UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS DA COMPUTAÇÃO

Guilherme Ocker Ribeiro

Stimming behavior detection for Autism Spectrum Disorder support

Florianópolis
2023

Guilherme Ocker Ribeiro

Stimming behavior detection for Autism Spectrum Disorder support

Dissertação a ser submetida ao Programa de Pós-Graduação em Ciências da Computação da Universidade Federal de Santa Catarina para a obtenção do título de mestre em Ciências da Computação.
Supervisor: Prof. Mateus Grellert, Dr.
Co-supervisor: Prof. Jônata Tyska Carvalho, Dr.

Florianópolis

2023

Ficha catalográfica gerada por meio de sistema automatizado gerenciado pela BU/UFSC.
Dados inseridos pelo próprio autor.

Ribeiro, Guilherme Ocker
Stimming behavior detection for Autism Spectrum Disorder support / Guilherme Ocker Ribeiro ; orientador, Mateus Grellert, coorientador, Jônata Tyska Carvalho, 2023.
90 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Ciência da Computação, Florianópolis, 2023.

Inclui referências.

1. Ciência da Computação. 2. machine learning. 3. computer vision. 4. autism spectrum disorder. 5. video event detection. I. Grellert, Mateus. II. Carvalho, Jônata Tyska. III. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Ciência da Computação. IV. Título.

Guilherme Ocker Ribeiro

Stimming behavior detection for Autism Spectrum Disorder support

O presente trabalho em nível de mestrado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Profa. Carina Friedrich Dorneles, Dra.
Instituição INE/CTC/UFSC

Prof. Claudio Machado Diniz, Dr.
Instituição UFRGS

Prof. Jônata Tyska Carvalho, Dr.
Instituição INE/CTC/UFSC

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de mestre em Ciências da Computação.

ICP-Edu
Documento assinado digitalmente
Marcio Bastos Castro
Data: 16/02/2024 09:11:40-0300
CPF: ***.707.720-**
Verifique as assinaturas em https://v.ufsc.br

Coordenação do Programa de
Pós-Graduação

GOV.BR
Documento assinado digitalmente
MATEUS GRELLERT DA SILVA
Data: 16/02/2024 00:18:45-0300
CPF: ***.664.940-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Mateus Grellert, Dr.
Supervisor:

Florianópolis, 2023.

AGRADECIMENTOS

Gostaria de agradecer, inicialmente o meu orientador, Prof. Mateus Grellert, pela orientação dedicada, paciência e expertise compartilhada ao longo deste percurso. Seu apoio nos percalços durante essa trajetória foram essenciais para atingir esse objetivo, sou imensamente grato por sua mentoria. Também estendo esses agradecimentos ao meu co-orientador Jônata Tyska Carvalho. Suas críticas construtivas, conselhos e incentivos foram muito importantes.

Aos colegas de pesquisa, principalmente ao Alexandre Soli, uma parte importante desse trabalho só foi possível com o teu apoio, muito obrigado. Juntos, construímos um ambiente colaborativo e inspirador.

À minha família e amigos, agradeço pelo constante apoio emocional, compreensão e principalmente incentivo ao longo desta trajetória. À Weduka pelo apoio e compreensão na busca deste sonho.

À minha esposa Juliane, pelo seu apoio incondicional durante todo esse tempo, permitindo foco nos momentos bons e perseverança nos momentos dificeis, seria impossivel sem estares ao meu lado.

Obrigado a todos que, de alguma forma, foram parte desta jornada.

“I want an AI-powered society because I see so many ways that AI can make human life better. We can make so many decisions more systematically or automate away repetitive tasks and save so much human time.”
(Andrew Ng, 2018)

RESUMO

Métodos de visão computacional não intrusivos que podem reconhecer comportamentos humanos complexos podem auxiliar no diagnóstico e tratamento de distúrbios neurológicos onde comportamentos repetitivos (stimming) são proeminentes, como o Transtorno do Espectro Autista (TEA). Métodos de aprendizado de máquina, especialmente aqueles relacionados à visão computacional, apresentam uma direção promissora para esse tipo de aplicação, no entanto, a eficácia desses métodos depende da existência de conjuntos de dados de boa qualidade. A construção de um conjunto de dados nesse domínio é desafiadora devido a questões de privacidade e à necessidade de conduzir experimentos principalmente com crianças. Portanto, propomos um conjunto de dados consolidado, anotado e padronizado chamado de Conjunto de Dados de Comportamento de Estimulação no Autismo (ASBD). Este conjunto de dados unificado pode permitir que pesquisadores tenham acesso a uma grande quantidade de dados, classificados e anotados para os comportamentos de estimulação, possibilitando uma abordagem mais rápida e eficiente ao trabalhar com esse tipo de coleção de vídeos. O Conjunto de Dados de Comportamento de Estimulação no Autismo (ASBD) está disponível abertamente e pode ser usado para o desenvolvimento de algoritmos de aprendizado de máquina que podem detectar padrões comportamentais associados ao transtorno do espectro autista e outras aplicações em que esse tipo de análise é importante. O conjunto de dados é composto por 165 clipes curtos anotados, com base em trechos de 155 vídeos distintos publicamente acessíveis no Youtube, com 154 sujeitos únicos de idades variando de crianças a jovens adultos, divididos em 48 mulheres e 106 homens. A duração média dos clipes é de 10,6 segundos. Os clipes anotados mostram não apenas o URL e a classe do vídeo, mas também o momento em que o comportamento relevante começa e sua duração. Exploramos como este conjunto de dados pode viabilizar um modelo de reconhecimento de ação de última geração na detecção de padrões de comportamento de estimulação em uma série de testes com subconjuntos do ASBD e técnicas de aprimoramento de dados, alcançando uma precisão de 76% no conjunto de dados de teste com o modelo ajustado finamente(*fine-tuning*) no ASBD aprimorado, o que representa uma melhoria de 34% e 19% em comparação com dois conjuntos de dados anteriores (SSBD e ESBD, respectivamente). Também demonstramos como este modelo treinado pode auxiliar uma arquitetura maior de detecção de eventos em sessões de terapia para crianças com TEA a obter melhores resultados globalmente, automatizando a detecção de ações de comportamento de estimulação, o que é implementado como uma aplicação web. Esta contribuição não apenas aprimora a compreensão dos comportamentos relacionados ao TEA, mas também fornece uma base para pesquisas e aplicações adicionais, especialmente na automação da detecção de comportamento de estimulação durante as sessões de terapia, auxiliando em uma abordagem mais eficiente para analisar seus resultados.

Keywords: aprendizado de máquina. visão computacional. transtorno do espectro autista. video. detecção de eventos.

ABSTRACT

Non-intrusive vision-assisted methods that can recognize complex human behaviors can help the diagnosis and treatment of neurological disorders where stimming behaviors are prominent, such as Autism Spectrum Disorder (ASD). Machine learning methods, especially those related to computer vision are a promising direction for this kind of application, however, the effectiveness of these methods depends on the existence of good-quality datasets. Building a dataset in this domain is challenging due to privacy matters and also to the need to conduct experiments mostly with children. Therefore, we propose a consolidated, annotated, and standardized Stimming Behavior Dataset. This unified dataset can enable researchers to have access to a larger pool of data, classified and annotated for the stimming behaviors, enabling a faster and more streamlined approach when working with this kind of video collection. The Autism Stimming Behavior Dataset (ASBD) is openly available and can be used for the development of machine learning algorithms that can detect behavioral patterns associated with autism spectrum disorder and other applications where this kind of analysis is important. The dataset comprises 165 annotated short-duration clips, based on excerpts from 155 distinct publicly accessible Youtube videos, with 154 unique subjects of ages spanning from toddlers to young adults and divided into 48 female and 106 male individuals. The average duration of the clips is 10.6 seconds. The annotated clips show not only the URL and class of the video but also the moment where the relevant behavior starts and its duration. We explore how this dataset can enable a state-of-the-art action recognition model in detecting the stimming behavior patterns in a series of tests with the subsets of the ASBD and data augmentation techniques, achieving 76% accuracy on the test dataset with the model fine-tuned on the augmented ASBD, which represents an improvement of 34% and 19% compared with two previous data sets (SSBD and ESBD, respectively). We also show how this trained model can help a larger architecture of event detection in therapy sessions for children with ASD achieve better results overall by automatizing the detection of stimming behavior actions, which is implemented as a web application. This contribution not only enhances the understanding of ASD-related behaviors but also provides a foundation for further research and applications, notably in automating stimming behavior detection during therapy sessions, helping a more streamlined approach to analyzing its outcomes.

Keywords: machine learning. computer vision. autism spectrum disorder. video. event detection.

RESUMO EXPANDIDO

Introdução

O Transtorno do Espectro Autista, ou simplesmente TEA, é considerado um transtorno do desenvolvimento complexo que influencia a capacidade de comunicar e aprender. Indivíduos com TEA costumam afetar não apenas a pessoa dentro do espectro, mas também sua família, amigos, sua comunidade e o sistema de saúde. Exigindo apoio e atenção mais intensos em comparação com indivíduos típicos, uma vez que podem ter sua capacidade cognitiva prejudicada. Familias que são compostas por pessoas dentro do espectro podem ter impactos na sua saúde mental e emocional, juntamente impactos negativos na sua vida financeira (ZEIDAN et al., 2022; KOŁAKOWSKA et al., 2017). Métodos de visão computacional não intrusivos que podem reconhecer comportamentos humanos complexos podem auxiliar no diagnóstico e tratamento de distúrbios neurológicos onde comportamentos repetitivos (stimming) são proeminentes como no TEA. Métodos de aprendizado de máquina, especialmente aqueles relacionados à visão computacional, apresentam uma direção promissora para esse tipo de aplicação, no entanto, a eficácia desses métodos depende fortemente da existência de bases de dados de boa qualidade. A construção de bases de dados nesse domínio é desafiadora por estar ligada a questões de privacidade e à necessidade de conduzir experimentos principalmente com crianças onde o TEA é mais comumente diagnosticado.

Objetivos

Propor uma base de dados unificada que possa servir de suporte para modelos de aprendizado de máquina. Além disso, buscamos propor uma arquitetura para um sistema baseado em inteligência artificial e técnicas de aprendizado de máquina que seja capaz de otimizar a análise realizada em vídeos de diagnóstico e terapias do TEA, buscando automatizar a detecção de interações relevantes entre os participantes, classificando-as e destacando esses momentos, reduzindo o tempo necessário para examinar manualmente o vídeo completo de cada sessão.

Metodologia

Foi conduzido um mapeamento sistemático no campo da detecção de eventos em vídeos gravados de terapias de indivíduos com TEA, fornecendo uma visão abrangente das técnicas e estruturas de aprendizado de máquina/visão computacional aplicadas nessa área. Percebendo a ausência de uma fonte consolidada, padronizada e devidamente anotada que pudesse ser usada para viabilizar novos modelos nesse campo, o banco de dados ASBD foi proposto.

Com a consolidação de três conjuntos de dados - SSBD (RAJAGOPALAN et al., 2013), ESBD (NEGIN et al., 2021), SSBD-Expandido (WEI et al., 2022) -, o ASBD também apresenta um trabalho de curadoria ao analisar cuidadosamente e anotar cada clipe. Essas anotações mostram não apenas o link do vídeo e a classe, mas também o instante de tempo relevante do comportamento, onde definimos o início e a duração do principal evento de *stimming*. Isso permite um acesso mais rápido e eficiente aos eventos relevantes, além de fornecer uma fonte padronizada (mesmos clipes, início e duração do evento e classe), na qual podem ser estabelecidas referências e com-

Table 1 – Consolidação dos resultados de validação e testes - Modelo VideoMAE *aprimorado* nos componentes do ASBD

Fine-tuning dataset	Validation acc.	Test acc. (ASBD dataset)
SSBD	0.32	0.35
SSBD Augm	0.60 (+0.28)	0.31 (-0.04)
ESBD	0.78	0.50
ESBD Augm	0.85 (+0.07)	0.56 (+0.06)
ASBD	0.68	**0.69**
ASBD Augm	0.68 ($\approx$0.0)	**0.76** (+0.07)

parações. O gênero do indivíduo e a unicidade também são especificados (gênero e unicidade são identificados pela descrição do vídeo, detalhes de áudio ou melhor suposição visual quando as duas primeiras não estão disponíveis).

Alem disso, atuando como outro componente em um sistema integrado maior, um modelo de última geração de reconhecimento de ações é treinado e avaliado neste conjunto de dados combinado com o objetivo de possibilitar a detecção de *stimming* no sistema, aprimorando sua eficiência na detecção de cenas relevantes de sessões de terapia gravadas para auxiliar profissionais em suas análises. Os clipes anotados mostram não apenas o URL e a classe do vídeo, mas também o momento em que o comportamento relevante começa e sua duração.

Resultados e Discussão

Foi realizado uma série de experimentos envolvendo o conjunto de dados unificado ASBD, resumidos na figura 1, onde as caixas cinzas representam o conjunto de dados proposto ABSD. Cada linha representa um experimento realizado, e nas colunas, especificamos as diferentes configurações, dados de treinamento, o modelo que está sendo treinado e os conjuntos de teste no qual seus resultados são comparados. Quatro experimentos são discutidos: no Experimento 1, selecionamos dois modelos (I3D e VideoMAE) pré-treinados no conjunto de dados *Kinectics-400* e os aplicamos diretamente ao conjunto de dados ASDB (sem *fine-tuning*); no Experimento 2, treinamos o modelo VideoMAE com conjuntos de dados existentes (SSBD e ESBD) e depois avaliamos seu desempenho em diferentes conjuntos de testes; o Experimento 3 envolve o treinamento do modelo VideoMAE com o conjunto de dados ASBD para avaliar se ele é capaz de melhorar a tarefa de detecção de *stimming*; finalmente, no Experimento 4, treinamos o VideoMAE com versões aprimoradas dos conjuntos de dados para melhorar ainda mais seu desempenho.

Os resultados consolidados do modelo VideoMAE pré-treinado no conjunto de dados Kinects-400 e ajustado finamente (*fine-tuning*) em cada subconjunto do ASBD e suas partes aprimoradas são apresentados na tabela 1. O melhor resultado alcançado, com uma precisão de 0,76%, com *fine-tuning* no conjunto de dados ASBD aprimorado, mostra um bom ponto de partida, onde modelos atualizados como o VideoMAE v2 (WANG, L. et al., 2023) e outros modelos inovadores que exploram aprendizado generativo e discriminativo podem melhorar ainda mais.

Considerações Finais

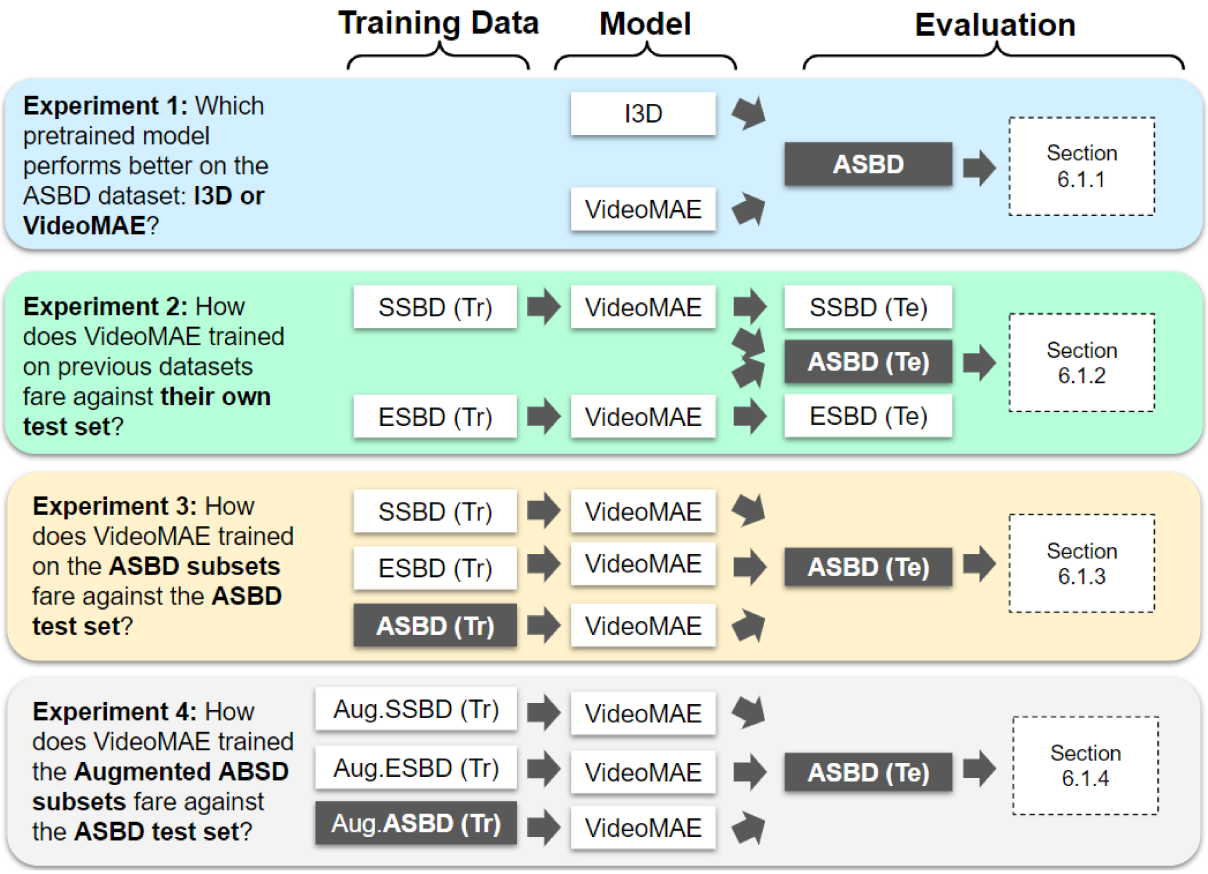


Figure 1 – Detail of the experiments

Source: the author

Neste trabalho, apresentamos uma base de dados de comportamentos repetitivos (*stimming*), consolidada, enriquecida e padronizada, que pode ser utilizada em pipelines de sistemas de visão computacional para apoiar o treinamento de modelos que visam auxiliar terapeutas e pais na análise do comportamento de crianças, em particular na identificação de comportamentos associados ao Transtorno do Espectro Autista (TEA). Buscamos com isso, permitir que modelos baseados em métodos de visão computacional e técnicas de aprendizado de máquina supervisionado alcançem resultados consideravelmente melhores quando utilizem um conjunto aprimorado e extenso nas suas fases de treino e teste do seu pipeline.

Avaliamos o desempenho de um modelo renomado de reconhecimento de vídeo, pré-treinado no conjunto de dados de ações humanas *Kinetics-400*, aplicando tecnicas de *fine-tunning* em nosso conjunto de dados e seus subconjuntos para entender como um conjunto de dados maior e curado se compara aos originais. Também analisamos como técnicas simples de aprimoramento (*augmentation techniques*) podem impactar no desempenho do modelo. Os modelos treinados no conjunto de dados ASBD e no conjunto aprimorado alcançaram os melhores resultados quando avaliados no conjunto de dados de teste fixado, com 69% e 76%, respectivamente.

Palavras-chave: machine learning. computer vision. autism spectrum disorder. video. event detection.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

Autism Spectrum Disorder or simply ASD is considered a complex developmental disorder that influences the ability to communicate and learn. ASD individuals tend to affect not only the person within the spectrum but also their family, friends, surrounding community, and the healthcare system. Demanding higher support and attention compared to typical individuals since they can have an impaired cognitive capacity, autistic households may have mental and emotional health impacted together with their financial lives (ZEIDAN et al., 2022; KOŁAKOWSKA et al., 2017).

Data from the Centers for Disease Control and Prevention (CDC)[1] shows that around 1 in 36 children are diagnosed within the spectrum in the United States, occurring in all racial, ethnic, and socioeconomic groups. It also shows that ASD is four times more prevalent among boys. Another recent study points out that ASD affects approximately 1/100 children worldwide (ZEIDAN et al., 2022). While still alarming, this reduced prevalence compared to the US study suggests the lack of proper diagnosis in some countries.

The rising number of cases, combined with the lifelong need for care and support that most individuals with ASD require in several aspects of their lives (healthcare, education, and social services), turns this disorder into a major societal concern. It can be associated with significant costs for the diagnosed individual, his/her family,

[1] https://www.cdc.gov/ncbddd/autism/data.html

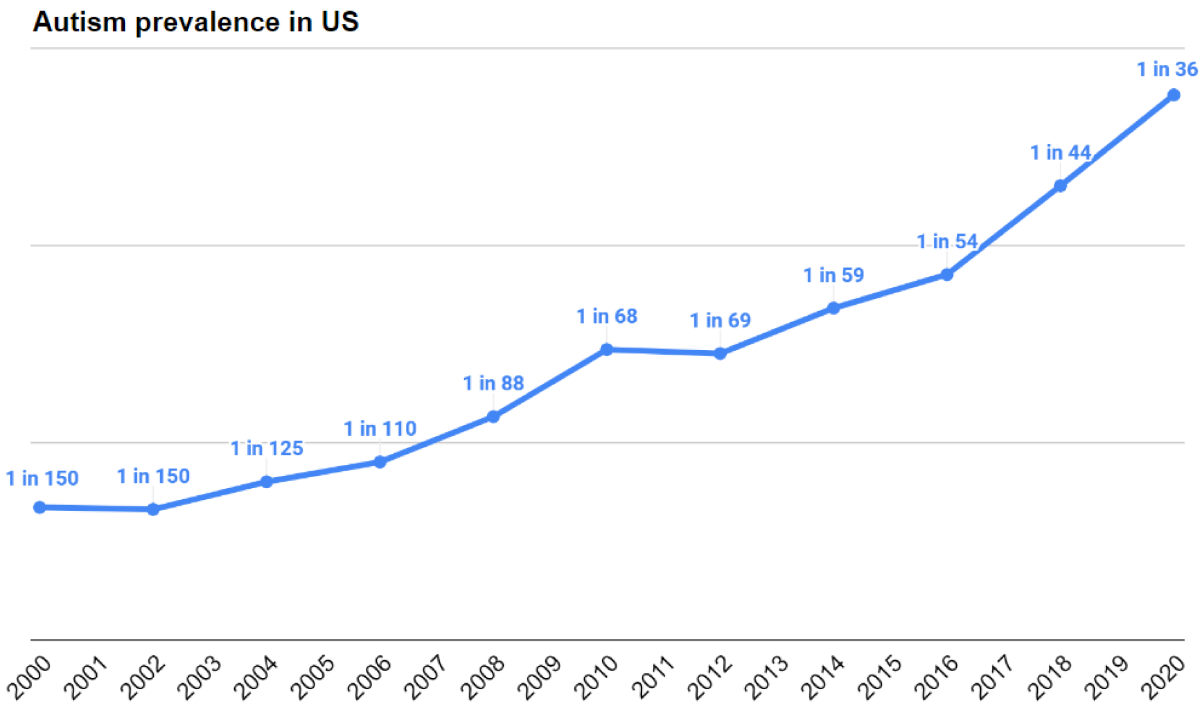


Figure 2 – ASD prevalence on the US over the years

Source: Adapted from (LOFTUS, 2023) - CDC/US

private and/or public health insurance systems, state financial aid programs, and more generally the whole society (KOŁAKOWSKA et al., 2017).

In terms of research, there are two areas that require attention: one is related to efficient means of diagnosis, while the other involves improving the effectiveness of therapeutic methods for already-diagnosed individuals. Currently, there is no prenatal test for autism that is reliable or widely used and since autism is mainly diagnosed based on behavioral symptoms after birth, there isn't a single biological or genetic marker that can indicate the predisposition or even the presence of the disorder reliably (MENTAL HEALTH (NIMH), 2023). Some researchers are studying several markers that could help the detection of prenatal autism, for example, studies found out that children within the autism spectrum have higher levels of certain antibodies and proteins in their blood that could be tested during the pregnancy in the mother's blood (CHEN et al., 2014; TALKOWSKI et al., 2012). According to Ramirez-Duque et al., (RAMIREZ-DUQUE; FRIZERA-NETO; BASTOS, 2018), early diagnosis is very important in improving social and cognitive functioning, and this is obtained mainly with the help of therapy sessions with professional workers trained to develop social and communication skills of children in the spectrum.

The therapy sessions commonly follow a method called Applied Behavior Analysis (ABA). ABA therapy is important not only because it allows professionals to work on mitigating harmful or antisocial behaviors, but it can also confirm uncertain diagnoses by observing behavioral cues related to ASD. While there is no ground rule that defines when ABA therapy must end, it is reported that some children might require several years(SHARMA; GONDA; TARAZI, 2018; LAI; LOMBARDO; BARON-COHEN, 2014). The sessions take place with one or more therapists in a room that usually contains toys to make the child feel comfortable and willing to interact. Each interaction or relevant event that takes place during sessions is annotated for posterior analysis, thus recording these sessions is a common practice to facilitate this task. However, the traditional ways of annotating are laborious and time-consuming, causing unwanted costs and delaying the diagnosis since clinicians often need numerous sessions as well as manually codifying the behaviors associated with ASD (stimming) to validate it.

Self-stimulatory behaviors, frequently called *stimming*, are stereotyped and repetitive behaviors that are often observed in individuals with ASD, the presence of these behaviors is often used as a diagnostic criterion or a metric for assessing therapy effectiveness (RAJAGOPALAN et al., 2013). However, their identification and quantification are typically done manually, which can be time-consuming and subjective. To help alleviate this issue, models based on computer vision techniques can be implemented to automatically detect these behaviors and even classify them by using certain predetermined characteristics. This can be achieved by training such models in labeled datasets where the model can learn how to recognize such patterns.

Considering this goal, many approaches of machine learning techniques have been proposed for helping the diagnosis and therapies of ASD in the past few years (PARIKH; LI, H.; HE, 2019; KOŁAKOWSKA et al., 2017; NOGAY; ADELI, 2020; RAHMAN et al., 2020; WANG, M.; YANG, 2022). The increasing number of works in this field coincides with the ASD increasing prevalence rate worldwide, its heterogeneous nature, and the increasing amount of data that is being generated by diagnosis and therapeutics methods used in ASD therapies and treatment methods.

The detection of symptomatic events or actions that can be related to ASD behaviors can help both the diagnosis and the therapeutic evolution of these individuals. Recently, through rapid development in the recognition methods of both images and videos based on computer vision methods, human recognition models have been applied and validated to achieve promising performance using various datasets, and have been deployed in several practical scenarios including video abnormal detection (DENG et al., 2022).

By detecting video events of behaviors such as gaze aversion, hand flapping, or head-banging, researchers can gain a better grasp of the underlying mechanisms of Autism and develop more effective interventions and treatments. This task is challenging and involves advanced techniques in computer vision and machine learning. In particular, it involves processing large amounts of video data that most of the time needs to be classified beforehand, extracting relevant features, and classifying events based on complex patterns and contexts.

A current challenge for investigating these kinds of applications is related to the fact that curated video data focused on autistic behaviors are most of the time unavailable publicly due to privacy and medical-patient confidentiality laws. Therefore, the advancement in this field relies on videos that are publicly shared on platforms such, for instance, YouTube. Some attempts at collecting and partially curating these videos have been part of previous research, most notably the Self-Stimulatory Behavior Dataset (SSBD) (RAJAGOPALAN et al., 2013), Expanded Stereotype Behavior Dataset (ESBD) (NEGIN et al., 2021) and Wei et al. Expanded SSBD (WEI et al., 2022) datasets. Of these three, only the SSBD provides a behavior time instant where it describes the start and duration where the stimming event occurs, leaving for the researchers that use the other datasets to *analyze and perceive* the instant where the event happens, generating discrepancies on the study of these behaviors and not providing a solid baseline for future analysis and works. The proposed unified dataset was carefully annotated on all clips, showing the start and duration of the stimming behavior, and enabling a more uniform approach when working and comparing the results on this kind of data, we also present an analysis of the predictions made on the annotated clips by a classic model and actions dataset to act as a baseline to future works.

1.1 OBJECTIVES AND METHODOLOGY

As previously discussed, computer vision can help assess children's actions and provide automated video coding for the stimming behaviors on the recorded video interventions. These interventions can aid mental health professionals in reducing the time it takes to diagnose ASD and notice changes between therapy sessions, providing children with ASD early therapeutic interventions and better treatment. However, current solutions for event detection of ASD therapy sessions lack proper, annotated datasets that encompass the wide variety of situations and environmental characteristics present in such sessions.

Therefore, the goal of this work is to propose a unified dataset which machine learning models can rely upon. Beyond that, we aim to propose an architecture for a system based on artificial intelligence and machine learning techniques that is capable of streamlining the analysis made on videos of ASD diagnosis and therapies, automating the detection of relevant interactions between the actors, classifying them, and putting those moments on evidence, cutting the time needed to peruse on the whole video of each session manually.

To achieve this goal, the following specific objectives were defined:

- Propose the Autism Stimming Behavior Dataset (ASBD) — A curated dataset, with uniform, unified and standardized characteristics, enabling a faster and more streamlined approach when working with this kind of data (Published in CBMS 2023).
- Present the architecture for an end-to-end system built to detect events of actors' interactions in ASD therapy session video recordings (Published in BRACIS 2022);
- Improve said system by proposing a model capable of automating the detection of stimming behavior;

A mapping review was conducted in the field of event detection on recorded therapy videos of ASD individuals, providing a broad vision of the machine-learning/computer vision techniques and frameworks applied in this area. Assessing the lack of a combined and curated source that could be used to enable new models in this field, the ASBD database was proposed. Acting as another tool in an integrated system, a state-of-the-art model of action recognition is trained and evaluated on this combined dataset aiming to enable stimming-behavior detection in the system, improving its efficiency in detecting relevant scenes of recorded therapy sessions to assist professionals in their analysis.

1.2 CONTRIBUTIONS OF THIS WORK

The following items summarize the scientific contributions developed during this work, including the ones directly related to this thesis as well as one that was part of a joint research effort:

- A literature review in event detection models, focusing on works that use machine-learning approaches on ASD video-recorded therapy sessions;
- Dataset - ASBD: Autism Stimming Behavior Dataset. Openly available on Github[2].
- Enable a model capable of automating the detection of stimming behavior.
- Evaluate said model in experiments on the novel proposed dataset and present comparisons against the related work.
- PlusMe web app: this research involves a collaboration with the European PlusMe project, a research effort involving therapy sessions of ASD children. In previous work of Soares (2023), a web application was developed to analyze therapy video sessions, and the solution of this work was included as a new feature. Openly available online[3].
- Publication A - (RIBEIRO; GRELLERT; CARVALHO, 2023): Stimming Behavior Dataset - Unifying Stereotype Behavior Dataset in the Wild - Presents the ASBD: Autism Stimming Behavior Dataset unifying and standardizing other stimming behaviors datasets built on "in the wild" configurations - Published at IEEE CBMS 2023.
- Publication B - (RIBEIRO et al., 2022): Event Detection in Therapy Sessions for Children with Autism. - Presents the architecture for an end-to-end system built to detect events of actors' interactions in ASD therapy session video recordings - Published at BRACIS 2022.
- Publication C - (CEOLIN et al., 2023): Adiposity and physical activity are among the main determinants of serum vitamin D concentrations in older adults: the EpiFloripa Aging Cohort Study. Published at Nutrition Research 2023.

1.3 TEXT ORGANIZATION

This work is organized as follows: Section 2 presents the background and concepts that are through the project; in Section 3 we discuss some of the related works on this field; Section 4 explores the dataset used in this work and the event detection

[2] https://github.com/OckerGui/Stimming-Behavior-Dataset
[3] https://github.com/OckerGui/therapy-aid-tool

architecture for ASD therapy sessions; Section 5 shows the analysis of the experiments we've achieved and discussion over it, ending in Section 5 with a conclusion and future works in this project.

2 BACKGROUND

This section describes the background concepts that will be explored over the project. First, ASD, its characteristics, and the repetitive behaviors that are highly associated with this disorder are viewed, next, the observational techniques that are used in these analyses are described, then computer vision and and its application in this specific field are discussed. Finally, we present the background on the YOLO model, a state-of-the-art real-time object detection system, and the VideoMAE model, a state-of-the-art action recognition algorithm.

2.1 AUTISM SPECTRUM DISORDER - ASD

Autism spectrum disorder (ASD) is defined as persistent deficits in social communication and social interaction that affect individuals across a wide spectrum, according to the Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-5) (EDITION et al., 2013). ASD encompasses a range of neurodevelopmental disorders characterized by impairments in social interaction, communication, repetitive behaviors, and restricted interests (MAENNER et al., 2021). It is considered a "spectrum" disorder because it encompasses a diverse range of symptoms and severity levels. While some individuals with ASD may have significant difficulties in daily life, others may exhibit remarkable talents and excel in specific areas.

ASD presents a range of symptoms that can vary from mild to severe in particular cases, also from skill to skill and child to child. These characteristics make diagnosis and therapy progress evaluation difficult and at the same time crucial for the effectiveness of the therapy (KOŁAKOWSKA et al., 2017). According to (ROGGE; JANSSEN, 2019), ASD is also related to many potential comorbidities such as epilepsy, attention problems, gastrointestinal problems, oppositional behavior, anxiety and depression, sleeping disorders, and feeding disorders. Other studies estimate that between 30 and 50% of individuals with ASD also have some kind of intellectual disability(ROGGE; JANSSEN, 2019). The diagnostic criteria for ASD have evolved over the years and are outlined in the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5), and the International Classification of Diseases, Tenth Revision (ICD-10). Early diagnosis and intervention are crucial for providing appropriate support and resources to individuals with ASD and their families.

The treatment and therapy approaches for individuals with ASD are highly individualized, taking into account the unique needs and strengths of each person (FRYE et al., 2022). Early intervention is widely recognized as a key component in improving outcomes for individuals with ASD. Evidence-based interventions often include Applied Behavior Analysis (ABA), speech and language therapy, occupational therapy, and social skills training. These therapies aim to address specific challenges related to commu-

nication, behavior, and social interaction. Additionally, various educational approaches, such as speech-generating devices, visual supports, and individualized education plans (IEPs), can be tailored to the child's needs (WONG et al., 2015). Medications may also be considered to manage specific symptoms or co-occurring conditions, but their use is determined on a case-by-case basis and typically involves consultation with a healthcare provider. Understanding and support from families, caregivers, and the community are essential in helping individuals with ASD thrive. Building an inclusive society that embraces neurodiversity is an ongoing goal, with advocacy groups and organizations working tirelessly to raise awareness and promote acceptance(HEALTH - OFFICE OF COMMUNICATIONS, 2021).

2.2 ASD THERAPY

Therapies for ASD encompass pharmacological and non-pharmacological interventions. Pharmacological treatments include psychostimulant drugs such as amphetamines, which are beneficial in improving comorbid hyperactivity and impulsivity in ASD patients, but provide little benefit on ASD core symptoms (SHARMA; GONDA; TARAZI, 2018). Although autism is rooted in biology, the most effective interventions so far are behavioral and educational; drugs have had only a minor role so far (LAI; LOMBARDO; BARON-COHEN, 2014).

Intervention, therapy, and support should be individualized and, if appropriate, multidimensional and multidisciplinary. The goals are to maximize an individual's functional independence and quality of life through development and learning, improvements in social skills and communication, reductions in disability and comorbidity, promotion of independence, and provision of support to families. Additionally, individuals should be helped to fulfill their potential in areas of strength (LAI; LOMBARDO; BARON-COHEN, 2014).

Complementary and integrative health has been garnering increased interest as supplementary non-pharmacological interventions for ASD. Those are diverse, and target a broad range of skills (cognitive, language, sensorimotor, and adaptive behaviors) via long-term intensive programs, and are grouped into applied behavior analysis (ABA) and structured teaching. The Early Start Denver Model is a further development from the ABA, in which a developmental framework and relationship aspects are emphasized. The early intensive behavioral intervention seems to enable the development of intelligence, communication, and adaptive function, and, to a lesser extent, language, daily living skills, and socialization. A shift from atypical to typical neurophysiology has been reported after 2 years of intervention with the Early Start Denver Model (LAI; LOMBARDO; BARON-COHEN, 2014).

Figure 3 – Stimming behavior - Frames (blurred to preserve the anonymity of identity) of videos showing different types of stimming behavior.

Source: the author

2.3 SELF-STIMULATING BEHAVIORS - STIMMING

The Diagnostic and Statistical Manual of Mental Disorders (ASSOCIATION, 2013) defines stimming behavior as "stereotyped or repetitive motor mannerisms" and is presented as one of the five key diagnostic criteria of autism spectrum disorder. According to Wei et al. (2022), these behaviors are often more prevalent in individuals inside the Autism spectrum than their typically developing peers. The reasons why individuals engage in self-stimulatory behavior are not clear and are likely to differ for different persons. It may be that the behavior provides sensory reinforcement or sensory stimulation to the individual, or the behavior may be used to regulate sensory input, either increasing stimulation or decreasing sensory overload (BOYD et al., 2010).

Stimming behaviors observed in individuals on the autism spectrum encompass a range of actions, which may vary from whole-body movements to more focused gestures. For instance, complete body stims like swaying or spinning directly influence the vestibular sensory system responsible for maintaining body balance and orientation. In contrast, there are stims that target specific sensory experiences without engaging the entire body. Examples of these include hand flapping, squinting, fixating on rotating objects (such as a ceiling fan), tactile exploration involving the stroking or rubbing of textured surfaces, olfactory stimulation through smelling objects, instances of head banging, and vocalizations like squealing. While some of these behaviors may disrupt learning or work activities, others, like head banging, can pose potential self-injury risks (CHARLTON et al., 2021; BOYD et al., 2010; RAJAGOPALAN et al., 2013).

2.4 OBSERVATION TECHNIQUES

Bertamini(2021)(BERTAMINI et al., 2021) points out that studying the early interactions of children is a core issue for research in infant typical and atypical development. Therefore, observational research is considered one of the main approaches in this area. Inside the clinical context, where ethical issues often put restrictions on the feasibility of controlled experiments, observational techniques are a keystone in developmental research in those individuals. These techniques are further emboldened by studies that showed similar results between behavioral observation and randomized controlled trials (RCTs) along with a set of quality criteria to strengthen observational results(ANGLEMYER; HORVATH; BERO, 2014). With that in mind, observational studies and RCTs can be used as complementary approaches to balance discovery and explanation, increasing generalizability to the wider population and generating data-driven hypotheses for subsequent confirmative designs.

Observational studies were leveraged by many techniques that were developed over the years across several large projects that aim to detect and classify human behavior (NIGAM; SINGH; MISRA, 2018). These studies mainly include cross-sectional, longitudinal and case-control designs, and have often been mistakenly considered as merely qualitative(BERTAMINI et al., 2021). On the other hand, recent techniques on observational analysis enable researchers to collect quantitative data and to employ more sophisticated computational approaches for the systematic observation of behavior, notably Hidden Markov Models (HMM), Artificial Neural Networks (ANN), Spatio-Temporal approaches among others(NIGAM; SINGH; MISRA, 2018). These techniques present a great opportunity that is highly relevant in the context of developmental and clinical research, where observational techniques have the great advantage of being almost completely non-invasive, or minimally invasive in many cases (BERTAMINI et al., 2021).

2.5 COMPUTER VISION

Computer vision has been helping those observational studies to achieve their results in the past decade via a multitude of ways (THEVENOT; LÓPEZ; HADID, 2018). Some of them include classification and segmentation of images(WU; LIU, Q.; LIU, X., 2019), object detection(ZHAO, Z.-Q. et al., 2019), video analysis(OPREA et al., 2020), facial expression analysis(ABDULLAH; ABDULAZEEZ, 2021) and behavioral patterns detection(LU; NGUYEN; YAN, 2020) to cite a few. Shabaz (2021) (SHABAZ et al., 2021), Abirami (2020) (ABIRAMI; KOUSALYA; BALAKRISHNAN, 2020) and Ramirez-Duque(2018) (RAMIREZ-DUQUE; FRIZERA-NETO; BASTOS, 2018) use deep learning techniques of Computer vision to analyze and extract relevant information from videos of daily routines or tests of individuals with ASD and then use that information to support diagnosis and therapy studies.

Figure 4 – Observation studies - With small to no invasion over the "known/safe" environment of the subjects, observation studies can be accomplished within recurrent therapy sessions or on handheld recorded videos, acting as a complementary approach to other studies.

Source: the author

According to observational therapies(BERTAMINI et al., 2021), the success of the therapy is directly connected to the feedback that the patient delivers. Feedback from an unrecorded session is limited to the therapist's local perception and memory, which is prone to misinterpretation and forgetfulness. A video-recorded session, on the other hand, allows the therapist to revisit events of interest noticed during the live session and also the perception of new events not detected before, thus generating a kind of record(HAILPERN et al., 2009).

In this record, each therapist is free to annotate the events of interest that make the most sense as an ASD professional, qualitatively and/or quantitatively, such as whether the child has had a good evolution concerning past sessions, how many times the child interacted with the therapist and the environment, how long the session lasted, among many others.

However, two interesting problems arise in this scenario, the first being the time spent on this manual analysis. Reviewing and taking notes of a therapy session could take more time than the session itself, which ends up being unfeasible for a large amount of data and/or when the responsible therapist treats other patients. And if the recording is not done by the same therapist who performed the session, the problem is the recording pattern. For one professional, certain events of interest are more important and descriptive than for others, so the type, format, and metrics of each record quickly disperse and become inconsistent from professional to professional.

2.6 VIDEO MAE

VideoMAE is a deep learning model introduced in the research paper "VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training" by Tong et al. (2022). This model extends the concept of masked autoencoders (MAE) to the domain of video analysis and has demonstrated state-of-the-art performance on various video classification benchmarks, namely kinetics-400 [1], Something-Something V2 [2], ucf101 [3] and hmdb51 [4]. The authors claim that the Masked autoencoder characteristic of this model can be well suited to achieve high performance on tasks that are dependent on small datasets and also that it is particularly designed for self-supervised video pre-training (SSVP).

The challenge of self-supervised learning in videos is achieved by proposing a customized video tube masking and reconstruction approach. This method effectively mitigates information leakage caused by temporal correlations during video reconstruction. Notable findings from the research include the ability of VideoMAE to maintain strong performance with a high proportion of masking (e.g., 90% to 95%), which is significantly higher than what is feasible with image-based models. This is attributed to the temporal redundancy present in video content, enabling a higher masking ratio. One of the key strengths of VideoMAE is its data efficiency. The model achieves impressive results, even when trained on very small datasets consisting of around 3,000 to 4,000 videos, without the need for additional data sources. This is partially attributed to the challenging nature of the video reconstruction task, which enforces high-level structure learning. Additionally, the research underscores the importance of data quality over data quantity in self-supervised video pre-training, emphasizing that domain shift between pre-training and target datasets is a significant consideration in the SSVP process.

2.7 YOLO ALGORITHM

Regarding object detection and classification, YOLO (You Only Look Once) is a state-of-the-art real-time object detection system(REDMON et al., 2016). YOLO's goal is to recognize items faster than traditional convolutional neural networks without sacrificing accuracy by glancing at the image only once and treating object detection as a single problem. The pipeline resizes the input before running it through a single convolutional neural network and thresholding the results based on the model's confidence level. To execute the detection, YOLO divides the image into many sub-regions and assigns five anchor boxes to each one. The likelihood of a certain object is calculated,

[1] https://paperswithcode.com/sota/action-classification-on-kinetics-400
[2] https://paperswithcode.com/sota/action-recognition-in-videos-on-something
[3] https://paperswithcode.com/sota/self-supervised-action-recognition-on-ucf101
[4] https://paperswithcode.com/sota/self-supervised-action-recognition-on-hmdb51

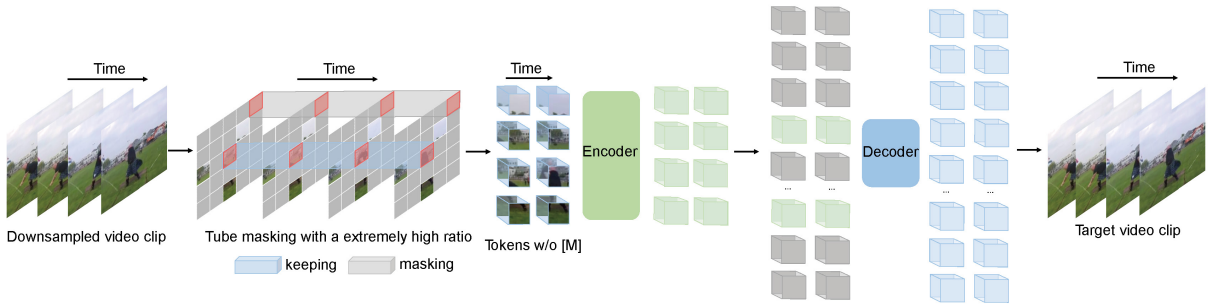


Figure 5 – VideoMAE overview - By masking random cubes and reconstructing the missing ones with an asymmetric encoder-decoder architecture, the model can exploit the high redundancy and temporal correlation in videos, allowing a more aggressive approach when masking enabling it to capture more useful spatiotemporal structures.

Source: (TONG et al., 2022)

and the zone with the highest probability is chosen. Since its original release, several iterations have been developed looking to improve its performance, access, and ease of use, currently on the 5th major iteration(YOLOv5)[5].

The network internal structure of YOLOv5 involves three distinct parts: backbone, neck, and output. The backbone is based on an incorporation of CSPNet which not only decreases the parameters and FLOPS (floating-point operations per second) but also the whole model size, increasing efficiency (WANG, C.-Y. et al., 2020). As its neck, path aggregation network (PANet) improves the propagation of low-level features together with adaptive feature pooling which links the feature grid and all feature levels. Finally, the head (Yolo layer) generates 3 different sizes (18×18, 36×36, 72×72) of feature maps to achieve multi-scale prediction, enabling the model to handle small, medium, and big objects(WANG, C.-Y. et al., 2020; XU et al., 2021).

The BottleneckCSP module extracting abundant information from images performs feature extraction on the feature map. Compared to other convolutional neural networks, the BottleneckCSP structure outperforms previous approaches by reducing the gradient information in CNN's optimization. The complexity of this structure occupies the entire network.

By adjusting the width and depth of the module, four different models can be obtained with different complexities. The models obtained in this way have been published in [16] as YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. For increasing the receptive field of the network and obtaining features of different scales, the SPP module is used.

The structure of YOLOv5 contains a bottom-up feature pyramid structure based on FPN. With this combined operation, while strong semantic features impart from top to bottom, feature pyramid caries positioning features from the bottom up. By combining features from different feature layers, the network improves the ability to recognize the

[5] https://docs.ultralytics.com/

targets with different scales. Consequently, the network ends with two types of outputs that classificate results and object coordinates. The structure of the YOLOv5 network can be seen in Fig. 1(XU et al., 2021).

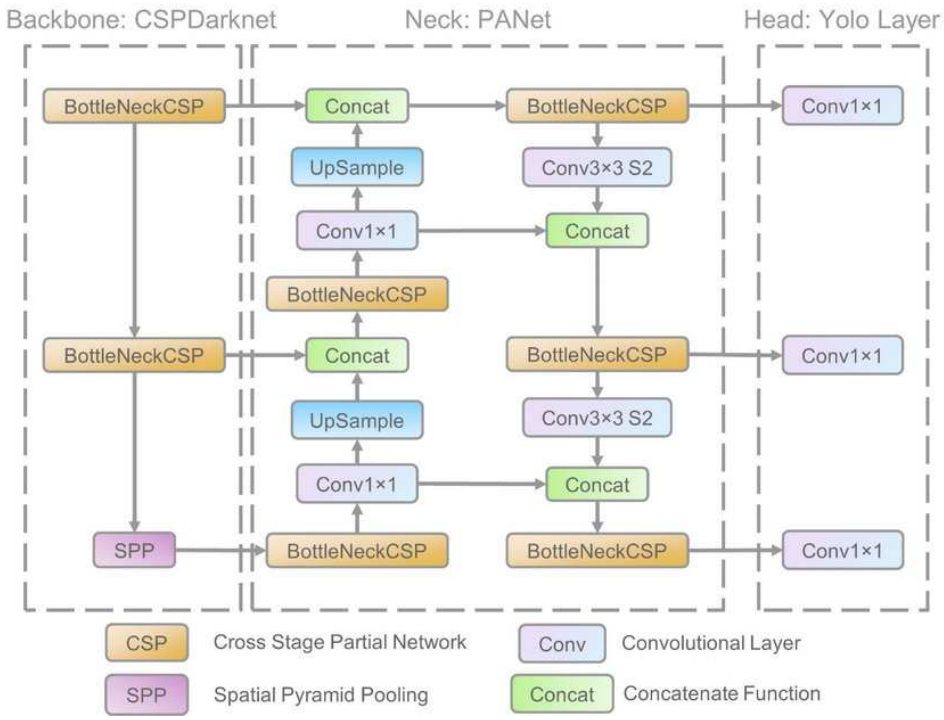


Figure 6 – YOLOv5 network architecture. Composed of three parts: (1) Backbone: CSP-Darknet, (2) Neck: PANet, and (3) Head: Yolo Layer. The data are first input to CSPDarknet for feature extraction and then fed to PANet for feature fusion. Finally, Yolo Layer outputs detection results (class, score, location, size).

Source: Xu et al. (2021)

2.8 DATA AUGMENTATION

Typically, deep learning models possess a huge number of parameters that need to be trained. The risk of overfitting with such big models is very high and big datasets with high variability are needed for networks to be able to generalize properly, achieving good results in the end. Data augmentation is a technique that can be used to generate more diverse training data. This process involves applying various transformations to the object of interest, in our case, videos, to create new samples for model training. In the video domain, these transformations can include temporal alterations like frame jittering, where frames are randomly dropped or duplicated, as well as spatial changes like random cropping and resizing. These techniques help the model become more invariant to minor variations in the input data, which can be particularly beneficial in action recognition tasks, where actions can vary in terms of speed, scale, and orientation (YUN et al., 2020; CAULI; REFORGIATO RECUPERO, 2022). For areas where datasets

for such models are less common or simply expensive to collect, data augmentation techniques can help enrich these, by simply generating more samples to achieve a more diverse set or by tackling specific problems like unbalanced classes by artificially generating new ones (YUN et al., 2020; CAULI; REFORGIATO RECUPERO, 2022).

3 LITERATURE REVIEW

In this chapter, we discuss a literature mapping review on the topic of event detection using machine-learning approaches in ASD video-recorded therapy sessions.

3.1 METHODS

This research strategy and methods were based upon the practices for the conduction of systematic map processes described by Petersen et al(PETERSEN et al., 2008) and Kitchenham et al(KEELE et al., 2007). Systematic map studies are aimed to avoid unnecessary duplication of effort and error in a somewhat too specific field before a more dense systematic review is done and can be used to identify opportunities and gaps in a set of primary studies. Also, as proposed by Pertersen(PETERSEN et al., 2008) within the characteristics of mapping reviews, quality assessment on the retrieved studies won't be executed.

3.1.1 Objective

The main objective of this study is to identify which machine learning techniques, models and datasets are being applied to detect, retrieve, and qualify events in videos of ASD therapy sessions.

3.1.1.1 Research questions -

Aimed to retrieve information regarding the goal defined above, the following research questions were defined:

Table 2 – Research questions for this mapping study

ID	Research Questions
RQ1	Which techniques/models are being applied to detect events in ASD therapy videos?
RQ2	Which framework or combination of?
RQ3	Which region/body parts are considered/focused on?
RQ4	There are techniques or metrics to quantify these detections?
RQ5	There are techniques or metrics to qualify these detections?

3.1.2 Databases and search string definition

3.1.2.1 Databases

The databases considered for this study were ACM[1], IEEExplore[2] and Scopus[3]. The Scopus database encompasses a much larger library since it can index several other databases, raising the reach of search queries executed on its search engine, therefore expanding the reach of this work. Also, Scopus allows a filter by topics or subject area, since the amount of results was far bigger without said filters, we decided to utilize it.

3.1.2.2 Search string -

The search string for this study was mainly composed with four fields in mind - **autism**, **machine learning**, **event detection**, and **videos**. For each of those fields, related keywords or acronyms were considered as well, finally composing the search string:

("event" OR "interactions") AND ("detection") AND ("video") AND ("machine learning" OR "ML") AND ("ASD" OR "autism")

This generic search string was then applied to the databases specified following the specific characteristics each search engine had (full search strings are on the Apendix8.1) and the results are discriminated in the table 3. The results considered in this study come from searches executed on June 23, 2020.

Table 3 – Database results

Database	Result
Scopus	891
IEEExplore	9
ACM	83
Total	965

3.1.3 Eligibility criteria definition

3.1.3.1 Inclusion criteria -

For the work to fall within the criteria we are looking for in this project, it must contain the all following characteristics:

1. application of machine learning techniques/models
2. in the context of ASD/Autism

[1] https://dl.acm.org/
[2] https://ieeexplore.ieee.org/Xplore/home.jsp
[3] https://www.scopus.com/search/form.uri#basic

3. with event/interaction detection on video recordings
4. published between 2018-2022 (included)
5. original works published on peer-reviewed process

3.1.3.2 Exclusion criteria -

Beyond that, if any work shows any of these characteristics, it will be excluded from this study:

1. duplicated studies
2. study out of context of autism or ASD
3. shows no application of machine learning
4. doesn't work with event/interaction detection on video recordings
5. study completely focused on facial or eye movement
6. study completely focused on emotion detection techniques
7. not written in the English language
8. works that are or compose books
9. somehow inaccessible through the UFSC academic network (in loco or VPN)
10. systematic literature review and mapping

3.1.4 Data Extraction Form - DEF

Table 4 shows the base structure and descriptions for each column of the Data Extraction Form utilized in this study, as well as the relevant research questions tiered with each column. *Annotate columns that are correlated to all RQs.

3.1.5 Selection workflow

The selection process executed in this study is presented in the figure 7 and is composed of 7 steps. First, the generic search string is specified for each target database and is executed. The retrieved results were consolidated on Zotero research assistant [4]. Then, .ris files with each source were generated by Zotero and imported to Rayyan tool [5]. Within Rayyan, the duplication detection and treatment were executed, followed by the inclusion and exclusion criteria assessment phase. In this phase, 1269 unique studies were systematically analyzed by their titles, keywords, and abstracts (TKA) against the inclusion and exclusion criteria, remaining 35 labeled on Rayyan as "TKA Selected". These 36 studies had their entire text evaluated against the inclusion and exclusion criteria, eliminating 22, resulting in 13 studies selected.

[4] https://www.zotero.org
[5] https://www.rayyan.ai/

Table 4 – Data Extraction Form Description

DEF Column	Description	RQ's
System Id	Main extraction id from Zotero - Unique for each study through all the process	–
DEF - ID	Id of the study on the DEF table	–
Title	Title of the study	–
RQ1 - ML Tec	Machine learning techniques/models applied in the study	RQ1
RQ2 - Frameworks	Which ML framework or combination of that were used to help the study	RQ2
RQ3 - Region	The focus of study to detect events or interaction on the body	RQ3
RQ4-Quantify	The study uses heuristics to quantify the number of events/actions that happen	RQ4
RQ5-Qualify	The study uses heuristics to qualify the detected events/actions	RQ5
Dataset (QTY/Len)	The size in quantity and length of each video used in the study - or a known public dataset	*
Dataset Label	Specify if the dataset have ground-truth annotations or labels	*
Diagnosis/therapy	The main aim of the study is to help detect/diagnose ASD or assist in therapy sessions and ASD evolution assessment	–
Notes	General notes of the collaborator/extractor related to perceived characteristics of the study	–
Publication Type	Type of venue where the study was published	–
Year	Year of publication	–
Country	Country of the Institution where the first author was affiliated when the study was published	–
Authors	Authors list	–
Database	Database source	–
Link	Link to the study (Retrieved within UFSC academic access - in loco or VPN)	–

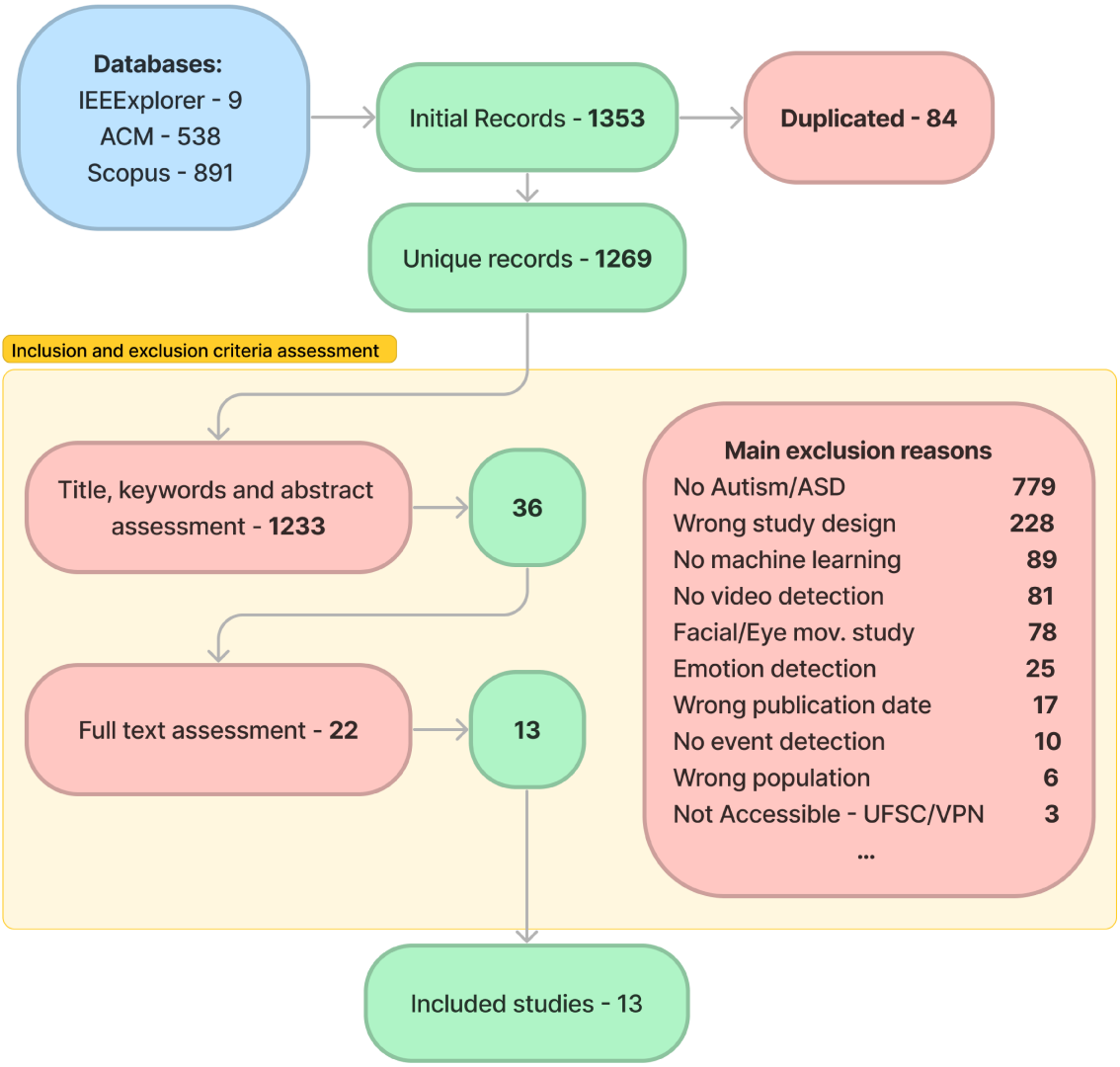


Figure 7 – Execution flow chart

Source: the author

3.1.6 DEF - Extraction

The selected studies were categorized and their information was structured into the Data Extraction Form table. This table is public available at Click me - DEF google Spreadsheet

3.1.7 Validity threats

This systematic mapping review was conducted following the practices described by Petersen et al(PETERSEN et al., 2008) and Kitchenham et al(KEELE et al., 2007), its protocol was revised by domain specialists aiming to mitigate any discrepancies on its construction, and its execution was conducted strictly following those protocol and guidelines, even though, some validity threats are known and described below.

Since this study was conducted mainly by its only author, the researcher bias is an existent threat as discussed by Petersen et al(PETERSEN et al., 2008). In the future, additional researchers will be added and this threat will be mitigated.

The protocol followed in this study, along with its criteria, strings, and data extraction form (DEF) are all reported in this study, its appendix, or available on demand. With this, we aim to mitigate the repeatability and descriptive validity threats inherent to secondary studies.

From the 35 studies selected at the "Title, keywords and abstract" step of our execution phase, 3 studies were inaccessible with the access rights inside our institution and their authors didn't respond in time for those to be considered in the next steps and they were excluded from the study. This scenario configures a threat since were not able to evaluate those studies.

The ACM's database search engine presented a much lower number of studies with a more narrow search string where it was built with synonyms linked by the operator "OR" and the main components of the generic search string linked with the operator "AND" like in the other database search strings. With a wider search string like the one used in this study, we aimed to mitigate the inconsistency on the search strings threat that this decision brought us by widening the query string and evaluating more studies in the execution phase.

We plan to execute forward and backward snowballing in the selected studies in the future to mitigate the threat associated with the choice of databases for this work.

3.2 RESULTS

Within this section, we aim to answer the research questions asked in the planning phase of this study. To improve the readability of both charts and tables we use an ID to identify the studies, please refer to the table 20 for the description of the studies, its references, and more data.

3.2.1 RQ1 - Which techniques/models are being applied to detect events in ASD therapy videos?

In figure 8 we can see that of the 14 studies that were selected, around 57% used CCN techniques or some combination with it, CNN models are known to detect and capture details on images or videos that other techniques weren't able to. It became widely known in the past few years as several models of this technique became popular and far more accessible through framework implementations (JAVED; PARK, 2020).

Also notable are the SVM techniques used in 50% of the works. Easier to use than CNN, it has been widely implemented in prior ML studies to classify individuals with and without ASD (ZHAO, Z. et al., 2021).

Highlighted in green, are Machine Learning techniques that fit under the umbrella of deep-learning models.

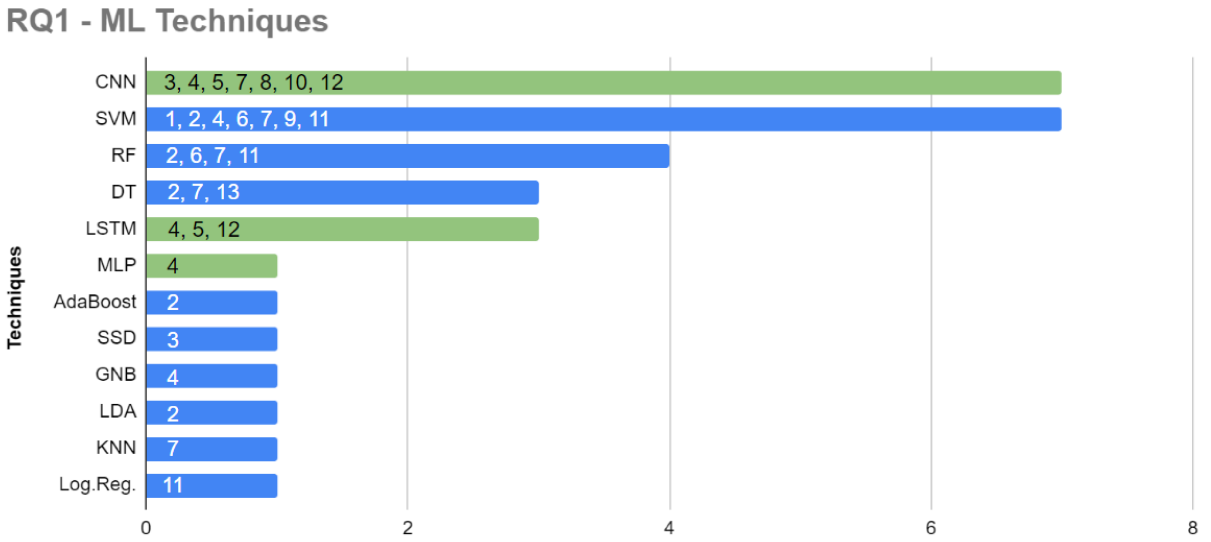


Figure 8 – RQ1 - Machine learning techniques

Source: the author

3.2.2 RQ2 - Which framework or combination of?

To support and enable the use of such often difficult models presented in the previous question, it is very common in computer science or multi-disciplinary studies to work with frameworks that implement these models. This question aims to perceive a distribution of these frameworks - Figure 9. We can notice a big prevalence of the OpenPose framework, used to extract skeletal data points from actors both in images and videos. The extracted data for this question were more focused on frameworks that enabled the authors to extract and analyze data related to visual or event points of interest.

Also, by combining the techniques and frameworks, we can see which frameworks have been used to support solutions with said technique - Figure 10.

3.2.3 RQ3 - Which region/body parts are considered/focused on?

As we discussed in the background section of this study, there are several signs to look for and ways to detect events associated with ASD individuals. In this question, we extracted data to show where the event detection studies have been focusing their efforts. As the frameworks used in these studies are closely related to the region, we combine these two data in a single figure 11. Here we separate posture from the whole body by analyzing if the study used more than one point on the head region, indicating a more complex set of heuristics associated with it.

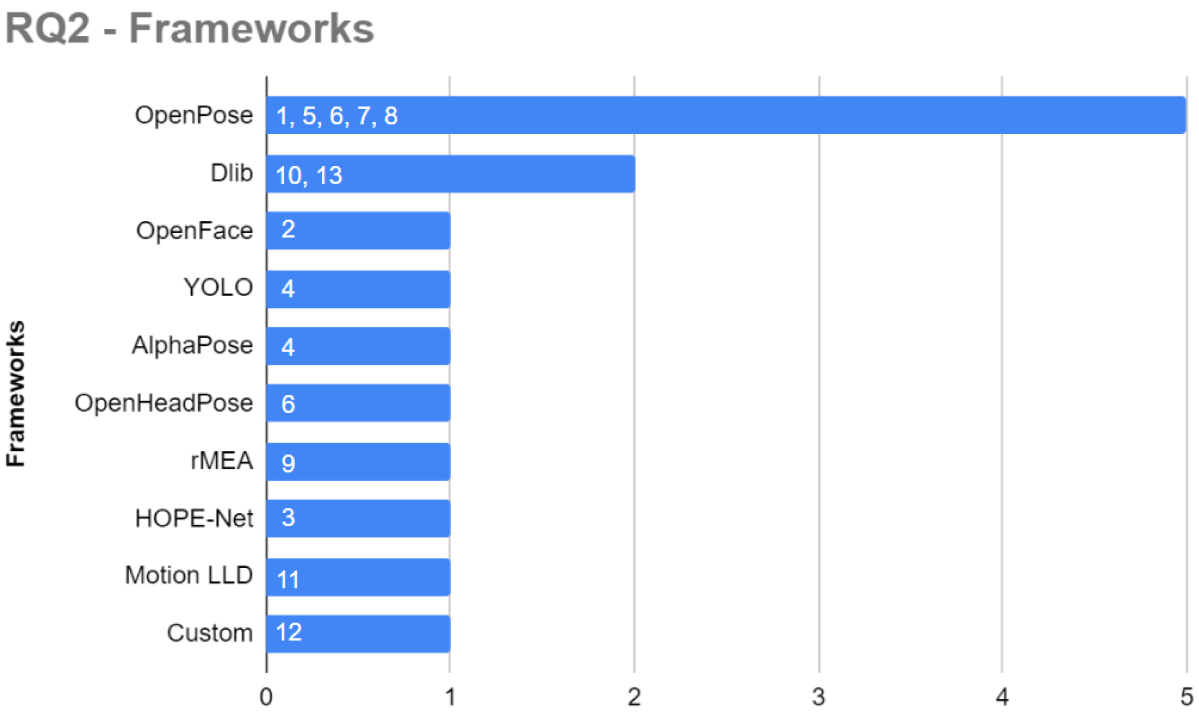


Figure 9 – RQ2 - Frameworks

Source: the author

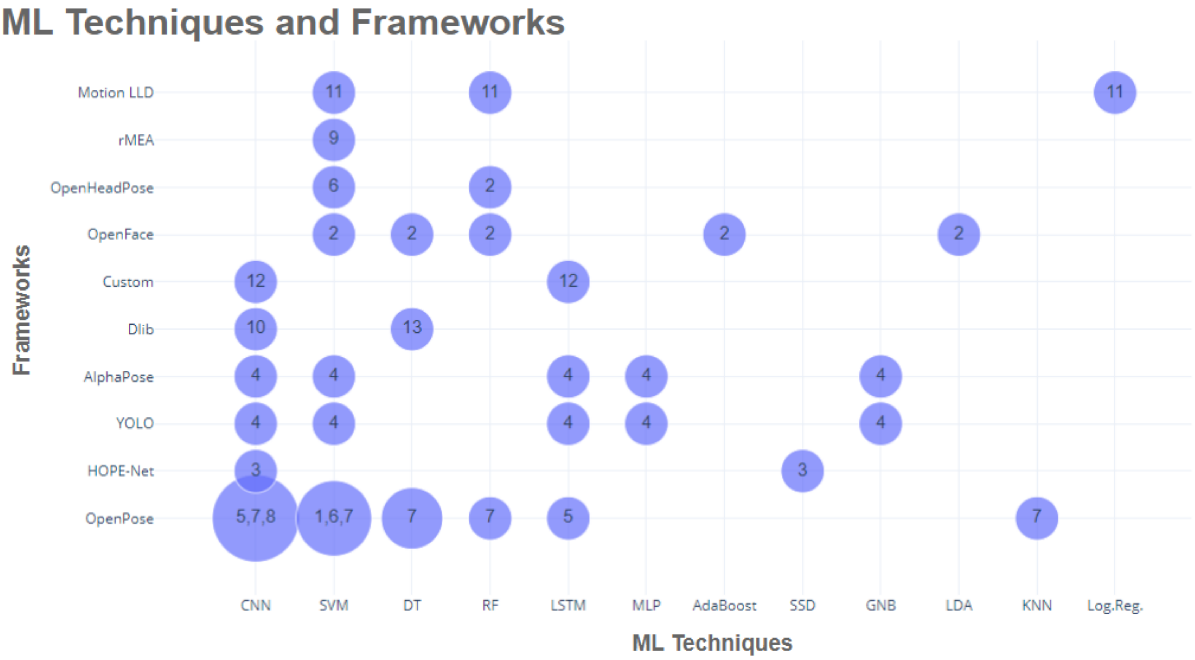


Figure 10 – Combination of ML techniques and frameworks used in the studies

Source: the author

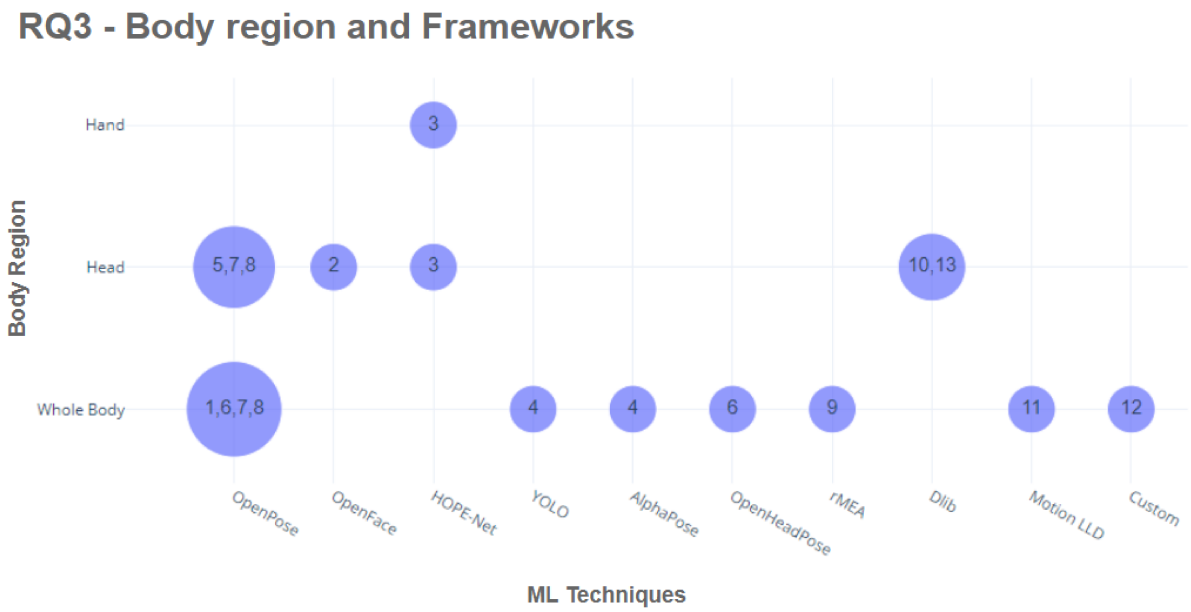


Figure 11 – RQ3 - Combination of body region and frameworks

Source: the author

3.2.4 RQ4 - There are techniques or metrics to quantify these detections?

All the studies that were selected applied some kind of heuristic that was able to count the amount of events that they were analyzing, however, not all of them were able to classify them as we can see in the next research question.

3.2.5 RQ5 - There are techniques or metrics to qualify these detections?

This question represented a more exploratory attempt to perceive if the selected studies were able to create complex heuristics that could detect and classify different actions or events through video analysis. To be able to do this, we observed if the studies could detect more than one type of event, like how a response to a call is different than a head-banging and arm-flapping, actions widely related to ASD individuals. In figure 13 we can see that slightly less than 70% of the studies were able to do so. Among the studies that were able to create these complex heuristics, we tried to perceive common categories between them. Those categories often represent common events often associated with ASD diagnosis or relevant to treatment, like head-banging - Figure 14.

3.2.6 Study focus

All selected studies categorized its solution as a tool to help professionals either detect atypical signs and support the diagnosis of ASD individuals - diagnosis - or to detect events to help understand and qualify how the treatments that were administrated

to the individuals were evolving - therapy. This distribution can be observed on the figure 15.

3.2.7 Study distribution

In the figure 16 we use the country associated with the institution of the main author as a factor of distribution of the studies around the world.

3.3 DISCUSSION

As an ample, diverse, and rapidly evolving area of computing sciences, machine learning presents a whole set of techniques and models that are changing and evolving each day. This allows an even more diverse set of problems to be solved but can pose as a challenge to multi-disciplinary teams to figure out which techniques or models can or should be used to deal with their study. Figure 8 enables us to visualize the techniques and answer the RQ1. CNN and SVM techniques being widely adopted, the first and more recent one as an emergent technique often associated with better results than the second, a simpler and widely used to solve classification problems for several years and are often used as a baseline for performance comparison of other techniques (ZHAO, Z. et al., 2021; NEGIN et al., 2021; JAVED; LEE; PARK, 2020). The use of more classical models like Decision Trees and ensemble models like Random Forests are predominant as comparison standpoints enabled by nowadays ease to use by implementations of popular frameworks (figure 9) like R-Studio [6], OpenPose [7] and Scikit-Learn [8] (NEGIN et al., 2021; GEORGESCU et al., 2019; WASHINGTON et al., 2021).

In figure 10 we can observe a diverse environment of combinations of frameworks and techniques, there were only 4 combinations that were explored by more than one author from a total of 110 possible mapped combinations.

Most of the studies selected analyzed the whole body or head region as part of its heuristics of detection, indicating a big effort to detect actions associated with gaze and whole-body responses. The most known signs that are mapped on the specialized literature (EDITION et al., 2013) are indeed associated with the face and head areas - Figure 11.

RQ4 was thought to evaluate if the studies were able to detect and quantify events from videos, a non-trivial task, but all the selected studies presented capable heuristics to do it.

On RQ5, the task difficulty is significantly higher as the capacity to qualify different sets of motions and events is far more complex. Even so, slightly less than 70% of

[6] www.rstudio.com
[7] https://github.com/CMU-Perceptual-Computing-Lab/openpose
[8] scikit-learn.org

the studies were able to do it. To cite a few of these implementations, Liu 2022 (LIU, J. et al., 2022) builds detection within time-to-respond heuristics that can detect and classify events based on if and how fast individuals respond to stimuli. Sayis 2020 (SAYIS; PARES; GUNES, 2020) study uses annotated video lapses to help identify key parts of the analyzed video and classify them in a way to identify if there are differences in nonverbal (body) behavior at an individual and interpersonal level between typical and atypical children. Javed 2020 (JAVED; LEE; PARK, 2020; JAVED; PARK, 2020) used pose and face data to model interactions that included eye gaze focus, vocalizations, smiling, self-initiated interactions, triadic interactions, and imitation. Each of those interactions has its own custom heuristics to be identified in the data, like computing when the gaze of the actor matches the position of the actor of the study. Negin 2021 (NEGIN et al., 2021) work builds up a database that enables custom heuristics to identify and classify a wide variety of events like spinning and hand-clapping. The study of all those heuristics will be further improved in future works.

A nearly even distribution between diagnosis and therapy focus shows that both the earlier identification of an atypical individual and the process of treatment are important steps that have been studied in the literature - Figure 15.

Despite being a worldwide issue(ZEIDAN et al., 2022), we observe that only the USA and China have more than one study between the selected papers, with 5 of the 14 studies having its main author associated with institutions in the USA - Figure 16.

3.4 CHAPTER SUMMARY

By following the guidelines specified in section 3, we aimed to build up knowledge of the state of the art related to the event detection methods and possibilities in therapy sessions for individuals with ASD, allowing the reproducibility and integrity evaluation within the threats identified. With an initial total of 1353 studies identified, systematically reduced to the 13 studies presented in the result section through a well-defined set of criteria.

We were able to identify and map a wide array of techniques, models, and frameworks capable of identifying, quantifying and even qualifying a complex set of events related to the ASD diagnosis and therapy.

By answering the research questions asked, we enable future works that can consider the models, frameworks, and characteristics that were evaluated by this study, allowing them to identify the strengths and explore the weaknesses found. Below we explore further the latest of those selected works presented above, allowing us to compare those studies against this one.

Zhong Zhao et al. (2021) experiment relies on OpenFace 2.0 lib [9] to extract head

[9] http://cmusatyalab.github.io/openface/

movement dynamics such as pitch (head nodding direction), yaw (head shaking direction), and roll (lateral head inclination) directions to compute two metrics, head rotation range (RR) and the amount of rotation per minute (ARPM). These data was collected from a sequence of yes/no questions from children with ASD and typical development who were encouraged to nod/shake their head while responding. To process the data and act as classifiers, a total of 6 ML models were used, namely SVM, linear Discriminant analysis (LDA), decision tree (DT), random forest (RF), and ensemble learning with boosting (ENS). They reported 92.11% of maximum classification accuracy with a DT classifier.

By annotating and classifying a series of home-made, publicly available, videos of children with ASD, culminating in a novel public dataset called ESBD, Negin et al. (2021) enable the construction of two different action recognition frameworks to analyze ASD behaviors and produce baseline results for the collected dataset. The first one is a descriptor-based framework based on a bag-of-visual-words approach and the second one receives pose sequences as input and models actions using an LSTM network. After that, a series of tests are made, comparing them against CNN-based deep learning methods. For each of the frameworks, an extensive series of tests with different feature extractors and models are made. Reportedly, the best combination of accuracy for each framework is 79% for Histogram of Optical Flow (HOF)+MLP on the descriptor-based framework and 61% on the Skeleton-LSTM combination for the Articulated Pose-based framework. These are compared against a 74% accuracy on a CNN model built upon a ConvLSTM(ResNet-152) based network combined with LSTMs and Softmax regressors for classification.

Aiming to train and test a head-banging detector, Washington et al. (2021) implemented a time-distributed convolutional neural network (CNN) using the Keras [10] Python library with a Tensorfow[11] backend. The skeletal key points in each frame of a home-made videos database (SSBD) are tracked and extracted using OpenPose lib, slightly modified to consider only the head region to tackle noise generated by home-made videos' predominant hidden body parts that are on occlusion. The extracted features are then fed into a long short-term memory (LSTM) neural network which is trained with Adam optimization(KINGMA; BA, 2014). They reported a mean F1-score of 90.77% between 3 cross-validation folds.

Jingjing Liu et al. (2022) points out the current limitation of the traditional clinical diagnosis methods, which rely on subjective observations, and advocates for the use of computer vision techniques to provide more objective and automated assessments. The article introduces a novel experimental protocol, "response-to-instruction" (RTI), designed to screen autism in toddlers by assessing their ability to disengage from toys

[10] https://keras.io/
[11] https://www.tensorflow.org/guide/

and respond to language instructions. To achieve this, a two-camera controlled setup is built and computer vision algorithms are employed, including hand detection and gaze estimation, and a database called the TASD (the ASD database) is established for comprehensive behavioral analysis.

Most of these works aim to assist in the detection of ASD characteristics with different approaches, but none of them are focused on how the therapy associated with each individual is progressing. By applying these techniques on an end-to-end system where videos are both the input and the output we are looking to detect and quantify ASD-related actions, enabling experts to distinguish typical from atypical individuals and also assess the therapy evolution in a faster and more decisive manner.

The YOLO family of algorithms is rapidly gaining importance in solutions that are based on computer vision as was discussed in the background section of this work. The only mention of any version of these algorithms on the observed literature (NEGIN et al., 2021), uses YOLOv3 (Table 5), which was vastly improved on its most recent iteration(YOLOv5), part of the framework that is used on this study. Beyond that, we plan to use and combine different heuristics of event detection that were implemented in these related studies as part of our framework.

Table 5 – Mapping review summary

Author	Goal	Tech.	Framework	Body Reg.	Public Database
Zhao (2021)	Diagnosis	DT	OpenFace	head movement	private
Negin (2021)	Diagnosis	SVM	HOG + YOLOv3	posture, head and hand movement	ESBD - on request
Washington (2021)	Diagnosis	CNN + LSTM	OpenPose	head movement	SSBD
Liu (2022)	Diagnosis	SSD + CNN	Hopenet + HKUST	hand movement gaze	private
This work	Diagnosis + Therapy	MAE	VideoMAE	whole body	ASBD

Computer vision techniques rely heavily on datasets as examples to help train models. These models learn from the data, enabling them to accurately identify specific actions. However, our analysis of the selected studies revealed a noticeable absence of publicly available datasets to support this process. Only one study(WASHINGTON et al., 2021) utilized public and shareable data that utilized the SSBD dataset proposed by Rajagopalan et al. (2013). Negin et al. (2021) proposed and utilized a meant-to-be public and shareable dataset similar to the SSBD, dubbed ESBD, but it was shared only upon request. Many of the studies mention the difficulty of compiling and sharing their datasets due to copyrights/privacy rights since those videos often present infant individuals and/or are collected in controlled and private environments like hospitals and clinics.

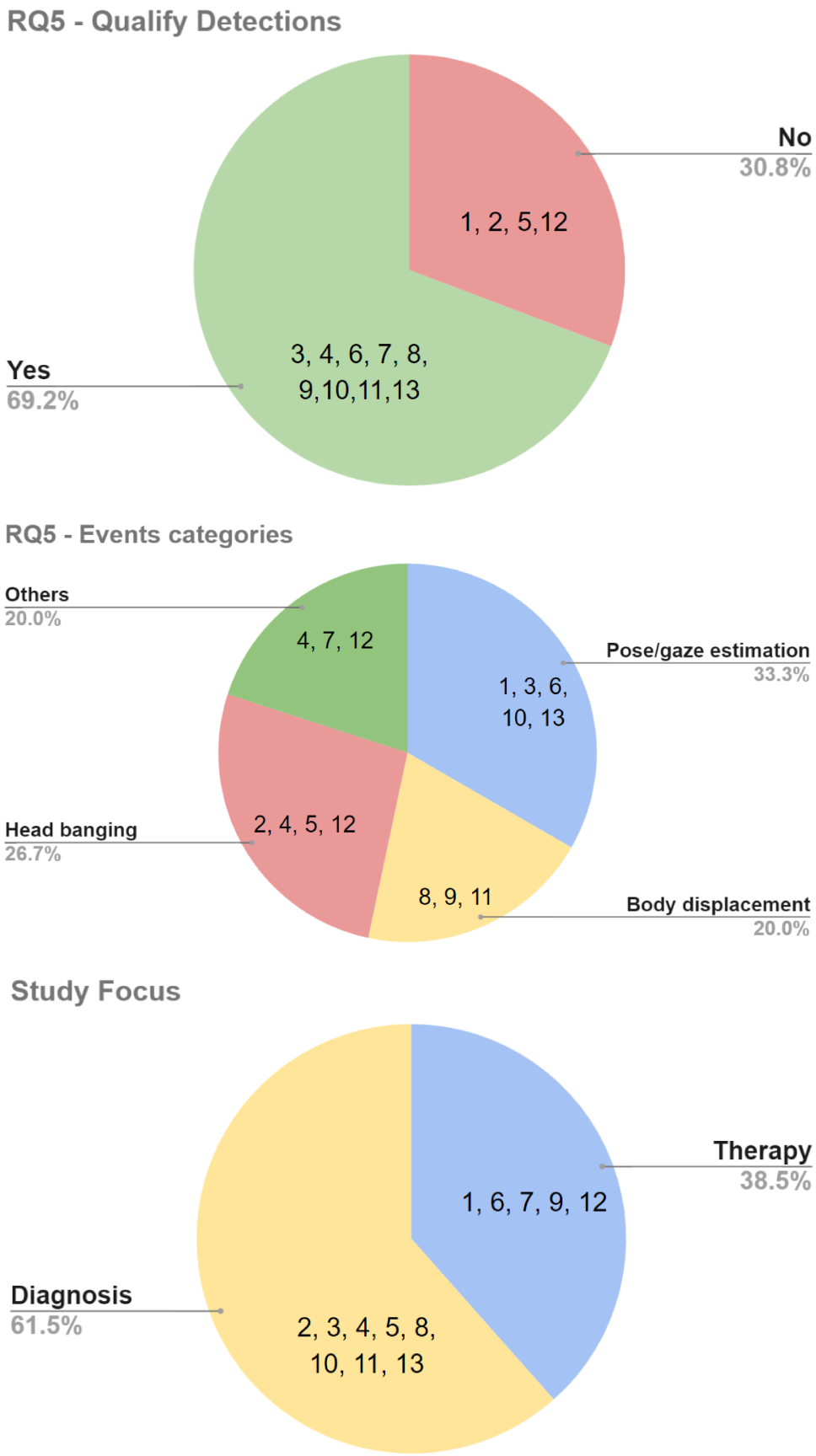


Figure 12 – Training loss, evaluation loss, and accuracy for the VideoMAE model fine-tuned on the subsets of ASBD

Source: the author

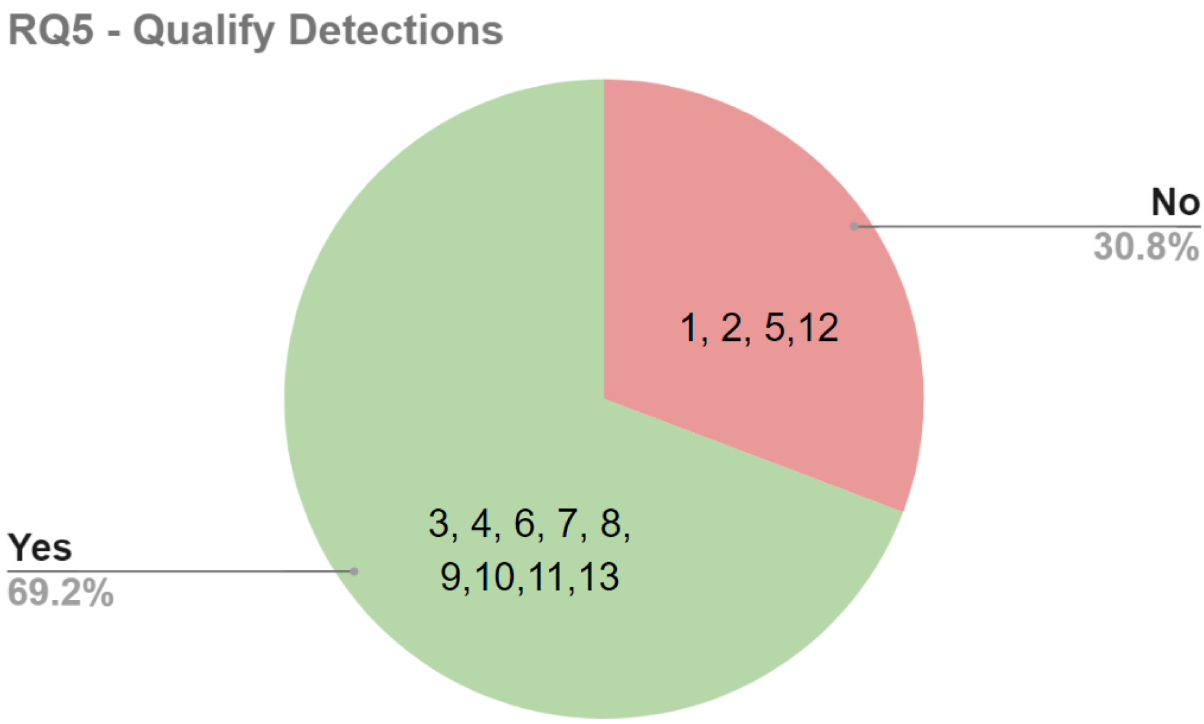


Figure 13 – RQ5 - Studies that categorize the detections

Source: the author

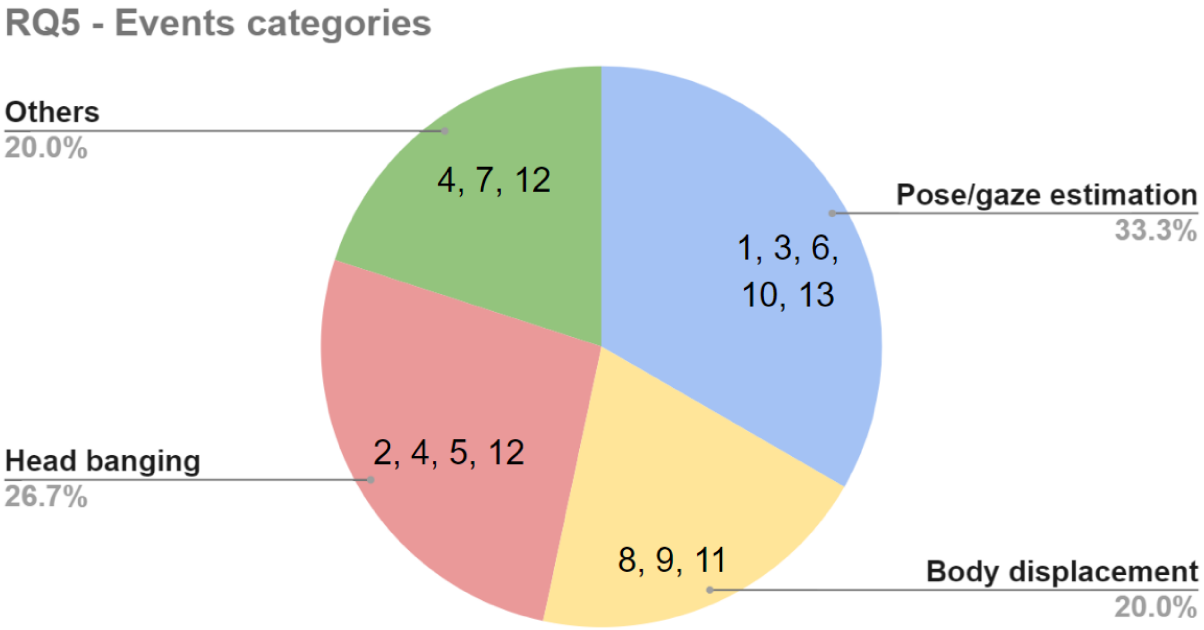


Figure 14 – RQ5 - Different categories of events

Source: the author

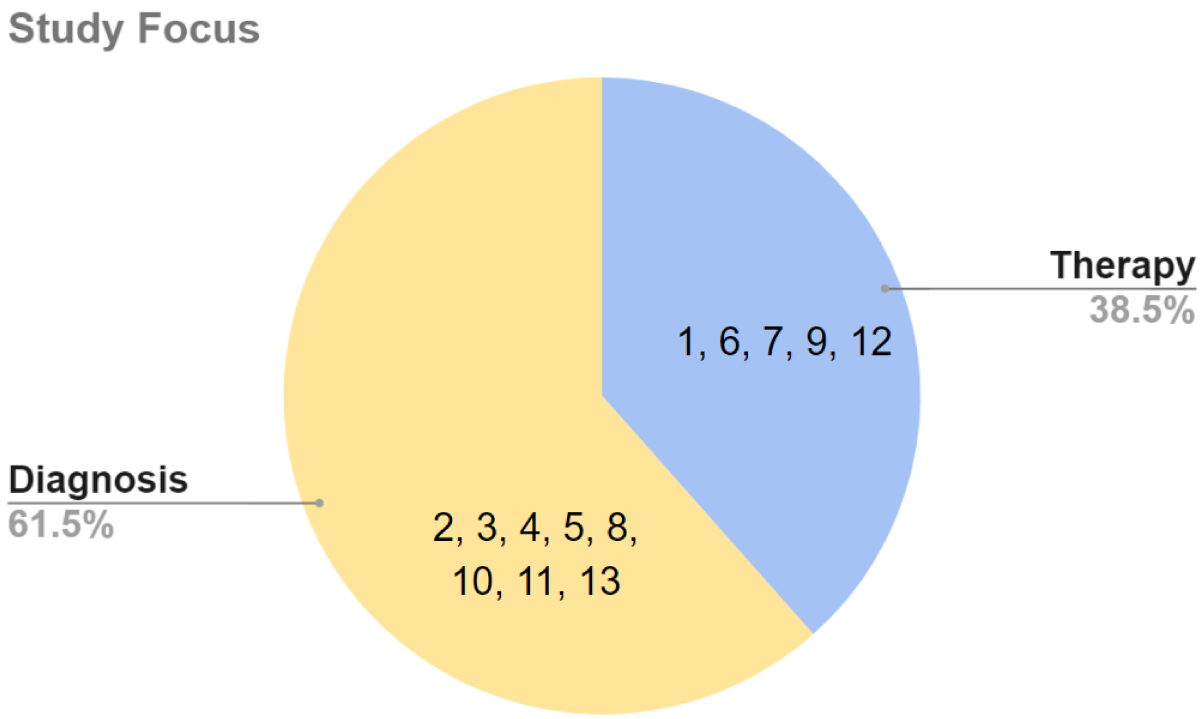


Figure 15 – Focus of the study

Source: the author

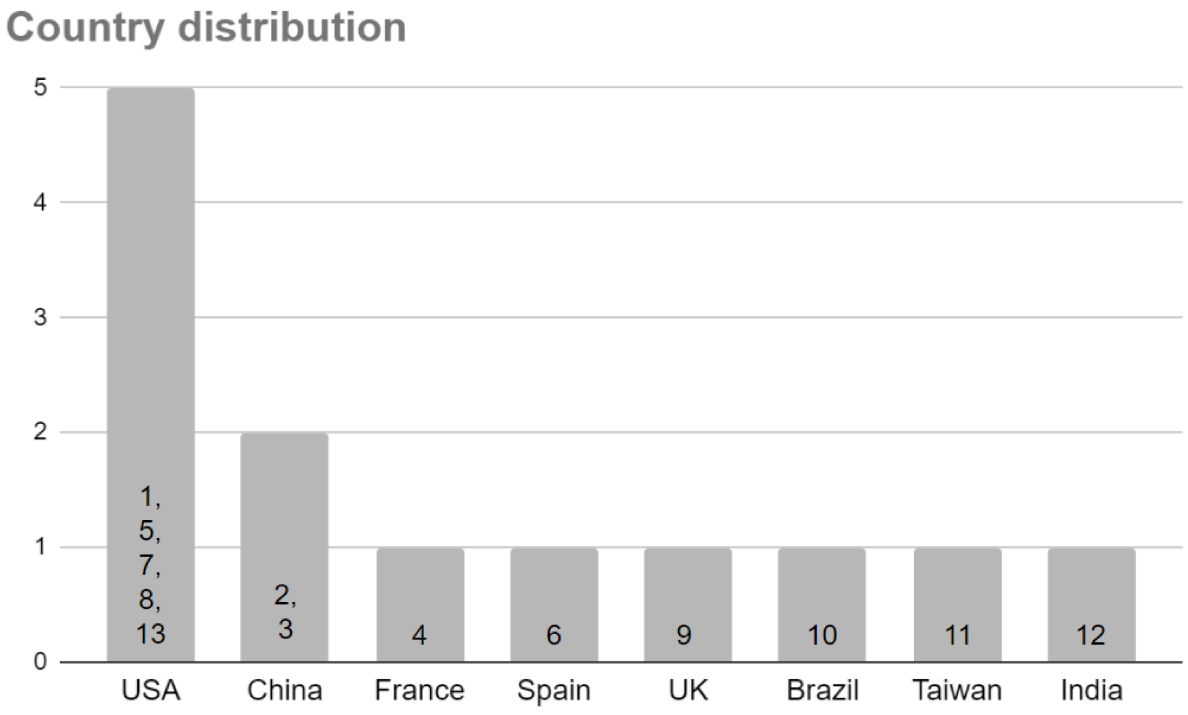


Figure 16 – Distribution of the selected studies by country

Source: the author

4 BEHAVIOR DATASETS

This section will present the three datasets that served as the basis for the ASBD proposal. The SSBD and Wei-SSBD ones are publicly available, whereas ESBD can be accessed by requesting access to the authors. The ESBD authors were contacted to request permission to use their dataset as part of our own extended version, to which they kindly agreed.

When proposing the first behavior dataset back in 2013, Rajagopalan et al. (RAJAGOPALAN et al., 2013) stated that, like the others *in the wild* datasets, behavior datasets could provide more domain-specific insights and often more clear and genuine behaviors since the actors are in an often known environment and are being filmed by their parents or caretakers, but also face challenges common to computer vision challenges in the task of automatic behavior analysis, associated with the uncontrolled environment characteristics of the videos, like camera motion, pose, illumination, cluttered background, video quality, occlusion, etc.

In addition to these challenges, the authors enumerate a few more that are specific to its domain:

1. Subtle behaviors: Detecting sporadic and brief stimming behaviors within videos, amidst other dominant activities.
2. Intensity and Continuity: Managing variations in behavior intensity and non-continuous occurrences.
3. Spatial Variance: Dealing with changing locations and potential interference from objects and people in the child's surroundings.
4. Social Cues: Considering the impact of interactions with caregivers or prompts on the child's behavior.
5. Mixing Behaviors: Addressing instances where the child transitions between different behaviors or combines them.
6. Object Influence: Separating the child's behavior from similar actions performed by associated objects.
7. Person Anxiety: Handling camera motions and reduced video quality caused by caregiver anxiety during intense stimming.
8. Context Stimming: Differentiating between the child's stimming and similar behaviors observed in external stimuli like TV programs or when the child is simply imitating said behavior.

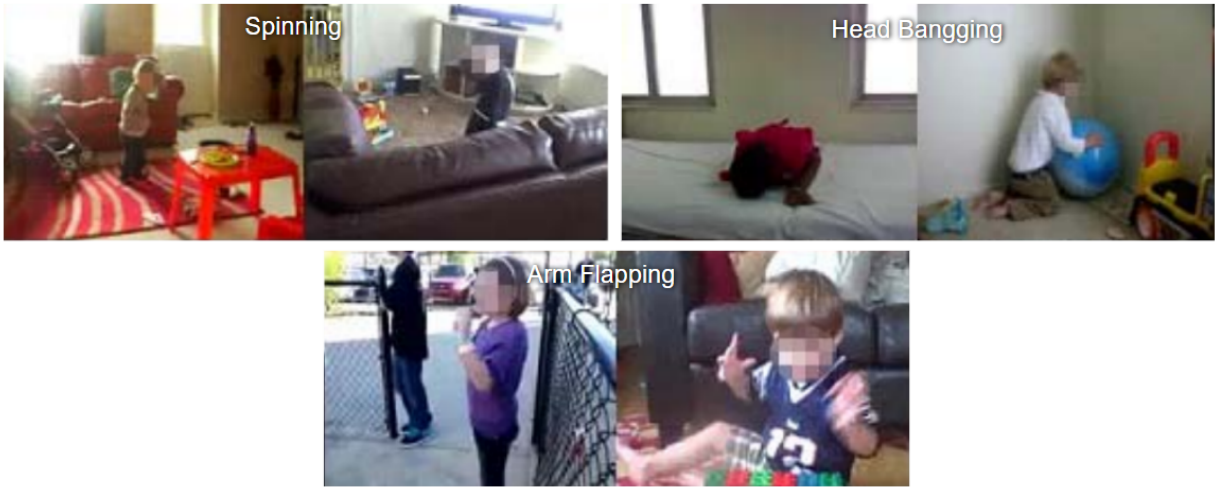


Figure 17 – Frames (blurred to preserve the anonymity of identity) of videos in all three stimming categories provided by (RAJAGOPALAN et al., 2013). The children exhibited different postures and were in different places. Moreover, seen are varying backgrounds, clutter, and multiple objects.

Source: (RAJAGOPALAN et al., 2013)

4.1 SELF-STIMULATORY BEHAVIOR DATASET (SSBD)

Rajagopalan et al. (RAJAGOPALAN et al., 2013) proposed in their work a dataset called Self-Stimulatory Behavior Dataset (SSBD) where videos recorded in *uncontrolled natural settings* can be used to help researchers enable their models to learn the "red flags" behaviors which prompt parents, caretakers, and therapists to search for specialized help. The authors were inspired by *in the wild* datasets like Labeled Faces In The Wild(HUANG et al., 2008) and Hollywood dataset (LAPTEV et al., 2008), which enabled big improvements in face recognition, human action recognition, and facial expression analysis fields and by MMDB dataset(REHG et al., 2013) that enabled studies of children dyadic behavior data.

The dataset consisted originally of 75 videos (59 still accessible at this moment), divided into three stimming behavior categories of 25 each (arm flapping, head banging, or spinning). The mean duration of the videos is 90 seconds and the minimum resolution for the videos is 320x240 pixels. Several extra attributes are available together with the dataset like body part, intensity, and modality. Sample snapshots of the dataset are shown in figure 17.

4.2 EXPANDED STEREOTYPE BEHAVIOR DATASET (ESBD)

Aiming to develop a framework capable of recognizing stereotype behaviors for early diagnosis of ASD, Negin et al. (NEGIN et al., 2021) propose a new dataset called ESBD -Expanded Stereotype Behavior Dataset, where two times more videos are collected together with an additional behavior (hand action) to complement the

Figure 18 – Frames (blurred to preserve the anonymity of identity) of videos in all three stimming categories provided by (NEGIN et al., 2021). Like SSBD, shows different backgrounds, camera angles, and clutter.

Source: (NEGIN et al., 2021)

three introduced by the SSBD dataset (arm flapping, head banging, or spinning). The hand action and arm flapping categories are distinct, where hand action focuses more on the finger actions and can help the detection of more actions than arm flapping.

In addition, the videos are labeled under the guidance of experts at the Child and Adolescent Psychiatrist Center at Ataturk University. Like the SSBD, the videos share the same *in the wild* characteristics, mostly captured by parents in daily living settings and there is no ground-truth information about the condition of the subjects in the videos, that is, there is no healthy vs. pathological annotation in the dataset.

There are no shared videos between the SSBD and ESBD and after manually clipping the target actions, there are 141 videos in total, a few samples are shown in figure 18, summing up to 108 subjects from which 76 are males and 32 are females. Since the dataset is not subject-oriented there are videos of different actions from the same subject. More detailed information about the dataset provided by the author can be found in Table 6. Even though the authors mention that the full dataset would be made public after the publication of the paper, the full and detailed dataset is still to be published by then, however, the author provided a list of the URLs and the main classification of the videos upon request by e-mail and gave the authorization to be made public by this work. The authors did not utilize the SSBD dataset in their paper, focusing the study on the novel ESBD dataset proposed by then.

4.3 WEI ET AL. EXPANDED SSBD (WEI-SSBD)

Wei et al. (WEI et al., 2022) start with SSBD and expand it by adding newly collected autism-related videos to form an extended dataset. They incorporated 12 new videos and removed 11 subjects that were considered noisy. The final version

Table 6 – Information about video quantity, frame number, and gender of the subjects in the ESBD dataset proposed by Negin et al. (NEGIN et al., 2021)

	Arm flapping	**Hand action**	**Head banging**	**Spinning**
Number of Videos	43	31	27	40
Min/Avg/Max number of frames	45/313/138	30/365/3828	30/258/1679	90/548/2545
Total number of frames	5938	11342	6982	21951
Male/Female	24/10	6/7	17/7	29/8

Table 7 – Number of videos of the original and the Extended SSBD dataset proposed by Wei et al. (WEI et al., 2022)

Number of videos	**Arm Flapping**	**Head banging**	**Spinning**
SSBD (original)	23	19	18
Extended SSBD	20	21	20

Figure 19 – Frames (blurred to preserve the anonymity of identity) of videos in all three stimming categories provided by (WEI et al., 2022).

Source: (WEI et al., 2022)

of the dataset used in their experiments is composed of 61 videos and 61 subjects, separated into 20, 21, and 20 videos for arm flapping, headbanging, and spinning, respectively. Table 7 shows a comparison made by the authors between the extended dataset and the original SSBD dataset. The classification introduced in the new videos was done by the authors themselves and a basic list composed of the video URL and each classification is publicly available in Github[1].

[1] https://github.com/Samwei1/autism-related-behavior/

5 PROPOSED SOLUTION

For the proposed solution, we explore the two main contributions of this work. First, we define the Autism Stimming Behavior Dataset (ASBD), a dataset that unifies and improves the usability of the already existing datasets. Using it we explore the capacity of a state-of-the-art video classification model (VideoMAE (TONG et al., 2022)) in capturing the video stimming behaviors characteristics and classify them accordingly. After that, we present and explore where this model, enabled by the ASBD, can fit and elevate the application of the event detection architecture for ASD therapy sessions.

5.1 AUTISM STIMMING BEHAVIOR DATASET (ASBD)

The SSBD dataset, published by Rajagopalan et al. (RAJAGOPALAN et al., 2013) dates back to 2013, and because it's composed basically of public links to Youtube shared videos, is subjected to degradation since these videos can be taken off without prior notice due to privacy or other reasons. Since its original publication, the list went from 75 videos to 59 accessible as we write this paper, a reduction of more than 20%.

A similar circumstance happened to ESBD, published by Negin et al. (NEGIN et al., 2021) in 2021. Originally with 118 Youtube links, only 97 are still accessible, showing a reduction of 17%.

The extended SSBD dataset, proposed by Wei et al. (WEI et al., 2022) in 2022, was built upon the SSBD original list and without prior access to the ESBD dataset videos, thus sharing most of its videos with these two datasets, adding only 9 distinct videos which are still available.

To alleviate the reduced number of videos, we combine these three datasets to create a consolidated one and also perform a work of curation by carefully analyzing and annotating each clip to show not only the video link and class but also the relevant behavior time instant, where we define the start and duration of the main stimming event, enabling a faster and streamlined access to the relevant events and providing a standard source (same clips, event start, duration and class) where baselines and comparisons can be built upon. The subject gender and the uniqueness of the subjects are specified as well (Gender and uniqueness are identified by video description, audio cues, or best visual guess when the first these are unavailable).

For the SSBD component, the original cuts and classification proposed by Rajagopalan et al. (RAJAGOPALAN et al., 2013) were maintained, even with a few longer clips that would later be considered noisy (WEI et al., 2022) and counter-producing for a action-recognition dataset (smaller videos with only one relevant action). We chose to maintain the original cuts to keep coherence among the other works that used the original clips for they research. For the ESBD(NEGIN et al., 2021) and Expanded SSBD(WEI

Table 8 – The Autism Stimming Behavior Dataset (ASBD) details

	Arm flapping	**Hand action**	**Head banging**	**Spinning**
Number of Videos	54	16	45	50
Unique Subjects	51	11	42	50
Average Stimming Duration	10.2s	7.5s	10.9s	11.7s
Male/Female	31/23	5/11	30/15	43/7

et al., 2022), since we had only the YouTube/google drive link of the video and the class associated to it, each single video clip were watched while looking for the most prominent cut that could explain the class that were originally attributed to it, by using our best guess to identify the underlying class (arm-flapping, head-banging, spinning and hand-action) and following the best practices for action-recognition datasets that aims to keep the video cuts short and concise, with and average of 10 seconds. By most prominent cut, we could say that we looked for the cuts where the action classes where dominant over other actions that could be happening, stronger or clearer movements that happens not much longer than a 10 seconds long window.

The Autism Stimming Behavior Dataset (ASBD) is comprised of 165 annotated short-duration clips, based on excerpts from 155 distinct publicly accessible YouTube videos, divided into 4 classes of stimming behavior (Arm flapping, Hand action, Head banging, and Spinning), with 154 unique subjects of ages spanning from toddlers to young adults and divided into 56 female and 109 male individuals. ASBD can be considered an 'in the wild' corpus, since it uses uncontrolled video scenes, as opposed to videos taken in lab environments or other scripted scenarios. The average duration of the clips is 10.6 seconds, with a maximum of 57 seconds and a minimum of 2 seconds. Other detailed characteristics like the distribution per category are shown in table 8 and the attributes available with their descriptions are available in table 9.

The dataset combines the strengths, challenges, and applications of other 'in the wild' datasets, enabling models to learn and recognize a wide variety of actions in conditions more similar to real-world scenarios, and the actions can be extrapolated from autism research to applications where the classes utilized could be relevant.

5.2 THE EVENT DETECTION ARCHITECTURE FOR ASD THERAPY SESSIONS

In the paper entitled "Event Detection in Therapy Sessions for Children with Autism" (RIBEIRO et al., 2022) we presented this architecture, where, considering the length and quantity of therapy sessions that an individual with ASD demands and the need for review of such recordings in search of relevant interactions for the therapy assessment, we proposed an end-to-end tool capable of providing an automated analysis of actors interactions in ASD therapy sessions - Figure 20. This work was further

Table 9 – The Autism Stimming Behavior Dataset (ASBD) attributes description

Attributes	Description
Source	Source dataset
Nro	Sequence number
Category	Stimming behaviour category
Videoclip	Unique videoclip name
URL	Reference website URL to the video
VideoDuration(s)	Whole video duration in seconds
EventStart(s)	The instant where the Stimming behaviour starts in seconds
EventEnd(s)	The instant where the Stimming behaviour ends in seconds
EventDuration(s)	The duration of the event in seconds
Availability	Determine if the video is publicly available in the Youtube or Google Drive (at time of publication)
Gender	Gender identified by video description, audio cues or best visual guess when the first 2 are unavailable
Unique Subject	Uniqueness of the subject based on visual appearance - Variability of dataset

expanded by Soares (2023) with a much richer analysis and methods. We present here the general architecture and where we aim to fit and explore new opportunities for stimming behavior detection.

5.2.1 PlusMe Dataset

The data used for training and testing the proposed solution are videos of therapy sessions for children with ASD. It was provided by the European project PlusMe [1] as supplementary material published in Sperati et al. (2020).

Typically, in these therapy sessions, there is a toddler, a caretaker, and a teddy bear interactive cushion, which is named after the aforementioned project, PlusMe(+me). PlusMe is an experimental interactive soft toy developed in collaboration with neurodevelopmental therapists. The shape, material, and functionality are designed with the concept of Transitional Wearable Companion (TWC) which is a “smart” companion toy, neither too complex as a standard robot nor too simple as a teddy bear(GIOCONDO et al., 2022).

During a session, the toddler interacts with the +me, the caretaker and the environment around (see Figure 20 (GIOCONDO et al., 2022)). According to the observation techniques mentioned earlier, these interactions and their characteristics are important factors that could serve as evidence of the cognitive and social development of the toddler.

[1] https://www.plusme-h2020.eu/

The dataset consisted of a total of 8 videos of different sessions with different toddlers. These videos were divided into 4 for the training phase, 2 for the validation phase, and finally 2 for the test. For each video in the training and validation phases, 600 frames were manually labeled by identifying the bounding boxes of each actor in each frame. For the test phase, 200 frames were labeled with the interactions among actors.

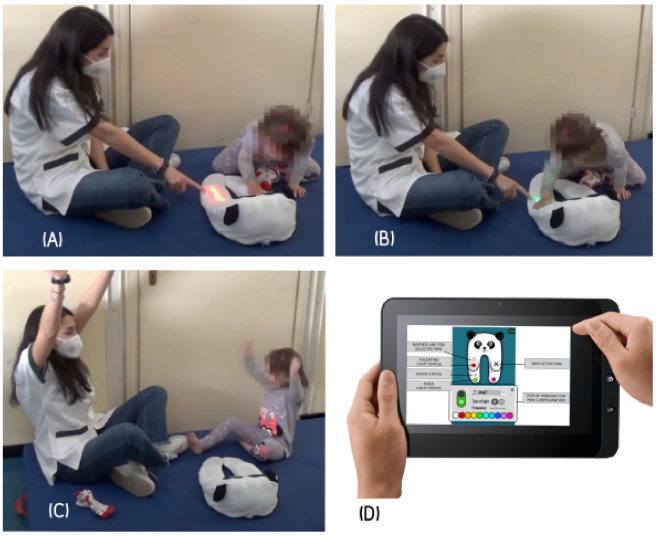


Figure 20 – An example of the experimental setting. A) the therapist points to the red blinking panda's paw, during the whack-a-mole activity; B) the child touches the paw, which reacts by changing color to green and emitting a brief song; C) child and therapist rejoice for the reward; D) the control tablet in the hand of the experimenter (in the same room)

Source: Giocondo et al. (2022)

5.2.2 The framework

The first step is to process the video-recorded session through a Computer vision tool that aims to detect, classify, and retrieve the bounding boxes delimiting the existing actors in each frame. The actors in this context are specifically the toddler, the caretaker, and the PlusMe device. These bounding boxes alone do not indicate any sort of interaction between actors, nor any notion of interaction length or type, only their position in each frame. Therefore, the second step is a bounding box event detection mechanism that processes information from the boxes' coordinates and identifies the overlap among them. Finally, based on the overlaps, an interaction timeline is built showing in which frames of the videos there are potential interactions. This functionality helps the therapists' team to rapidly identify the highlighted video snippets where interactions between the actors potentially occur, streamlining the process of video

assessment, saving time, and reducing the overall costs of ASD research, diagnosis, and therapy - Figure 21.

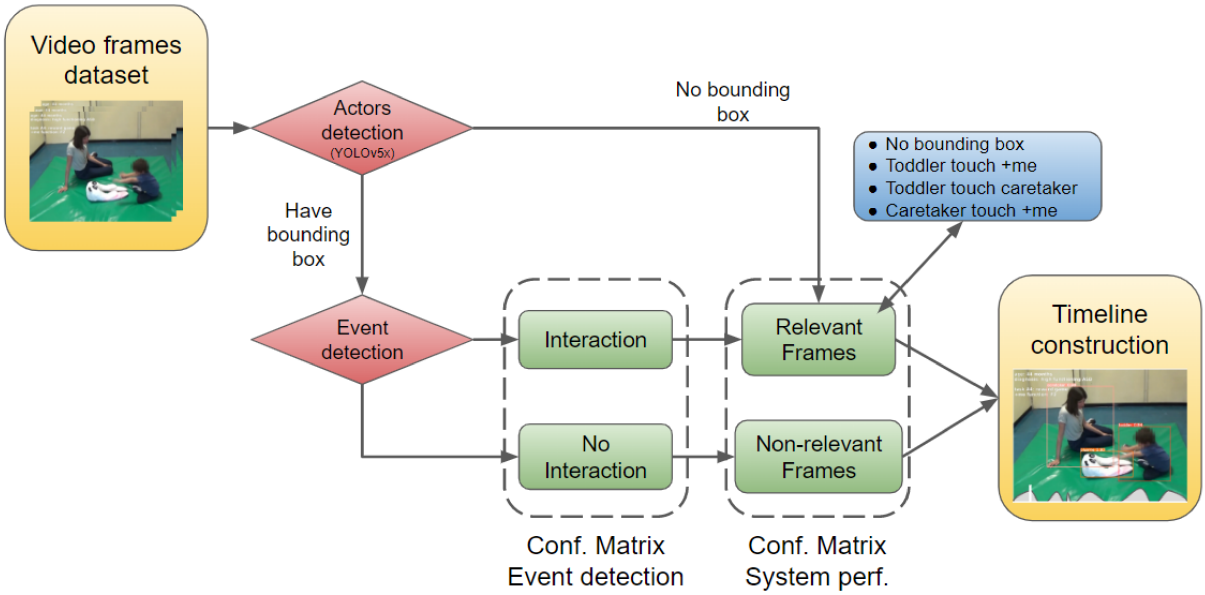


Figure 21 – Initial proposed solution outline. First, the content of a recorded therapy session is loaded into a Computer vision tool, generating a bounding box for each actor. Next, a heuristic-based event detection mechanism outputs relevant information about interactions.

Source: the author

5.2.2.1 Actors detection tool - YOLOv5

For the task of actors detection, we employed the YOLOv5 model available from the Ultralytics repository [2] under the GNU General Public License v3.0, and implemented with the open-source machine learning framework PyTorch 2 [3]. The only modifications concerning the default hyper-parameters were changing the batch size to 1 and the image size to 256x256 pixels. Such modifications had to be done due to memory limitations of the training environment (GPU) detailed later. The network was trained using transfer learning with the default YOLO network (trained on COCO dataset [4]), and the training dataset.

The result of this step is a model capable of processing any video within the trained characteristics and detecting the actors, correctly classifying them, and generating a correspondent bounding box for each frame. The bounding boxes output, adjusted to a pixel-coordinate format is then used as an input for our next step.

[2] https://github.com/ultralytics/yolov5
[3] https://pytorch.org
[4] https://cocodataset.org/

5.2.2.2 Bounding box event detection

Based on the actors bounding boxes detected in the previous step, we use an event detection algorithm built upon the IoU (Intersection over Union) operation between each bounding box coordinates, generating information on the "perceived interaction" between the actors. The IoU operation is largely used to evaluate how two bounding boxes interact and is even used inside the YOLOv5 algorithm to evaluate and further improve its own prediction(JOCHER et al., 2022; REDMON et al., 2016). The event detection tool at this current implementation detects the intersection in a binary form of a given frame. Since a simple overlap (big or small) can't characterize an interaction by itself, we treated this as a "predicted interaction". Such predictions are then evaluated against the "ground truth" to evaluate the accuracy of the event detection tool. For simplicity, in this iteration of the system, the interaction between the actors is defined by any overlap between their bounding boxes. Once again, the results of this step are adjusted accordingly and then used as input for the next and final step of the proposed solution.

5.2.2.3 Event detection timeline construction

Finally, we use the detected bounding box interaction information from the previous step, group them accordingly, and merge this information as an overlay on the original video, highlighting the actors' interaction frames and enabling a faster evaluation of the interaction between them in the recorded video therapy.

5.2.2.4 Stimming behavior detection

As an add-on to the framework proposed in the paper (RIBEIRO et al., 2022), further developed by Soares (2023) and enabled by the ABSD dataset (RIBEIRO; GRELLERT; CARVALHO, 2023), we aim to fine-tune a model able to capture and classify stimming behavior of the videos uploaded to the event detection architecture mentioned earlier, adding a new layer of confidence in the process to diagnose ASD and help the underlying therapy sessions by pointing out the classes and quantity of stimming behaviors detected through the videos - Figure 22.

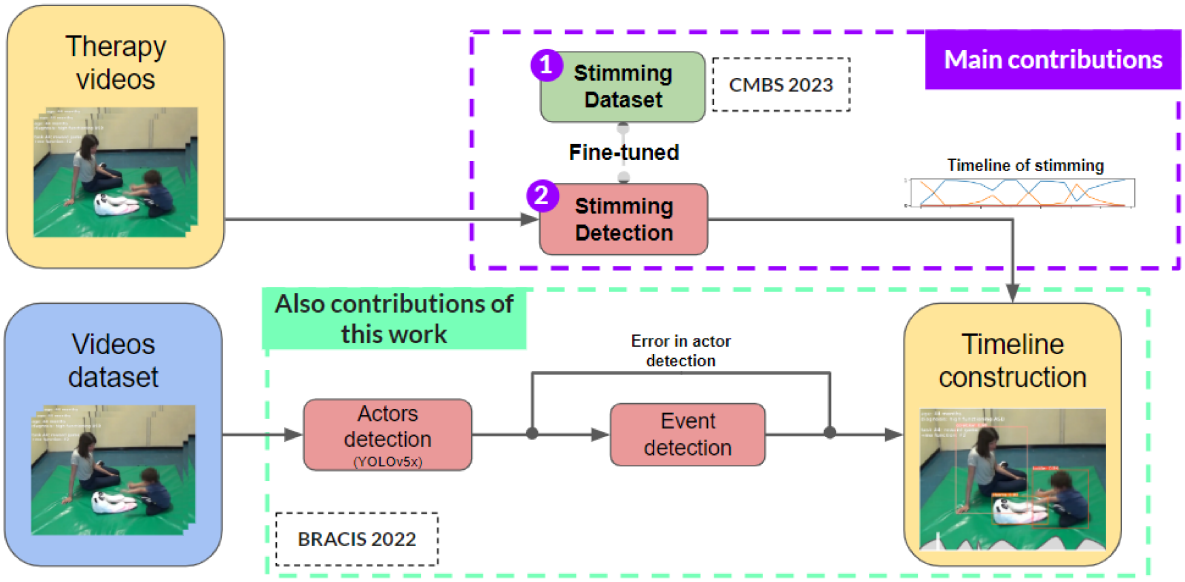


Figure 22 – Updated solution outline, with the Stiming Detection add-on component.

Source: the author

6 EXPERIMENTS AND DISCUSSIONS

This chapter is divided into two sections. The first one, Section 6.1.5, is related to the main contribution of this thesis, which is the ASBD augmented dataset for stimming behavior detection. The section presents a series of experiments that evaluate and compare the proposed dataset with alternative solutions. In Secion 6.2, we introduce the stimming behavior detection model (VideoMAE) in web application that was developed by a previous study from our group.

6.1 EXPERIMENTS ON ASBD

In this section, we make a series of experiments involving the unified ASBD dataset. These experiments are shown in the figure 23 and detailed in the following sections. The gray boxes represent the proposed dataset ABSD. Each row represents an experiment made, and in the columns, we specify the different configurations on, training data, the model that is being trained on, and the evaluation test where its results are compared against. Four experiments are discussed: in Experiment 1, we select two models (I3D and VideoMAE) pre-trained on the Kinectics400 dataset and employ them in the ASDB dataset directly (no fine-tuning); in Experiment 2, we train the VideoMAE model with existing datasets (SSBD and ESBD) and then assess their performance on different test sets; Experiment 3 involves training the VideoMAE model with the ASBD dataset in order to evaluate if it is able to improve the stimming detection task; finally, in Experiment 4, we train the VideoMAE with augmented versions of the datasets in order to further improve their performance. For details regarding the training phase parameters, please refer to Section 6.1.2. At the end of this section, a summarizing discussion is presented, and the best results obtained in this work are compared with related solutions from the literature.

6.1.1 Comparing I3D and VideoMAE pre-trained on Kinetics-400

6.1.1.1 I3D model pre-trained on Kinetics-400

Carreira and Zisserman (CARREIRA; ZISSERMAN, 2018) proposed in 2017 a groundbreaking large-scale dataset (Kinetics-400) for action recognition in videos together with a new model called the Two-Stream Inflated 3D ConvNet (I3D) which uses both spatial and temporal information to improve action recognition accuracy. It helped to advance the field of action recognition in videos and became a reference to the area at the time.

This classic action recognition model pre-trained in the Kinetics-400 dataset[1] is capable of recognizing 400 action classes in a wide array of situations and shares

[1] https://paperswithcode.com/paper/the-kinetics-human-action-video-dataset

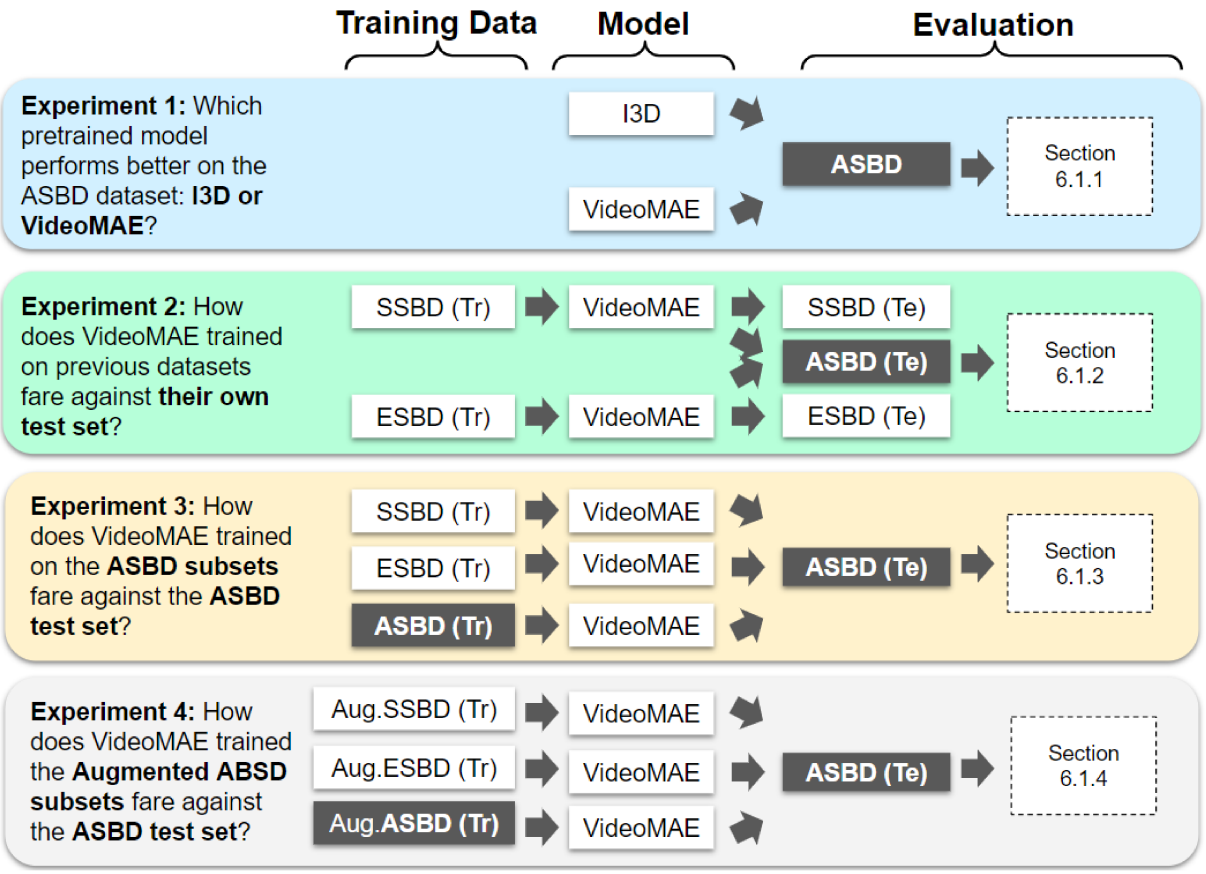


Figure 23 – Detail of the experiments

Source: the author

some classes that could characterize the classes that we are looking to identify as stimming behavior. Also, The Kinetics-400 dataset was chosen because it shows a good relevance when searching for datasets used in video research[2]. A few of those classes, as discussed below, are directly correlated to classes that we are aiming to classify within the ASBD.

When presenting our dataset to the I3D model[3] we can notice that a few classes become apparent in explaining our desired classification classes even without any kind of training within it, like 'ice skating' and 'dancing ballet' for our 'spinning' category reaching an average prediction of around 42% and 38% respectively, 'tapping guitar' and 'finger-snapping' for our 'hand action' category with 60% and 38%. Headbanging classes share the same name in both datasets but still, the model didn't recognize it as the primary class for the category, reaching 48% of the average predictions while the 'brushing hair' class achieved 51,5% of average probability and the arm flapping category was identified primarily as 'sign language interpretation' class with 47% probability, in Kinetics-400 as seen in table 10.

As we can see, the categories that we are looking for in the autism stimming be-

[2] https://paperswithcode.com/datasets?mod=videos&page=1
[3] https://github.com/deepmind/kinetics-i3d

Table 10 – Predictions for the ASBD on I3D model pre-trained on Kinetics-400 database

Spinning		**Head banging**	
ice skating	42.3%	brushing hair	51.4%
dancing ballet	38.5%	headbanging	47.9%
gargling	33.9%	crawling baby	38.5%
faceplanting	29.1%	pushing cart	33.3%
drop-kicking	27.0%	brushing teeth	31.2%
Hand Action		**Arm Flapping**	
tapping guitar	59.9%	sign lang.	47.1%
finger snapping	37.9%	air drumming	36.5%
playing poker	37.5%	jogging	29.9%
washing hands	32.8%	brushing teeth	27.4%
clapping	28.7%	finger snapping	26.8%

havior study aren't well explained by the classes that exist in one of the most prominent action recognition datasets and there is a need to update models by using machine learning pipelines to achieve the desired classification.

6.1.1.2 VideoMAE model pre-trained on Kinetics-400

Looking to see how the more recent VideoMAE video classification algorithm behaves in the same circumstances, we use the model pre-trained in the very same Kinetics-400 and present it again to our dataset, the results are presented in table 11. When comparing against the I3D model tested earlier, it seems that the videoMAE model couldn't capture enough characteristics to be able to identify humane relatable actions within the range of actions that exist in the kinetics classes, where actions like the relatable "headbanging" class fell to only 22% when comparing to the 48% observed in the I3D model. Nevertheless, we aim to explore further the videoMAE capacity with fine-tuning techniques and see how able it can be to classify the main ASD stimming behavior classes that we are studying.

6.1.2 VideoMAE model pre-trained on Kinetics-400 - fine-tuned on ASBD and its subsets

6.1.2.1 VideoMAE fine-tuning configurations

Looking to explore the differences of each subset of the ASBD and how much impact the unification of the datasets can have in the recognition of the stimming behavior actions, we fine-tuned the VideoMAE model pre-trained on Kinetics-400 on each of the 2 main subsets, SSBD and ESBD and finally with the whole ASBD dataset.

The amount of videos when taking the subsets into consideration and its 4 underlying classes is somewhat small, even with a unified dataset as shown in table 8. With that in mind, a rough 80% - 10% - 10% split approach on the training, validation,

Table 11 – Predictions for the ASBD on VideoMAE model pre-trained on Kinetics-400 database

Spinning		**Head banging**	
bending back	63.1%	clean. windows	73.3%
pushing cart	39.7%	using computer	65.4%
gargling	30.5%	climbing ladder	28.2%
playing frisbee	24.4%	eating chips	22.3%
yoga	22.8%	headbanging	22.1%
Hand Action		**Arm Flapping**	
eat. spaghetti	34.8%	brushing teeth	37.5%
massaging feet	24.0%	eat. watermelon	35.8%
drumm. fingers	23.1%	news anchoring	34.6%
shaking head	18.2%	crawling baby	27.1%
sign lang.	16.7%	reading book	24.5%

Table 12 – Training, evaluation, and test splits detail on each subset and the four action classes presented in the ASBD

	Arm flapping			**Hand action**			**Head banging**			**Spinning**		
	Tr	**Val**	**Te**	**Tr**	**Val**	**Te**	**Tr**	**Val**	**Te**	**Tr**	**Val**	**Te**
SSBD	17	2	2	-	-	-	15	2	2	15	2	2
ESBD	26	3	3	12	2	2	16	2	2	23	3	3
Wei e-SSBD	1	-	-	-	-	-	3	2	1	2	-	-
ASBD	44	5	5	12	2	2	34	6	5	40	5	5

and test sets was made taking into consideration data leakage between the sets since there are a few videos that show actions of the same subjects, the final distribution is shown in the table 12.

The VideoMAE model was built over the Transformers library of the Hugging face framework[4] using the default values except the batch size and epochs parameters described in the table 13. The internal value "max steps" that is used in the following charts and tables is calculated using the equation(1), where the batch size and epochs are used to determine the number of steps that are needed to make the whole training set go through the network "n" times defined by the initial "epoch" value. Similarly, the "test clips" quantity is also a variable associated with the preprocessing pipeline used in the bigger ASBD model, where a few techniques of video preprocessing are applied when dealing with each sample of the test dataset. All the tests presented in the next sections were executed using this "fixed" test dataset - unless explicitly mentioned - (68 distinct clips cut from the 17 video samples in the test dataset) looking to maintain fairness on the analysis when comparing models trained in the different datasets. The exceptions are those tests when we compare each subset of the test dataset and how they would fare in a bigger test dataset (SSBD vs ASBD (figure 24) and ESBD vs ASBD

[4] https://huggingface.co/docs/transformers

Table 13 – Model and dataset configuration - VideoMAE experiments on the ASBD dataset and its subsets

	SSBD	**ESBD**	**ASBD**
Training videos	47	77	130
Validation videos	6	10	18
Test videos	6	10	17
Test clips	21	43	68
Batch size	12	12	12
Epochs	50	50	50
Max steps	150	300	500

(figure 25)).

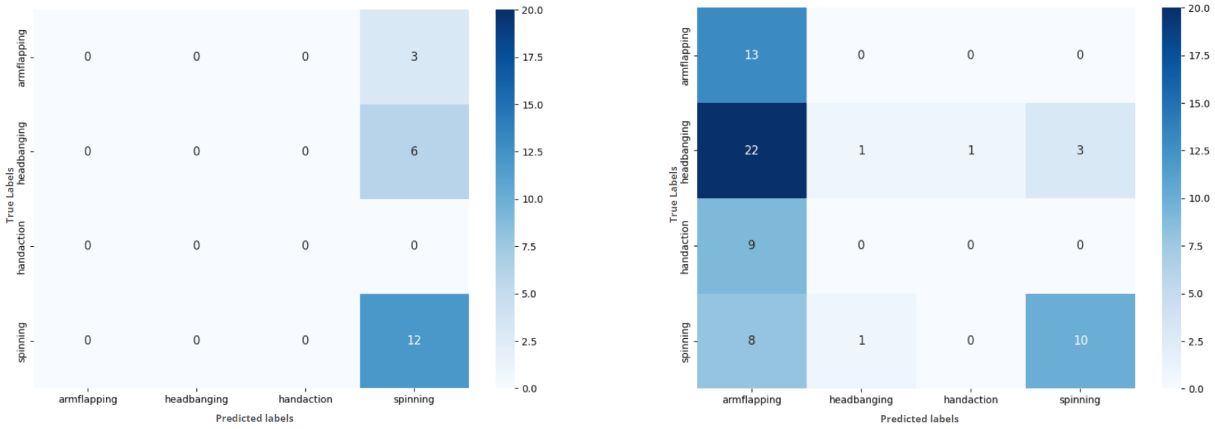
$$max\_steps = (\frac{num\_train\_videos}{batch\_size}) * num\_epochs \quad (1)$$

We used the Google Colab[5] environment on its premium plan to access an "A100 GPU" setup that is capable of accommodating videos in batches of 12 on its 40 GB of RAM. Also, the epochs configuration value was used upon verification that the training and evaluation loss were decoupling, meaning that the model became overfitted for the current data and configuration.

6.1.2.2 VideoMAE on the subsets of ASBD - on their own training and test sets

The limitations of each subset on its own can be further observed if we analyze how the fine-tuned model performs on the limited test set of each subset of ASBD. In the figure 24, we present the confusion matrices of the VideoMAE model fine-tuned only in the SSBD dataset when being tested against its smaller 6 videos test set (21 clips) and the bigger 17 videos of the ASBD test set (68 clips). We can observe an apparent bias towards the spinning class existing on the smaller dataset, raising the hypothesis that the model learned that by classifying most of the samples in the spinning class since, that way, it could achieve the best result in the reduced training samples. When making the same analysis on the ESBD dataset, shown on the figure 25, we can see an improvement in the correct classification of all classes but most prominent in the headbanging class, showing a good complement between the test sets and that the model is indeed slightly better than what could be verified if only the ESBD would be used on its own.

[5] https://colab.google/

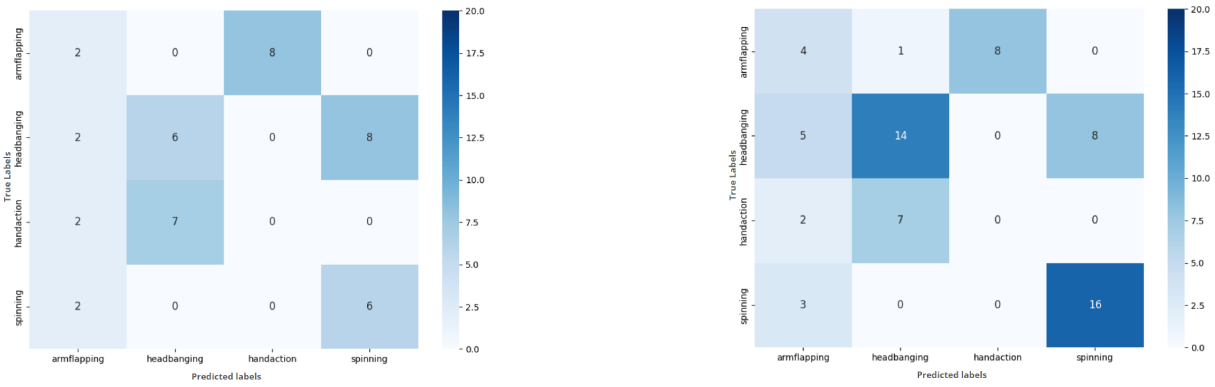


(a) Experiment 2 - VideoMAE model trained and tested only on the SSBD dataset

(b) VideoMAE model trained on SSBD and tested in the ASBD Test Dataset

Figure 24 – Comparison evaluating how the VideoMAE model trained only on the SSBD dataset would perform in its own test dataset and the bigger ASBD test dataset

Source: the author



(a) VideoMAE model trained and tested only on the ESBD dataset

(b) VideoMAE model trained on ESBD and tested in ASBD Test Dataset

Figure 25 – Experiment 3 - Comparison evaluating how the VideoMAE model trained only on the ESBD dataset would perform in its own test dataset and the bigger ASBD test dataset

Source: the author

Table 14 – Validation and test results - VideoMAE experiments on the ASBD dataset and its subsets

	SSBD	**ESBD**	**ASBD**
Best validation Loss	1.18	0.62	1.66
Best validation Accuracy	0.32	0.78	0.68
Test Loss	0.96	1.61	0.79
Test Accuracy	0.35	0.50	0.69

6.1.3 VideoMAE on the subsets of ASBD - on their own training sets and on a unified test set (ASBD (Te))

As we can observe in the charts presented in figure 26, which shows the evolution of the training loss, evaluation loss, and evaluation accuracy over the training steps, the model trained on the ESBD subset managed to reach the best accuracy metric on the training process, but as can be seen in the table 14, when presented to the same test set, the ASBD managed a far better result in predicting the stimming behavior actions.

These behaviors can be further analyzed in the confusion matrices of those models presented in figure 27. The model trained on the SSBD dataset appears to strongly overfit towards the arm flapping class, while the model trained only on the ESBD dataset didn‘t manage to capture well that class, spreading its predictions over every other class. Beyond that, the ESBD trained model managed to get a good grasp of both the spinning and head-banging actions achieving good accuracy in those classes. The model trained in the bigger ASBD dataset achieved the best overall accuracy, being able to correctly identify most of the arm flapping, head-banging, and spinning samples, while still looking slightly biased towards the arm flapping class when dealing with the head banging and hand action classes. Every model had a hard time when dealing with the smaller of the classes, hand-action, probably associated with the far smaller training dataset samples since it was a class proposed only by the later ESBD dataset.

6.1.4 ASBD data augmentation

Cauli and Reforgiato Recupero (2022) explored video data augmentation techniques focused on deep learning models, they mention that when the focus is on human action recognition or prediction, the skeletal animation of the subjects is needed to simulate the motion. Also in domain randomization methods, camera motion must be taken into account and the variation in textures, illumination, and object shapes must be constant or coherent throughout the entire video sequence. Taking that into consideration, we applied 2 simple techniques for temporal data augmentation, aiming to improve the number of videos used in the training dataset, horizontal flipping, and brightening, some samples are shown in figure 29.

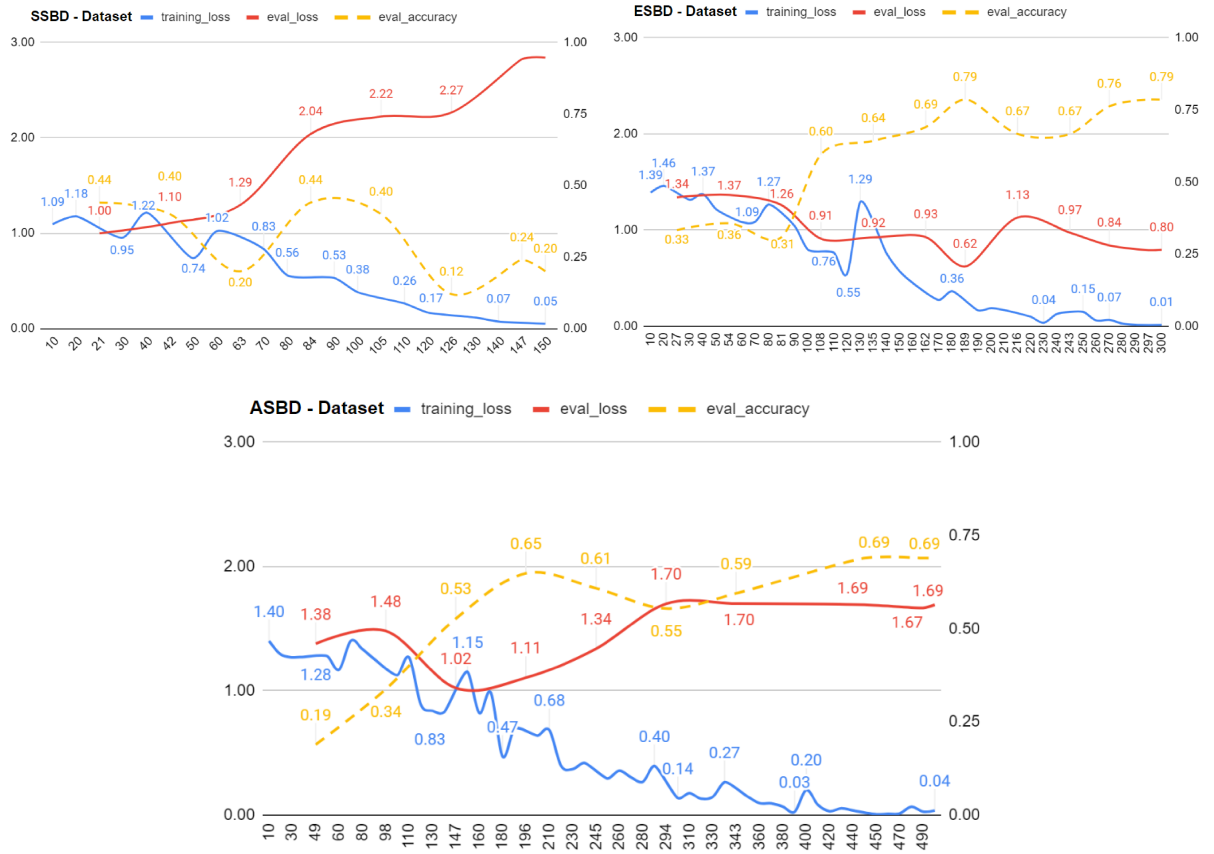
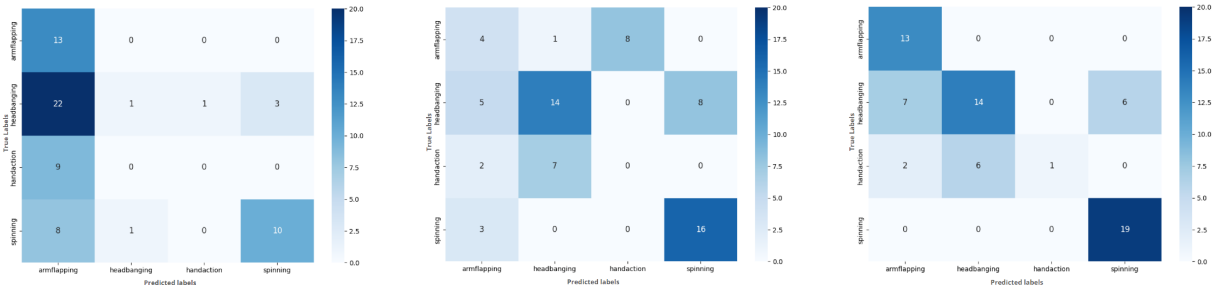


Figure 26 – Training loss, evaluation loss, and accuracy for the VideoMAE model fine-tuned on the subsets of ASBD

Source: the author



(a) VideoMAE model trained on SSBD and tested in ASBD Test Dataset

(b) VideoMAE model trained on ESBD and tested in ASBD Test Dataset

(c) VideoMAE model trained and tested on ASBD

Figure 27 – Experiment 1 - Confusion matrix of the models trained on each subset of the ASBD dataset and tested in the same test data of the ASBD dataset

Source: the author

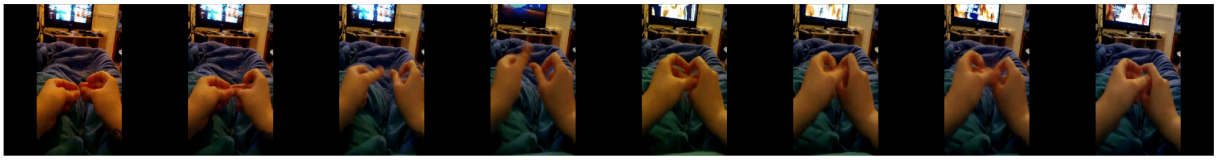
With those techniques, we can effectively double the training dataset by applying horizontal flipping and brightening to each of the training videos, and we can check the new number of training videos and max steps within the table 15.

Even though the augmentation techniques used were somewhat simple, they

(a) Sample v_HeadBanging_22, clip 0, correctly classified in all datasets

(b) Sample v_HeadBanging_24, clip 2, correctly classified on ESBD fine-tuned model, miss-classified on ASBD and SSBD models as Spinning

(c) Sample v_Spinning_25, clip 1, correctly classified in all datasets



(d) Sample v_HandAction_22, clip 4, miss-classified in all datasets as headbanging (correctly in augmented ASBD and ESBD)

Figure 28 – Analysis of a few samples of videos with their respective labels and predictions on different datasets

Source: the author

Table 15 – Model and dataset configuration - VideoMAE experiments on the augmented dataset

	SSBD Augm	**ESBD Augm**	**ASBD Augm**
Training videos	94	154	260
Validation videos	6	10	18
Test videos	6	10	17
Test clips	21	43	68
Batch size	12	12	12
Epochs	50	50	50
Max steps	350	600	1050

Figure 29 – Augmentation samples - All training videos were horizontally flipped and brightened by a factor of 1.5.

Source: the author

managed to improve the performance on both the evaluation and test datasets of the models trained in all subsets of ASBD, improving the test accuracy of the ASBD in 0.7 point as we can see in the table 16.

In figure 30, we can observe that the results and trend lines of training loss, evaluation loss, and accuracy are somewhat similar to the results observed with the models trained on the ASBD. On the training and evaluation sets, again, the model trained on augmented ESBD outpaced both SSBD and the more complete ASBD, but it also underperforms on the test set. The model fine-tuned on the augmented ASBD managed to achieve a 0.76% accuracy on the test set, while the ESBD one increased its accuracy to 0.56% and the SSBD one had a performance decrease to 0.35%.

The confusion matrices of the classification of the video test samples by these models on the ASBD test dataset are presented in figure 31. We can notice that the model trained on the augmented SSBD subset lost the ability to correctly classify the spinning class, that was present on the original model, missing all the samples of that class. Also had a significant increase in the bias toward the arm flapping class, but managed to increase the accuracy on the head-banging class, something that was pretty much nonexistent in the original model. The model trained on the augmented ESBD dataset kept good performance in detecting the spinning and head-banging classes while increasing the bias of miss-classifying some samples of the headbanging

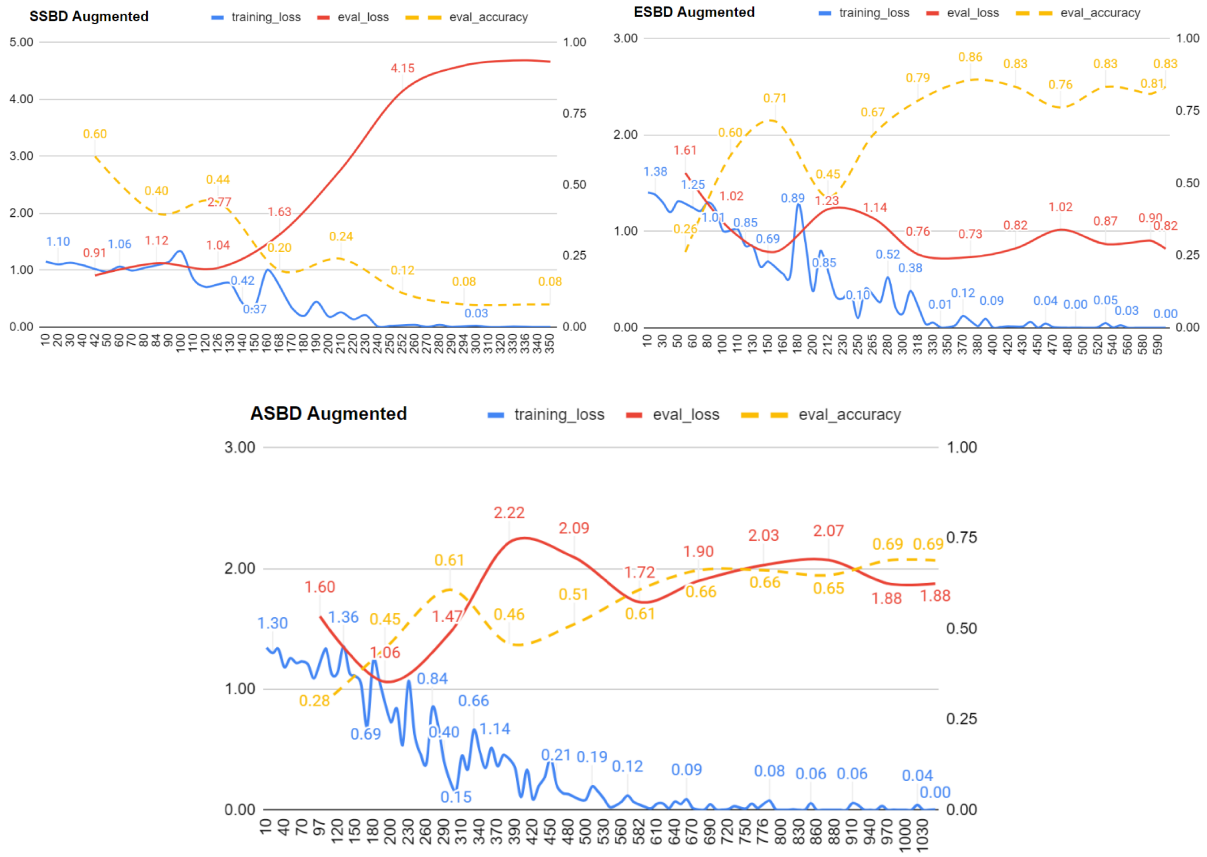


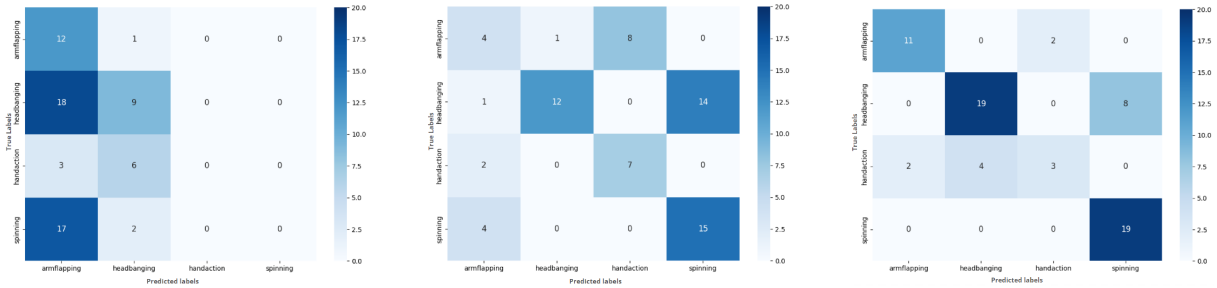
Figure 30 – Training loss, evaluation loss, and accuracy for the VideoMAE model fine-tuned on the subsets of ASBD Augmented by video augmentation techniques

Source: the author

Table 16 – Validation and test results - VideoMAE experiments on the augmented dataset

	SSBD Augm	ESBD Augm	ASBD Augm
Best validation Loss	0.90	0.73	1.87
Best validation Accuracy	0.60	0.85	0.68
Test Loss	0.90	2.47	1.00
Test Accuracy	0.31	0.56	0.76

class towards the spinning category, on a good point, the new samples managed to help it correctly capture the characteristics of the hand-action class, that wasn't present on the original model. Finally, on the model trained with the full ASBD dataset, we can see that the ability to correctly classify the spinning and head-banging classes was kept while improving a pretty much nonexistent capacity to detect the hand action class. More complex techniques like video mixing and temporal cropping(CAULI; REFORGIATO RECUPERO, 2022) are to be explored in future works.



(a) VideoMAE model trained on the augmented SSBD and tested in ASBD Test Dataset

(b) VideoMAE model trained on the augmented ESBD and tested in ASBD Test Dataset

(c) VideoMAE model trained and tested on the augmented ASBD

Figure 31 – Confusion matrix of the models trained on each subset of the augmented ASBD dataset and tested in the same test data of the ASBD dataset

Source: the author

Table 17 – Consolidated validation and test results - VideoMAE model fine-tuned on the subsets of ASBD

Fine-tuning dataset	Validation acc.	Test acc. (ASBD dataset)
SSBD	0.32	0.35
SSBD Augm	0.60 (+0.28)	0.31 (-0.04)
ESBD	0.78	0.50
ESBD Augm	0.85 (+0.07)	0.56 (+0.06)
ASBD	0.68	**0.69**
ASBD Augm	0.68 (≈0.0)	**0.76** (+0.07)

6.1.5 Summary of the Experiments and Related Work Comparison

Following the specific objective presented on the section 1.1, the consolidated results of the VideoMAE model pre-trained on the Kinects-400 dataset and fine-tuned on each subset of the ASBD and its augmented counterparts are shown on the table 17. The best result achieved, of 0.76% accuracy, fine-tuned on the ASBD augmented dataset, shows a good starting point where updated models like the VideoMAE v2((WANG, L. et al., 2023)) and other novel models that explore generative and discriminative learning could improve even further. It's important to notice also that when comparing the results of other studies in table 18, and the very study that presents the VideoMAE model that is utilized in this work, the difference in resources utilized. Both Deng et al. (2022) and Tong et al. (2022) report the utilization of several GPUs in their studies, where the results reported in this work were enabled by a single GPU and time constraints.

Several authors(RAJAGOPALAN et al., 2013; NEGIN et al., 2021; WEI et al., 2022; DENG et al., 2022) used subsets of this proposed dataset in their own models and pipelines, and their main results are shown in table 18.

Rajagopalan et al.(RAJAGOPALAN et al., 2013) mentions that the inherent peri-

Table 18 – Comparison of the previous works that used subsets of the proposed dataset.

Author	Dataset	Model	Best result
Rajagopalan et al. (2013)	SSBD	STIP + BOW + Harris3D	50.7%
Negin et al. (2021)	ESBD	Hist. conc. + HOF-BOVW + MLP	79%
Wei et al. (2022)	Ext. SSBD	I3D + MS-TCN	83%
Deng et al. (2022)	SSBD + ESBD	VST+L	95.47%
This work	**ASBD**	**VideoMAE**	**76%**

odic motion of these behaviors has not been learned properly by their model resulting in low accuracy and mentions that more robust techniques could do so.

Negin et al.(NEGIN et al., 2021) comments that, in their dataset, the descriptor-based framework together with the histogram combination technique is more effective in maintaining temporal information when compared with other techniques. Using the same set of features, Bag Of Visual Words (BOVW) methods with Multi-layer Perceptron (MLP) classifier achieve relatively higher accuracy compared to the Neural Networks (NN) based methods tested (3DCNN and ConvLSTM). They mention that this could be attributed to the nature of activities in the collected dataset. Unlike action datasets, the convolutional filters encountered difficulties in extracting useful information from the background to recognize the activities.

Wei et al.(WEI et al., 2022) cite that the best result is achieved with an Inflated 3D Convnet and Multi-Stage Temporal Convolutional Networks (I3D + MS-TCN), reaching a 0.83 Weighted F1-score for classification of the three autism-related actions presented in his extended SSBD, outperforming existing methods.

Deng et al.(DENG et al., 2022) propose a novel Video Swin Transformer model that is enriched with Vision-language models based on the CLIP (Contrastive Language-Image Pretraining) text encoder, it elevates considerably the results previously achieved.

Even though the VideoMAE model on this training configuration didn't achieve the best results as shown in the table 18, we believe that the experiments on this model fine-tuned on the proposed ASBD dataset presented in this section shows that the unified dataset could improve those models even further by providing a bigger and consolidated dataset.

6.2 EVENT DETECTION WEB APPLICATION FOR ASD THERAPY SESSIONS

For the event detection architecture we gonna explore some of experiments the done on its main original components - Actors detection tool - Events detector - Timeline construction (RIBEIRO et al., 2022) and the proposed add-on to the original framework

where the VideoMAE model enabled by the ASBD dataset can be utilized.

6.2.1 Actors detection tool - YOLOv5

The actors' detection tool used was the extra large model YOLOv5x, which produces better results in nearly all cases but has more parameters [6]. This model was trained for 189 epochs, in a batch size of 1 and image size of 256x256 due to GPU limitations. We used an NVIDIA GeForce GTX 1650 GPU with 4 Gb VRAM for the training model. Other hyperparameters like SGD optimizer and initial learning rate of 0.01 were used as they were provided by default. The training took around 16 hours. The best results were found around 100 epochs (specifically the 89th epoch) achieving a precision of 0.94, recall of 0.82, and mAP@0.5 of 0.90.

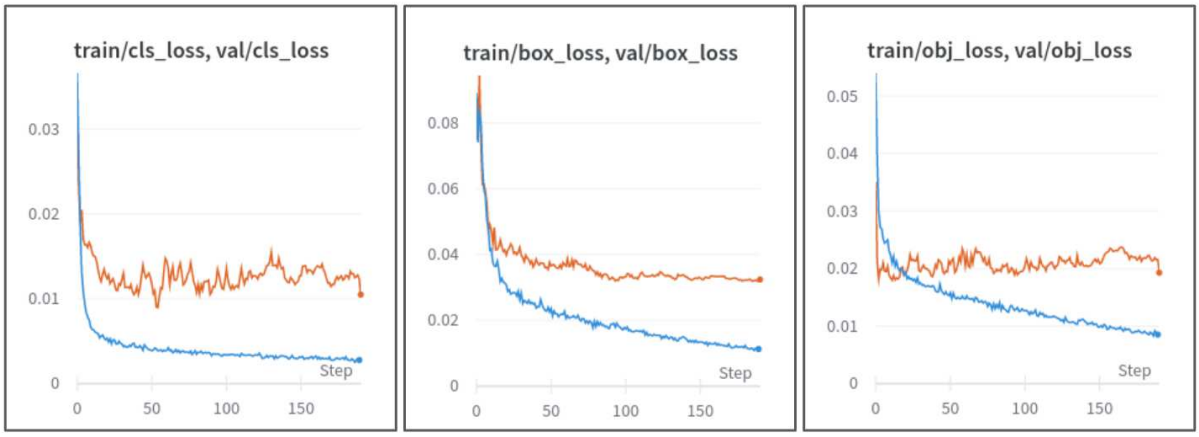


Figure 32 – YOLOv5 model main statistics, Precision, recall, and mAP through the training epochs timeline.

Source: the author

The training evolution of the YOLO model can be seen in figure 32. YOLO loss function is applied to both the training set (4 videos) and validation set (2 videos), totalizing 600 frames fed into the neural network, and is composed of three parts:

1. box_loss — bounding box regression loss (Mean Squared Error) - Represents how well the algorithm can locate the center of an object and how well the predicted bounding box covers an object.
2. obj_loss — objectness loss (Binary Cross Entropy) - A measure of the probability that an object exists in a proposed region of interest. If the objectivity is high, this means that the image window is likely to contain an object.
3. cls_loss — the classification loss (Cross-Entropy) - How well the model could predict the class of the object correctly.

[6] https://github.com/ultralytics/yolov5/wiki/Tips-for-Best-Training-Results

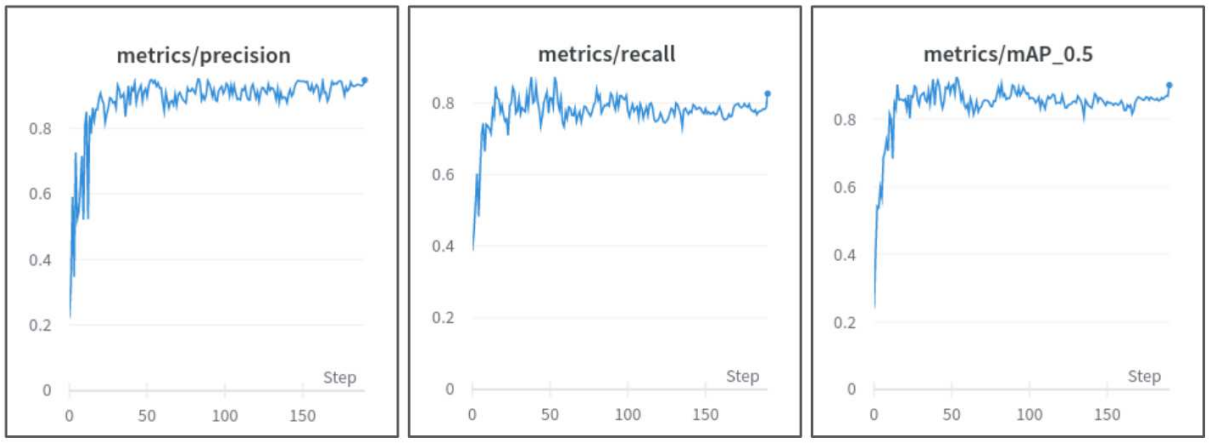


Figure 33 – YOLOv5 model main statistics, Precision, recall, and mAP through the training epochs timeline.

Source: the author

Also, in figure 33, we have the precision/recall of the model and finally the mean average precision (mAP). The main component of mAP is the average precision metric (AP), which in turn, is composed of both precision and recall that are calculated for each class (caretaker, toddler, and plusme). The average precision (AP)((2)) is the mean of the precisions after each relevant image is retrieved. Where Pr is precision at r'th element in the image. After calculating each element in the image, the summation is divided by recall (R). Average precision is a measure that combines recall and precision for ranked retrieval results. The Mean Average Precision((3)) is the arithmetic mean of the average precision values for an information retrieval system over a set of n query topics(ATAŞ; VURAL, 2021).

$$AveragePrecision = \frac{\sum_n P_r}{R} \tag{2}$$

$$mAP = \frac{1}{n}\sum_n AP_n \tag{3}$$

6.2.2 Events detector

Both Sperati et al. (2020) and Giocondo et al. (2022) explored interactions between actors that happened through the action of actually touching something or someone. There are various ways to detect touches between objects and people. One way is to directly detect the event, generating bounding boxes for each type of touch, such as around regions where a child is touching the PlusMe bear or the therapist.

A simple way that we can explore touch between bounding boxes in object detection (performed by YOLOv5) is using a simple heuristic: "Whenever there is any touch or overlap between a pair of bounding boxes, it will be considered as an interaction

between that pair of objects." This way, the detection of an interaction is actually a prediction based on the overlap of the bounding boxes.

Since the events detection tool predicts the existence of interactions of the bounding boxes, we can evaluate these predictions by manually asserting if there was, in fact, an interaction at the given frame (ground truth). Then we can assert the accuracy of the events detection tool in detecting the real events and pointing out metrics like false positives and false negatives (Figure 34) and perceive improvements in these metrics by fine-tuning the event detection algorithm (RIBEIRO et al., 2022; SOARES, 2023).

This way, to analyze the performance of the bounding box overlap/interaction heuristic, we should ignore situations where there is a missing bounding box. However, to assess the performance of the tool as a whole (the system), we should consider them. With prior knowledge of this particular dataset and to properly evaluate the events detector algorithm, we applied a filter beforehand and sent all the frames where the actors' detection tool failed to provide a bounding box for any of the actors directly to the end of the system. These filtered frames are annotated as relevant for the final analysis and bypass the event detector algorithm entirely. Out of the 200 frames used for this analysis, 19 of them contained detection errors, while the remaining 181 did not. Therefore, since every frame is a part of the system's output, the analysis of the system includes all 200 frames, but the analysis of the detector only considers the 181 frames that did not contain errors.

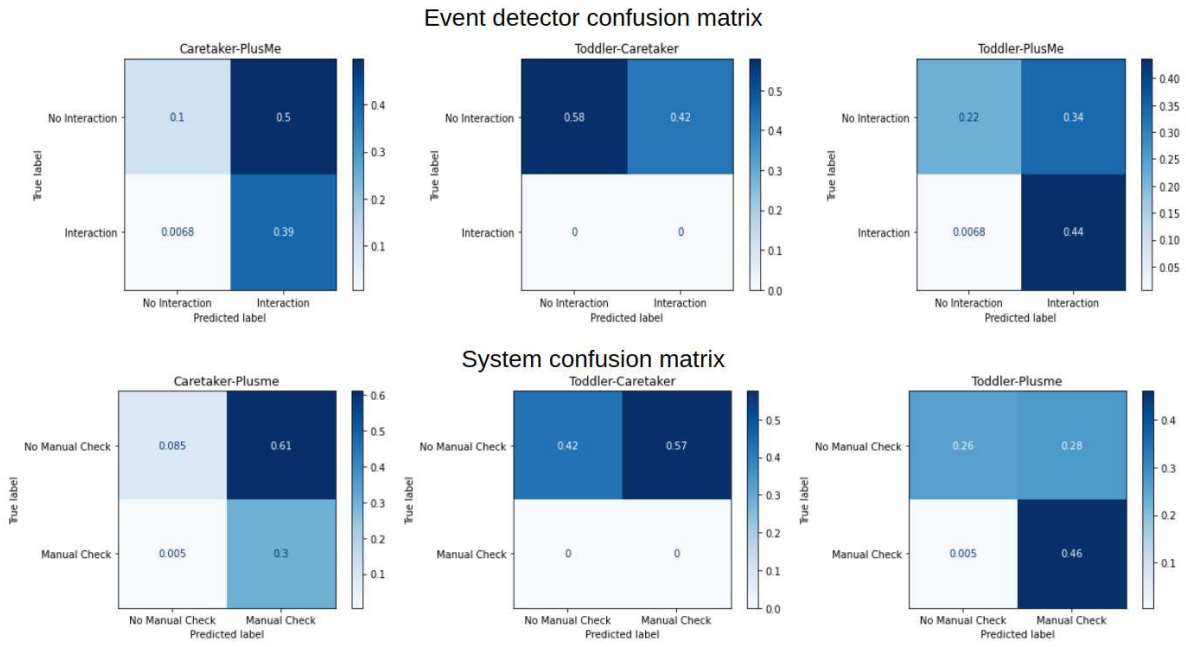


Figure 34 – Confusion matrix for event detector and the system as a whole.

Source: Ribeiro et al. (2022)

6.2.3 System prediction performance

Since both the event predictor and the system as a whole work with predictions and real values (ground-truth - manually assessed), we can create confusion matrices for each interaction pair concerning "a frame with an interaction" and consider "a frame as relevant". This way, we are evaluating how good the predictor is at using the bounding box overlap heuristic for a pair of interactions and how good the system is at indicating to the therapist where they should focus their attention in conducting their medical analysis (Figure 34).

6.2.4 Video length reduction

We extracted a sample of 100 frames from the 2 videos reserved for test purposes (not used in training or validation of the model). For these videos, the reduction of the frames was computed for each interaction detected between the classes and finally, we can evaluate the reduction of the videos when considering only the relevant frames.

Table 19 – Video results - Reduction by interaction between classes

Video	Frames	Ct&Td	Ct&+me	Td&+me	All interactions
Video 1	100	34,0%	0,0%	35,0	0,0%
Video 2	100	51,4%	17,8%	18,8	17.8%

6.2.5 Detection of stimming behavior on PlusMe videos

Soares (2023) expanded the initial architecture proposed in our paper (RIBEIRO et al., 2022) with a web application where the healthcare professionals could, by themselves, upload videos to the framework and receive insights accordingly when considering the actor and events detection mentioned earlier, with the possibility to record those actions in sessions, identify subjects and assess how they are evolving through the therapy. Within this web application, we aim to use the same upload method and analyze the videos submitted to both the event detections and the detection of stimming behavior of the subjects35. With these detections, we enrich the analysis that could be made within the submitted video and introduce the possibility of detecting and perceiving how those detections evolve during the therapy sessions.

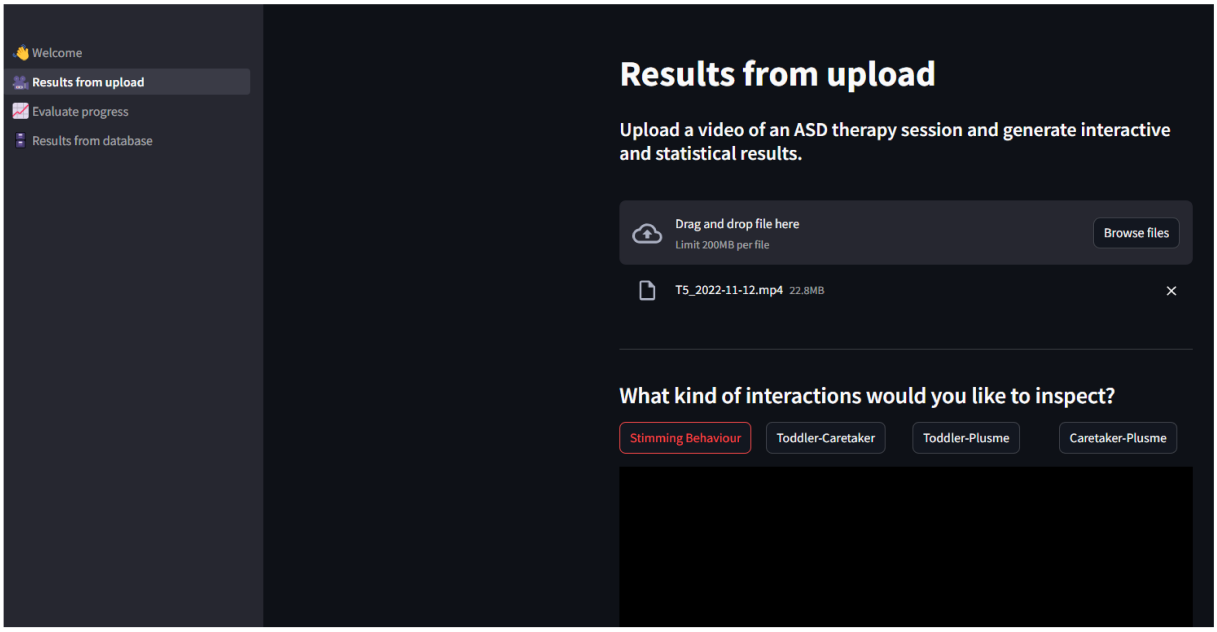


Figure 35 – The detect stimming behaviour function on the web application

Source: the author

7 CONCLUSION AND FUTURE WORKS

7.1 CONCLUSION

Most computer vision techniques are heavily reliant on datasets as essential training examples to develop and fine-tune models. Whether it's detecting faces in images, identifying traffic signs in self-driving cars, or recognizing gestures in sign language, the more diverse and comprehensive the dataset, the better equipped the model becomes to accurately perceive and interpret visual information. Consequently, the quality, quantity, and diversity of the training data are pivotal factors in determining the ultimate performance and generalizability of computer vision models. In our study, we identified the lack of publicly available datasets that depict the stimming behavior very present in individuals within the ASD spectrum, which can cause difficulties or even hinder the development of models that could rely on them.

In this work, we introduce a composed, enriched, and standardized Stimming Behavior Dataset that can be utilized in computer vision systems pipelines to support the training of models that aim to help clinicians and parents analyze children's behaviors, and in particular help to identify behaviors associated with Autism Spectrum Disorder, since computer vision methods based on supervised machine learning techniques achieve considerable better results when bigger datasets are used on the trained phase of its pipeline process. We assess the performance of a renowned video recognition model, pre-trained on the Kinetics-400 human action dataset, by fine-tuning it on our dataset and its subsets to perceive how the bigger and curated dataset compares to the original ones. We also analyze how simple augmentation techniques could impact the model's performance. The models fine-tuned on the ASBD dataset and the augmented counterpart achieved the best results on the fixed test dataset with 0.69% and 0.76% respectively. The Autism Stimming Behavior Dataset is an important step in enabling researchers that aim to apply action recognition models on the identification of the action classes that are present in the diagnosis and therapy of individuals currently within the autism spectrum since it can streamline the process of building a dataset, classifying it, manually curating the videos in search of the relevant time where the behaviors start and finish and details like if the individuals are the same. It also helps to standardize the video clips, making comparing studies made with this dataset more reliable on reproducibility and relevance of the results achieved. With the experiments made, reaching 0.75% accuracy on the augmented ASBD dataset, we hope to provide a solid baseline on how the ASBD dataset can perform on newer models of action recognition. The article that describes this dataset was presented and published at the 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS) with the title "Stimming Behavior Dataset - Unifying stereotype behavior dataset in the wild"(RIBEIRO; GRELLERT; CARVALHO, 2023).

We also describe a framework that aims to streamline the analysis of whole videos that could assist in decision-making in the therapy of patients with Autism Spectrum Disorder (ASD). By utilizing a model capable of detecting actors in the submitted videos and employing heuristics on the detected boundaries, we can predict the relevancy of the frames by indicating the existence of interactions between the actors, pointing out those can help alleviate the amount of work needed to analyze and annotate these videos. The proposed solution has the potential to reduce manual checks by up to 51.4%, resulting in a substantial decrease in the workload for health experts, researchers, therapists, and specialists. Additionally, it highlights the potential benefits for researchers, therapists, and specialists in automating the identification of features and events in video-recorded therapy sessions, ultimately saving them time in the process. This framework and its experiments were first detailed, submitted, accepted, and presented at the 11th Brazilian Conference on Intelligent Systems (BRACIS 2022) with the title "Event detection in therapy sessions for children with Autism"(RIBEIRO et al., 2022).

The framework to detect events was further explored and expanded on the work of Soares (2023). A result of his work is a web application that can process uploaded videos on the process of actor detection, event prediction, and relevant events on an interactive timeline. As an add-on to the whole framework, we enable the framework to leverage the best-trained model on our experiment section to be able to process the submitted videos and detect stemming behaviors along with the other statistics, increasing the range of analysis that the framework can make.

7.2 FUTURE WORKS

In future work, we plan to add more videos to the dataset by collaborating with clinics that treat ASD and maintain a form for curated public URL links to be submitted as entries to the dataset, helping alleviate the intrinsic nature of "in the wild" datasets where videos became restricted over time.

Also, we aim to explore how novel models that explore Generative and Discriminative Learning simultaneously applied to computer vision will fare when applied to the challenges present in this ASBD dataset.

BIBLIOGRAPHY

ABDULLAH, Sharmeen M Saleem; ABDULAZEEZ, Adnan Mohsin. Facial expression recognition based on deep learning convolution neural network: A review. **Journal of Soft Computing and Data Mining**, v. 2, n. 1, p. 53–65, 2021.

ABIRAMI, SP; KOUSALYA, G; BALAKRISHNAN, P. Activity recognition system through deep learning analysis as an early biomarker of ASD characteristics. In: INTERDISCIPLINARY Approaches to Altering Neurodevelopmental Disorders. [S.l.]: IGI Global, 2020. P. 228–249.

ANGLEMYER, A; HORVATH, HT; BERO, L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. **Cochrane Database of Systematic Reviews**, John Wiley & Sons, Ltd, 2014. ISSN 1465-1858.

ASSOCIATION, American Psychiatric. **Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5(TM))**. Hardcover. [S.l.]: American Psychiatric Publishing, 27 May 2013. ISBN 978-2123100119.

ATAŞ, Kubilay; VURAL, Revna Acar. Detection of Driver Distraction using YOLOv5 Network. In: 2021 2nd Global Conference for Advancement in Technology (GCAT). [S.l.: s.n.], 2021. P. 1–6.

BERTAMINI, Giulio; BENTENUTO, Arianna; PERZOLLI, Silvia; PAOLIZZI, Eleonora; FURLANELLO, Cesare; VENUTI, Paola. Quantifying the Child–Therapist Interaction in ASD Intervention: An Observational Coding System. **Brain Sciences**, MDPI AG, v. 11, n. 3, p. 366, 2021.

BOYD, Brian A.; BARANEK, Grace T.; SIDERIS, John; POE, Michele D.; WATSON, Linda R.; PATTEN, Elena; MILLER, Heather. Sensory features and repetitive behaviors in children with autism and developmental delays. **Autism Research**, Wiley, n/a–n/a, 2010.

CARREIRA, Joao; ZISSERMAN, Andrew. **Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset**. [S.l.: s.n.], 2018. arXiv: 1705.07750 [cs.CV].

CAULI, Nino; REFORGIATO RECUPERO, Diego. Survey on Videos Data Augmentation for Deep Learning Models. **Future Internet**, v. 14, n. 3, 2022. ISSN 1999-5903.

CEOLIN, Gilciane; MATSUO, Luísa Harumi; OCKER, Guilherme; GRELLERT, Mateus; D'ORSI, Eleonora; VENSKE, Débora Kurrle Rieger; MOREIRA, Júlia Dubois. Adiposity and physical activity are among the main determinants of serum vitamin D concentrations in older adults: the EpiFloripa Aging Cohort Study. **Nutrition Research**, Elsevier, v. 111, p. 59–72, 2023.

CHARLTON, Rebecca A.; ENTECOTT, Timothy; BELOVA, Evelina; NWAORDU, Gabrielle. "It feels like holding back something you need to say": Autistic and Non-Autistic Adults accounts of sensory experiences and stimming. **Research in Autism Spectrum Disorders**, Elsevier BV, v. 89, p. 101864, Nov. 2021.

CHEN, L.S.; XU, L.; DHAR, S.U.; LI, M.; TALWAR, D.; JUNG, E. Autism spectrum disorders: a qualitative study of attitudes toward prenatal genetic testing and termination decisions of affected pregnancies. **Clinical Genetics**, Wiley, v. 88, n. 2, p. 122–128, 2014.

DENG, Andong; YANG, Taojiannan; CHEN, Chen; CHEN, Qian; NEELY, Leslie; OYAMA, Sakiko. **Problem Behaviors Recognition in Videos using Language-Assisted Deep Learning Model for Children with Autism**. [S.l.]: arXiv, 2022. Available from: `https://arxiv.org/abs/2211.09310`.

EDITION, Fifth et al. Diagnostic and statistical manual of mental disorders. **Am Psychiatric Assoc**, v. 21, p. 591–643, 2013.

FRYE, Richard E; ROSE, Shannon; BOLES, Richard G.; ROSSIGNOL, Daniel A. A Personalized Approach to Evaluating and Treating Autism Spectrum Disorder. **Journal of Personalized Medicine**, MDPI AG, v. 12, n. 2, p. 147, 2022.

GEORGESCU, Alexandra Livia; KOEHLER, Jana Christina; WEISKE, Johanna; VOGELEY, Kai; KOUTSOULERIS, Nikolaos; FALTER-WAGNER, Christine. Machine Learning to Study Social Interaction Difficulties in ASD. **Frontiers in Robotics and AI**, Frontiers Media SA, v. 6, 2019.

GIOCONDO, Flora; FAEDDA, Noemi; CAVALLI, Gioia; SPERATI, Valerio; OZCAN, Beste; GIOVANNONE, Federica; SOGOS, Carla; GUIDETTI, Vincenzo;

BALDASSARRE, Gianluca. Leveraging Curiosity to Encourage Social Interactions in Children with Autism Spectrum Disorder: Preliminary Results Using the Interactive Toy PlusMe. In: CHI Conference on Human Factors in Computing Systems Extended Abstracts. New Orleans, LA, USA: Association for Computing Machinery, 2022. (CHI EA '22).

HAILPERN, Joshua; KARAHALIOS, Karrie; HALLE, James; DETHORNE, Laura; COLETTO, Mary-Kelsey. A3: HCI Coding Guideline for Research Using Video Annotation to Assess Behavior of Nonverbal Subjects with Computer-Based Intervention. **ACM Trans. Access. Comput.**, Association for Computing Machinery, New York, NY, USA, v. 2, n. 2, 2009. ISSN 1936-7228.

HEALTH - OFFICE OF COMMUNICATIONS, National Institutes of. **What are the treatments for autism?** [S.l.]: US Department of Health and Human Services, 2021. Available from: `https://www.nichd.nih.gov/health/topics/autism/conditioninfo/treatments#f2`.

HUANG, Gary B; MATTAR, Marwan; BERG, Tamara; LEARNED-MILLER, Eric. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In: WORKSHOP on faces in'Real-Life'Images: detection, alignment, and recognition. [S.l.: s.n.], 2008.

JAVED, Hifza; LEE, WonHyong; PARK, Chung Hyuk. Toward an Automated Measure of Social Engagement for Children With Autism Spectrum Disorder—A Personalized Computational Modeling Approach. **Frontiers in Robotics and AI**, Frontiers Media SA, v. 7, 2020.

JAVED, Hifza; PARK, Chung Hyuk. Behavior-Based Risk Detection of Autism Spectrum Disorder Through Child-Robot Interaction. In: COMPANION of the 2020 ACM/IEEE International Conference on Human-Robot Interaction. Cambridge, United Kingdom: Association for Computing Machinery, 2020. (HRI '20), p. 275–277.

JOCHER, Glenn et al. **ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference**. [S.l.]: Zenodo, 2022. Available from: `https://zenodo.org/record/6222936`.

KEELE, Staffs et al. **Guidelines for performing systematic literature reviews in software engineering**. [S.l.], 2007.

KINGMA, Diederik P.; BA, Jimmy. **Adam: A Method for Stochastic Optimization**. [S.l.]: arXiv, 2014. Available from: https://arxiv.org/abs/1412.6980.

KOŁAKOWSKA, Agata; LANDOWSKA, Agnieszka; ANZULEWICZ, Anna; SOBOTA, Krzysztof. Automatic recognition of therapy progress among children with autism. **Scientific Reports**, Springer Science and Business Media LLC, v. 7, n. 1, 2017.

LAI, Meng-Chuan; LOMBARDO, Michael V; BARON-COHEN, Simon. Autism. **The Lancet**, v. 383, n. 9920, p. 896–910, 2014. ISSN 0140-6736.

LAPTEV, Ivan; MARSZALEK, Marcin; SCHMID, Cordelia; ROZENFELD, Benjamin. Learning realistic human actions from movies. In: IEEE. 2008 IEEE Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2008. P. 1–8.

LIU, Jingjing et al. Early Screening of Autism in Toddlers via Response-To-Instructions Protocol. **IEEE Transactions on Cybernetics**, v. 52, n. 5, p. 3914–3924, 2022.

LOFTUS, Yolande. Autism Statistics You Need To Know in 2023. **Autism Parenting Magazine**, Sept. 2023. Available at: https://www.autismparentingmagazine.com/autism-statistics/ (Accessed: November 5th, 2023).

LU, Jia; NGUYEN, Minh; YAN, Wei Qi. Deep learning methods for human behavior recognition. In: IEEE. 2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ). [S.l.: s.n.], 2020. P. 1–6.

MAENNER, Matthew J. et al. Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2018. **MMWR. Surveillance Summaries**, Centers for Disease Control MMWR Office, v. 70, n. 11, p. 1–16, 2021.

MENTAL HEALTH (NIMH), National Institute of. **Autism Spectrum Disorder**. 2023. Available from: https://www.nimh.nih.gov/health/topics/autism-spectrum-disorders-asd. Visited on: 10 Mar. 2023.

NEGIN, Farhood; OZYER, Baris; AGAHIAN, Saeid; KACDIOGLU, Sibel; OZYER, Gulsah Tumuklu. Vision-assisted recognition of stereotype behaviors for early

diagnosis of Autism Spectrum Disorders. **Neurocomputing**, Elsevier BV, v. 446, p. 145–155, 2021.

NIGAM, Swati; SINGH, Rajiv; MISRA, A. K. A Review of Computational Approaches for Human Behavior Detection. **Archives of Computational Methods in Engineering**, Springer Science and Business Media LLC, 2018.

NOGAY, Hidir Selcuk; ADELI, Hojjat. Machine learning (ML) for the diagnosis of autism spectrum disorder (ASD) using brain imaging. **Reviews in the Neurosciences**, De Gruyter, v. 31, n. 8, p. 825–841, 2020.

OPREA, Sergiu; MARTINEZ-GONZALEZ, Pablo; GARCIA-GARCIA, Alberto; CASTRO-VARGAS, John Alejandro; ORTS-ESCOLANO, Sergio; GARCIA-RODRIGUEZ, Jose; ARGYROS, Antonis. A review on deep learning techniques for video prediction. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE, 2020.

PARIKH, Milan N; LI, Hailong; HE, Lili. Enhancing diagnosis of autism with optimized machine learning models and personal characteristic data. **Frontiers in computational neuroscience**, Frontiers, v. 13, p. 9, 2019.

PETERSEN, Kai; FELDT, Robert; MUJTABA, Shahid; MATTSSON, Michael. Systematic mapping studies in software engineering. In: 12TH International Conference on Evaluation and Assessment in Software Engineering (EASE) 12. [S.l.: s.n.], 2008. P. 1–10.

RAHMAN, Md; USMAN, Opeyemi Lateef; MUNIYANDI, Ravie Chandren; SAHRAN, Shahnorbanun; MOHAMED, Suziyani; RAZAK, Rogayah A, et al. A Review of machine learning methods of feature selection and classification for autism spectrum disorder. **Brain sciences**, Multidisciplinary Digital Publishing Institute, v. 10, n. 12, p. 949, 2020.

RAJAGOPALAN, Shyam; DHALL, Abhinav; GOECKE, Roland; RAJAGOPALAN, Shyam. Self-stimulatory behaviours in the wild for autism diagnosis. In: PROCEEDINGS of the IEEE International Conference on Computer Vision Workshops. [S.l.: s.n.], 2013. P. 755–761.

RAMIREZ-DUQUE, Andrés A.; FRIZERA-NETO, Anselmo; BASTOS, Teodiano Freire. Robot-Assisted Diagnosis for Children with Autism Spectrum Disorder Based on

Automated Analysis of Nonverbal Cues. In: 2018 7th IEEE International Conference on Biomedical Robotics and Biomechatronics (Biorob). [S.l.: s.n.], 2018. P. 456–461.

REDMON, J.; DIVVALA, S.; GIRSHICK, R.; FARHADI, A. You Only Look Once: Unified, Real-Time Object Detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA, USA: IEEE Computer Society, 2016. P. 779–788.

REHG, James et al. Decoding children's social behavior. In: PROCEEDINGS of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2013. P. 3414–3421.

RIBEIRO, Guilherme Ocker; GRELLERT, Mateus; CARVALHO, Jonata Tyska. Stimming Behavior Dataset - Unifying Stereotype Behavior Dataset in the Wild. In: 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS). [S.l.: s.n.], 2023. P. 225–230.

RIBEIRO, Guilherme Ocker; SOARES, Alexandre Soli; CARVALHO, Jônata Tyska; GRELLERT, Mateus. Event Detection in Therapy Sessions for Children with Autism. In: XAVIER-JUNIOR, João Carlos; RIOS, Ricardo Araújo (Eds.). **Intelligent Systems**. Cham: Springer International Publishing, 2022. P. 221–235.

ROGGE, Nicky; JANSSEN, Juliette. The Economic Costs of Autism Spectrum Disorder: A Literature Review. **Journal of Autism and Developmental Disorders**, Springer Science and Business Media LLC, v. 49, n. 7, p. 2873–2900, 2019.

SAYIS, Batuhan; PARES, Narcis; GUNES, Hatice. Bodily Expression of Social Initiation Behaviors in ASC and Non-ASC Children: Mixed Reality vs. LEGO Game Play. In: COMPANION Publication of the 2020 International Conference on Multimodal Interaction. Virtual Event, Netherlands: Association for Computing Machinery, 2020. (ICMI '20 Companion), p. 140–149.

SHABAZ, Mohammad; SINGLA, Parveen; JAWARNEH, Malik Mustafa Mohammad; QURESHI, Himayun Mukhtar. A Novel Automated Approach for Deep Learning on Stereotypical Autistic Motor Movements. In: ADVANCES in Medical Diagnosis, Treatment, and Care. [S.l.]: IGI Global, 2021. P. 54–68.

SHARMA, Samata R.; GONDA, Xenia; TARAZI, Frank I. Autism Spectrum Disorder: Classification, diagnosis and therapy. **Pharmacology & Therapeutics**, v. 190, p. 91–104, 2018. ISSN 0163-7258.

SOARES, Alexandre Soli. **Uma ferramenta para auxílio a terapias do transtorno do espectro autista usando aprendizado de máquina**. [S.l.: s.n.], 2023.

SPERATI, Valerio et al. Acceptability of the transitional wearable companion "+ me" in children with autism spectrum disorder: a comparative pilot study. **Frontiers in psychology**, Frontiers Media SA, v. 11, p. 951, 2020.

TALKOWSKI, Michael E. et al. Sequencing Chromosomal Abnormalities Reveals Neurodevelopmental Loci that Confer Risk across Diagnostic Boundaries. **Cell**, Elsevier BV, v. 149, n. 3, p. 525–537, 2012.

THEVENOT, Jérôme; LÓPEZ, Miguel Bordallo; HADID, Abdenour. A Survey on Computer Vision for Assistive Medical Diagnosis From Faces. **IEEE Journal of Biomedical and Health Informatics**, v. 22, n. 5, p. 1497–1511, 2018.

TONG, Zhan; SONG, Yibing; WANG, Jue; WANG, Limin. **VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training**. [S.l.: s.n.], 2022. arXiv: 2203.12602 [cs.CV].

WANG, Chien-Yao; LIAO, Hong-Yuan Mark; WU, Yueh-Hua; CHEN, Ping-Yang; HSIEH, Jun-Wei; YEH, I-Hau. CSPNet: A new backbone that can enhance learning capability of CNN. In: PROCEEDINGS of the IEEE/CVF conference on computer vision and pattern recognition workshops. [S.l.: s.n.], 2020. P. 390–391.

WANG, Limin; HUANG, Bingkun; ZHAO, Zhiyu; TONG, Zhan; HE, Yinan; WANG, Yi; WANG, Yali; QIAO, Yu. **VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking**. [S.l.: s.n.], 2023. arXiv: 2303.16727 [cs.CV].

WANG, Minxiao; YANG, Ning. OTA-NN: Observational Therapy-Assistance Neural Network for Enhancing Autism Intervention Quality. In: 2022 IEEE 19th Annual Consumer Communications and Networking Conference (CCNC). [S.l.: s.n.], 2022. P. 1–7.

WASHINGTON, Peter; KLINE, Aaron; MUTLU, Onur Cezmi; LEBLANC, Emilie; HOU, Cathy; STOCKHAM, Nate; PASKOV, Kelley; CHRISMAN, Brianna;

WALL, Dennis. Activity Recognition with Moving Cameras and Few Training Examples: Applications for Detection of Autism-Related Headbanging. In: EXTENDED Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. [S.l.]: ACM, 2021.

WEI, Pengbo; AHMEDT-ARISTIZABAL, David; GAMMULLE, Harshala; DENMAN, Simon; ARMIN, Mohammad Ali. **Vision-Based Activity Recognition in Children with Autism-Related Behaviors**. [S.l.]: arXiv, 2022. Available from: `https://arxiv.org/abs/2208.04206`.

WONG, Connie et al. Evidence-Based Practices for Children, Youth, and Young Adults with Autism Spectrum Disorder: A Comprehensive Review. **Journal of Autism and Developmental Disorders**, Springer Science and Business Media LLC, v. 45, n. 7, p. 1951–1966, 2015.

WU, Hao; LIU, Qi; LIU, Xiaodong. A review on deep learning approaches to image classification and object segmentation. **Comput. Mater. Continua**, v. 60, n. 2, p. 575–597, 2019.

XU, Renjie; LIN, Haifeng; LU, Kangjie; CAO, Lin; LIU, Yunfei. A Forest Fire Detection System Based on Ensemble Learning. **Forests**, v. 12, p. 217, 2021.

YUN, Sangdoo; OH, Seong Joon; HEO, Byeongho; HAN, Dongyoon; KIM, Jinhyung. **VideoMix: Rethinking Data Augmentation for Video Classification**. [S.l.: s.n.], 2020. arXiv: `2012.03457 [cs.CV]`.

ZEIDAN, Jinan; FOMBONNE, Eric; SCORAH, Julie; IBRAHIM, Alaa; DURKIN, Maureen S; SAXENA, Shekhar; YUSUF, Afiqah; SHIH, Andy; ELSABBAGH, Mayada. Global prevalence of autism: a systematic review update. **Autism Research**, Wiley Online Library, v. 15, n. 5, p. 778–790, 2022.

ZHAO, Zhong; ZHU, Zhipeng; ZHANG, Xiaobin; TANG, Haiming; XING, Jiayi; HU, Xinyao; LU, Jianping; QU, Xingda. Identifying Autism with Head Movement Features by Implementing Machine Learning Algorithms. **Journal of Autism and Developmental Disorders**, Springer Science and Business Media LLC, v. 52, n. 7, p. 3038–3049, 2021.

ZHAO, Zhong-Qiu; ZHENG, Peng; XU, Shou-tao; WU, Xindong. Object detection with deep learning: A review. **IEEE transactions on neural networks and learning systems**, IEEE, v. 30, n. 11, p. 3212–3232, 2019.

8 APPENDICES

8.1 APPENDICES

8.1.1 Selected studies references and data

This is a digest of the DEF spreadsheet data that was constructed within this study. The full DEF table is public available at Click me - DEF google Spreadsheet

Table 20 – Selected studies

DEF-ID	Reference	Publication Type	Year	Country	Database
1	Heath 2019	CONF	2019	USA	IEEE
2	Zhao 2022	JOUR	2022	China	Scopus
3	Liu 2022	JOUR	2022	China	Scopus
4	Negin 2021	JOUR	2021	France	Scopus
5	Washington 2021	CONF	2021	USA	Scopus
6	Sayis 2020	CONF	2020	Spain	Scopus
7	Javed 2020-1	JOUR	2020	USA	Scopus
8	Javed 2020-2	CONF	2020	USA	Scopus
9	Georgescu 2019	JOUR	2019	UK	Scopus
10	Ramírez-Duque 2019	JOUR	2019	Brazil	Scopus
11	Chen 2019	JOUR	2019	Taiwan	Scopus
12	Manocha 2019	JOUR	2019	India	Scopus
13	Liu 2018	CONF	2018	USA	Scopus

8.1.2 Machine Learning techniques glossary

Table 21 – Machine Learning techniques glossary

Term	Definition
AdaBoost	Adaptive Boosting
CNN	Convolutional Neural Networks
DT	Decision Trees
ELM	Extreme learning machine
GNB	Gaussian Naive Bayes
KNN	K-Nearest Neighbor
LDA	Linear Discriminant Analysis
LSTM	Long short-term memory
Log.Reg.	Logistic Regression
MLP	Multilayer Perceptron
RF	Random Forest
SSD	Single Shot MultiBox Detector
SVM	Support vector machines

8.1.3 Search strings

Scopus -	(event OR interactions) AND (detection) AND (video) AND ("machine learning" OR ml) AND (asd OR autism) AND (LIMIT-TO (PUBYEAR , 2022) OR LIMIT-TO (PUBYEAR , 2021) OR LIMIT-TO (PUBYEAR , 2020) OR LIMIT-TO (PUBYEAR , 2019) OR LIMIT-TO (PUBYEAR , 2018)) AND (LIMIT-TO (SUBJAREA , "COMP") OR LIMIT-TO (SUBJAREA , "MEDI") OR LIMIT-TO (SUBJAREA , "PSYC"))
IEEE Xplore -	(“All Metadata”: interactions video ASD machine learning) OR (“All Metadata”: interactions video autism machine learning) OR (“All Metadata”: detection video autism machine learning) OR (“All Metadata”:event detection video autism machine learning)
ACM -	[All: interaction] AND [All: detection] AND [All: video] AND [All: asd] AND [All: machine] AND [All: learning] AND [All: interaction] AND [All: detection] AND [All: video] AND [All: autism] AND [All: machine] AND [All: learning] AND [All: event] AND [All: detection] AND [All: video] AND [All: asd] AND [All: machine] AND [All: learning] AND [Publication Date: (01/01/2018 TO 12/31/2022)]