2022). Such models aim to create an adequate representation of patient speech which will allow to increase performance on ASR tasks. The work of Korzekwa et al. (Korzekwa et al., 2019) presents an attempt to jointly detect dysarthria and recognize the speech, achieving 93% accuracy in dysarthria detection. However, the dataset used in the work (Kim et al., 2008) was relatively small (15 dysarthric and 13 control speakers) and did not include participants with ataxia. Given the features of ataxic speech discussed above, the approach of Weston et al. (Weston et al., 2021) can potentially be used in ataxic dysarthria quantification. The authors built a neural network aiming at representing prosody features, such as rhythm, articulation, pitch, and timbre, which are very relevant in ataxic dysarthria. However, the model and data it was trained on are not publicly released for fully investigating this opportunity.

One limitation of the present study is the absence of assessment of speech rhythm, which is particularly relevant for evaluating mild stages of ataxic dysarthria. Previous studies assessed speech rhythm abnormalities in conversational speech
through the normalized Vowel Pairwise Variability index, computed as the mean average absolute difference between the durations of vowel-to-vowel intervals (Grabe & Low, 2002; Low, 1998). Future studies should address this, taking into account that such a measure, if to be computed automatically, requires automatic segmentation of speech into utterances. Models for dysarthric speech representing prosody may help with this challenge.

Another limitation of the presented approach is its language dependency. The wav2vec2 model was pre-trained on English LibriSpeech corpus (Panayotov et al., 2015), fine-tuned on English phonetic annotations on the TIMIT dataset (Garofolo et al., 1993), and as such cannot be directly ported for ataxia assessment in other languages. Recent developments on cross-lingual pre-training and the representation of the entire international phonetic alphabet in the predictions (Conneau et al., 2020), with or without fine-tuning on a phonetic corpus of a particular language (Malmsten et al., 2022; Xu et al., 2021), can be exploited to make the approach here proposed language-agnostic.

Additionally, the relatively small longitudinal cohort is a limitation of the current study. Larger longitudinal datasets will be necessary to further evaluate the sensitivity of this approach for quantifying disease change.

As a potential future direction, assessment of the entropy per vowel can be tried with models specialized in paralinguistic speech. Such an approach will require a sentence reading task to make the stimulus universal for all participants and will
provide a clue into which vowels in which positions are harder for ataxic speakers to articulate, which may become itself an early biomarker of ataxic dysarthria. Another future direction would be to include perceptual assessments of ataxic speech by speech professionals (speech therapists or phoneticians), who may be able to detect subtle changes in speech that are not captured by the BARS speech assessment. American Speech-Language-Hearing Association recommends assessing motor speech disorders on phonatory/respiratory (in phonation, reading, and conversation), oral agility (diadochokinesis), and speech intelligibility (on phoneme, word, sentence, and conversation levels). Comparison of AVE and MISD with both perceptual (by speech professionals) and acoustic measurements at these levels may provide additional insight into the physiological nature of those metrics.

**Conclusions**

Average vowel entropy, a proxy measure of intelligibility based on recent advancements in automatic speech recognition, was found to be informative in assessing the severity of ataxic dysarthria. Mean intensity standard deviation in vowel segments, extracted in the same processing pipeline, was shown to be able to capture early signs of speech impairment. Evidence provided by longitudinal analysis shows promise of average vowel entropy measure as a potential outcome measure in longitudinal clinical trials. These results show the potential clinical value of modern automatic speech recognition tools.

Continuous EEG monitoring for seizure diagnostics is a common practice in Neonatal Intensive Care Units (NICU) since newborns may not exhibit clinical signs of seizures. Even though the monitoring is continuous, screening of the EEG recording is done visually by an epileptologist once in several hours. Such an approach inevitably leads to delays in seizure detection and treatment, hence increasing the seizure burden and worsening outcomes. In this chapter, we describe an automatic deep learning-based seizure detection method, built from scratch. Our proposed model highlights the EEG channels where the seizure is likely occurring using method based on multi-instance learning. It is also independent of the number of EEG channels, thus facilitating its use in different centers using different EEG hardware.

We justify the model’s portability to new facilities by demonstrating the model high performance on an independent dataset which was not used in the training process. We additionally demonstrate the agreement of the model with the human rater on a per-channel level as an indicator of the interpretability of the model. Finally, we show that the model’s agreement with a human rater is only slightly lower than between the two human raters. These results show the potential of the proposed model for seizure detection and automatic evaluation of the seizure burden, which can serve as a biomarker of the neurodevelopmental outcome.
**Introduction**

Seizures during the neonatal period are a common emergency in NICUs. After a perinatal hypoxic-ischemic event, 30-60% of infants develop seizures (Kharoshankaya et al., 2016; Nash et al., 2011). Fifteen percent of infants with seizures die and an additional 50% experience significant disability, including cerebral palsy, intellectual disability, and future epilepsy (Lai et al., 2013; Ronen et al., 2007). In newborns, clinical seizures symptoms can be extremely subtle or not exist at all, thus requiring electroencephalographic (EEG) monitoring for seizure identification (Wietstock et al., 2016). At leading and major medical centers, seizure detection currently relies on a clinical neurophysiologist reviewing continuous EEG recordings at standard intervals (at Duke University Hospital currently every 4-6 hours) to identify seizures in the preceding time period. Because seizure screening occurs once every several hours, treatment delays are inevitable. This issue motivates the development of a continuous monitoring solution to decrease time to seizure identification and treatment as timely intervention is critical for positive outcomes. Given recent advances in automated seizure detection (Ansari et al., 2019; O'Shea et al., 2020; Temko et al., 2011a), the goal of creating machine learning software tools to automatically detect seizures and help clinicians to make decisions now seems more achievable than ever (Mathieson et al., 2016a,b; Temko et al., 2015).
To study our proposed learning framework, we focus on two sources of data. Recently, the Helsinki University Hospital has released a NICU dataset of neonatal seizures with three distinct raters ("Helsinki dataset" from now on) (Stevenson et al., 2019). Additionally, we have built a dataset from our own historical cache of patients, yielding 31 additional individuals, to build and evaluate the proposed methods ("Duke dataset"). Having these two datasets allows us to evaluate methods in the context of two centers’ data and additionally evaluate how well the learned algorithms generalize to a new center, an important consideration in deployment.

This application comes with several important considerations. A first issue is that the data suffers from severe label imbalance, i.e., low proportion of seizure events, a noted issue in training machine learning models (Johnson & Khoshgoftaar, 2019). Additionally, the training data comes from several patients, each with highly varying levels of seizure rates (in the Duke dataset it varies from 0.09% to 24%). We propose to address this challenge as a group-label imbalance problem (controlling for class imbalance individually per patient, referred in our case as ‘Patient-Class imbalance’), and explore best data subsampling practices for training in this scenario.

Second, training datasets typically only provide “weak labels,” meaning that only periods of time containing seizures are labeled without specifying the EEG channel exhibiting the seizure. However, as mentioned above, seizures in neonates are not typically whole-brain events and are often localized to individual brain regions,
meaning that the seizure only appears in some of the measured EEG channels.

Therefore, we would like a method to help localize a seizure to specific channels. Recent work has begun to utilize weak labels in CNNs (Ansari et al., 2019; O’Shea et al., 2020), but has yet to focus on effective localization, applicable to the task at hand. This goal is twofold: (i) we would expect that building this information into the method would improve performance; and (ii) downstream implementations would almost certainly require manual verification, and highlighting EEG channels which presumably exhibit seizures could accelerate this process. We address this challenge by building an attention-based Multi-Instance Learning (MIL) framework (Ilse et al., 2018). The MIL framework (Kraus et al., 2016; Wang et al., 2018) is used to handle weak labeling, whereas the attention mechanism is used to highlight channels of interest for classification. We go further, evaluating the highlighting done by attention mechanism through comparing it with human per-channel seizure annotation to find out whether network “sees” the same thing as human does.

A third critical consideration is that previous studies have shown good performance metrics on in-house datasets (Ansari et al., 2019; Tapani et al., 2019; Temko & Lightbody, 2016), and only one study so far evaluated the results on an external dataset (O’Shea et al., 2020). However, in neonatal seizures, the inter-rater agreement is often relatively low (Stevenson et al., 2015; Stevenson et al., 2019). We explore both the
pure AUC from our predictive metrics, but also evaluate how well the chosen algorithm would replicate a doctors’ analysis using a variety of approaches and thresholds.

**Technical significance**

The proposed attention-MIL framework can help localize which channels are likely to indicate seizures, which we validate empirically. While previous studies had considered the weak labeling problem (Ansari et al., 2019; O'Shea et al., 2020), recovering seizure channels from weak labels can help accelerate downstream deployment and human validation. Additionally, we explore how the group-class imbalance affects the proposed algorithms during training. We provide additional metrics to explore how the algorithm matches human decision making at different parameter settings; while there are reports on matching algorithm and human performance (Temko et al., 2011b), prior studies with neural networks have focused primarily on AUC (O'Shea et al., 2017, 2020). Finally, it is rare in the deep learning literature in this field to have a true second dataset collected in a different context; we posit that these results help reveal the true deployment utility of the learned algorithms.

**Clinical relevance**

Intermittent review of continuous EEG recordings by a neurophysiologist inevitably leads to delays in seizure identification and treatment. A prior survey of neurophysiologists and neurointensivists showed that the frequency of reviewing EEGs varies widely: only 5% of surveyed physicians reviewed EEGs continuously, while 75%
reviewed it two or more times per day (Gavvala et al., 2014). A similar survey demonstrated that 50% of responders reviewed EEGs two times a day or less (Abend et al., 2010). Higher seizure burden is independently associated with worse neurodevelopmental outcomes, both for hypoxic ischemic encephalopathy patients (Glass et al., 2009; Kharoshankaya et al., 2016), as well as in other pediatric critical care situations (Payne et al., 2014). In the NICU, the paucity of clinical signs suggestive of seizure in neonates results in most (if not all) seizures being identified on EEG after which the clinical team caring for the infant is informed. Decreasing time to seizure identification and treatment is therefore essential for reducing seizure burden and potentially improving clinical outcomes. The benefits of a fully continuous monitoring system are clear, as an automated detection system could flag potential seizures and lead to more timely seizure treatment. Such a system would need to capture most seizures and be highly specific since systems with high false positives are frequently ignored. A key component of this continuous monitoring system is the development of a reliable automatic detection procedure; in this chapter, we present a machine learning approach to the automatic detection problem based upon datasets from two centers.

**Generalizable insights about machine learning in the context of healthcare**

Transferring models that apply to EEG data is difficult due to differences in equipment and clinical protocols used to perform data collection. While electrodes are
usually placed according to international standards (e.g., the 10-20 placement system (American Encephalographic Society, 2014)), deployed systems differ between centers (e.g., different numbers of electrodes), yielding different dimensionalities of data. This is a challenge when transferring models between centers, hindering real-world applicability. Therefore, we focus on learning machine learning models that are robust to such differences, and we evaluate its multi-center capabilities by evaluating on data from multiple centers and electrode layouts. By comparing the automatic evaluation with doctor evaluations, we revealed the necessity to tune thresholds to specific data sources rather than solely considering AUC. Furthermore, for a high-stakes decision, we assert that it is critical to underpin the decision in an interpretable manner to facilitate human review. To address this challenge, our proposed model is agnostic to the amount of channels and highlights those channels that are likely to exhibit seizure activity for a given timeframe. We then evaluate how our highlighting system matches with human interpretation.

**Cohort**

**Data collection and annotation**

**Duke dataset**

Patients aged <30 days who received continuous EEG (cEEG) monitoring between 2012 and 2019 were first identified through the EEG database system utilized by Duke University Medical Center (Natus Neurowoks). Medical records were then
manually reviewed and infants who were concurrently undergoing therapeutic hypothermia while being monitored on EEG were selected. A total of 154 patients were identified, 45 of whom developed seizures during cEEG monitoring, as assessed by an experienced epileptologist. After exclusion of corrupt files, cEEG data of 31 infants with seizures were available. This study of human subjects was approved by the Duke Health Institutional Review Board (Pro00100420).

Among 31 infants retained in the dataset, 42% (n=13) were female with a median gestational age of 39 weeks (Inter-Quartile Range (IQR) 38-40) at time of birth. Median time from birth to EEG placement was 9 hours (IQR 5-11). EEG recordings started at onset or soon after initiation of therapeutic hypothermia and recordings continued until 24 hours after rewarming. There were more seizures typically in the beginning and they decreased in frequency later in the recordings, yet entire recordings, regardless of therapeutic hypothermia phase, were used for algorithm development and training.

An experienced epileptologist from Duke University Medical Center annotated the dataset. Annotation for each seizure was provided in a separate table marking the beginning and end time of the seizures with 1 second resolution. In total, the dataset contained 2320 hours of recording with 50.81 hours of annotated seizures.

Summary of the Duke dataset is provided in Table 8 and Figure 9.
Figure 9: Histogram of seizure rate per patient in the Duke dataset (left) and the Helsinki dataset (right) on a log-scale. Red dotted line is prevalence of seizures over entire dataset (2.2% for the Duke dataset, 18.1% for the Helsinki dataset)

Since the system is intended to monitor all at-risk individuals, it is critical to maintain low false alarm rate. It is especially important on patients without seizures so that the system would not get ignored by practitioners. For an additional evaluation of our algorithm on such patients, we utilized 10 out of 154 newborns who underwent therapeutic hypothermia but did not develop seizures. This subset of patients had a median gestational age of 39 weeks (IQR 37-40) at time of birth and their median time from birth to EEG placement was 9 hours (IQR 5-12), and duration of each recording was 24 hours.

Subsample of the Helsinki dataset for cross-dataset validation

To get a better understanding of the generalizability of an algorithm, it is important to evaluate it in a variety of environments. For that purpose, we used data and annotations from the Helsinki dataset (Stevenson et al., 2019). We selected patients
that had seizures by consensus of 3 raters (total of 39 patients, 53% (n=21) female, 41% (n=16) male, gender not provided for 2 patients). Median gestational age for this subsample was 39 weeks (IQR 38-40). Summary for the subsample is provided in Table 9 and Figure 9.

Table 8: Summary of seizure amount, duration, and total recording duration in the Duke dataset. The dataset consists of 31 patients with continuous EEG recordings (minimum duration of 24 hours), which is typical of the multiple-day seizure monitoring protocols utilized in many NICU settings.

<table>
<thead>
<tr>
<th>Amount of seizures</th>
<th>Total hours</th>
<th>Total seizure hours</th>
<th>Seizure rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total 1778</td>
<td>2320.00</td>
<td>50.81</td>
<td>-</td>
</tr>
<tr>
<td>Mean 71.95</td>
<td>35.31</td>
<td>2.41</td>
<td>2.81%</td>
</tr>
<tr>
<td>Std 71.95</td>
<td>35.31</td>
<td>2.41</td>
<td>4.91%</td>
</tr>
</tbody>
</table>

Table 9: Summary of seizure amount, duration, and total recording duration in a subset of 39 patients from the Helsinki dataset who had seizures by consensus.

<table>
<thead>
<tr>
<th>Amount of seizures</th>
<th>Total hours</th>
<th>Total seizure hours</th>
<th>Seizure rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total 343</td>
<td>60.12</td>
<td>10.91</td>
<td>-</td>
</tr>
<tr>
<td>Mean 8.80</td>
<td>1.54</td>
<td>0.28</td>
<td>18.60%</td>
</tr>
<tr>
<td>Std 11.2</td>
<td>0.71</td>
<td>0.38</td>
<td>21.09%</td>
</tr>
</tbody>
</table>

Data extraction

To make results comparable with existing literature, all the data in this paper was extracted and preprocessed with the routine outlined in (Temko et al., 2011a) using the publicly available code from (Tapani et al., 2019).

EEG electrode setup for Duke dataset was based on the international 10-20 placement system modified for neonates as recommended by the American Clinical Neurophysiology Society Guidelines (Kuratani et al., 2016; Shellhaas et al., 2011). EEG
recordings were initially collected with a sampling frequency of 256Hz, using 9 electrodes. As is standard practice, bipolar derivations (differences between time-series from neighboring electrodes) were computed, resulting in the following 12 data channels (a.k.a. the ‘double banana’ montage): Fp1-C3, C3-O1, Fp2-C4, C4-O2, Fp1-T3, T3-O1, Fp2-T4, T4-O2, T3-C3, C3-Cz, Cz-C4, C4-T4. Notch filtering (at 60 Hz for the Duke dataset, and at 50 Hz for the Helsinki dataset), high-pass filtering at 0.5 Hz, low-pass filtering at 16 Hz and down-sampling to 32 Hz was performed. Then data for each patient was split into subsequent 8-second chunks with 4 seconds overlap (referred from here on as epochs). Any period with data losses in the recording (a small minority of data) was removed.

**Feature choices**

For the $i$-th patient after preprocessing we had $N_{ep,i}$ epochs with dimension $(N_c, 256)$ where $N_c$ is number of bipolar channels (12 for the Duke dataset, 18 for the Helsinki dataset), and 256 is the amount of timepoints per 8 seconds on 32 Hz downsampled data. For the deep learning approaches, here developed and investigated, this data format was directly used. The Support Vector Machine approach here tested relies on human-engineered features, so each epoch of data was converted to 55 features per channel. These features follow (Temko et al., 2011a), and are representative of frequency domain, time domain, and information theory-based characteristics of the signals.
Methods

In this work we compare two novel deep learning approaches and one classical approach (SVM), and investigate how the choice of data balancing techniques influences overall performance over the algorithm. We use AUC on leave one patient out cross-validation (LOO CV) as the main performance metric to evaluate how our performance generalizes to new individuals. Furthermore, we explore how well best performing algorithms generalize using cross-center validation. Specifically, we use the publicly available Helsinki dataset (Stevenson et al., 2019) and take the best performing model on full Duke dataset and evaluate its performance on the Helsinki dataset. We additionally assessed the performance of a publicly available SVM model pre-trained on the Helsinki dataset (Tapani et al., 2019; Temko et al., 2011a) on Duke dataset. Finally, we analyze how well one of our models can identify seizure activity per channel using only per-epoch labels for training, and how performance is associated with inter-rater agreement.

Machine learning models

Deep learning models

Our methods are based on the Convolutional Neural Network due to its widespread success in signal processing task. We primarily focused on two architectures. These architectures use a per-electrode (or per-channel) feature extractor with weights shared across all electrodes. Our feature extractor is based on Inception blocks (Szegedy et al., 2015) for their multi-scale filtering. We hypothesize that this structure might help
in classification due to the evolution of seizures in frequency. In preliminary experiments we saw an improvement in performance of this architecture over the standard CNN filter approach. After adapting the Inception block to one-dimensional data, our feature extractor had 8,514 trainable parameters. Below, we discuss the structure of our two proposed networks, which can be seen in Figure 10.

In our first deep learning network (DL1), the outputs of the per-channel feature extractor are concatenated and then passed through a dense layer. Since the number and order of channels is fixed, using a dense layer overall helps the classification since the channels are not independent (at least because channels are bipolar derivations of raw electrode signals); however, this should be carefully addressed since seizure activity appears in different channels for different patients.

In contrast, in our second deep learning network (DL2), the output of the per-channel feature extractor is passed through an attention-MIL layer, as outlined in Ilse et al. (2018). We built upon this framework with the intention that channels exhibiting seizures should be given more weight, which could both improve modeling and facilitate communication of the results. After the attention layer, a weighted average of the features is passed through a dense layer. Thus, this model is agnostic to channel interaction, facilitating portability to any channel layout. This is a desired feature for a generalizable seizure detection algorithms, since EEG setup can vary in different NICUs (Ansari et al., 2019), and such configuration allows the model to be used in different
NICUs without retraining, and also to jointly learn from multi-center weakly labeled datasets. This is also critical in our cross-center validation, because the two centers use different electrode layouts. To be more specific, the attention-MIL layer in DL2 takes as an input a bag of \( \{ h_k \} \) features \( h_k \in \mathbb{R}^{1 \times 48} \) in our case, \( k = 1 \ldots N_c \) (with \( N_c \) the number of channels), and outputs

\[
z = \sum_{k=1}^{N_c} a_k h_k ,
\]

where

\[
a_k = \exp \left( w^T \tanh (Vh_k^T) \right) \frac{1}{\sum_{j=1}^{N_c} \exp (w^T \tanh (Vh_j^T))},
\]

and \( w \in \mathbb{R}^{L \times 1} \), and \( V \in \mathbb{R}^{L \times 48} \). \( w \) and \( V \) are the learned weights, and \( L \) is the inner dimension of the attention-MIL layer. For this work we selected \( L=32 \).
Figure 10: Graphical schema of the two deep learning architectures for seizure predictions and a schema for the feature extractor. DL1 model (a), DL2 model (b). Both models share same per-channel feature extractor module. Feature extractor weights are the same for all channels. The last number in each block in the feature extractor schema is the amount of filters. Strides are equal to one if not stated otherwise.
Critically, attention-MIL weights can be used as a proxy for whether channels exhibit seizure activity, and help clinicians understand “where to look at,” i.e., which channels contributed most to the detection of seizure.

In total, DL1 and DL2 had 27,011 and 11,683 trainable parameters respectively. In each experiment we trained a network for 25,000 steps with a batch size of 256.

We implemented our DL models in the Keras framework (Chollet et al., 2015) with TensorFlow GPU backend, and run them on a desktop with 6-core i7 Processor with 64Gb of RAM and GeForce 1080 Ti GPU. Both code and pre-trained DL2 model on Duke dataset are available at https://github.com/dyisaev/seizure-detection-neonates.

**Classical ML models - support vector machines**

To compare the proposed and studied deep learning approaches with classical ML approaches, we selected a model which has shown good results in previous publications (Tapani et al., 2019; Temko et al., 2011a). We replicated the exact procedure of feature extraction using publicly available code (Tapani et al., 2019), training the model and predicting seizure. The model used radial basis function SVM based on 55 features (Temko et al., 2011a). The model trains on 55x1 features, representing 8-second recording segment per channel. It takes advantage of strong labels, combining only data from channels marked as ‘seizure’ in seizure samples and data from random channels from non-seizure segments during training. The model predicts seizure per-epoch if at least one channel exhibits seizure; predictions are done per channel, smoothed with moving average of 3
consecutive segments, and finally overall time segment prediction is done by max-pooling per-channel predictions.

**Data balancing approaches**

Previous literature suggests the detrimental effect of class imbalance on CNN performance (Buda et al., 2018), and no work so far has fully explored the influence of class balancing on the classification performance in neonatal seizures. Moreover, imbalance in seizure burden varies across patients (see Figure 9). Thus, we tested each method with 3 types of balancing approaches: No balancing (simply subsampling all available training epochs); Class balancing (keeping the proportion of classes (labels of seizure/non-seizure) in each minibatch equal); Patient-Class balancing (keeping the proportion of (Patients x Classes) partitions equal in each training minibatch). While Class balancing addresses the problem of algorithms seeing much more negative (non-seizure) than positive (seizure) examples, there is still a problem of algorithms seeing much more positive examples from patients with high seizure burden in this approach. Patient-Class balancing intends to address that and is expected to provide better generalization.

**Post-processing**

We can use post-processing procedures to reduce short false positive periods and link together longer seizures, providing a slight boost to AUC. We explicitly specify in
the reported results if post-processing was used, which includes probability reweighting (to adjust for true class prevalence in the dataset) and transforming the outputs to improve robustness (see Appendix B).

**Results**

**Evaluation approach/Study design**

We selected area under the receiver operating curve (AUC) on leave-one-patient-out (LOO) cross-validation as a main measurement of model performance. We also explored the influence of post-processing, so we estimated performance in 2 ways. First, we assessed AUC when prediction is evaluated on each epoch of LOO patient’s data. Second, we applied the post-processing procedures for the best performing model with different thresholds for computing AUC and evaluated AUC on each second of the LOO patient’s data.

For cross-center validation, we selected best performing model on Duke dataset and the publicly available SVM model trained on the Helsinki dataset (referred as SVMr in (Tapani et al., 2019)). We computed AUC per patient on the Helsinki dataset for 39 patients that had seizures by consensus. However, the 3 raters of the Helsinki dataset disagreed on precise beginnings and ends of seizure periods regions, thus we used for computing only the regions where all 3 raters agree. This is directly comparable with AUCs reported in previous work (O'Shea et al., 2020).
We also assessed Cohen’s κ between the proposed algorithm output and a human rater for the Duke dataset (or a consensus of 3 raters for the Helsinki dataset), as well as the sensitivity and specificity dependance on the selected decision threshold given our data and algorithm.

To evaluate how well our best algorithm (trained only on patients with seizures) performs on patients without seizures, we computed specificity and number of seizures detected on a previously unseen set of 10 patients’ recordings from NICU of Duke University Medical Center deemed as non-seizure by the same epileptologist who annotated the Duke dataset. All recordings were 24 hours long. This requires selection of the decision threshold, which we set as the probability of positive class over the entire training dataset (see Appendix B for derivations).

To assess how well the attention-MIL mechanism of DL2 model captures seizure channels, we performed AUC analysis of attention-MIL scores on Duke dataset for seizure epochs. We computed the AUC value between the scores and human annotation in two settings: (a) per channel and epoch (‘Attention AUC’) - each individual channel was assigned a positive or negative label based on the epileptologist per-electrode labels, and AUC was calculated using the channel-specific prediction (i.e., prediction if attention only used that channel); and (b) per epoch (‘Attention AUC per epoch’) - if at least one channel exceeding the decision threshold in an epoch is deemed a seizure by the human rater, then we consider the epoch as true positive, and we compute true and false positive
rates. To the best of our knowledge, this was the first quantitative assessment of how well the deep learning algorithm trained using weak (per-epoch) labels was able to provide per-channel annotations.

**Results on machine learning approaches on different balancing techniques**

Results of different balancing techniques and their influence on the deep learning approaches are summarized in Table 10. To measure significance of difference between each pair of approaches we performed Wilcoxon paired signed-rank test (Wilcoxon, 1945), see Appendix B. Class Balancing approach on DL2 model outperformed all other approaches, resulting in AUC of 0.950. Note that the SVM approach does not operate on weak labels, and so is limited by the availability of per-channel labels.

**Table 10: Results of different balancing approaches and their influence on the performance on Duke dataset (average AUC on leave one patient out cross-validation). Standard deviation (SD) is shown in parentheses. Results do not include post-processing routine.**

<table>
<thead>
<tr>
<th>Model</th>
<th>No Balancing</th>
<th>Class Balancing</th>
<th>Patient-Class Balancing</th>
</tr>
</thead>
<tbody>
<tr>
<td>DL1</td>
<td>0.933 (0.055)</td>
<td>0.923 (0.070)</td>
<td>0.911 (0.086)</td>
</tr>
<tr>
<td>DL2</td>
<td>0.923 (0.057)</td>
<td><strong>0.950 (0.041)</strong></td>
<td><strong>0.943 (0.051)</strong></td>
</tr>
<tr>
<td>SVM</td>
<td>0.822 (0.063)</td>
<td>0.772 (0.061)</td>
<td>0.765 (0.058)</td>
</tr>
</tbody>
</table>

To further explore the influence of post-processing on the results, we performed post-processing on the best performing model (Class-balanced DL2). With post-processing the model achieved the average AUC of 0.970 (SD: 0.033).
Results on cross-dataset validation

The results for cross-dataset AUC presented in Table 11 were achieved including the post-processing routine, which was the same for both datasets. The drop in performance between datasets was less for DL2 than for SVMr, showing significant promise for DL2 to generalize to new centers.

We note that the SVMr shows significantly higher performance than the SVM trained on our own data. Again, the SVM does not operate on weak labels, and so is limited by the availability of per-channel labels, which was higher in the Helsinki dataset. Additionally, the Helsinki dataset labeled positive and negative channels on the montage (bipolar derivations) whereas Duke dataset labeled individual electrodes that was expanded to the montage. This difference in labeling could explain this performance difference because the algorithms actually operate on the montage.

Table 11: Results of cross-dataset validation as measured by average AUC on per-patient evaluation of models. Evaluation on the same dataset is done via Leave One Patient Out cross-validation. Uncertainties given are the SD over patients.

<table>
<thead>
<tr>
<th>Model</th>
<th>Trained On</th>
<th>Duke dataset</th>
<th>Helsinki dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>DL2 (Class balance)</td>
<td>Duke dataset</td>
<td>0.970 (0.033)</td>
<td>0.925 (0.099)</td>
</tr>
<tr>
<td>Pre-trained SVM (SVMr)</td>
<td>Helsinki dataset</td>
<td>0.826 (0.117)</td>
<td>0.923 (IQR 0.869–0.990)*</td>
</tr>
</tbody>
</table>

* Data provided in Tapani et al. (2019)

Association of per-patient AUC scores with inter-rater agreement

We wanted to evaluate how well our proposed method works relative to a typical rater. To do this, we calculated the average inter-rater agreement for each patient using Cohen’s κ on the Helsinki dataset. We then compared this value to the AUC calculated
on each patient, shown in Figure 11. It is clear from the picture that as Cohen’s $\kappa$ grows, both AUC grows and variability in AUC reduces. In other words, when human raters agree with each other, we largely agree with them. Spearman’s $\rho$ correlation coefficient between average $\kappa$ and AUC is 0.56 ($p<0.001$), showing a strong statistical relationship.

![Figure 11: Scatterplot of average Cohen’s $\kappa$ of inter-rater agreement vs cross-dataset AUC on patients with consensus seizures/non-seizures (on the Helsinki dataset).](image)

**Agreement between the algorithm and a human rater**

Using a threshold of 0.022 (corresponding to a 0.5 threshold corrected for the prevalence in the Duke dataset), we calculated the agreement with a human rater on Duke dataset using Cohen’s $\kappa$, and our algorithm gave a median value of 0.517 with an IQR of 0.313-0.671 on the per-patient agreement with human rater on Duke dataset. Median value was 0.59 with an IQR of 0.119-0.769 of the agreement with a consensus of 3 raters on the Helsinki dataset. Because these values are dependent on the chosen threshold, we wanted to evaluate how much the choice of threshold impacts the achieved
performance. We visualize the median and IQR of Cohen’s $\kappa$, sensitivity and specificity compared to a varying decision threshold, in Figure 12.

It is clear from the graph that optimal thresholds are different for the two datasets. If we select an optimal threshold based on Cohen’s $\kappa$ on the Duke dataset (green dotted vertical line, Figure 12 (top row, left image)) we get a serious drop in sensitivity and specificity on the Helsinki dataset.

**Validation on non-seizure patients**

We next evaluated the false positive rate on patients where an epileptologist did not mark any seizures. Specifically, we used 10 patients, each of which had 24-hours of non-seizure recordings, from Duke University Medical Center. At the chosen threshold level, the algorithm flagged 589 seizures, with median of 42 seizures per recording (IQR 25.5-84.5). Median specificity per patient was 0.98 (IQR 0.94-0.99). Median duration of the detections was 30 seconds (IQR 28-32.5), which, given that 16 seconds is a collaring length in post-processing, provides a ‘raw’ detection length of 3-4 consecutive epochs. These results show the importance of decision threshold selection, post-processing, and adapting the algorithm to background noise (Temko et al., 2013). The level of false positives is, of course, related to the seizure threshold, as can be seen more broadly in Figure 12.
Figure 12: Ranges of Cohen’s $\kappa$, sensitivity, and specificity as decision threshold changes for DL2 Class balanced model. Top row: Results on the Duke dataset LOO; Three bottom rows: Results on the Helsinki dataset, for rater 1, rater 2, and rater 3 respectively. Red dotted vertical line - threshold corresponding to a 0.5
threshold corrected for the true prevalence (Duke dataset). Green dotted vertical line - empirical optimal threshold based on Cohen’s $\kappa$ in the training sample (Duke dataset). Threshold values are provided on a log scale.

**Attention network visualization**

Finally, we evaluated the performance of the attention mechanism of the DL2 network, which is summarized in Table 12. It is worth noting that for the Duke dataset per-channel annotations were provided per electrode, while for the Helsinki dataset per-channel annotations were provided per bipolar derivations. Thus, if an electrode was marked as ‘seizure’ in an epoch, then we considered all bipolar derivations including that electrode as seizures. As a result of this, different electrode annotations lead to different amounts of bipolar derivations considered seizures (e.g., for Fp1 two channels were marked, while for C3 four channels were marked).

**Table 12:** Attention network performance of DL2 (Class balance) model on the Duke dataset and the Helsinki dataset. Measured by averaged AUC (SD in parentheses). Computation of “Attention AUC” used agreement between each thresholded score and human annotations per channel per epoch; “Attention AUC per epoch” uses agreement of at least one channel score exceeding threshold with human annotation.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Attention AUC</th>
<th>Attention AUC per epoch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duke dataset</td>
<td>0.811 (0.096)</td>
<td>0.927 (0.058)</td>
</tr>
<tr>
<td>Helsinki dataset</td>
<td>0.701 (0.107)</td>
<td>0.807 (0.167)</td>
</tr>
</tbody>
</table>

To provide a qualitative measure on how the attention network works, Figure 13 summarizes how weights are distributed in the attention network in seizure/non-seizure samples for one of the patients for the entire recording. We also visualize the output of the network during the beginning and end of a seizure event in F. For this patient, all
seizures were focused on leads O1 and O2, as annotated in the Duke dataset. While the algorithm and rater agreed on the general location of the seizure, there was disagreement on the exact start and end location.

Figure 13: Average attention scores across all samples for one of the patients from Duke dataset. Ticks marked red - ground truth, provided by epileptologist (for all seizures of this patient, epileptologist marked O1 and O2 as electrodes where seizures are visible). Attention AUC for this patient was 0.88.

**Ablation study**

We performed an additional ablation study to determine the impact of the attention layer on overall prediction. We removed the attention layer and performed simple averaging of per-channel features after the distributed feature extractor layer, with the other hyperparameters held constant. The attention layer provided marginal improvement of AUC (0.945 with ablation of attention vs 0.950 for full network). This observation indicates that classification power of our approach comes mostly from the
feature extractor selected. Regardless, the utility of the attention layer is useful for communicating the results, as demonstrated in Figure 14.

Figure 14: Examples of seizure predictions. (Top) Beginning of a seizure in a patient from Duke dataset. The green dotted line marks the beginning from the epileptologist. The magenta dotted line marks the beginning decided by network. The colored background intensity corresponds to how much attention weight is given to each channel at each time segment. (Bottom) End of the same seizure. The green
dotted line marks the end as labeled by the epileptologist. The network deems the whole segment as a seizure, and most of the weight for its decision is coming from channel C3-O1, which is also deemed the relevant channel by the epileptologist.

**Discussion**

**Deep learning models and balancing strategies**

It is well-established that deep learning models suffer under significant label imbalance. Few studies on adult epilepsy explicitly took into account data imbalance (Wu et al., 2020; Yuan et al., 2017), and most of the previous studies on data imbalance in CNN training were focused around CNN for image classification (Johnson & Khoshgoftaar, 2019). In our data, the seizure prevalence per patient is highly variable (from 0.08% to 24.3% in the Duke dataset), so we hypothesized that addressing the data imbalance might be a crucial issue in algorithm performance. However, we found only slight changes in performance due to the varying data balancing strategies. Part of this may be due to the methods evaluated; in our study we explore only data-level methods, comparing no balancing, class balancing (same amount of seizures/non-seizures per batch, also known as ‘class-aware sampling’ (Shen et al., 2016)), and Patient-Class balancing (same amount of seizures/non-seizures both per batch and per patient in batch). The structure of features in the DL2 model (both per-channel extracted features and weighted average of per-channel features have 48 dimensions) could facilitate a variety of additional data-level approaches (e.g., SMOTE (Chawla et al., 2002)). We did find highly variable performance with different neural network structures. In our second Deep Learning
model (DL2), we proposed an approach that is electrode-number agnostic; that is, it can work on different devices and electrode layouts without retraining the network. In this network, the class balancing approach worked the best. We hypothesize that this is because the Patient-Class balancing over-weighted less common seizures from low seizure-prevalence patients. However, the variability in the models shows that the results on balancing are inconclusive.

Because the post-processing approaches are dependent on the probability estimates, we want to have proper probability estimates. However, these balancing schemes get rid of the class prior and must be corrected to give proper probabilistic estimates. In our case, we have done this by post-scaling of output probabilities (Buda et al., 2018; Lawrence et al., 2012; Zhou & Liu, 2006). This post-scaling could be combined with other calibration approaches (e.g., Platt scaling) to get accurate probabilities.

**Algorithm-rater and inter-rater agreement**

Our results on agreement in the Duke dataset LOO setting and the Helsinki dataset indicate that high AUC values are not enough for the deep learning seizure detection algorithm to be immediately transferable to clinical practice. Our algorithm reaches median 0.517 agreement with human rater on the Duke dataset and median 0.59 with consensus of 3 raters on the Helsinki dataset, as compared to 0.807 (IQR 0.540-0.913) of Cohen’s κ averaged across 3 pairs of human raters on the Helsinki dataset. We can see that agreement between the algorithm and human raters is worse than agreement between the
3 human raters. It’s also evident that variability on cross-dataset prediction for Cohen’s $\kappa$ is much higher. This may be due to different prevalence of seizures in the Helsinki dataset compared to Duke dataset (Stevenson et al., 2015; Vach, 2005) and the generalization error. (Tapani et al., 2019) approached the problem of agreement between algorithm and human annotation as agreement between 3 raters (2 humans and an algorithm). They reported that Fleiss’ $\kappa$ (Fleiss & Cohen, 1973) dropped if one human rater is replaced by the algorithm. The need to search for a decision threshold, and to decide the costs of false detections and false negatives (misses), gives an intuition that cost-sensitive learning (Ling & Sheng, 2008) may be another approach to address class imbalance, where cost can be either fixed (Wang et al., 2018) or learned (Khan et al., 2018).

Note that there appears to be some gain possible from personalizing the threshold, meaning that we may need to build strategies to calibrate to individuals. This avenue could be explored through a meta-learning approach.

In addition to the $\kappa$ metric, metrics that evaluate agreement on a per-event basis could be used to further assess the clinical feasibility of the algorithm. For example, analyzing the positive (seizure) agreements, negative (non-seizure) agreements and disagreements between algorithm and raters, as proposed in (Stevenson et al., 2015), done across entire recording or per-hour could be used. These metrics will be addressed in future work.
Interpretability of the results

In high-stakes decision-making, many people are rightfully wary of black-box decisions (Rudin, 2019). In our scenario, we view this system as a support tool where any predicted positives could be reviewed more quickly. In such a scenario, it would facilitate chart review to have the system be as descriptive as possible. While our attention-based system does not produce interpretable filters, it can easily highlight relevant channels and time periods for a clinician to review.

As the Attention AUC on at least one channel detected as seizure is 0.927 (SD 0.058) on Duke dataset, we consider that this system could help decrease evaluation time. While the system performance drops down to 0.807 (SD 0.058) when evaluated on the Helsinki dataset, this implies that the system is still robust to true domain shifts and can be increasingly fine-tuned. It is also important to mention that while helping to highlight the relevant channels, attention mechanism does not add to classification power of the model, which can be seen from ablation study.

While other approaches have considered weak labels, the system by (O'Shea et al., 2020) was an ensemble of 3 networks with prediction averaged from three outputs. While undoubtedly improving performance of the model, it significantly constrains the interpretability. Another CNN-based system (Ansari et al., 2019), while using weak labeling, did not provide interpretations of channel importance due to network architecture.
Conclusions

In this work we provided an assessment of how different models and balancing methods influence learning in neonatal seizure detection from EEG. We proposed a model that provides a level of importance to each of the channels - a proxy to whether a channel exhibits seizure activity or not. This model is portable to an EEG dataset with an arbitrary amount of channels without need for adjustment or retraining, and can provide decreased checking time for use in a secondary evaluation by a doctor. To our knowledge, we also provided the first assessment of agreement between human raters and deep learning algorithm for detecting neonatal seizures. The system, to date, has shown excellent AUC; however, we do not exactly mimic doctor behaviors towards labeling, and the estimate Cohen’s $\kappa$ values were comparatively low, showing room to further improve the algorithm. Future work will attempt to increase this value by focusing on improved learning strategies, additional data integration, and individualizing to a patient, e.g., by meta-learning.
Chapter 6. Combining EEG and behavioral measures in autism studies

Autism is characterized by early attentional differences that often precede the hallmark symptoms of social communication impairments. Development of novel measures of attentional behaviors may lead to earlier identification of children at risk for autism. In this chapter we introduce a behavioral measure, Relative Average Look Duration (RALD), indicating attentional preference to different stimuli, such as social versus nonsocial stimuli; and then study its association with neurophysiological activity. We show that (1) autistic and neurotypical children differ in their both (absolute) Average Look Duration (ALD) and RALD to stimuli during an EEG experiment, with the most pronounced differences in looking at social stimuli; and (2) associations between looking behaviors and neurophysiological activity, as measured by EEG, are different for autistic children versus neurotypical. Even when autistic children show attentional engagement to social content, our results suggest that their underlying brain activity is different from neurotypical children. Therefore, the study described in this chapter, introduces RALD as a new measure of social/nonsocial attentional preference in ASD and demonstrates the value of incorporating attentional variables measured simultaneously with EEG into the analysis pipeline.
Introduction

Autism spectrum disorder (ASD) is characterized by early attentional differences that often precede the hallmark symptoms of social communication impairments and restricted and repetitive behaviors (Elsabbagh et al., 2013; Elsabbagh et al., 2011; Werner et al., 2000). Attentional processes such as orienting, disengagement from and sustaining attention to relevant stimuli (Chawarska et al., 2012; Elsabbagh et al., 2013; Keehn et al., 2013; McPartland, Webb, et al., 2011), and the ability to share attention (Dawson et al., 2004; Mundy & Acra, 2012) are foundational for the development of social abilities and social communication. Research has demonstrated deficits in all of these domains of attention in infants and children with ASD (Dawson et al., 2004). As such, screening and diagnosis place particular emphasis on these behaviors; and early interventions target these attentional processes to facilitate the acquisition of social and communication skills (Dawson et al., 2010; Kasari et al., 2006; Landa et al., 2011). In this study, we investigate the associations between attention and simultaneously recorded neurophysiological signals in children with autism. Our results suggest that even when autistic children show attentional engagement to social content, their underlying brain activity is different from neurotypical children.

A distinctive sign of autism is robust differences in the amount of attention directed toward social versus nonsocial stimuli, documented across the lifespan and reported as early as 6 months of age in infants who later develop autism (Chawarska et
neurophysiological recordings (e.g., Event-Related Potentials (ERP) (Luck, 2014) and spontaneous EEG) and looking behavior paradigms (e.g., via habituation (Colombo & Mitchell, 2009) and gaze (Haider et al., 2017), which are often measured with eye-tracking technology (Sasson & Elison, 2012), but can also use standard computer vision (Bovery et al., 2018) have been widely used in autism research, few studies have reported reliable and robust results that combine these measures and jointly analyze them. Our study aims to fill in this gap, investigating the associations between looking/attentional behavior and neurophysiological patterns, registered simultaneously in a synchronized fashion.

In infant studies, a common way to assess social attention is via diverse “habituation” paradigms (Colombo & Mitchell, 2009; Jones et al., 2017; Jones et al., 2016; Webb et al., 2010). For example, in some studies, static stimuli (faces or objects, i.e., social and non-social images) are presented to the participants. Then, look durations, defined as time between initial look at the stimulus and look away, are recorded. From such measures of look duration, various statistics can be derived, including time to habituate (formally, decline attention), peak look duration, mean look duration, and/or the number of looks (Hendry et al., 2018; Jones et al., 2016; Webb et al., 2010). Webb et al. (2010) showed that toddlers with more severe ASD symptoms based on the Autism Diagnostic Observation Schedule (ADOS, Lord et al. (2012)) took significantly longer to
habituate to faces than objects (houses). Further, the children with severe ASD took significantly longer to habituate to faces than groups of neurotypical toddlers, as well as toddlers with less severe ASD, developmental delay (DD), siblings of autistic children, and siblings of neurotypical children. In another study (Jones et al., 2016), infants who later developed autism showed shorter look durations to faces than objects, and their peak time to look to faces happened later than in infants who did not develop autism, suggesting that infants at risk for autism attend to faces differently than neurotypical infants.

A common method/technology used to measure and analyze looking behavior (or gaze) is eye-tracking (Sasson & Elison, 2012). A broad body of literature, including some of the above mentioned papers, has offered insights into visual social attention using this method (see Guillon et al. (2014) for a review). Eye-tracking studies are commonly used to assess patterns of attention to dynamic stimuli (Chawarska et al., 2012; Shic et al., 2011), e.g., movies or changing images, where both social (i.e., people) and non-social cues (e.g., toys) are presented and compared. The dynamic nature of the stimuli used in many of these studies is more ecologically valid than studies with static stimuli. In a study of toddlers with ASD, Chawarska et al. (2012) showed that the difference between total looking time at faces/objects becomes apparent only when the child was viewing child-directed speech and the actress was making eye contact, referred to as a dyadic bid. In such a scene, the ASD group showed diminished attention
to the face and the mouth while their attention to toys in the same scene was increased. This suggests that ASD toddlers have difficulties holding attention to the face particularly in dyadic bid situations. The same experiment, when conducted with 6-month old infants at risk for ASD (Chawarska et al., 2013), demonstrated reduced attention by infants who later developed ASD to the overall social scene, as well as to a person and her face, without any significant difference by types of activities shown (dyadic bids, joint attention, moving toys).

While eye-tracking and habituation studies provide us with rich data on attentional behavior patterns to social/non-social stimuli, it is also of interest to understand underlying neurophysiological activity during attention that might help explain differences in social attention. For example, using an ERP paradigm, Dawson et al. (2012) showed that neurotypical young children, as well as autistic children receiving an early developmental intervention (Early Start Denver Model (Dawson et al., 2010)), showed increased activity in the theta band and decreased activity in the alpha band while attending to static faces, while ASD children who received treatment as usual showed the opposite pattern (greater activity in theta and less in alpha in response to static toy stimuli). While an extensive body of literature exists for ERP biomarkers of social attention in ASD (Dawson et al., 2005; McPartland, Wu, et al., 2011), fewer studies have focused on measures of spontaneous EEG and their relationship to attention to social dynamic stimuli. In an experiment with neurotypical infants and preschool
children that used dynamic stimuli (child-directed speech, manipulating toys, and visual attention to bubbles), theta band power was found to increase during emotionally stimulating conditions (child-directed speech and manipulating toys) as compared to a baseline (bubbles stimulus), supporting previous evidence of a relationship between theta power and attentional states (Aftanas et al., 2004; Maulsby, 1971). In another study with dynamic stimuli (Orekhova et al., 2014), an EEG connectivity analysis of the data from infants watching social/nonsocial videos was performed, however, the actual type of stimuli was not a factor influencing the EEG results.

Given the potential role of attention in influencing underlying neurophysiological activity, it is of interest to simultaneously measure both attention and EEG and jointly consider them for analysis. This was done in several EEG studies in which the participant’s behavior was videotaped synchronously with EEG recording (Murias, Major, Compton, et al., 2018; Orekhova et al., 2014; Orekhova et al., 2006; Stroganova et al., 1998), and subsequently looking behavior (looks at the screen), as well as motor behavior (significant motion), and/or emotional behavior (crying, excessive smiling) was coded off-line. However, in those studies, the coded participant’s behavior was used only for data preprocessing (Murias, Major, Compton, et al., 2018; Orekhova et al., 2006; Stroganova et al., 1998), or group comparisons of the behavioral variables (Orekhova et al., 2014), and joint analysis of EEG and behavior was not explicitly performed. To the best of our knowledge, in the area of ASD research, only one study so
far performed simultaneous eye-tracking (as a proxy to attention) and EEG analysis in an experiment with joint attention in a small sample of high-functioning autistic children (Billeci et al., 2017). The authors reported a positive correlation between cumulative fixation duration on face and beta and gamma band relative power. Our study uses a different age range and introduces significant additional analysis of both the attentional (as a measure of relative responses to social vs non-social stimuli) and the EEG signal, as well as their interaction.

In the present study, we explore a method to jointly study EEG and synchronized looking behavior during the same experiment. We introduced a new measure, Relative Average Look Duration (RALD), a normalized measure indicating attention preference when comparing different stimuli, i.e., social vs non-social, allowing a within-subject comparison of relative attention directed to different stimuli. This allows for an individual measure of attentional preference. Furthermore, young children with autism are not compliant with instructions to sit quietly without any video stimuli while spontaneous EEG are collected. Thus, it is considered standard in the field of autism research not to include a video-free baseline but rather to examine how the brain responds to different types of stimuli during spontaneous EEG recording (Webb et al., 2015); the proposed RALD naturally addresses this challenge. As mentioned above (McPartland, Webb, et al., 2011; Webb et al., 2010), previous studies have shown differences in preference to social as compared to nonsocial stimuli in children with
ASD. Here we develop a new measurement of relative/preferential attention and jointly analyze it with EEG. First, we examined how autistic (ASD) and typically developing (TD, or neurotypical in the following sections) groups differ in their RALD at two videos displaying complex dynamic social and nonsocial audiovisual stimuli, as well as a neutral, less complex video (bubbles cascading) that did not involve sound. Second, we investigated how RALD correlates with brain activity, as reflected in the relative power spectral density (RP) of the EEG signal in four frequency bands. While the joint study of EEG and looking behavior was previously used in EEG artifact correction (Plöchl et al., 2012) and in the studies of human reading (Dimigen et al., 2011; Hollenstein et al., 2018; Kretzschmar et al., 2013; Notaro & Diamond, 2018), to the best of our knowledge only one study made an attempt to jointly analyze looking behavior and EEG in relation to social attention (Billeci et al., 2017). Our study goes further, proposing to use RALD as a measure of preferential attentional behavior and study its association with EEG.

**Methods**

**Participants**

All caregivers/legal guardians of participants gave written, informed consent, and the study protocol was approved by the Duke University Health System Institutional Review Board. Methods were carried out in accordance with institutional, State, and Federal guidelines and regulation.
ASD participants

Participants were 31 children with ASD (23 males, 8 females) between 28 and 81 months of age (mean=55.3, SD=14.8). Children with ASD were part of a single site, prospective, randomized, double-blind, parallel group study of placebo versus a single intravenous autologous or allogeneic, unrelated cord blood (CB) infusion in ASD children aged 2-7 years. The trial was conducted under IND #15949. Only data from the baseline visit, which were collected before infusions, were used in this analysis. Clinical diagnosis of ASD was based on the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5), and established by expert clinicians using the Autism Diagnostic Observation Scale (ADOS-2, Lord et al. (2012)) and the Autism Diagnostic Interview, Revised (ADI-R, (Le Couteur et al., 2003)). Additional inclusion criteria included (1) stability on current medications for at least 2 months prior to the infusion, (2) participants and parents/guardians were English speaking, and (3) availability of autologous umbilical cord blood unit or ≥4/6 HLA-matched allogeneic unrelated umbilical cord blood unit from the Carolinas Cord Blood Bank. Exclusion criteria included (1) a history of prior cell therapy, (2) use of intravenous immunoglobulin (IVIG) or other anti-inflammatory medications (with the exception of NSAIDs), (3) known genetic syndrome (e.g., Fragile X), presence of dysmorphic features, pathogenic mutation or copy number variation associated with ASD, and/or other significant medical and/or psychiatric comorbidity, (4) obvious physical dysmorphology, (5) an
uncontrolled seizure disorder, (6) significantly impaired renal or liver function, (7) known active CNS infection, evidence of uncontrolled infection, and/or HIV positivity, (8) family unwilling or unable to commit to study-related assessments, and/or (9) clinically significant abnormalities in complete blood count. The mean Full Scale IQ of ASD study participants was 80.4 (SD = 21.9) based on the Mullen Scales of Early Learning Composite Score (Mullen, 1995) or Differential Ability Scales Second Edition (DAS-II, Elliott (2007)).

**Typically developing participants**

Children who did not have a diagnosis or suspected diagnosis of ASD were recruited from the community and the Duke Center for Autism and Brain Development research registry to participate in a study of preschool age children with and without autism. A randomly chosen subset of these children (N=31) who were age matched to the ASD participants were included in the current analyses. Participants were 31 children (14 males, 17 females) between 39 and 71 months of age (mean=53.3, SD=10.5). Children were eligible to be in the TD control group if they had scores on the Strengths and Difficulties Questionnaire (SDQ) that were within the normal range for all scales. The SDQ is a parent-report screening tool for measuring internalizing and externalizing difficulties in children (Goodman, 1997; Goodman, 2001; Goodman et al., 2000).

Exclusion criteria for this group included having (1) a biological sibling or parent diagnosed with ASD or developmental delay (DD), (2) a genetic disorder (e.g., Fragile
X), (3) vision or hearing problems, (4) a significant motor impairment (e.g., cerebral palsy), (5) chronic or acute medical illness, and (5) a seizure in the last year, a seizure disorder, or being on medication for seizures. The mean Full Scale IQ of TD study participants was 114.3 (SD=13.5) based on the Mullen Scales of Early Learning Composite Score or Differential Ability Scales Second Edition (DAS-II). In order to control for IQ differences between the two groups, IQ was used as a covariate in the analyses.

**EEG measures**

**Protocol**

Continuous EEG was recorded while the participant watched three video stimuli which were each shown twice (total of 6 videos, 6 minutes). Video content was dynamic stimuli consisting of a woman singing nursery rhymes while she gestured (“Social,” video 1), brightly colored dynamic toys that made noise (“Toys,” low social content, video 2), and bubbles cascading across the screen with no auditory content (“Bubbles,” low social and audio content, video 3), see Figure 15 for corresponding screenshots. The order of social and toys videos was counterbalanced to eliminate any potential order effects, and bubbles was always shown last. During the experiment, two behavioral assistants accompanied the child. They ensured that standard conditions were in place during each experiment, including dimming the lights, seating participants in their parent’s lap in a comfortable armchair 65 inches from the monitor, and redirecting
participants in instances of movement and/or poor attention to the videos. The child’s face was recorded from a camera beneath the screen synchronized with the EEG. It allowed post-session editing of periods of inattention. EEG data were recorded from 124 channels with reference to Cz using a Hydrocel Geodesic Sensor Net and Net Amps 400 amplifier (Electrical Geodesics, Eugene, Oregon). Data were collected using Netstation 4.5.6 with a sampling rate of 1000 Hz.

Figure 15: Screenshots of three stimuli used in the study. Left – Social, Center – Toys, Right - Bubbles stimulus.

EEG pre-processing and data attrition

Data were processed with Matlab 2014a, using the open source Fieldtrip (Oostenveld et al., 2011) and EEGLAB (Delorme & Makeig, 2004) toolboxes for all operations on EEG data. Data were filtered with a 1-100 Hz bandpass filter and a 60 Hz notch filter. Participant videos were inspected for gross inattention to the video and movement artifact, and these time points were removed from EEG analyses. For each participant’s data persistent bad channels that were deviant in 33% of trials (identified with Fieldtrip function ft_rejectvisual based on within channel variance and kurtosis) were interpolated using spline interpolation, as implemented in the Fieldtrip function
ft_repairchannels. Interpolation was chosen to keep consistent datasets across participants. Amount of interpolated channels was between 4 and 21 for ASD participants and 6 and 24 for TD participants. Data were decomposed using Second Order Blind Identification (SOBI) as implemented in EEGLAB (Belouchrani et al., 1993; Delorme & Makeig, 2004). Topographic maps of SOBI components were inspected and electrooculogram (EOG) and electromyogram (EMG) components were removed. Forty one-second epochs from each of two presentations of the stimulus with minimal movement contamination were retained (again using ft_rejectvisual function). Data were then re-referenced to the common average as laid out in Nunez and Srinivasan (2006) using Fieldtrip ft_preprocessing function. Finally, a fast Fourier transformation (FFT) was performed on the rectangular windowed time series. For each of the 3 stimulus conditions presented twice, the presentation with the least amount of movement artifact was chosen for analysis.

**EEG data analysis**

As a result of preprocessing step, EEG data is a 3-dimensional array of voltages, with dimensions 40x124x1000 (40 one-second epochs, 124 channels, 1000 samples per second). Three scalp regions of interest (Regions: Frontal, Central, and Posterior), per McEvoy et al. (2015), were used in the analysis. Twelve channels covering the left hemisphere, right hemisphere, and midline were included in each of the three regions (Murias, Major, Compton, et al., 2018). Per each participant and each condition, average
Power Spectral Density (PSD) from 40 artifact-free seconds of EEG recording for each channel was binned into four power bands: theta (5–7 Hz), alpha (8-10 Hz), beta 1 (11-20 Hz), and beta 2 (21-30 Hz) (Murias, Major, Compton, et al., 2018). Relative Power Spectral Density (RP) was calculated by dividing PSD in each band by the total signal power between 3 and 30 Hz for each channel, resulting in four RP values per channel during each video stimulus. RP per region was then calculated by averaging values from twelve channels within the region. Additionally, log-ratio of RP between pairs of video stimuli V1 and V2 was computed, as well as log-ratio of Theta/Beta power ratio,

\[
LR_{V1-V2,\text{band},ROI} = \log\left(\frac{RP_{V1,\text{band},ROI}}{RP_{V2,\text{band},ROI}}\right),
\]

\[
LR_{TBR,V1-V2,ROI} = \log\left(\frac{RP_{V1,\text{theta},ROI}/(RP_{V1,\text{beta1},ROI} + RP_{V1,\text{beta2},ROI})}{RP_{V2,\text{theta},ROI}/(RP_{V2,\text{beta1},ROI} + RP_{V2,\text{beta2},ROI})}\right).
\]

**Measurement of Relative Average Look Duration**

By comparing within-subject relative average look duration to nonsocial versus social stimuli, each participant’s attention to one type of stimulus serves as a “baseline” for comparing that participant’s attention to the second type of stimulus. Attention was coded as a binary signal based on the same video recording used for pre-processing EEG data. The following summary features were extracted from this binary signal:

1. Total looking duration (\(T\text{Look}_{V}\)) – total amount of time child was watching the screen when video stimulus V was presented;
(2) Average look duration ($ALD_v$) - This attention/looking variable was computed as follows for each video stimulus $V$:

$$ALD_v = \frac{TLook_v}{\#Number\ of\ looks\ at\ the\ screen\ during\ the\ video\ presentation}$$

By this variable we are providing a measurement of the intermittent behavior of attention.

Relative average look duration ($RALD_v$) – This core new measurement was computed as

$$RALD_{V1,V2} = \frac{ALD_{V1} - ALD_{V2}}{ALD_{V1} + ALD_{V2}}.$$ 

This measure can be considered as a measure of engagement in one type of video ($V1$), treating another as baseline level ($V2$). As can be seen from the formula, $RALD_{V1,V2}$ takes values in the range of [-1, 1]. For example, full engagement in Social video and disengagement in Toys video means $RALD_{Social,Toys} = 1$, and, vice versa, $RALD_{Social,Toys} = -1$ means full engagement in Toys video and disengagement in Social video.

**Statistical analysis methods**

To evaluate the ability to distinguish the ASD from the TD group using the attention measures, two-way ANCOVA was performed for $ALD_v$, with categorical predictors of Group (ASD/ TD) and video type (Bubbles/Social/Toys), and IQ, Sex and Age as covariates. Then, one-way ANCOVA was performed for $RALD_{V1,V2}$ with predictor
of Group and same covariates. In the order of increasing complexity, we applied the following models, each time testing whether a new predictor significantly increases explanatory power of the model:

\[
ALD \sim IQ \quad \text{(Eq. 1a)}
\]
\[
ALD \sim Sex + IQ \quad \text{(Eq. 1b)}
\]
\[
ALD \sim Age + Sex + IQ \quad \text{(Eq. 1c)}
\]
\[
ALD \sim Group + Age + Sex + IQ \quad \text{(Eq. 1d)}
\]
\[
ALD \sim Group + VideoType + Age + Sex + IQ \quad \text{(Eq. 1e)}
\]
\[
ALD \sim Group + VideoType + Group:VideoType + Age + Sex + IQ \quad \text{(Eq. 1f)}
\]

A univariate linear modeling approach was used to search for possible associations between the proposed baselined attention measures \((RALD_{Social,V2})\) and EEG signal features, namely log-ratio of \(RP\) (relative power) and \(TBR\) (Theta/Beta ratio). In other words, \(LR_{(Social-V2(choice of non-social))}\) and \(LR_{(TBR,Social-V2(choice of non-social))}\) were treated as dependent variables, while \(RALD_{Social,V2}\), Group, and interaction between \(RALD_{Social,V2}\) and Group were taken as predictors, with Age, Sex, IQ treated as covariates. This is all expressed in the following equation:

\[
(LR_{Social-V2,band,Region}, \text{or } LR_{TBR,Social-V2,Region}) = \beta_0 + \beta_1(RALD_{Social,V2}) + \\
\quad \beta_2 Group + \beta_3 Group \times RALD_{Social,V2} + \beta_4 Age + \beta_5 Sex + \beta_6 IQ \\
\]  
\text{(Eq. 2)}

In all models we were interested in the effect of \(RALD_{Social,V2}\) and interaction term \((Group \times RALD_{Social,V2})\) on the dependent variable. False Discovery Rate (FDR, Benjamini
and Hochberg (1995)) correction was applied to p-values corresponding to these 2 regression coefficients from all models of the above type. Initially FDR correction was done for 60 tests, since not only LR was used as a dependent variable, but also RP_{Social}. However, results for RP_{Social} were not directly interpretable. So for the sake of reproducibility we recomputed FDR for 30 tests. Nothing changed in the significance of the results, only FDR-corrected p-values slightly changed.

For those models that proved significant on RALD or Group*RALD interaction we tested whether adding the interaction term significantly improves explanatory power of the model, by applying 3 models in the order of increasing complexity:

\[
(LR_{Social\_v2.band.Region, or LR_{TBR, Social\_v2.Region}}) = \beta_0 + \beta_2 Group + \beta_4 Age + \beta_5 Sex+\beta_6 IQ \\
\text{(Eq. 3)}
\]

\[
(LR_{Social\_v2.band.Region, or LR_{TBR, Social\_v2.Region}}) = \beta_0 + \beta_1 (RALD_{Social,v2}) + \beta_2 Group + \beta_4 Age + \beta_5 Sex+\beta_6 IQ \\
\text{(Eq. 4),}
\]

and model from Eq. (2).

**Results**

**Average look duration**

First, and for the sake of completeness, we report initial results on ALD, since it forms the basis of RALD, which is the main subject of this work. Patterns of attention during each type of stimuli for both groups can be seen on Figure 16 (a). Our sequential tests for explanatory power revealed that it increases by adding Group (Eq. 1d), Video
type (Eq. 1e) and Group×Video type interaction (Eq. 1f) (p<0.05, p<0.001, p<0.05 on F-tests respectively), even after controlling for differences in IQ. We observed a strong effect of Group, such that ASD children exhibited shorter look durations than the TD group (F\textsubscript{1,177}>35.39, p<0.001); and of video stimulus type, with both groups most engaged in Toys and least engaged in Bubbles (F\textsubscript{2,177}>16.04, p<0.001). Further, an interaction effect was significant (F\textsubscript{2,177}>3.21, p<0.05), suggesting that relative level of engagement between Social, Toys and Bubbles is different in ASD group and TD group; In the ASD group, decreased ALD was most evident while viewing the Social video, which was not the case for the TD group. See Figure 16 (b) for details. This effect also becomes evident when considering RALD measure below.

**Relative average look duration**

The ability of 3 different RALD\textsubscript{v1,v2} measures (contrasting Social vs. Toys, Social vs. Bubbles, and Toys vs. Bubbles videos) to distinguish ASD from TD group was explored. To this end, one-way ANCOVA models with Sex, Age and IQ confounding variables were exploited. RALD\textsubscript{Social,Bubbles} and RALD\textsubscript{Social,Toys} (which can be considered Social vs. Non-Social) demonstrated significant ability to separate between the groups (F\textsubscript{1,57}>4.43, p<0.04 and F\textsubscript{1,57}>10.50, p<0.002 respectively), while RALD\textsubscript{Toys,Bubbles} (two non-social stimuli) did not (F\textsubscript{1,57}>0.50, p<0.48).
**EEG measures results**

We selected $RALD_{Social,Tags}$ for combined analysis with EEG since it better separates groups (see above), and being a relative measure it eliminates potential effects of baseline mood, excitation, or drowsiness at the day of experiment. It also follows independent findings about the value of changes/differences (RALD) contrary to absolute behaviors (McPartland, Webb, et al., 2011; Webb et al., 2010).
Figure 16: Visual attention measurements. (a) Each line (31 lines total) on each one of the six images represents one participant’s attention during the course of 120 seconds (2 video repetitions of 60 seconds length). (b) Behavior of $ALD_v$ for each video type. Asterisks mark level of significance on two-sample t-test between two groups within each video type. * $p < 0.05$, **** $p < 0.0001$. 
When taking the log-ratio of $RP_{Social}$ and $RP_{Toys}$ (which is a measure of relative brain activation while viewing social versus nonsocial stimuli; see (Dawson et al., 2012)), different patterns of associations between $RALD_{Social,Toys}$ and EEG measures become evident. Specifically, $LR_{Social-Toys,Theta,Central}$ and $LR_{Social-Toys,Theta,Posterior}$ had a positive association with $RALD_{Social,Toys}$ in the TD group and negative association with $RALD_{Social,Toys}$ in the ASD group at the level of $p<0.1$, while inverse pattern of association was observed in Beta 1 Frontal band ($p<0.05$).

The log-ratio of Theta-Beta Ratio ($LR_{TBR,Social,Toys}$) measure showed significant positive association with $RALD_{Social,Toys}$ (in TD group, all regions) and significant negative association (in ASD group, Posterior region), or a tendency to negative association (in ASD group, Frontal and Central regions), as can be seen from confidence intervals shown in Figure 17 and Table 13. Our sequential tests for explanatory power demonstrated that simply adding $RALD_{Social,Toys}$ to the model containing only covariates and Group (Eq. 4 compared to Eq. 3) does not improve the explanatory power of the model regardless of region and frequency band. However, adding $RALD_{Social,Toys}$ and interaction Group $\ast$ $RALD_{Social,Toys}$ (Eq. 2 compared to Eq. 3) increases explanatory power in all models that we reported significant above ($p<0.05$ on F-test).

While our primary hypothesis involved comparing brain activity across two audiovisual conditions that differed in social versus nonsocial content, we also carried out similar analyses comparing the social and bubbles conditions and the toys and
bubbles conditions. In this case, the two conditions differed not only in content but also level of stimulation because the bubbles condition did not involve audio. No significant results were found.

Figure 17: Relations between $RALD_{Social,Toys}$ and $LR_{Social-Toys,Theta}$, $LR_{Social-Toys,Beta1}$, and $LR_{TBR,Social-Toys}$ in Posterior Region for TD and ASD groups. Log-Ratio (LR) takes on relative powers of frequency bands as arguments. Results are typical for all the Regions.
Table 13: Associations of log-ratio of EEG Relative Power in Social and Toys videos and \( LR_{TBR, Social-Toys} \) and RALD. EEG measurements were aggregated to three regions, frontal (F), central (C), and posterior (P). Spectral power was binned into four frequency bands: Theta (5–7 Hz), Alpha (8–10 Hz), Beta 1(11–20 Hz), Beta 2 (21–30 Hz). Associations in bold are significant. In “Confidence Interval” section in case at least one association is significant, positive associations are marked in bold underline with italic, negative associations in bold underline, and no association in bold.

<table>
<thead>
<tr>
<th>log((RP_{Social}/RP_{Toys}))</th>
<th>( RALD_{Social,Toys} )</th>
<th>Group (* RALD_{Social,Toys})</th>
<th>Confidence Interval for ( \beta ) coefficient at ( RALD_{Social,Toys} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region</td>
<td>( \beta )</td>
<td>( p)-val (FDRcorr)</td>
<td>( \beta )</td>
</tr>
<tr>
<td>Theta F</td>
<td>0.2304</td>
<td>0.150</td>
<td>-0.3122</td>
</tr>
<tr>
<td>Theta C</td>
<td>0.2950</td>
<td>0.070</td>
<td>-0.3881</td>
</tr>
<tr>
<td>Theta P</td>
<td>0.2911</td>
<td>0.094</td>
<td>-0.4389</td>
</tr>
<tr>
<td>Alpha F</td>
<td>0.1220</td>
<td>0.527</td>
<td>-0.0027</td>
</tr>
<tr>
<td>Alpha C</td>
<td>0.0108</td>
<td>0.982</td>
<td>0.2416</td>
</tr>
<tr>
<td>Alpha P</td>
<td>-0.1715</td>
<td>0.405</td>
<td>0.3648</td>
</tr>
<tr>
<td>Beta1 F</td>
<td>-0.3124</td>
<td>0.041</td>
<td>0.4514</td>
</tr>
<tr>
<td>Beta1 C</td>
<td>-0.1401</td>
<td>0.275</td>
<td>0.2130</td>
</tr>
<tr>
<td>Beta2 P</td>
<td>-0.2460</td>
<td>0.137</td>
<td>0.4019</td>
</tr>
<tr>
<td>Beta2 C</td>
<td>-0.3480</td>
<td>0.204</td>
<td>0.3734</td>
</tr>
<tr>
<td>Beta2 P</td>
<td>-0.2036</td>
<td>0.275</td>
<td>0.1314</td>
</tr>
<tr>
<td>LR_{TBR} F</td>
<td>0.5609</td>
<td>0.041</td>
<td>-0.7426</td>
</tr>
<tr>
<td>LR_{TBR} C</td>
<td>0.4599</td>
<td>0.042</td>
<td>-0.5829</td>
</tr>
<tr>
<td>LR_{TBR} P</td>
<td>0.5824</td>
<td>0.042</td>
<td>-0.8842</td>
</tr>
</tbody>
</table>

Discussion

In order to illustrate the value of jointly studying attentional behavior and EEG, we first investigated a metric for average look duration (ALD), defined as the average length of separate looking periods (intermittent attention) to a complex, dynamic stimulus. We found that, compared to age matched neurotypical children, children with autism have shorter average look durations to both social and nonsocial complex
dynamic audiovisual stimuli. ALD is the building block for a newly proposed measure of relative average look duration (RALD) to different stimulus types; RALD to the social compared to nonsocial stimuli exhibited differential associations with neurophysiological measures for the autistic and neurotypical groups. These results indicate that the neural systems that mediate relative differences in sustained attention to social versus nonsocial stimuli are not the same for children with autism versus neurotypical children. Additionally, adding an interaction term, thus accounting for differential associations in neurotypical and autistic children, significantly increased explanatory power of the model even after controlling for group differences in IQ and sex. These results therefore support the idea that autistic and neurotypical children process social and nonsocial stimuli differently and that combining simultaneously recorded attentional behavioral data and EEG data adds explanatory value in understanding these differences.

Group differences in ALD were most robust when the children were viewing the social stimulus, further supporting the use of RALD to capture differential attention between different stimulus types. That is, autistic children had shorter look duration for all stimuli types, but the effect was most pronounced in the social condition. When the relative measure (RALD) was examined, it was found that the contrast between social and nonsocial stimuli ($RALD_{Social,Toys}$, $RALD_{Social,Bubbles}$) distinguished the neurotypical and autism groups, while the contrast between two nonsocial stimuli ($RALD_{Toys,Bubbles}$) did not
yield group differences. Autistic deficits in sustained attention appear to be strongest when social content is involved, highlighting a context-specific difference in attention. This may distinguish autism from other disorders of attention, including attention-deficit/hyperactivity disorder (ADHD) and schizophrenia. Our findings are consistent with other studies that have shown differences between autistic and neurotypical children in total looking time and peak look duration in the context of social attention (Chawarska et al., 2012, 2013; Hendry et al., 2018; Klin et al., 2002; Shic et al., 2011). However, these studies are limited in that they did not measure average look duration and only focused on maximum and total durations.

As the contrast between the Social and Toys stimuli for average look duration \((RALD_{Social,Toys})\) was most robust, we used it for subsequent investigation of its relationship with simultaneously recorded EEG, where the measure of interest was relative EEG power during the social as compared to nonsocial stimuli. Analysis revealed that correlations between RALD for social versus toys stimuli and underlying patterns of EEG differed for children with autism versus neurotypical children. For neurotypical children, as average look duration to social stimuli relative to toys increased, central and posterior EEG theta power while viewing social versus toy stimuli also increased. This is consistent with previous studies that have found that frontal theta band activity increases when individuals pay attention to multi-sensory stimuli involving auditory and visual input (Keller et al., 2017), particularly since the social...
stimuli were more complex (e.g., involving language) than the dynamic toy stimuli. Furthermore, reduced frontal beta power while viewing social versus toy stimuli was associated with increased attention duration to the social relative to the nonsocial video. Studies have shown that working memory encoding is associated with a transient reduction in beta power (see review by Hanslmayr and Matuschek (2014)). The social video involves the actress speaking and gesturing to the child which might have invoked working memory processes.

Different patterns emerged in the autism group. Increased posterior theta power and decreased frontal and posterior beta power while viewing the social versus toy stimuli was associated with shorter average look durations to the social video relative to the nonsocial video. Taken together, we conclude that neurotypical children with preferential attention to social vs nonsocial stimuli exhibit an expected brain response while watching the social stimuli that is characterized by high levels of theta power and low levels of beta power across the scalp. ASD children, however, appear to show the opposite effect or no association at all. This suggests that even when autistic children show preferential attentional engagement with social content, their underlying brain activity is not the same as the neurotypical children. Given the relatively small sample size and the fact that we did not make a priori predictions regarding these associations for the autism group, replication with a larger sample size is needed.
Our findings, showing main differences between autistic and neurotypical children in the associations of looking behavior and EEG in theta and beta bands, prompted us to study a metric, typically used in ADHD research – Theta-Beta ratio (TBR) (Arns et al., 2012; Gloss et al., 2016; Lenartowicz & Loo, 2014), which measures an increase in theta power relative to a decrease in beta power. Indeed, TBR log-ratio was positively associated with $RALD_{Social,Toys}$ for neurotypical children across all scalp regions, while a negative association existed for ASD children across the posterior region. There has been evidence for use of TBR as a biomarker for ADHD, but the exact neural basis of the TBR is still poorly understood (see Lenartowitz and Loo (2014) and Jeste et al. (2015) for reviews). However, no previous research to our knowledge has looked into TBR as a measure of brain response to social/nonsocial stimuli, especially in children with autism. Interestingly, it is estimated that 37-85% of the autism population has comorbid ADHD (Leitner, 2014). While both disorders involve disruptions in attention, individuals with ADHD show more pronounced deficits in sustained attention than those with ASD (Davis & Kollins, 2012; Johnson et al., 2007). Since TBR is atypical in both ADHD and ASD, we may be probing attentional brain circuitry that is commonly disrupted in both disorders. More work is needed to understand the similarities and differences in brain functioning and attentional behaviors between ASD and ADHD participants.

Previous research has studied EEG activity during social and non-nonsocial videos in clinical trials for children with autism. In a study comparing a TD group and
two behavioral intervention models (Early Start Denver Model (Dawson et al., 2010) and community intervention), Dawson et al. (2012) reported increased log-ratio (Faces vs Objects) theta power in the TD group as well as the ESDM group, while the opposite pattern was observed in the group that received community intervention. In the present study we found that, for the neurotypical group, increased theta power during the social stimulus and log-ratio (Social vs Toys) was associated with increased preference to social videos, consistent with Dawson’s findings in the neurotypical and ESDM groups. Increases in theta power have been implicated in the allocation of greater attentional and cognitive resources (Dawson et al., 2012). Furthermore, Murias et al. (2018) found that higher baseline beta power was predictive of changes in the Vineland Socialization subscale score in an open-label trial testing the efficacy of umbilical cord blood for children with ASD. It is clear from these studies that theta and beta power are viable, modifiable biomarkers for autism, and the current results provide additional evidence that these brain markers are associated with the ability to sustain attention, which involves development of inhibitory and executive functioning skills. Data shows that there is similar association in autistic children in our study (log-ratio of beta power increases as engagement in social stimulus increases), while this pattern is not showing up in neurotypical children. This supports the idea that autistic and neurotypical children process social and nonsocial stimuli differently, and combining behavioral and EEG data is a way to reveal this difference.
A potential weakness in our study was the fact that TD and ASD groups differed not only in terms of an ASD diagnosis but also in cognitive ability and sex distribution, with the ASD group having a mean lower IQ and more male participants compared to the TD group. All results accounted for this group difference by including IQ and sex as covariates. However, future work should include comparisons between groups with similar cognitive ability and sex distribution. In addition, the study used a relatively small sample size which limited statistical power; nevertheless, many findings were still robust enough to be statistically significant.

In the presented work we proposed a method to analyze looking behavior during synchronized spontaneous EEG recording. We showed that the measures of looking behavior that incorporate not only total attention time but also average look duration, differentiate between neurotypical and autistic children and are associated with differential patterns of EEG activity in neurotypical and autistic children. Future work will aim at combining this measure with EEG signal features for improving assessment of autism spectrum disorder.
Chapter 7. Computer vision analysis for labeling inattention during EEG recordings with visual stimuli

As was shown in the previous chapter, there is value in combining EEG and visual attention in the studies of autism, which can lead to novel biomarkers for this condition. Synchronized EEG and video recordings are typical in experimental settings, and the data collected from the video cameras are a source for manual annotation. Currently, detailed coding to determine the participant’s attention is typically a labor-intensive process that relies on subjective human judgment.

In this chapter, we propose a machine learning model trained on computer vision features such as head pose, gaze, and face landmarks to detect participants’ visual attention. Variability of children’s behavior during the EEG acquisition requires algorithm adaptation to a new participant and review of the results by the human. We train the model on an initial dataset and then adapt it with a small subset of data per participant. We show that, compared to the non-adapted model, after training on additional 2560 labeled frames (equivalent to 85.3 seconds) of the video of a new participant, the performance of the algorithm substantially improves.

Our results demonstrate the feasibility of automatic tools to detect inattention during EEG recordings, and their potential to reduce the subjectivity and time burden of human coding. Accelerating and making this process more objective will allow to scale
up the combined EEG and behavior studies by involving multiple research centers and increasing the sample sizes.

The tool for model adaptation and human review will be made publicly available to the research community.
Introduction

Electroencephalogram (EEG) is a widely used method for studying brain-behavior relations. A typical EEG recording session includes visual and/or auditory tasks, which can be presented in an event-related potential (ERP) paradigm or during continuous EEG recording. Collecting data using visual tasks in children is significantly more challenging due to their reduced ability to continuously pay attention to visual stimuli (DeBoer et al., 2007; Thierry, 2005). A meta-analysis by Stets et al. (2012) reports that studies involving visual tasks in infants have significantly higher attrition rates than auditory or combined visual and auditory tasks. While reports of attrition rates in different studies vary (Bell & Cuevas, 2012; DeBoer et al., 2007; Stets et al., 2012), a general recommendation is designing tasks that will be engaging for children (i.e., facilitating the maintenance of visual attention (Bell & Cuevas, 2012)). To facilitate visual attention children may be asked to provide a behavioral response (e.g., press a button (Ellis & Nelson, 1999; Todd et al., 2008)), or an experimenter may gently redirect a child to the screen when noticing signs of disengagement (Isaev, Major, et al., 2020; Murias, Major, Davlantis, et al., 2018; Todd et al., 2008).

Removing segments of the data where a participant did not look at the screen is the first stage of data processing in recordings with visual stimuli. Typically, researchers either code the participant’s attention on-line (Dawson et al., 2012; Ellis & Nelson, 1999; Orekhova et al., 2006) by pressing a button which sends a marker to the EEG recording
when the participant was not attending to the stimulus, or by recording the video of
participant’s behavior synchronously with the EEG recording and marking periods of
inattention post-hoc by reviewing the video (Isaev, Major, et al., 2020; Murias, Major,
Compton, et al., 2018). This is a burdensome manual process requiring significant time
and effort. It is also highly subjective, for example the annotator might only see the
participant’s face and must guess whether the participant’s gaze is directed to the area
inside or outside of the screen. Subjectivity on this first stage of data processing poses an
obstacle for EEG studies, in particular for multi-center ones, since control for EEG data
quality in multi-center studies is critical (Kaiser et al., 2021).

In addition to its value for data curation, information about inattention periods
can be useful for creating clinical biomarkers. There is evidence of alterations in
orienting, disengagement from, and sustaining attention to relevant stimuli in the early
autistic development (Elsabbagh et al., 2013; Keehn et al., 2013; McPartland, Webb, et al.,
2011; Werner et al., 2000), which undoubtedly influences the amount of inattention
during the EEG study. Though a typical study of continuous EEG/ERP excludes from
analysis time periods where the participant is not engaged with the visual stimulus
(Orekhova et al., 2014; Orekhova et al., 2006; Stroganova et al., 1998), inattentiveness
during EEG in social/nonsocial stimuli can be a measure that distinguishes autistic and
neurotypically developing populations, used alone or in conjunction with EEG power
features (Isaev, Major, et al., 2020).
Conventional eye-tracking technologies can address the problem of detecting inattention. Presenting stimulus on the eye-tracker screen and simultaneously recording eye-tracking and EEG (see Ahtola et al. (2017) for the example of the setup) allows to detect whether the participant is watching the screen or not. For example, a study by Maguire et al. (2014) proposed using an eye-tracker synchronized with EEG to present an “attention-getter” animation in an experiment with children 6-8 years old. They reported increased retention of EEG data compared to the condition where children were asked to provide a behavioral response (button pressing). However, the usage of an eye-tracker just for detecting inattention is a too expensive solution and not scalable.

By contrast, computer vision analysis (CVA) on data recording with off-the-shelf cameras provides scalable tools to objectively measure children behavior. This is largely enabled by the progress in face detection and estimation of facial landmarks, head pose, and gaze (Baltrusaitis et al., 2018; Krafa et al., 2016; Lugaresi et al., 2019; Torre et al., 2015). In non-EEG settings, these tools have been able to detect head turns in response to name (Perochon et al., 2021), and capture patterns of gaze in a low-cost setting without additional calibration (Chang et al., 2021; Erel et al., 2022). For example, iCatcher (Erel et al., 2022) is a publicly available supervised deep learning model trained to classify infants’ gaze into three categories (‘left’, ‘right’, and ‘away’) based on the facial appearance. In the work of Qian et al. (2022), supervised machine learning in
combination with CVA approaches was applied for blink and head movement artifacts
detection in a minimally constrained portable EEG setting.

In this work, we develop a combination of CVA and supervised machine
learning model to detect inattention periods during the EEG recordings. This is
computed from the videos of the child’s head and upper body, captured synchronously
with EEG and with simple off-the-shelf cameras. We hypothesized that automatic CVA
codes of eye gaze coordinates, head pose descriptors (pitch, yaw, and roll), and nose
landmarks could reliably detect periods of visual distraction from the screen using a
supervised machine learning model. At the same time, we propose minimally involving
human annotators in the fine-tuning of the model and post-processing of the results
using a graphical user interface (GUI). Minor human involvement is critical since head
poses and facial expressions of children vary a lot in clinical populations, justifying the
need and opportunity for tuning the pre-trained model for new participants. Recent
work, iCatcher, provides evidence that the least agreement between human annotators
and automatic models occurs on the label ‘looking away from the screen’ (Erel et al.,
2022). The postprocessing stage gives an opportunity for additional quality control of
inattention periods proposed by the model. The overall proposed pipeline still
significantly reduces coding time and subjectivity. We therefore train the model on an
annotated dataset of 23 children and then adjust it to a new child by labeling a limited
amount of randomly selected additional frames on the new video. We will openly share
online the GUI for the video and CVA features inspection, model retraining, and predictions post-processing.

**Method**

**Participants**

Participants were twenty-three children (16 males), ranging from 49 to 95 months of age who were part of a study funded by the National Institute of Health. The ethnic and racial composition of the sample was as follows: White, 17; Black, 0; Asian, 2; Other and mixed race, 4; Hispanic, 4. All 23 children in the sample had autism spectrum disorder (ASD), and 11 children had comorbid attention deficit/hyperactivity disorder (ADHD) in addition to ASD.

All caregivers/legal guardians of participants gave written, informed consent and the study protocol was approved by the Duke University Health System Institutional Review Board. Methods were carried out in accordance with institutional, State, and Federal guidelines and regulations.

*Recording synchronized video and EEG*

Continuous EEG was recorded as participants were presented with three video stimuli, followed by Event-Related Potential (ERP) protocol with the presentation of faces and houses, audio oddball, and visual evoked potentials (VEP) stimuli. One or two clinician assistants were present in the room during the EEG recording to ensure the quality of the session and to redirect participant’s attention back to the screen in case
they were distracted. EEG data were recorded from 124 channels with reference to Cz using a Hydrocel Geodesic Sensor Net and Net Amps 400 amplifier (Electrical Geodesics, Eugene, Oregon). Data were collected using Netstation 4.5.6 with a sampling rate of 1000 Hz. The child’s face was recorded from a Basler ACE acA1300-30uc camera below the screen synchronized with the EEG. The camera resolution was 1296x966 pixels and frame rate was 30 fps. To synchronize camera and EEG, an in-house software code was used, based on the Basler pylon library” and Cedrus StimTracker hardware device used to set markers on the EEG recording. A diagram of the recording setup is shown in Figure 18.

Figure 18: EEG Recording setup. Video from the camera is recorded on the Video Recording Computer, which sends a marker to the EEG Recording Computer via Cedrus Stimtracker every 100 frames. This allows to synchronize between the EEG and video recordings.

**Extracting CVA features**

To extract the CVA features, we used an in-house code involving 3 steps: (a) face detection and disambiguation, (b) extraction of landmarks and head pose angles, and (c) gaze estimation. The raw set of extracted features per frame included nose $x$ (horizontal) and $y$ (vertical) coordinates in the frame, gaze $x$ and $y$ coordinates in the presentation screen plane, and head pose angles (pitch, yaw, and roll).
**Face detection and disambiguation.** Code for face detection and disambiguation used the `face_recognition` python library†† based on the `dlib` C++ library (King, 2009). Every time the algorithm detected more than one face on the video (which happened either due to ambiguity of face detection – one face was detected twice, or when another person, e.g., clinician assistant entered the frame), the algorithm showed the frame with a bounding box and prompted the user to select the correct participant’s face.

**Extraction of landmarks and head pose angles.** After the faces were detected, an algorithm for facial landmark extraction based on the `intraface` software library (Torre et al., 2015) was applied to the detected faces. As a result, facial landmark pixel coordinates, as well as pitch, yaw, and roll head pose angles were obtained.

**Gaze estimation.** The iTracker software (Krafka et al., 2016) was used for gaze estimation, providing gaze x and gaze y coordinates in the screen plane. Even though iTracker was trained to predict gaze coordinates on a mobile device screen for the frames captured from mobile device frontal camera, we used the output of iTracker as a proxy for gaze coordinates in the presentation screen plane. The software package is modular and this component can be easily replaced by others as preferred by the user.

Since the `intraface` library is not available to the general public at the moment, for the convenience of potential users we will make publicly available an alternative

†† https://pypi.org/project/face-recognition/
processing pipeline which consists of our original face estimation and disambiguation code, and a code for landmarks, head pose and gaze extraction using the popular OpenFace software package (Baltrusaitis et al., 2018).

Data attrition

Due to pauses between EEG/ERP recordings where the behavior of participants was significantly different, inattention detection was restricted only to the periods during the actual recordings, and the training set for the machine learning (ML) model included only data from frames inside those periods. Frames where the face could not be detected (hence there was no information on landmarks and head pose) were excluded from the analysis as well.

Data preprocessing

Since inattention is invariant to the direction (it happens both when participants looks to the right or to the left, turn the head up or down, etc.), each feature was transformed into a positive (‘plus’) and negative (‘minus’) version,

\[
feature_{plus} = \max(0, feature - \text{median}(feature)),
\]

\[
feature_{minus} = \text{abs} \left( \min(0, feature - \text{median}(feature)) \right)
\]

The final set of features for the analysis is reported in Table 14.
Table 14: List of input features per frame for the machine learning model.

<table>
<thead>
<tr>
<th>Feature name</th>
<th>Feature description</th>
</tr>
</thead>
<tbody>
<tr>
<td>noseX&lt;sub&gt;plus&lt;/sub&gt;</td>
<td>Nose coordinates</td>
</tr>
<tr>
<td>noseX&lt;sub&gt;minus&lt;/sub&gt;</td>
<td></td>
</tr>
<tr>
<td>noseY&lt;sub&gt;plus&lt;/sub&gt;</td>
<td></td>
</tr>
<tr>
<td>noseY&lt;sub&gt;minus&lt;/sub&gt;</td>
<td></td>
</tr>
<tr>
<td>gazeX&lt;sub&gt;plus&lt;/sub&gt;</td>
<td>Gaze coordinates</td>
</tr>
<tr>
<td>gazeX&lt;sub&gt;minus&lt;/sub&gt;</td>
<td></td>
</tr>
<tr>
<td>gazeY&lt;sub&gt;plus&lt;/sub&gt;</td>
<td></td>
</tr>
<tr>
<td>gazeY&lt;sub&gt;minus&lt;/sub&gt;</td>
<td></td>
</tr>
<tr>
<td>yaw&lt;sub&gt;plus&lt;/sub&gt;</td>
<td>Head pose angles</td>
</tr>
<tr>
<td>yaw&lt;sub&gt;minus&lt;/sub&gt;</td>
<td></td>
</tr>
<tr>
<td>pitch&lt;sub&gt;plus&lt;/sub&gt;</td>
<td></td>
</tr>
<tr>
<td>pitch&lt;sub&gt;minus&lt;/sub&gt;</td>
<td></td>
</tr>
<tr>
<td>roll&lt;sub&gt;plus&lt;/sub&gt;</td>
<td></td>
</tr>
<tr>
<td>roll&lt;sub&gt;minus&lt;/sub&gt;</td>
<td></td>
</tr>
</tbody>
</table>

After preprocessing the features, the participant identifier was one-hot encoded and added to the feature list. This allowed to learn a separate bias term in the first layer of the trained neural network, resembling the design of mixed models.

**Data labeling**

Data for all 23 participants was labeled by one of the co-authors using the Elan software. Four participants were randomly selected for independent annotation by another co-author. Both annotators did not participate in data analysis. Annotators labeled data using the recorded video as ‘gaze off screen’ (GOS) if the participant looked away from the screen, and/or as ‘head turn’ (HT) if the participant turned their head. For the purpose of inattention detection, a frame was labeled as ‘inattention’ if it either was
labeled as a head turn or gaze off screen. Agreement on inattention labels between independent annotators was assessed with Cohen’s κ (Cohen, 1960).

Training and evaluating machine learning model

Given the frame-by-frame preprocessed data as an input, we trained a multi-layer perceptron (MLP) model with 2 hidden layers (layer dimensions 512 and 14 were selected empirically) (Hastie et al., 2001). The target variable was inattention label per each frame with cross-entropy as a cost function. Adam optimizer was used for model training (Kingma & Ba, 2015). We used weighted sampling for model training to allow each batch to have close to equal amount of positive and negative samples (inattention and attention respectively). Models were trained in the pytorch framework (Paszke et al., 2019). To evaluate the model performance, we assessed average precision (AP, also known as area under precision-recall curve), area under the ROC curve (AUC), and maximal Cohen’s κ (MK) between the human annotator and the machine learning predictions per participant across different thresholds. Evaluation was done using the leave-one-subject-out cross-validation (LOSO CV) method.

Transfer learning: adjusting ML model to a new subject

Our adaptation approach involved selecting a batch of 128 frames (corresponding to 4.270 s) for labeling and training for 20 iterations on a newly labeled data at each epoch of additional training. To evaluate the performance of this approach, we assessed the three metrics defined in the previous section, considering both
sequential (where frame features and labels are sampled into the batch sequentially from the beginning of the video, which resembles how human would look through the dataset and label it), and random frame sampling approaches. We additionally assessed the maximum of median Cohen’s $\kappa$ across distribution, and computed the respective prediction threshold at epochs 5, 10 and 20, which correspond to 21.3, 42.6 and 85.3 additionally labeled seconds of data per subject. The exact algorithm was as follows:

1. Set $N = 128$ (the batch size).
2. Create empty dataset for labeled data.
3. Set $\text{Epoch} = 0$.
4. Predict probabilities of sample being positive in each frame.
5. If approach is Random sampling, randomly sample $N$ frames into the batch from the participant’s data.
6. If approach is Sequential sampling, sample next $N$ frames from the beginning of the participant’s data into the batch.
7. Remove frames included in the batch from the participant’s data.
8. Add batch to the labeled dataset‡‡
9. Train for 20 iterations on labeled dataset
10. Compute AP, AUC, and MK
11. Set $\text{Epoch} += 1$.

‡‡ For training in LOSO CV framework we used the labels from the dataset for the subject the algorithm was being adapted to.
12. If Epoch == 50: stop


**Agreement measurements between model and human and between two humans**

One of the metrics of quality assessment was Cohen’s $\kappa$. To compare maximal median $\kappa$ value between the model and the human annotator with human agreement level, we randomly selected 4 participants and performed independent labeling by another annotator. Then we computed Cohen’s $\kappa$ to measure agreement between both human annotators. We additionally computed Cohen’s $\kappa$ between the model prediction on a threshold level corresponding to maximal median $\kappa$ at epoch 20 and a consensus annotation of the two human raters (in a consensus annotation frame is labeled ‘inattention’ only if both annotators labeled it as such, otherwise frame is labeled ‘attention’).

**Graphical User Interface for visualizing and retraining the model**

We created a web-based GUI which allows for visualizing the data, labeling the data frame-by-frame and re-training the model, and post-processing of the data (see Figure 19 for screenshot). The tool with installation and usage instructions will be made publicly available.
Figure 19: A: Visualization of CVA features together with the video of the participant (here blurred to protect privacy). B: Interface for labeling the frames.

Results

Dataset statistics

The original full dataset consisted of 566,043 frames. After dropping frames where the face or gaze were not detected, 535,539 frames were retained (5.38% of invalid frames), with mean and standard deviation of amount of frames per participant of 23,284 and 6,193 respectively. The total frames labeled as inattention across the retained dataset was 79,629 (14.86% of the dataset). The percentage of inattention frames and total amount of frames per participant can be found in Table 15.

Transfer learning results

The results of transfer learning can be seen in Table 18 and Figure 20. The sequential sampling approach performed substantially worse than the random sampling approach. Median AP, AUC and MK were 0.850, 0.963, 0.744 respectively at the start of the training (no adaptation to the participants yet). By epoch 20, median AP was 0.963,
AUC 0.989, and MK 0.889 on random sampling approach as compared to median AP 0.613, AUC 0.858, and MK 0.553 in sequential sampling approach.
Table 15: Total frames per participant and percentage of frames labeled as inattention.

<table>
<thead>
<tr>
<th>Participant #</th>
<th>% Inattention frames</th>
<th>Total frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>PT1</td>
<td>4.39</td>
<td>26,513</td>
</tr>
<tr>
<td>PT2</td>
<td>4.87</td>
<td>26,238</td>
</tr>
<tr>
<td>PT3</td>
<td>33.89</td>
<td>21,760</td>
</tr>
<tr>
<td>PT4</td>
<td>22.98</td>
<td>26,431</td>
</tr>
<tr>
<td>PT5</td>
<td>30.99</td>
<td>25,353</td>
</tr>
<tr>
<td>PT6</td>
<td>3.43</td>
<td>26,616</td>
</tr>
<tr>
<td>PT7</td>
<td>28.03</td>
<td>5,811</td>
</tr>
<tr>
<td>PT8</td>
<td>7.27</td>
<td>26,529</td>
</tr>
<tr>
<td>PT9</td>
<td>34.93</td>
<td>5,153</td>
</tr>
<tr>
<td>PT10</td>
<td>21.58</td>
<td>16,425</td>
</tr>
<tr>
<td>PT11</td>
<td>16.62</td>
<td>26,597</td>
</tr>
<tr>
<td>PT12</td>
<td>12.96</td>
<td>19,114</td>
</tr>
<tr>
<td>PT13</td>
<td>5.22</td>
<td>26,523</td>
</tr>
<tr>
<td>PT14</td>
<td>38.32</td>
<td>22,738</td>
</tr>
<tr>
<td>PT15</td>
<td>23.40</td>
<td>25,883</td>
</tr>
<tr>
<td>PT16</td>
<td>14.26</td>
<td>26,452</td>
</tr>
<tr>
<td>PT17</td>
<td>20.02</td>
<td>25,361</td>
</tr>
<tr>
<td>PT18</td>
<td>9.80</td>
<td>26,359</td>
</tr>
<tr>
<td>PT19</td>
<td>3.18</td>
<td>25,228</td>
</tr>
<tr>
<td>PT20</td>
<td>15.19</td>
<td>26,134</td>
</tr>
<tr>
<td>PT21</td>
<td>3.03</td>
<td>25,819</td>
</tr>
<tr>
<td>PT22</td>
<td>17.53</td>
<td>26,204</td>
</tr>
<tr>
<td>PT23</td>
<td>5.44</td>
<td>26,298</td>
</tr>
<tr>
<td>Total</td>
<td>14.86 (79,629 frames)</td>
<td>535,539</td>
</tr>
</tbody>
</table>
Figure 20: Performance metrics on different sampling/adaptation methods. Average precision, Maximal Cohen’s $\kappa$ and AUC per each epoch. Line color is median, and shaded area is interquartile range per each epoch.

Cohen’s $\kappa$ analysis.

Cohen’s $\kappa$ at different levels of prediction threshold for both sampling approaches (random and sequential) at epochs 5, 10, and 20 are shown in Figure 21. Thresholds at the highest median $\kappa$ and the corresponding median $\kappa$ values are shown in Table 16. The highest median $\kappa$ ranges between 0.792 and 0.888 in the random sampling approach, and between 0.223 and 0.440 in the sequential one.

Table 16: Thresholds and Cohen’s $\kappa$ levels at highest median value of $\kappa$ in the two sampling approaches at epochs 5, 10, 20.

<table>
<thead>
<tr>
<th>Sampling approach</th>
<th>Epoch</th>
<th>Threshold</th>
<th>Median Cohen’s $\kappa$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random sampling</td>
<td>5</td>
<td>0.254</td>
<td>0.792</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.446</td>
<td>0.837</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.280</td>
<td>0.888</td>
</tr>
<tr>
<td>Sequential sampling</td>
<td>5</td>
<td>0.01</td>
<td>0.223</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.04</td>
<td>0.307</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.254</td>
<td>0.440</td>
</tr>
</tbody>
</table>
Figure 21: Median (thick line) and Interquartile Range (shaded area) of Cohen’s $\kappa$ at different threshold levels at epochs 5, 10, and 20.

**Agreement between model and human coder and between two human coders**

A second independent annotator labeled videos from four participants, which in total accounts for 74,543 frames or 13.9% of the data. It took the second annotator approximately 22 hours to label the data, resulting in average of 1.06 seconds per frame. Cohen’s $\kappa$ values between the two human annotators ranged between 0.548 and 0.844. Agreement between model and consensus annotation increased with each epoch of additional training and was in the ranges [0.685 – 0.938] at epoch 5, [0.730-0.947] at epoch 10, and [0.831 – 0.961] at epoch 20 (see Table 17).
Table 17: Agreement level (Cohen’s κ) between human annotators and between the model and consensus annotation at epochs 5, 10 and 20.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PT1</td>
<td>0.584</td>
<td>0.685</td>
<td>0.730</td>
<td>0.831</td>
</tr>
<tr>
<td>PT9</td>
<td>0.727</td>
<td>0.841</td>
<td>0.910</td>
<td>0.939</td>
</tr>
<tr>
<td>PT10</td>
<td>0.548</td>
<td>0.730</td>
<td>0.825</td>
<td>0.826</td>
</tr>
<tr>
<td>PT16</td>
<td>0.844</td>
<td>0.938</td>
<td>0.947</td>
<td>0.961</td>
</tr>
</tbody>
</table>

**GUI for visualizing and preprocessing pipeline**

We developed a web-based GUI which may be used for reviewing the CVA features of the video, additional labeling of frames and retraining the model, and post-processing of the data, including setting the model decision threshold and rejection of falsely detected inattention events. We will make publicly available a pipeline for data preprocessing based on in-house code for face detection and OpenFace framework for head pose and gaze estimation (Baltrusaitis et al., 2018).
Table 18: Average precision, AUC, and Maximal Cohen’s κ percentiles at different epochs with two sampling/adaptation alternatives. The random sampling approach outperforms the sequential sampling one on all three metrics on each listed epoch.

<table>
<thead>
<tr>
<th>Sampling approach</th>
<th>Epoch</th>
<th>Average precision %-ile</th>
<th>AUC %-ile</th>
<th>Maximal Cohen’s κ %-ile</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>50%</td>
<td>25%</td>
<td>75%</td>
</tr>
<tr>
<td>Random sampling</td>
<td>0</td>
<td>0.850</td>
<td>0.712</td>
<td>0.919</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.907</td>
<td>0.829</td>
<td>0.945</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.930</td>
<td>0.883</td>
<td>0.967</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.963</td>
<td>0.931</td>
<td>0.980</td>
</tr>
<tr>
<td>Sequential sampling</td>
<td>5</td>
<td>0.433</td>
<td>0.270</td>
<td>0.703</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.532</td>
<td>0.397</td>
<td>0.778</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.613</td>
<td>0.449</td>
<td>0.810</td>
</tr>
</tbody>
</table>
Discussion

In this work we proposed a method for detection of periods of inattention to visual stimulus in EEG recordings. The tool is based on the CVA of participants’ behavior videos synchronously recorded with EEG. We outlined a data processing pipeline, including face and facial landmarks detection, head pose computation, and gaze estimation. We proposed a MLP model for predicting inattention from these CVA features, and random sampling as a means for fine-tuning the model for each participant. We made publicly available a GUI that allows for visualization of the CVA features, model fine-tuning, prediction thresholds adjustment, and results post-processing.

The proposed random frame sampling approach for model adaptation to the participant outperforms the sequential sampling approach. Compared to the initial non-finetuned model prediction, the model trained on additional 2560 labeled frames (equivalent to labeling only about 85 seconds of the video) significantly improved performance. On the other hand, sequential frame sampling performance decreases a lot in the initial 5 epochs (see Figure 20), then gradually improves, but does not get close to the performance of random the sampling approach. The reasons behind this include the strong temporal correlation of the features, hence low variability in the new input data,
and the rare occurrence of inattention (prevalence of inattention is 14.86%), causing the absence of positive labels in many batches.

In agreement with previous studies (Erel et al., 2022), we have found that agreement on inattention labeling from participants’ behavior videos by human coders was in the ‘moderate’ to ‘substantial’ ranges in 3 out of the 4 subjects, and in the ‘perfect’ range only for one subject (McHugh, 2012). However, when model performance was compared to the consensus annotation between humans, the minimal agreement was already in the ‘substantial’ range after labeling 640 additional frames, and in the ‘perfect’ range in all 4 subjects after labeling 2560 frames. As such, the proposed model tends to agree with human annotators where the human annotators agree among themselves, pointing to a more objective assessment of inattention.

Labeling inattention is a challenging task for human, likely because annotators have to guess the boundaries of the stimulus presentation screen. The provided GUI tool allows for visualizing the raw CVA features together with the participant’s video, also enabling coders to label frames for the fine-tuning or postprocessing stage. When the annotator needs to make a decision on ambiguous frame, they can play the video to compare the frame in question with neighboring frames, which may help to better evaluate whether the participant attends to the screen.
Our results show that the proposed approach can help to label data more efficiently. Given that labeling takes about 1.06 seconds per frame, the need to label only about 2560 frames for a high quality labeling can significantly reduce time and effort.

Modularity of the tool we developed allows to use any input/output compatible CVA pipeline and machine learning model, while keeping the same GUI. The initial model can be retrained as the amount of labeled data increase.

Using the same prediction model and tool for discarding inattention periods may facilitate multi-center studies by unifying the data preprocessing pipeline. Another way to facilitate multi-center studies is to perform preprocessing and labeling of the data in each center separately, and then share only the CVA features and annotations for training of the model with larger amounts of data. Such approach helps to preserve privacy of the data in each center, allowing to share only specific de-identified CVA features.

A limitation of the study is the absence of a published model and our original full preprocessing pipeline. The reason for it is the removal of the *intraface* library (Torre et al., 2015) from the public access. We provide the code for an alternative preprocessing pipeline predicting the same features based on the publicly available OpenFace library, and the model structure and the interface that needs to be implemented for it to be fully integrated into the GUI.
A potential future direction is working with missing data. CVA could not detect the face in 5.38% of the frames in our dataset, likely due to either extreme angles of the head with respect to the camera or because of face occlusions. Future studies may attempt to associate these periods with attention/inattention to the screen by using imputation/interpolation methods.

**Conclusion**

We presented a low-cost scalable approach to inattention detection during EEG recordings using computer vision analysis, and made publicly available a tool for visualization, model fine-tuning, and post-processing of the system’s results. We also made publicly available an example of computer vision analysis pipeline which can be used in future studies. We showed that fine-tuning the model on small amounts of new data by labeling the data on a per-frame basis substantially increases the model performance. Our work demonstrates that computer vision analysis is a feasible way for detecting inattention in EEG studies.
Chapter 8. Conclusion

In this dissertation, we focused on methods to create digital biomarkers and outcome measures. Following this goal, we used several methodologies for extracting interpretable measures of behavior and brain activity from deep learning models, either by utilizing a model pre-trained on a publicly available dataset or training the model from scratch on a clinical dataset. This approach showed its effectiveness in the behavioral analysis of caregiver-child interaction (chapter 3), assessment of cerebellar ataxia severity and progression (chapter 4), and neonatal seizure detection (chapter 5). We proposed a novel behavioral metric for attentional preference in the EEG setting (relative average look duration) and showed the value of the combined analysis of EEG and behavior by demonstrating different associations of EEG signal power and simultaneously captured visual attention in children with autism and neurotypical children (chapter 6). The value of this work on combined analysis of EEG and behavioral data inspired the development of the machine learning-based approach to automatic detection of visual attention periods on video recordings synchronized with EEG. This approach would scale up EEG and behavior studies by accelerating the visual attention annotation in EEG experiments (chapter 7).

Besides promising results, in the end, we want to discuss several crucial challenges that we have faced during this work. Candidate metrics for biomarkers from
pre-trained neural networks required an extensive quality check due to the dataset shift between the model training set and the clinical dataset to which the model is applied. In chapters 3 and 4 we exploited the output of the deep learning models pre-trained on publicly available non-clinical datasets and showed that they can provide interpretable clinical measures of behavior. The intuition behind the proposed measures of caregiver responsiveness (chapter 3) and Average Vowel Entropy (chapter 4) was to extract the most robust features under the shift of the distribution between the training dataset and the clinical dataset. In both cases, we analyzed the quality of the extracted derivative metrics.

In the case of the caregiver responsiveness, two humans annotated the extracted ‘reaching to the toy’ events on a subset of the data and the specificity of the detections was computed. Clusters with different caregiver responsiveness levels were revealed, and the stability of the clusters was assessed. The revealed clusters were stable and associated with the clinical and caregiver-reported measures. However, the specificity of ‘reaching to the toy’ detections was moderate, implying the limitation for the interpretation of the responsiveness measure. In the case of Average Vowel Entropy as an ataxia biomarker, high specificity was guaranteed by the design of the metric (fundamental frequency (F0) was independently extracted by another software and only the segments that were predicted as vowels and had valid F0 across the entire segment
were retained). The test-retest reliability of the metric was high, securing further interpretations of this metric. These two examples demonstrate the importance of quality checking when building behavioral biomarkers from models pre-trained on generally available non-clinical datasets.

Generalizability is another quality measure when building biomarkers for clinical purposes directly from machine learning models outputs and training models from scratch. In chapter 5, we proposed a method for seizure detection and showed that it has a good performance on a separate dataset with different amounts of EEG channels, indicating the potential to generalize to another EEG facility, enable multi-center neonatal EEG studies, and build biomarkers of seizure burden in neonates.

Our work on seizures and inattention detection (chapters 5 and 7) faced another problem of the clinical data – disagreement between human raters on the events in question (in our case seizures and inattention periods). Disagreement can come both from missing the events or from disagreeing on the beginning and the end of the event in time-series data. In both works, the designed algorithms had a lower agreement with individual raters than with consensus between raters, pointing to a more objective detection of the events by the algorithm. In future studies, additional measures can be evaluated such as event-based precision and recall measures.
Another essential problem for the implementation of machine learning methods in clinical and research environments is the selection of an operational threshold for the best performance. In our work, we used two different approaches to address this problem. One is selecting an operational threshold based on the best median agreement between an algorithm and a human annotator on a training dataset. Another approach is described in chapter 7, where the user is allowed to dynamically select the threshold and visually explore the results together with raw signal features.

Finally, small sample sizes and limited diversity of the samples need to be addressed in future studies on biomarkers, which is relevant to all studies presented in this work. Further advancements in standardizing the data processing, and reducing the burden of human labeling, including those suggested in this dissertation, can help to increase sample sizes and facilitate multi-center collaboration.

Overall, our work has shown the feasibility of building digital behavioral and EEG biomarkers using pre-trained deep learning models and outlined the future directions for combining EEG and behavior. Currently, in a new lab-based experiment, we are working on the concurrent recording of EEG and behavior from the tablet device. In the future, this work can be expanded beyond the clinic. The advancement of consumer-grade EEG may at some point make it possible to concurrently record EEG and behavior at home. Caregiver-child interactions and any speech data may be
recorded at home using smartphones. This may greatly expand the size of the samples and the diversity of participants, making the resulting digital biomarkers more robust and reliable. Eventually, this progress will lead to remote diagnostics, risk assessment, and monitoring of the treatment progress, thus significantly reducing the healthcare costs, mitigating the lack of specialists, and making healthcare more accessible to the broader population.
Appendix A. Additional materials to chapter 3.

Bias of CVA methods based on participant race

To determine whether CVA methods are biased based on participant race, we analyzed whether percentage of dropped frames is associated with race or the interaction of race:person\textsuperscript{56}. Since the frames are dropped based on the confidence of the OpenPose landmark detection algorithm, it can serve as a proxy for performance quality of the algorithm for different races. We performed two-way ANOVA with race and person as factors. Race factor and interaction of race and person were not significant when all races were taken into the analysis (for race factor $F(4,146)=0.48$, $p=0.75$, for race:person interaction factor $F(4,146)=1.126$, $p=0.35$), nor when only comparing White/Caucasian and Black/African American subgroups (for race factor $F(1,122)=0.14$, $p=0.71$, for race:person interaction factor $F(1,122)=1.00$, $p=0.32$).

Association of amount of dropped frames with diagnostic group

Mean percentage of dropped frames across diagnostic groups per child was 8.1, 2.8, 5.1, 5.8; per caregiver: 7.6, 13.1, 10.5, 12.1 for Control, ADHD, ASD, ASD+ADHD respectively.

\textsuperscript{56} Caregiver or child
To assess whether the difference in dropped frames between the diagnostic groups was significant, we performed two-way ANOVA with group and person (caregiver/child) as factors. While the main effect of the person factor was significant (F(1,146)=8.00, p = 0.005), neither group, nor interaction between group and person factors were significant (F(3,146)=0.054, p=0.98 for group, F(3,146)=0.557, p=0.64 for group:person).

**Association of amount of continuous reaching events (CREs) with demographic variables**

We additionally analyzed whether the amount of detected continuous reaching events was associated with demographic variables, including age, sex, maternal education level, and race and ethnicity of the child. We performed two-way ANOVA with demographic variable and person as factors. Amount of CREs was associated with child’s sex, with boys exhibiting more CREs than girls (both together with caregivers; F(1,152)=10.23, p = 0.002). The amount of CREs was also associated with maternal education level, exhibiting difference in mean number of CREs across different levels of education (F(5,144)=2.470, p=0.04), however there was no steady trend of increase or decrease of amount of CREs as maternal educational level increases. Amount of CREs was not associated with age, race, and ethnicity of the child. Interaction between person and demographic variable was not significant in all models.
Stability of clusters discovered via Mixed Markov Model fitting

To assess the reliability of the two clusters revealed by fitting a Mixed Markov Model, we performed an alternative clustering via a model-based distance method (Ghassempour et al., 2014) and a resampling-based reliability analysis (Levine & Domany, 2001). In the first approach we followed the algorithm described in (Ghassempour et al., 2014). We fitted a Markov Chain with four states (No RE, Child RE, Caregiver RE, Both RE) to each of the 78 time series. Then likelihoods of each time series under each fitted model were computed, providing a likelihood matrix of size 78x78. Distances were computed via the symmetric Kullback-Leibler divergence,

\[ D_{KL}(i,j) = \sum_{t=1}^{78} P(T_t|\lambda_i) \log \frac{P(T_t|\lambda_i)}{P(T_t|\lambda_j)} \]

and

\[ Distance(i,j) = \frac{1}{2} (D_{KL}(i,j) + D_{KL}(j,i)) \]

where \( T_k \) is \( k \)-th time-series and \( \lambda_k \) is \( k \)-th Markov Model parameters (transition matrix). Then based on the distance matrix, Partitioning Around Medoids method was used to cluster the data into two clusters.

Seventy out of seventy-eight subjects were clustered identically in the original method (based on Mixed Markov Models) and the method described above. Silhouette
score (Rousseeuw, 1987) of the method above was 0.56, and silhouette score of the original clustering computed on the distances from the method above was 0.46, providing evidence of clusters existence.

Another independent assessment of cluster reliability was performed based on subsampling 60 out of 78 RE time series 100 times. At each subsample, we computed average agreement of cluster labels of each possible pair of time series with cluster labels of the same pair in original clustering, and then averaged agreement scores across all 100 subsamples. Such an assessment provides a value in the range between 0 and 1, where 1 indicates perfect agreement. Average agreement score was 0.89 (SD=0.17, median=0.98), which indicates high cluster stability.
Table 19: Pairwise associations between all study variables. Chi-squared statistics (X) are provided for pairs of categorical variables, F-statistics (F) from one-way ANOVA for pairs of categorical – continuous variables, Spearman’s rank correlation (rho) for pairs of ordered categorical - continuous variables, and Pearson’s correlation (R) for continuous - continuous variable pairs. VCom – Vineland Communication Score, VSoc – Vineland Socialization Score, VIQ – Verbal IQ.

<table>
<thead>
<tr>
<th>Age</th>
<th>Sex</th>
<th>Ethnicity</th>
<th>Race</th>
<th>Maternal education level</th>
<th>VCom</th>
<th>VSoc</th>
<th>Verbal IQ</th>
<th>Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F(3,74) = 4.34 p = 0.007</td>
<td>X(3, N=78) = 3.02 p = 0.24</td>
<td>X(12, N=78) = 12.80 p = 0.022</td>
<td>X(15, N=78) = 15.85 p = 0.039</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F(1,76) = 2.10 p = 0.15</td>
<td>F(1,76) = 4.44 p = 0.038</td>
<td>F(4,73) = 2.02 p = 0.1</td>
<td>rho = 0.02 p = 0.85</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X(1,N=78) = 0.14 p = 0.7</td>
<td>X(4,N=78) = 5.23 p = 0.26</td>
<td>X(5,N=78) = 4.85 p = 0.43</td>
<td>F(1,76) = 3.12 p = 0.081</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X(4,N=78) = 1.97 p = 0.74</td>
<td>X(5,N=78) = 13.73 p = 0.017</td>
<td>F(1,76) = 0.53 p = 0.47</td>
<td>F(1,76) = 1.04 p = 0.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X(20,N=78) = 24.4 p = 0.23</td>
<td>X(4,73) = 2.084 p = 0.092</td>
<td>F(4,73) = 1.451 p = 0.226</td>
<td>F(4,73) = 3.24 p = 0.017</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>rho = 0.526 p &lt; 0.001</td>
<td>rho = 0.350 p = 0.002</td>
<td>rho = 0.419 p &lt; 0.001</td>
<td>X(5,N=78) = 6.73 p = 0.240</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R = 0.76 p &lt; 0.001</td>
<td>R = 0.80 p &lt; 0.001</td>
<td>F(1,76) = 9.52 p &lt; 0.003</td>
<td>VCom</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R = 0.65 p &lt; 0.001</td>
<td>F(1,76) = 5.86 p = 0.018</td>
<td>VSoc</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F(1,76) = 10.53 p = 0.002</td>
<td>Verbal IQ</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 20: Model parameters (degrees of freedom, required effect size) of Models M1-M10.

<table>
<thead>
<tr>
<th>Model</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
<th>M5</th>
<th>M6</th>
<th>M7</th>
<th>M8</th>
<th>M9</th>
<th>M10</th>
</tr>
</thead>
<tbody>
<tr>
<td>w.-g. df</td>
<td>76</td>
<td>74</td>
<td>73</td>
<td>70</td>
<td>76</td>
<td>75</td>
<td>73</td>
<td>72</td>
<td>71</td>
<td>69</td>
</tr>
<tr>
<td>Age</td>
<td>b.-g. df</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>f</td>
<td>0.32</td>
<td>-</td>
<td>0.32</td>
<td>0.32</td>
<td>-</td>
<td>0.32</td>
<td>-</td>
<td>0.32</td>
<td>0.32</td>
</tr>
<tr>
<td>Group</td>
<td>b.-g. df</td>
<td>-</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>f</td>
<td>0.38</td>
<td>0.38</td>
<td>0.38</td>
<td>-</td>
<td>-</td>
<td>0.38</td>
<td>0.38</td>
<td>0.38</td>
<td>0.38</td>
</tr>
<tr>
<td>Age*Group</td>
<td>b.-g. df</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>f</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.38</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cluster</td>
<td>b.-g. df</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>f</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.32</td>
<td>0.32</td>
<td>0.32</td>
<td>0.32</td>
<td>0.32</td>
</tr>
<tr>
<td>Age*Cluster</td>
<td>b.-g. df</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>f</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.32</td>
</tr>
<tr>
<td>Group*Cluster</td>
<td>b.-g. df</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>f</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.38</td>
</tr>
</tbody>
</table>

w.-g. df – within-group degrees of freedom
b.-g. df – between-group degrees of freedom
f – Cohen’s f effect size
Table 21: Statistics of ANOVA models M1-M5.

<table>
<thead>
<tr>
<th>Model</th>
<th>M1: ~ Age</th>
<th>M2: ~ Group</th>
<th>M3: ~ Age + Group</th>
<th>M4: ~ Age + Group + Age:Group</th>
<th>M5: ~ Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>VCom</td>
<td>VSoC</td>
<td>VIQ</td>
<td>VCom</td>
</tr>
<tr>
<td>Age</td>
<td>F</td>
<td>8.16</td>
<td>4.77</td>
<td>8.41</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Pr(F)</td>
<td>0.006</td>
<td>0.032</td>
<td>0.005</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>η²</td>
<td>0.1</td>
<td>0.06</td>
<td>0.1</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Cohen's f</td>
<td>0.33</td>
<td>0.25</td>
<td>0.33</td>
<td>-</td>
</tr>
<tr>
<td>Group</td>
<td>F</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>13.5</td>
</tr>
<tr>
<td></td>
<td>Pr(F)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>η²</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>Cohen's f</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.74</td>
</tr>
<tr>
<td>Age*Group</td>
<td>F</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Pr(F)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>η²</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Cohen's f</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cluster</td>
<td>F</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Pr(F)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>η²</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Cohen's f</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Age*Cluster</td>
<td>F</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Pr(F)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>η²</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 22: Statistics of ANOVA models M6-M10.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent variable</td>
<td>VCom</td>
<td>VSoC</td>
<td>VIQ</td>
<td>VCom</td>
<td>VSoC</td>
</tr>
<tr>
<td>Age</td>
<td>F</td>
<td>9.42</td>
<td>5.11</td>
<td>9.63</td>
<td>-</td>
</tr>
<tr>
<td>Pr(&gt;F)</td>
<td>0.003</td>
<td>0.027</td>
<td>0.002</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Eta-squared</td>
<td>0.11</td>
<td>0.06</td>
<td>0.11</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cohen's f</td>
<td>0.35</td>
<td>0.26</td>
<td>0.36</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Pr(&gt;F)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Eta-squared</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.35</td>
<td>0.34</td>
</tr>
<tr>
<td>Cohen's f</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.74</td>
<td>0.73</td>
</tr>
<tr>
<td>Age*Group</td>
<td>F</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Pr(&gt;F)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Eta-squared</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Variable</td>
<td>Cluster</td>
<td>F</td>
<td>Pr(&gt;F)</td>
<td>Eta-squared</td>
<td>Cohen's f</td>
</tr>
<tr>
<td>----------</td>
<td>---------</td>
<td>-------</td>
<td>--------</td>
<td>-------------</td>
<td>-----------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>&lt;0.001</td>
<td>0.004 0.022</td>
<td>0.002 0.016</td>
</tr>
<tr>
<td></td>
<td>Age*Cluster</td>
<td>F</td>
<td>Pr(&gt;F)</td>
<td>Eta-squared</td>
<td>Cohen's f</td>
</tr>
<tr>
<td></td>
<td>Group*Cluster</td>
<td>F</td>
<td>Pr(&gt;F)</td>
<td>Eta-squared</td>
<td>Cohen's f</td>
</tr>
<tr>
<td></td>
<td>Adjusted $R^2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

178
Appendix B. Additional materials to chapter 5.

Post-processing: probability reweighting

Our derivations for adjustments of the classifier output probability given the prevalence of positive class follow the lines of (Elkan, 2001; Pozzolo et al., 2015). We can model our sampling strategy as follows: Let $s$ be a Bernoulli variable defining whether an epoch is taken into the training sample or not, $y$ a label taking two values (1 for seizure, 0 for non-seizure), $X$ an epoch. Then

$$(s|y=i) \sim \text{Bernoulli}(\beta_i), i = 0,1$$

Also, let $p(y = i) = \pi_i$.

$p(s|X,y) = p(s|y)$ since only label is important to make a decision whether an epoch is taken into the training subset for class balanced training. The strategy where we take equal amount of seizures and non-seizures into the training sample can be defined as

$$\beta_0 \pi_0 = \beta_1 \pi_1$$

or

$$\frac{\beta_0}{\beta_1} = \frac{\pi_1}{\pi_0} = \beta$$

If $p(y = 1|x, s = 1)$ is the output of a classifier trained on the balanced set, then by Bayes Theorem we can write the following:
where \( p(y = 1|x) \) is the probability of seizure in the original unbalanced model, which we are looking for. Let us denote \( p = p(y = 1|X) \), \( p_s = p(y = 1|X, s = 1) \), so we can rewrite Eq. (B1) as

\[
    p_s = \frac{\beta_1 p}{\beta_1 p + \beta_0 (1 - p)} = \frac{p}{p + \beta(1 - p)}
\]

Rewriting the formula we get

\[
    \frac{p}{p + \beta(1 - p)}
\]

The last formula gives us an adjustment of probability that should be done on the output of our algorithm.

For later post-processing steps, we need to choose the threshold. According to Bayes decision theory, if we deem classification cost of correct examples as 0, the threshold

\[
    \tau = \frac{l_{1,0}}{l_{1,0} + l_{0,1}}
\]

where \( l_{i,j} \) is the cost incurred in deciding \( i \) when the correct class is \( j \). Without prior knowledge, we select \( l_{1,0} = \pi_1 \) and \( l_{0,1} = \pi_0 \) as the threshold (Pozzolo et al., 2015), so the threshold becomes \( \tau = \pi_1 \). This corresponds to an operating point of 0.5 of balanced classifier.
Post-processing: transforming the outputs to improve robustness

Post-processing was done following (Tapani et al., 2019) and took adjusted output probability per epoch as an input. Since annotations were done by human rater on a per-second basis, post-processing had an upsampling step (converting per-epoch probabilities to per-second probabilities). Since epochs were 8-seconds long with 4-seconds overlap, each epoch prediction was transformed into 4 seconds prediction in upsampling.

The post-processing steps were as follows: a) median filtering of 3 consecutive epochs prediction probabilities; b) upsampling per-epoch predictions back to per-second resolution; c) removing all predictions labeled as ‘seizure’ which were less than 10 seconds long; d) “collaring” (extending each seizure prediction by 8 seconds in both directions). When computing AUC, steps a) and b) were performed before applying the decision threshold and making a binary 0/1 decision, and steps c) and d) were performed each time after applying the decision threshold.

Significance tests of difference in performance of DL models

To assess the difference in performance, as measured by AUC on leave-one patient out cross-validation, we applied Wilcoxon paired signed-rank tests each pair of models on each balancing approach. Results are shown in Table 23.
Table 23: P-values of Wilcoxon signed-rank tests between leave-one-patient out AUCs on each level of balancing of each deep learning model.

<table>
<thead>
<tr>
<th>Model (balancing)</th>
<th>DL1 (None)</th>
<th>DL1 (Class)</th>
<th>DL1 (Patient-Class)</th>
<th>DL2 (None)</th>
<th>DL2 (Class)</th>
<th>DL2 (Patient-Class)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DL1 (None)</td>
<td>-</td>
<td>0.07454</td>
<td>0.00816</td>
<td>0.17014</td>
<td>0.01025</td>
<td>0.18919</td>
</tr>
<tr>
<td>DL1 (Class)</td>
<td>-</td>
<td>-</td>
<td>0.13132</td>
<td>0.86000</td>
<td>0.00070</td>
<td>0.00209</td>
</tr>
<tr>
<td>DL1 (Patient-Class)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.58321</td>
<td>0.00002</td>
<td>0.00017</td>
</tr>
<tr>
<td>DL2 (None)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.00004</td>
<td>0.00036</td>
</tr>
<tr>
<td>DL2 (Class)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.00449</td>
</tr>
<tr>
<td>DL2 (Patient-Class)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

182
References


abnormalities via mobile phone video and machine learning. *Scientific Reports, 10*(1), 18641. https://doi.org/10.1038/s41598-020-75661-x


189

*Scientific Reports, 8*(1), 1-7. 
https://doi.org/10.1038/s41598-018-35215-8

https://doi.org/10.1016/j.jaac.2012.08.018

https://doi.org/10.1542/peds.2009-0958

https://doi.org/10.1037/0012-1649.40.2.271

https://doi.org/10.1207/s15326942dn2703_6


https://go.exlibris.link/Q0rzMymD


Hanslmayr, S., & Matuschek, J. (2014). Entrainment of Prefrontal Beta Oscillations Induces an Endogenous Echo and Impairs Memory Formation. *Current Biology, 24*(8), 904-909. https://doi.org/10.1016/j.cub.2014.03.007


196


206


215


Torre, F. D. I., Chu, W. S., Xiong, X., Vicente, F., Ding, X., & Cohn, J. (2015, 4-8 May 2015). IntraFace. 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG),


217


https://doi.org/10.1016/j.neucom.2018.05.083


https://doi.org/10.1016/j.patcog.2017.08.026


Studies: The Autism Biomarkers Consortium for Clinical Trials [Methods].


on Empirical Methods in Natural Language Processing: System Demonstrations Online.


Zhang, J., Xing, Y., Ma, X., & Feng, L. (2017). Differential Diagnosis of Parkinson Disease, Essential Tremor, and Enhanced Physiological Tremor with the Tremor Analysis
of EMG. *Parkinson’s Disease, 2017, 1597907.*
https://doi.org/10.1155/2017/1597907

https://doi.org/10.1109/TKDE.2021.3070203


https://doi.org/10.1121/10.0008579

https://doi.org/https://doi.org/10.1016/B978-0-12-801608-4.00001-3
Biography

Dmitry Isaev was born in November, 1984 in Ulianovsk, USSR. He attended Moscow State University from 2001 to 2006 and graduated with a specialist degree (analog of bachelor’s and master’s) in Applied Math and Informatics. After working for 6 years as a software developer, project manager, and business analyst, he returned to academia and attended the Computational Linguistics program at the Higher School of Economics, Moscow. In 2014 he received a master’s degree in Computational Linguistics and continued working in the Higher School of Economics at the Center for Language and Brain on projects related to brain neuroimaging and electrophysiology in language disorders. In 2015 Dmitry joined Imaging Genetics Center of the University of Southern California to work on brain MRI imaging projects. In 2017 he joined the lab of Guillermo Sapiro at Duke University to pursue the Ph.D. degree in Biomedical Engineering. Dmitry’s research interests lie in applying machine learning methods to various health data, specifically in the areas of computer vision, medical imaging, and electrophysiology.

Dmitry is an amateur piano and harmonica player. Together with his wife, Roza, they had many happy hours attending classical music concerts at Duke’s Baldwin Auditorium.