Multi-View Digital Representation Of Social Behaviours In Children And Action Recognition Methods

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

This research reported in this thesis was supported by the award of a Research Training Program scholarship to the PhD Candidate.

Carter Sun

Principal Supervisor

Professor Alistair McEwan

Auxiliary Supervisors

Professor Adam Guastella

Professor Philip De Chazal

Dr Eleni Andrea Demetriou

Faculty of Engineering

The University of Sydney

2024

Statement Of Originality

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Signature:

Name: Carter Sun

Date: 28/02/2024

Authorship Attribution Statement

Chapter 3 of this thesis contains intellectual property disclosures currently filed, and part of the contents have been submitted for journal publication.
I co-designed the study with the co-authors, and was in charge of data extraction, analysis, validation, and visualisation and wrote the drafts of the manuscript.

Chapter 5 of this thesis contains contents published as:

[Sun, Carter, et al. "Statistical Validation of An Automated Method for Calculating Time Domain Heart Rate Variability on The QT Dataset". IEEE EMBC Conf. 2023.] Available at: 10.1109/EMBC40787.2023.10341110. A copy of this paper is presented in the appendix. I co-designed the study, and was in charge of data extraction, analysis, validation, and visualisation and wrote the drafts of the manuscript. Data collection was conducted by the co-author R.T.

The poster abstract [Sun, Jian, et al. "MEASUREMENT OF BODY FAT IN PREMATURE AND TERM INFANTS BY NEAR-INFRARED SPECTROSCOPY". PSANZ. July 2020.] presented in the appendix of this thesis has been published. Available at: https://doi.org/10.1111/jpc.14832.
I co-designed the study, and was in charge of data collection, analysis, and wrote the drafts of the manuscript.

Signature:

Name: Carter Sun

Date: 28/02/2024

As supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Supervisor Name: Professor Alistair McEwan

Signature:

Date: 28/02/2024

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Professor Alistair McEwan and Professor Adam Guastella, for their unwavering guidance and support throughout the challenging yet rewarding journey of my doctoral study. Their tremendous mentorship has played a crucial role in shaping my academic growth and navigating the challenges encountered along the way. These opportunities have also allowed me to gradually understand the current state of science and technology innovations in clinical medicine, such as psychology and rehabilitation.

I would also like to express my sincere appreciation to Associate Professor Wanli Ouyang and Associate Professor Luping Zhou for their dedicated guidance and expertise. Their guidance enabled me to quickly catch up on the essential skills required in the merging fields of computer vision, and data mining of complex clinical data. In particular, their continuous supervision and feedback during the COVID-19 lockdown periods were instrumental in ensuring the continued progress of my research.

I would also like to thank all my mentors and colleagues from the Clinic for Autism and Neurodevelopmental (CAN) at the Brain and Mind Centre, and the School of Biomedical Engineering. This includes my auxiliary supervisors: Professor Philip de Chazal and Doctor Eleni Andrea Demetriou. In addition, I want to express my gratitude to the clinicians in the group, Emma Thomas, Rinku Thapa, and Doctor Zahava Ambarchi for their support in data collection and annotations, as well as other colleagues and collaborators: Professor Amit Lampit, Doctor Kelsie Boulton, Doctor Andrew Goodwin, Doctor Haifeng Zhao, Ayesha Sadozai and Nina Perry. They have provided invaluable support in their respective areas of expertise, ensuring the smooth progress of my PhD research.

A special thanks to my parents, family members and friends, who have given me endless help and support. They have always stood by my side and given me the encouragement, understanding and moral support to overcome the difficulties in my life. Without their support, I would not have grown and achieved so much thus far.

Table Of Contents

List Of Figures

List Of Tables

List Of Supplementary Materials

Co-authored Research Papers During Candidature

Wearable augmentative and alternative communication (wAAC): a novel solution for people with complex communication needs.

Haifeng Zhao, Petra Karlsson, Darryl Chiu, **Carter Sun**, Omid Kavehei, Alistair McEwan

Virtual Reality, https://doi.org/10.1007/s10055-023-00818-8

My contribution: Participated in data collection, data analysis and manuscript editing.

Executive function in children with neurodevelopmental conditions: A systematic review and meta-analysis.

Ayesha K. Sadozai, Eleni A. Demetriou, **Carter Sun,** Amit Lampit, Kelsie A. Boulton, and Adam J. Guastella

(The paper is under second round of review in "Nature Human Behaviour")

My contribution: Participated in the peer review process as the second reviewer and implemented R coding for the meta-analysis.

Attention and Executive Delays in Early Childhood: A Meta-analysis of Neurodevelopmental Conditions.

Dabin Lee, Kelsie A. Boulton, **Carter Sun**, Natalie L. Philips, Martha Munro, Fiona Kumfor, Eleni A. Demetriou, Adam J. Guastella

(The paper is under review)

My contribution: Implemented R coding for meta-analysis and data visualisation.

Burden of paediatric hospitalisations to the health care system, child and family: A systematic review of Australian studies (1990-2022).

Eleni Andrea Demetriou, Kelsie Boulton, Rinku Thapa, Michael Bowden, John Gilroy**, Carter Sun**, Adam Guastella

Lancet-Regional Western Pacific, https://doi.org/10.1016/j.lanwpc.2023.100878

My contribution: Data visualisation.

List Of Abbreviations

ADOS-2	Autism Diagnostic Observation Schedule 2nd edition
AI	Artificial Intelligence
ASD	Autism Spectrum Disorder
AUC	Area Under the Curve
AU	Action Unit
BPM	Beats Per Minute
CI	Confidence Intervals
CNN	Convolutional Neural Network
DOR	Diagnostic Odds Ratio
DT	Decision Tree
ECG	Electrocardiogram
EEG	Electroencephalogram
ESCS	Early Social Communication Scales
FFT	Fast Fourier Transform
FN	False Negative
FP	False Positive
GLM	General Linear Model
Grad-CAM	Gradient-weighted Class Activation Mapping
HF	High Frequency
HMM	Hidden Markov Model
HOG	Histogram of Oriented Gradients
HRV	Heart Rate Variability
IBI	Inter-beat-interval
K-NN	K-Nearest Neighbour
KS	One-sample Kolmogorov-Smirnov
Lasso	Least Absolute Shrinkage and Selection Operator
LDA	Linear Discriminant Analysis
LF	Low Frequency
LSTM	Long Short-Term Memory
MCC	Matthews Correlation Coefficient

MeanNN	Mean Of R-R Intervals in Milliseconds
MHR	Mean Heart Rate
MSBR	Multi-modal Social Behaviour Recognition
NB	Naïve Bayesian
NN50	The Number of Pairs of Successive R-R Intervals That Differ by More Than 50ms
NSR	Normal Sinus Rhythm
NVIQ	Nonverbal Intelligence Quotient
NVDQ	Nonverbal Developmental Quotient
PCA	Principal Component Analysis
PEEC	Pose-Dependent Egocentric Eye Contact
PiCNN	Pose-implicit Convolutional Neural Networks
pNN50	The Proportion of Successive R-R Intervals That Are Larger Than 50ms
PPG	Photoplethysmography
QUADAS-2	The Quality Assessment of Diagnostic Accuracy Studies-2
RBF	Radial Basis Function
RF	Random Forest
RMSSD	Root Mean Square of Successive R-R Interval Differences
ROC	Receiver Operating Characteristics
RSA	Respiratory Sinus Arrhythmia
SD	Standard Deviation
SDNN	Standard Deviation of R-R Intervals in Milliseconds
SROC	Summary Receiver Operating Characteristics Curve
SVM	Support Vector Machine
TD	Typical Developing/ Neurotypical Control
TN	True Negative
TOST	Two One-Sided Test
TP	True Positive
VB-MAPP	Verbal Behaviour Milestones Assessment and Placement Program
VMM	Visible Markov Model
VR	Virtual Reality
WHO	World Health Organization

Abstract

Autism spectrum disorders (ASD) affect at least 1% of children globally. It is partially defined by social behaviour delays in eye contact and joint attention during social interaction and with evidence of reduced heart rate variability (HRV) under static and social stress environments. Currently, no validated artificial intelligence or signal processing algorithms are available to objectively quantify behavioural and physiological markers in unrestricted interactive play environments to assist in the diagnosis of ASD.

This thesis proposes that social behavioural and physiological markers of children with ASD can be objectively quantified through a synergistic digital approach from multi-modal and multi-view data sources. First, a novel deep learning (DL) framework for social behaviour recognition using a fusion of multi-view and multi-modal predictions is proposed. It utilises true-colour images and moving trajectory (optical flow) images extracted from fixed camera video recordings to detect eye contact between children and caregivers in free play while elucidating unique digital features of eye contact behaviour in multiple individual social interaction settings. Moreover, for the first time, a support vector machine model with feature selection is implemented along with statistical analysis, to identify effective facial features and facial orientations for use in identifying ASD during joint attention episodes in free play. Furthermore, a customised NeuroKit2 toolbox was validated using the open-source QT database and a clinical baseline social interaction task. This toolbox facilitates the automated extraction of HRV metrics and allows between-group comparisons in physiological markers.

The work highlights the importance of developing explainable algorithms that objectively quantifying multi-modal digital markers. It offers the potential for the use of digitalised phenotypes to aid in the assessment of ASD and intervention in naturalistic social interaction.

Chapter 1: Introduction

This research incorporates the concept of multi-modal synergy and focuses on the development of automated algorithms to analyse objective digital behavioural and physiological markers in naturalistic social interaction scenarios between multiple individuals to assist in ASD assessments. The innovative mining of digital features from video recordings in assessing social interaction skills aids in determining clinically meaningful features for classification. The multi-view recording approach employed in this research facilitates the evaluation of social interaction skills when child participants are not restricted in their movements to perform natural responses to the environment. It also ensures a comprehensive analysis of the social behaviours of child participants from multiple angles when movements block some camera views. In addition, automated social behaviour recognition and biomarker analysis algorithms help reduce labour costs and save hours of workload for clinicians in manually coding the social skills presented by the child participants during assessments. Automation also helps reduce the risk of bias and inconsistency that human errors may introduce.

This chapter outlines the background information on the topics discussed in this thesis, presents the motivation and summarises the overall methodology. Further, it presents the thesis argument, research problem statements and research questions and concludes with an outline of the structure of each chapter in answering each research question.

1.1 Research Motivation

ASD affects at least 1% of children worldwide and is characterised by delayed social interaction and communication challenges, with many exhibiting restrictive and repetitive behaviours [1]. In clinical practice, social skill assessments and interventions for patients with ASD often require clinicians to interact with the patients using standardised structured or unstructured plays [2, 3]. During these assessments, clinicians observe the patients' social responses and gain insight into patient's overall social skills. After each assessment session, clinicians would manually conduct a detailed analysis of social behaviours presented by the patients, which typically lasts about an hour. These standardised assessment tools sometimes delay diagnosis because many neurodevelopmental disorders have similar presenting features [4, 5], and some symptoms may not be present in early developmental stages [6, 7]. Other assessment modes, such as subjective caregiver reports and surveys, could introduce bias, resulting in diagnostic delays [8].

Moreover, in clinical trials and diagnostic settings, patients' subjective placebo response can introduce biased outcomes and result in delayed intervention [9]. Therefore, an objective and automated behaviour detection algorithm could assist clinicians in transferring subjective observational coding into quantifiable digital features for quantitative analysis. In addition, a combination of behavioural and physiological marker analysis can provide a more rapid and comprehensive overview of ASD symptoms and assist clinicians in making intervention plans.

1.2 Problem Statements

A key phenotypic characteristic of children with ASD is a lack of understanding and responsiveness to the surrounding social environment. The recognition of subjective biases [10] often introduced in assessment processes draws attention to the inclusion of artificial intelligence (AI) in ASD assessment [11]. This is due to its advancements in investigating digitalised phenotypes, offering the potential for objective markers to support assessments and interventions. Further, encouraging parental cooperation in AI intervention paradigms can effectively stimulate natural responses in children and improve their social skills [12]. However, current research on the use of AI to assess individuals' spontaneous social responses is limited to the evaluation of eye gaze moving patterns, facial expressions, and physiological responses within an experimental setting featuring restricted activity [11, 13]. In addition, the AI systems proposed in the current literature have not been widely used in clinical practices, as there is a lack of ethical guidance and consensus on acceptable performance thresholds and a lack of information on underlying mechanisms and acceptable operating conditions based on which these systems could be deployed.

In studies aimed at analysing atypical social behaviours during ASD assessments and interventions, current research lacks practical algorithms to detect participants' natural responses during free play. The challenge is partly due to the unpredictable nature of human social interaction. For instance, changes in head orientation during social interaction could block camera views from tracking faces, and facial detection algorithms may not be able to assign the same face identification number when multiple participants are in the same scene. Consequently, the automatic detection of natural responses in social interaction poses significant challenges.

On the other hand, conventional mechanisms for social attention analysis through visual stimuli, such as pre-recorded videos featuring actors or avatars to direct participants' social attention, typically require eye trackers or built-in cameras on electronic devices to record eye gaze and facial responses. However, these approaches require limiting body movements to record changes in facial information continuously. Subsequently, machine learning (ML) algorithms were trained on these extracted features and tested on newly collected data to assess their reliability in analysing differences in the visual responses to video stimuli between ASD and non-ASD individuals.

Moreover, physiological changes reflect psychological reactions and have been studied for emotion recognition and ASD assessments. In these studies, external devices, such as electroencephalography (EEG) or functional Magnetic Resonance Imaging (fMRI), are often employed for brain activity assessments. Electrocardiography (ECG) is used to analyse the electrical activity of the heart. HRV metrics are derived from recorded ECG signals to provide information about the heart's rhythm, heart rate, and cardiac functions. However, significant body movements introduce motion artifacts and increase the risk of collecting poor-quality physiological signals. Notably, no reliable automation algorithm has been translated into clinical services for cleaning long-term ECG signal recordings collected in complex heart conditions.

Furthermore, with the growing research interest in detecting social behaviours during diagnostic assessments, behaviour detection becomes a great challenge when patients, clinicians, and caregivers are all recorded in the same camera frame to form multi-individual interactions. In these studies, video recording devices have been used extensively to evaluate the social skills exhibited by the participants during clinical assessments and interventions. Clinicians typically review audio and video signals collected during the sessions to assess participants' abilities to engage in joint social interaction plays and manually code any social

behaviours presented throughout the assessment sessions to calculate the frequency and duration for subsequent quantitative analysis.

While recent studies applying AI for behavioural assessments in video recordings have been promising, most of these studies have focused primarily on participants' behaviours [11, 13]. Fewer studies have assessed social interactions among multiple individuals (the participants, caregivers or assessors) [14]. Moreover, many studies have opted for external devices to analyse children's social behaviours, such as emotion recognition and attention detection [15]. These approaches typically require manual feature extraction or model development using open-access data. These approaches greatly limit the opportunity to analyse the natural social behaviours of children in a free play, unrestricted movement setting, and cannot directly study some essential social skills, such as direct eye contact and joint attention, which require interaction between at least two social partners.

This thesis addresses these challenges by proposing a multi-view, multi-modal digital feature learning mechanism to analyse behavioural phenotypes presented during social interaction assessments. Firstly, a DL framework is introduced to recognise unpredictable behavioural patterns of direct eye contact between child participants and assessors captured from multiple camera views. These assessments involve recording multiple individuals, including child participants, assessors, and caregivers. During the assessments, caregivers are seated in the trial room and instructed not to interfere with children's natural responses. The social interaction assessments are solely based on video recordings and wearable physiological monitoring.

Secondly, ML models are trained based on facial features extracted from children's facial responses to joint attention events occurring in free play. Statistical mechanism analysis accompanies this process to examine potential discriminatory digital phenotypes for classifying ASD from typical controls (TD). The choice of using ML algorithms for facial feature analysis is informed by the wealth of domain knowledge available and its strengths in analysing structured features. The feature extraction processes in our proposed analytical framework are fully automatic, eliminating the need for handcrafted facial or body movement features.

In addition, extending beyond video processing for behaviour assessments, the use of physiological signals to reflect physical responses and emotions in social interaction assessments have been constrained by a lack of effective automatic signal processing methods for investigating distinguishable biomarkers in complex free play settings. An automated ECG signal processing toolbox is first customised to clean recorded signals and derive HRV metrics. The toolbox is then validated against human-annotated ECG signals collected at rest and during exercises. This includes an open-source public dataset, the MIT-QT dataset, and our clinical dataset. Furthermore, the toolbox is applied to investigate changes in HRV between free play and baseline tasks in the social interaction assesses, providing insights into intergroup differences between ASD and TD groups.

1.3 Aims And Research Questions

Considering the identified challenges in objectively quantifying social competence presented by multiple participants during face-to-face social interaction assessments through video recordings, and recognising the lack of validated automatic physiological signal processing toolboxes for analysing physiological marker changes in free play, this thesis aims to pioneer assistive technologies to recognise digital phenotypic features of social behaviours and to explore potential behavioural and physiological markers that could aid with ASD diagnosis. The following are the main research questions addressed in this thesis:

1. Could a deep learning model identify key social cues in a natural, free play environment?
2. Could facial features help classify ASD from TD children using joint attention episodes presented in free play?
3. Do physiological biomarkers, such as ECG signals collected in natural social interaction settings, reveal differences between children with ASD or the TD group?

1.4 Thesis Argument

This thesis argues that the strengths of signal processing and AI algorithms in extracting objective, quantifiable, synthesised digital representations of behavioural and physiological markers can be used to assess social interactions between participants and clinicians during ASD assessments and contribute to ASD diagnosis.

1.5 Background

1.5.1 Behaviour-based Assessments Using AI

In light of the rapid advancements in AI algorithms, particularly in pattern recognition, object detection and action recognition tasks, integrating these automatic technologies into the medical domains is becoming increasingly prevalent. These algorithms leverage extensive data labelling and training on visual information, such as images and videos, offering objectivity that compensates for the subjective biases inherent in human observations [16-19].

AI algorithm development in ASD research has predominantly focused on facial detection and recognition [20], emotion recognition, and gaze-shifting pattern analysis [13-15]. The fundamental principle underlying the use of deep neural networks is to process complex feature patterns of the target objects. For example, in facial feature analysis, the initial steps before training classification models involve face and landmark detection. Recent studies recommend using the OpenFace2.0 toolkit invented in 2018 by Baltrusaitis et al. [21], who significantly contributed to facial feature extraction. OpenFace2.0 has found extensive use in ASD research, covering a broad range of facial analysis tasks such as face detection, facial landmark detection, head pose estimation, facial action unit analysis, and gaze estimation. Similarly, the IntraFace software introduced by Xiong et al. [22] has also been deployed for facial landmark detection, head pose estimation and facial expression tasks. For instance, Cai et al. [23] proposed a convolutional neural network (CNN) framework to explore the application of temporal information from head-related features of children extracted from OpenFace2.0 in ASD diagnosis using a home video at the time of parent-initiated joint attention and achieved over 95% accuracy. Hanna et al. [24] found that individuals diagnosed with ASD were less able to imitate emotions than the TD group, exhibiting slower and less accurate responses. Gaze shifting on video stimuli was found to be a potential biomarker in ASD diagnosis [25-27].

ML algorithms are preferred for data mining structured features to find discriminative patterns for classification tasks [28]. Wei et al. [29] reviewed the existing ML algorithms developed for ASD classification in eye-tracking studies. It was found that support vector machine (SVM) and Random Forest (RF) are the algorithms applied in most studies, and a pooled accuracy of 88% using SVMs and 71% using RFs is achieved. The application of ML models presented by Hyde et al. [30] had findings that were consistent with other literature. The study found that supervised learning algorithms developed based on generic data and neuroimaging, especially binary classifiers, are often used in ASD diagnostic studies.

DL algorithms could also be applied for behaviour recognition [17, 18]. Simonyan and Zisserman [31] introduced in 2014 a two-stream framework for behaviour recognition in video recordings, laying the foundation for developing multi-modal spatial-temporal feature fusion algorithms. Later in 2018, Wang et al. [32] improved the two-stream framework by allowing sparse sampling video segments for action recognition, where the temporal segment network was introduced. Kohli et al. [33] trained a two-stream deep learning network on the NTU dataset with seven types of actions related to ASD diagnosis.

1.5.2 Physiological Biomarker

In ASD research, the synergistic integration of physiological signals analysis with behavioural and cognitive assessments has emerged as a pivotal methodology for objectively quantifying human responses to social cues. As it directly reflects the physiological states of individuals on the autism spectrum, a synergistic approach comprehensively presents the underlying intricacies between neurobiological processes and observable social behaviours. Technology and devices designed for collecting those physiological signals have been widely utilised in research environments and clinical settings for early diagnosis, interventions, and clinical services that improve clinical outcomes.

The application of non-invasive structural and functional neuroimaging embedded with AI technologies has greatly contributed to advancing the understanding of brain activity and functional connectivity [34]. William et al. [35] suggested that resting-state EEG signals may reflect atypical cognitive development. Duffy et al. [36] proposed that EEG coherence-based phenotype could show significant differences between ASD and TD groups. Despite these advancements, current neuroimaging studies primarily focus on analysing brain region activations, particularly during tasks where participants' movements are restricted. This is because the practical application of these commercially available devices is limited to use in controlled environments due to hardware constraints.

Physiological biomarker studies, however, could also be performed with focus on the electrical activity of the heart. For instance, changes in heart rate and HRV calculated from ECG signals reveal the influence of the autonomic nervous system on cardiac conditions [37]. Recent studies indicate that cardiovascular deregulation characterised by increased heart rate and decreased autonomic activity have been found in individuals with ASD [38]. Cardiorespiratory dysrhythmias in children with ASD are associated with reduced cardiac vagal activity [39]. HRV reflects autonomic dysfunction, with reductions observed in high-frequency components and respiratory sinus arrhythmia (RSA) among the ASD population [40, 41]. The heterogeneity in HRV reactivity is observed in different social tasks and medications undertaken during assessments [41].

Current commercial products have facilitated the collection of ECG data across all age groups with widespread clinical applications in hospital settings. This allows ECG signals to become a primary source of physiological biomarker detection tools. As ECG has become an essential healthcare resource, the development of compact ECG devices has improved over the past decades from 12-electrode at the bedside to 2-electrode that can be worn on the move. In recent years, the demand for wearable sensors has further fuelled the development of embedded ECG signal acquisition. With the development of portable ECG devices, research and clinical services could explore the heterogeneity of HRV changes during social interactions in the ASD population [42, 43].

1.6 Thesis Outline

This thesis consists of six chapters, including an introduction (chapter 1), a systematic review and meta-analysis (chapter 2), three empirical studies (chapters 3-5), and a conclusion chapter (chapter 6). The thesis structure is illustrated in Figure 1.1.

1) Chapter 1 declares the research gaps, questions, and background information addressed in this thesis.
2) Chapter 2 conducts a systematic review and meta-analysis to summarise existing AI applications in ASD diagnosis and social behaviour analysis within social interaction contexts, with a focus on features such as facial information. The systematic review critically evaluates AI algorithms and feature modalities studied in the current literature during social interaction assessments. These findings are then categorised based on types of social interactions, revealing research gaps that can be addressed in this thesis. The meta-analysis rigorously evaluates the diagnostic test accuracy and odds ratio of applying AI models to ASD diagnosis. The findings strongly advocate the need to automate social behaviour detection exhibited by individuals with ASD in natural social interaction settings, especially in assessments involving multiple participants. Furthermore, it also emphasises that analysing these social biomarkers in conjunction with an ASD diagnosis using AI algorithms reduces subjective biases and labour costs. This chapter also discusses the acceptance of model performance thresholds for the AI algorithms proposed in each eligible study, exploring their potential deployment in clinical environments.
3) Chapter 3 addresses the first research question by developing a multi-view, multi-modal social interaction recognition framework that recognises direct eye contact behaviour between the child participants and the assessors during a free play task of social interaction assessment. The framework was built on deep learning models to extract feature patterns of eye contact from sparsely sampled true colour RGB images and stacks of movement trajectory features. By adopting a multi-modal feature learning mechanism, the framework enables deep learning models to recognise key social cues from a naturalistic social interaction context. It highlights the potential for objectively quantifying social behaviours to assist with clinical assessments.
4) Chapter 4 addresses the second research question through an empirical study. As joint attention is a pivotal social skill in ASD diagnosis, and the ability to convey and understand facial information is an essential social cue in assessments, this study focuses on investigating children's natural facial response characteristics during joint attention in a free play task with the assessor. The study aims to explore the potential applications of these facial features recorded from multiple facial orientations in ASD classification, compensating for current approaches that emphasise the need for frontal facial features. Given the nature of dependence and non-linear relationships among facial features, and the data imbalance between joint and non-joint attention episodes in naturalistic social settings, this study addresses these challenges using machine learning algorithms incorporating high-dimensional kernels for classification tasks. With the incorporation of feature dimensionality reduction techniques, the study examines facial features collected from different facial orientations to discriminate the ASD population from the TD group. The study adopts a comprehensive evaluation approach, combining a data-driven machine learning modelling approach with mutual validation through statistical methods to understand the underlying mechanisms of the selected facial features in ASD classification.
5) Chapter 5 develops a customised automatic ECG signal processing and HRV metrics calculation toolbox to address the third research question. Firstly, the adaption of the

NeuroKit2 toolbox is subjected to statistical equivalence tests using two ECG datasets with expert manual labelling to verify its consistency with HRV calculations derived from those manually inspected signals. The two datasets include: 1) the open-source MIT-QT database collected at rest and during exercise in the adult population, validating time-domain HRV metrics provided in the dataset, and 2) the baseline social interaction task from our clinical dataset of children aged 3-12 years used in this thesis, validating both the time- and frequency-domain HRV metrics, to assess the usability of the customised toolbox applied in diverse populations and heart conditions. The equivalence between metrics computed by the automated algorithm and manually cleaned ECG data is deemed critical for the toolbox to be valid in clinical practice. The validated algorithm is then applied to analyse changes in HRV in the social interaction assessments and determine the differences in HRV metrics between the ASD and TD groups.

6) Chapter 6 concludes the thesis, summarising the primary outcomes supporting the thesis argument, discussing the contribution of this thesis to the literature, and offering reflections on future work in applying multi-modal, multi-view feature learning strategies to assist with clinical assessments.

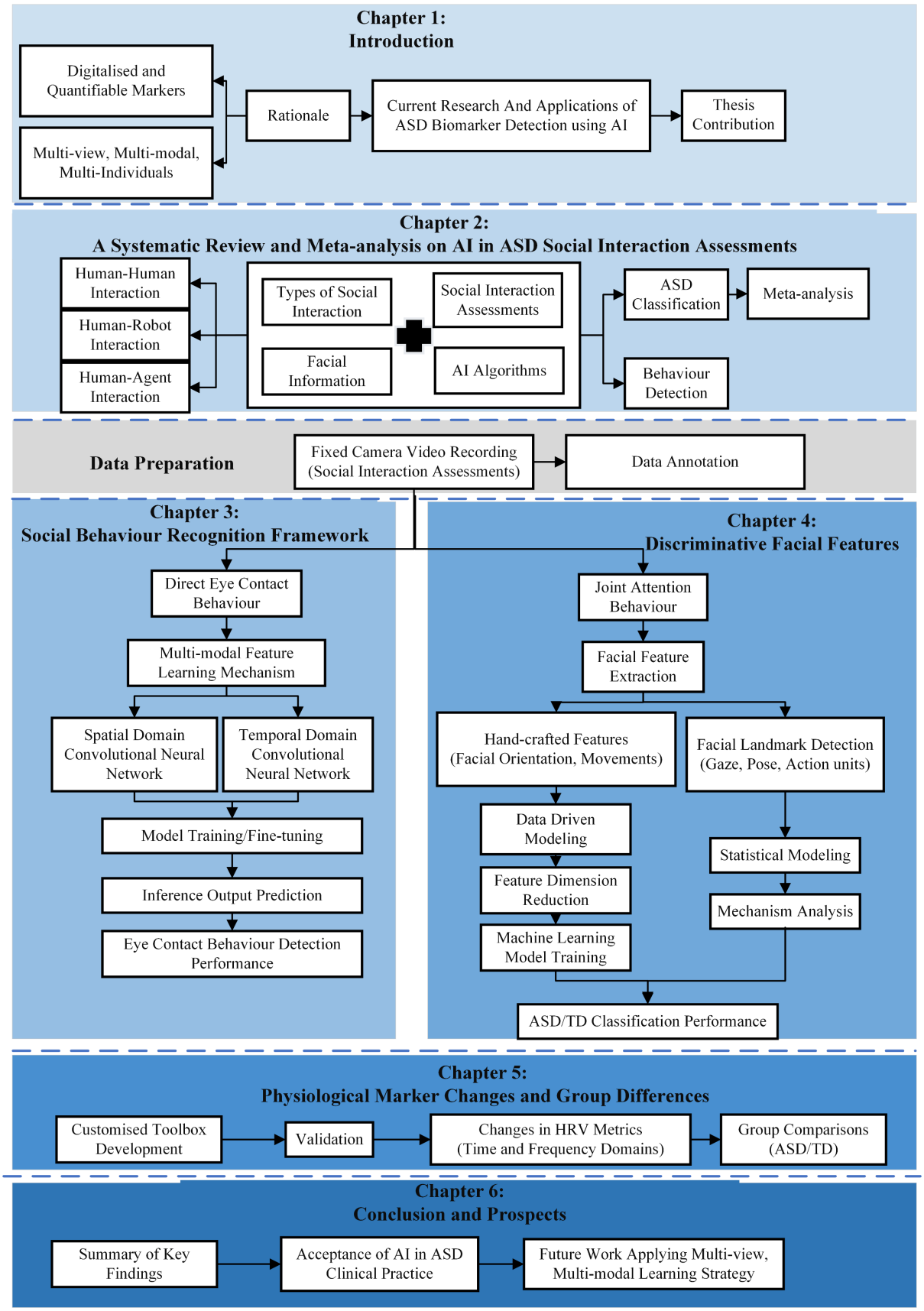


Figure 1.1. Thesis Structure Flow Chart.

Chapter 2: Identifying Autism Spectrum Disorder Using Artificial Intelligence Algorithms With A Focus On Social Interaction And Facial Information - A Systematic Review And Meta-Analysis

The literature on the use of AI technologies to assess emotional expression and atypical brain activities for ASD diagnosis has been extensively researched. However, little is known about using AI as an assistive technology for assessing social interaction skills and social competence in clinical settings. This chapter conducts a systematic review of the current literature on machine learning and deep learning algorithms for the assessment of ASD and intervention in social interaction settings. In considering that behavioural markers, such as facial and gaze responses in social interaction, are critical sources of information in psychological assessments, this research has focused on digitised information containing facial features. The systematic review summarises feature modalities and algorithms designed for social behaviour recognition and the use of these digital features in ASD diagnosis. In addition, a meta-analysis of studies reporting diagnostic test accuracy was conducted to evaluate the effectiveness of AI in assisting ASD diagnosis. Furthermore, this chapter discusses the gaps in existing research and provides future research directions on the use of automated algorithms for ASD diagnosis in social interaction settings. Lastly, this chapter provides the theoretical basis for the empirical studies presented in chapters three, four and five.

2.1 Introduction

Autism spectrum disorders (ASD) are a group of neurodevelopmental conditions that significantly impact social, communication and behavioural development [1, 44]. Over 1% of children worldwide are affected by ASD [1]. Despite the advancement of gold-standard assessment tools, quantifying social behaviours remains challenging due to subjective biases and time-consuming task in current observational coding schemes. Moreover, the overlap of behavioural phenotypes in ASD with other mental disorders further complicates clinical assessments [45, 46]. Current clinical diagnosis often integrates observational screening tools with parent/caregiver reports to reflect social competencies. However, this approach falls short in evaluating real-world naturalistic interactive environments, often leading to delayed diagnosis and missed early intervention opportunities. Therefore, the implementation of objective and quantifiable mechanisms for social competence assessments, ASD diagnosis and interventions is crucial to enhancing clinical diagnosis processes, reducing delays, and optimising intervention outcomes.

To assist the assessment process, digital features extracted from multi-modal data sources have the potential to perform objective data mining based on pattern recognition frameworks of AI models [47]. Familiar sources of digital features may include data obtained from video recordings, wearable sensors, neuroimaging, and genetic and molecular information. These digital features may play a critical role in standardising data collection procedures with greater objectivity, reducing human error or subjective bias in clinical assessments, and complementing traditional statistical approaches by analysing high-dimensional linearly non-separable features. With the increasing research in AI algorithms to explore ASD diagnostic boundaries, understanding underlying structural and functional responses associated with ASD criteria may be the basis for digital diagnosis and intervention in real-world clinical settings [34, 48, 49]. However, there remains ongoing debate on the ethical use of AI and the feasibility and acceptability of using AI in clinical practice [50].

There has been no current evaluation of the existing clinical evidence for using digital AI tools to identify ASD or ASD-like phenotypes to further inform this debate [51]. Existing studies exploring the use of AI to quantify digital features for the assessment of ASD has explored atypical behavioural features [13], including gaze patterns [52], emotion recognition [53], and motor movements [54], with a particular focus on facial features [55].

This systematic review and meta-analysis examined the existing literature on ASD research. Specifically, the aim of the systematic review was to identify AI techniques suitable for use in natural social interaction settings to support an ASD diagnosis, specifically incorporating facial features. This review also assessed the diagnostic test accuracy of machine learning and deep learning algorithms utilising facial information as digital phenotypical features in ASD diagnosis during social interaction assessments.

Further, the meta-analysis aimed to 1) determine the overall effects and heterogeneity of different input modalities and experimental tasks employed in ASD diagnosis through the use of AI. 2) The types of hardware setups and digital feature extraction algorithms applied in various types of social interactions, including interpersonal interaction, human-to-agent interaction and human-to-robot interaction for assessments and intervention applications.

2.2 Methods

A systematic review of published peer-reviewed studies was conducted to analyse the existing literature on the application of AI algorithms (including machine learning and deep learning) to assess natural social behaviours in a clinical setting and on the use of behavioural phenotypes, including facial features, as biomarkers for the diagnosis (categorising individuals with ASD and those in the TD group). The effectiveness of AI models trained for ASD diagnosis using digitised facial information in social interactions and its consistency with clinical assessment outcomes were further assessed through meta-analysis.

2.2.1 Protocol And Registration

This review was registered on PROSPERO with the registration number CRD42021249905. The guidelines of the preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA) [56, 57] were followed.

2.2.2 Eligibility Criteria

To assess the utility of applying AI algorithms in ASD studies, participants were required to have a formal diagnosis of ASD. If a study also included a TD group, the corresponding participants' demographic information was also extracted. The reference standards defined in this review were considered gold-standard diagnostic assessment tools used to classify the diagnostic status of each participant, for example, the use of ADI-R and ADOS-2 for ASD assessment [58]. The machine learning and deep learning algorithms reported in the eligible studies were referred to as index tests. Each reported model prediction result was used to check the performance of the index test against clinical diagnoses using the gold standard diagnostic assessment tool (the reference standards). Eligible studies also reported on experimental tasks incorporating a social interaction component to meet the aims of the review to analyse the feasibility and effectiveness of existing technologies in assessing natural social interactions between co-located social partners. The interactor could be a human, a social robot or a computer agent (such as video stimuli or avatar) who attracted the participants' attention through social interaction. Studies of any design (longitudinal, cross-sectional) in any setting were included, provided they met the inclusion criteria with no lower year limit. There were no limitations in terms of years, language, and type of publications to ensure no bias in identifying the relevant literature. The period of inclusion is determined by the search engine and includes all articles published to date.

2.2.3 Information Sources And Search

Studies were sourced from Medline, Embase, Scopus, Web of Science Core Collections, IEEE Xplore, and ACM Digital Library through the Sydney University Library website. The timeline was set for all articles that meet the search criteria on each database up until the 16th of January 2023. All the papers found through the above databases were exported to EndNote and Covidence for abstract screening on the 16th of January 2023. Search terms used in each database were based on four key concepts: 1) inclusion and focus on groups diagnosed with ASD or ASD-like population who later received formal ASD diagnosis, 2) use of artificial intelligence, 3) inclusion of facial and eye gaze measures, and 4) conducted in a social context. Detailed search terms are presented in Supplementary Table 1. A snowballing approach was conducted when full-text screening of the identified articles was completed.

2.2.4 Study Selection

The articles found from each database were first imported into EndNote, and duplicates were removed. The remaining studies were then exported from EndNote into Covidence for further duplicate removal and preparation for screening. Two authors (C.S. and A.S.) reviewed the study abstract and assigned a category label for future references to each article during the screening process based on the inclusion and exclusion criteria.

Inclusion Criteria

1. Including peer-reviewed articles only for data quality assurance; 2. The main study must include participants diagnosed with ASD in all age groups; 3. The experiments must be in a social interaction setting; 4. Machine learning or deep learning algorithms applied in the articles must include quantitative analysis evaluation methods; 5. Including only empirical studies presenting original data ensures methodological consistency and transparency and reduces potential bias from derived data in other review articles.

Exclusion Criteria

1. Animal studies or models developed based on animal experiments were excluded, as this review focuses on assessing human naturalistic reactions; 2. Not using machine learning or deep learning approaches; 3. Studies where the methods section did not explain how the models were trained were excluded due to the lack of transparency and replicability; 4. Studies were excluded from the analysis if their results section lacked evaluation of model performance for ASD classification or social biomarker recognition, indicating a deficiency in transparency.

In the preliminary screening process, conference papers were excluded if they did not include detailed full-text, to maintain the quality and reliability of the included literature, minimise bias, and ensure the integrity of the findings. Following screening based on title and abstract, the remaining studies were transferred to full-text review to confirm whether they met the inclusion criteria and select articles suitable for data extraction. The full-text articles were shared with all reviewers through EndNote. Conflicts were resolved by a third independent reviewer (E.D. or A.M. or A.G.).

Data extraction was completed by two reviewers (C.S. and A.S.) using a standardised data extraction template. While doing the full-text screening, references cited by the article authors were also reviewed for additional abstract screening. This way, the risk of bias was minimised, and potentially useful studies were further screened. Articles that meet inclusion criteria and reported comparison results of their proposed index tests with their reported reference standards for ASD diagnosis were included in the meta-analysis. The corresponding authors of the eligible article were contacted if the article did not have sufficient data to conduct the meta-analysis. Two authors provided the required data to be considered in the meta-analysis. Those eligible studies that still lack the data needed for meta-analysis were included in the systematic review .

2.2.5 Data Collection Process

The same reviewers (C.S. and A.S.) extracted the data from each included article. Data extraction tables were pre-defined to extract the raw data from included papers.

2.2.6 Definitions For Data Extraction

Variables extracted in the eligible studies included: the author of each study, year of publication, research focus, population (formal diagnosis of ASD, high-risk group with subsequent confirmed diagnosis of ASD, TD), ASD diagnostic tool (reference standards) and type of assessment (diagnostic, screening), other scales used, autism severity level, type of control group, participant age (mean, standard deviation) or age range, male distribution, sample size, IQ tests used, IQ selected tool, IQ type, IQ composite scores (mean and standard deviation), study design (type of experiments, experimental tasks, experiment duration, device used, interaction type), feature extraction algorithm (face detection, facial and head pose, movement, gaze, vocal, responsiveness), input modality, type of feature inputs, and type of output classes, model settings (datasets used, cross-validation and data augmentation management), classification algorithms (index tests), thresholds and performance measures (accuracy, precision, sensitivity (or recall), specificity, negative predictive value, positive predictive value, area under the curve (AUC), F1 score, Matthews correlation coefficient (MCC)).

An outcome is considered a true positive (TP) when it has been confirmed by field experts or by a gold standard diagnostic tool and the proposed index tests successfully predict it. An outcome is defined as a true negative (TN) when confirmed as a comparison group for TP. A false negative (FN) is defined when the actual outcome is positive but predicted negative. False positive (FP) is defined when the actual outcome is negative but predicted positive. The binary outcomes extracted from diagnostic test accuracy studies are the ASD (TP) and TD (TN) groups. The data extraction also included social biomarkers identified using machine learning or deep learning algorithms, such as “eye contact” (TP) and “no eye contact” (TN); “attention” (TP) and “distraction” (TN); “initiation of joint attention gaze shift detection” (TP) and “no joint attention detected” (TN); “social engagement” (TP) and “no social engagement” (TN). Other relevant social biomarkers, such as performance scores for responses to joint attention (non-binary outcomes) and responsiveness scores for assessment tasks, are also included in the systematic review. However, meta-analyses on diagnostic accuracy were conducted for binary classification of ASD and TD only.

2.2.7 Risk Of Bias Assessment And Applicability

The Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) tool [59] was used to assess the risk of bias and applicability. QUADAS-2 assesses each study in four domains: patient selection, index test, reference standard, and the flow and timing of patients in the study. Judgments of risk of bias and level of concerns were categorised as LOW/HIGH/UNCLEAR, and the tool was also used to help determine the level of risk of bias by completing further detailed sub-questions (Yes/No/Unclear). The total score was plotted as a bar chart indicating the risk of bias and applicability concerns. The same two independent reviewers (C.S. and A.S.) assessed the risk of bias and applicability using the QUADAS-2 tool, and discrepancies were discussed with other co-authors to reach a consensus.

2.2.8 Diagnostic Accuracy Measures

In the primary meta-analysis, contingency tables were prepared for the outcomes of each machine learning or deep learning algorithm to calculate proportional data: sensitivity and specificity and their corresponding 95% confidence intervals. Each element of the table (TP, TN, FP, FN) was extracted directly from the included studies or provided by the corresponding authors.

2.2.9 Synthesis Of Results

A systematic review of all eligible studies was first conducted. In the case where eligible studies reported different outcome results, when more than one combination of input modalities was explored, or when multiple algorithms (index tests) were used for ASD and TD classification or social behaviour recognition, each set of comparison data was considered a separate outcome. The eligible studies can be broadly grouped into binary outcomes and multiple outcomes. The studies that included binary outcomes with sufficient data (specificity and sensitivity, and contingency table) and designated the classification of ASD and TD are included in the meta-analysis.

2.2.10 Quantitative Meta-Analysis

Quantitative analyses were conducted, and the results were displayed using bar graphs, showing the number of outcome measures (observations) reported for each combination of classification algorithms and input data modalities. These were then grouped by different types of social interactions (human-human, human-agent, and human-robot interaction).

The *mada* and *metafor* packages recommended in the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy [60] were used for quantitative meta-analysis of the extracted data in R (version 4.2.2). Univariate analyses were performed for sensitivity, specificity, and diagnostic odds ratio (DOR), which were presented in the form of forest plots. A bivariate analysis was performed using a Summary Receiver Operating Characteristics (SROC) curve to determine the pooled diagnostic test accuracy. SROC curve was a plot of the sensitivity value (true positive rate) and 1-specificity (false positive rate) to find the trade-off between sensitivity and specificity. A multi-variate meta-regression analysis was conducted to investigate the potential sources of heterogeneity across the studies, such as the type of experimental tasks administrated. The analysis was performed on the DOR of each outcome, and effect sizes were pooled using proportional data with 95% of Clopper-Pearson confidence intervals (CIs) of a random effects model for each individual study. The inverse-variance pooling methods were applied to combine results from studies with varying sample sizes in the meta-analyses. The τ-squared variance was computed using the DerSimonian-Laird estimator to evaluate the between-study variances. If any of the values of the contingency table (TP/TN/FP/FN) is zero, a continuity correction of 0.5 is applied to avoid zero-counting problems. Various heterogeneity measures were reported to quantify the variability of the extracted data (Cochran's Q, I^2, τ^2).

To further investigate subgroup effects, subgroup analyses (or moderator analyses) were also conducted separately for machine learning models (index tests), input modalities and experimental tasks, and interactive effects were also examined (index tests and modality, a combination of all three moderators). Subgroup analyses were also performed using univariate DOR from random-effects models and weighted individual outcome measures for each study using the general inverse variance method. In addition, publication bias (with the trim-and-fill method) and robust variance analyses were also conducted to count small study effects and explore potential impact due to repeated measures reported on the same participants within studies.

2.3 Results

2.3.1 Study Selection

The literature search (completed on the 16th of January 2023) identified 36,226 abstracts, of which 10,665 were screened after removing 25,557 duplicates. There were 10,368 abstracts excluded as they did not meet the inclusion criteria at the abstract screening stage. The remaining 301 full-length articles, including four additional studies found during snowballing, were screened by the two independent reviewers, resulting in 27 studies fulfilling the inclusion criteria for the systematic review. Of the 27 eligible studies, fourteen focused on using machine learning or deep learning algorithms in ASD classification, suitable for a meta-analysis of diagnostic test accuracy. However, only seven studies provided sufficient data to conduct the analysis. The remaining studies were excluded from the meta-analysis because neither the sensitivity and specificity nor the two-by-two contingency table were provided. The detailed screening process is illustrated in the PRISMA 2020 flow diagram in Figure 2.1.

2.3.2 Study Characteristics

Table 2.1 summarised the key study characteristics of each eligible study, including the research focus, sample sizes, experimental tasks, interaction type, type of feature inputs, machine learning and deep learning algorithms (index tests), and dataset management procedures. More detailed information on assessment tools (reference standards), type of assessment, and demographic characteristics ((autism severity level, age, male participant number, IQ) were shown in Supplementary Table 2. Information on the types of experiments and experimental duration were summarised in Supplementary Table 3; input features, algorithms, and model performance were shown in Supplementary Table 4. Lastly, Supplementary Table 5 provided an overview of the feature extraction algorithms and toolboxes utilised in the eligible studies.

There were a total of 2,270 (1,138 ASD, 1,132 TD) participants' data reported in the eligible studies, aged from early childhood to early adolescence (0-13 years), and an adult group aged up to 32 years. Five studies did not provide information on the age the participants included. Reference standards, including gold standards diagnostic tools (ADOS/ADOS-2, DSM-5/DSM-IV-TR, ADI-R) and screening tools (AQ Questionnaire, M-CHAT, SCQ) for ASD diagnosis, were reported in 19 studies, and eight studies did not provide this information.

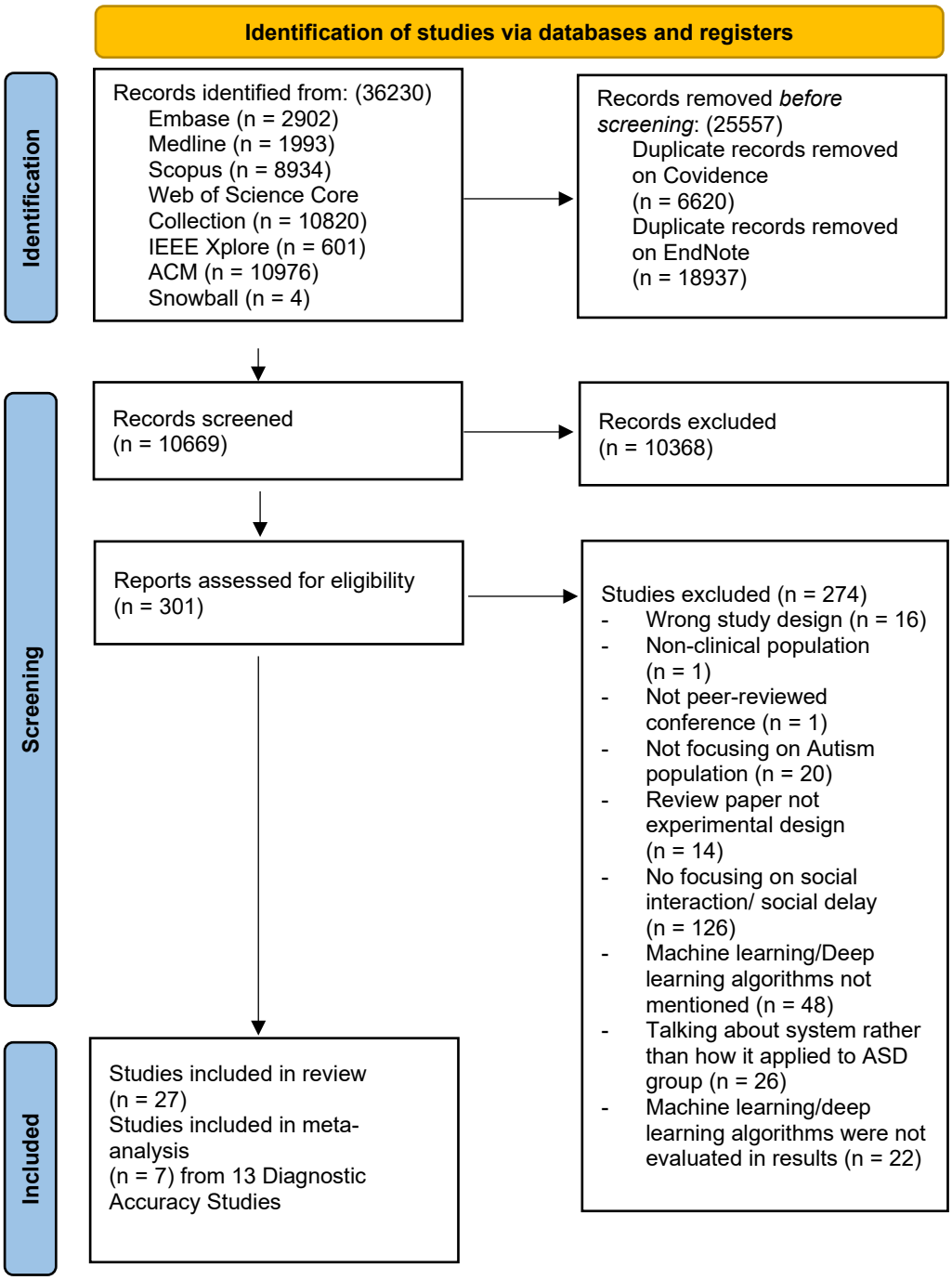


Figure 2.1. Literature search strategy flow diagram.

Table 2.1. Overview of eligible studies included in the systematic review.

Authors	Research Focus	Sample Sizes	Experimental Tasks	Interaction Type	Type of Feature Inputs	Machine Learning Algorithms	Dataset Management
C. Tang et al., 2020 [61]	ASD/TD Prediction	ASD: n=40, TD: n = 43	Still-Face Process	Human-Human Interaction	Head Movement Features, Facial Appearance Features, Vocal Features	SVM, K-NN	subject-independent 10-fold cross-validation, 90% training and 10% testing. **Threshold:** 0.7-0.8 for sensitivity and specificity, PPV (OR precision) > 0.5 acceptable
H. Drimalla et al., 2019 [62]	ASD/TD Prediction	ASD: n= 37, TD: n = 44	Simulated Interaction Task	Human-Human Interaction	Facial Expression, AU Intensity, Gaze Angle, Audio	SVM (radial basial function kernel), Random Forest, Stacked CNN, Pooled CNN	nested cross-validation (an outer leave-one-out cross-validation loop and an inner 3-fold cross-validation loop for hyperparameters tuning)

(Continued)

Table 2.1. *(Continued)*

Authors	Research Focus	Sample Sizes	Experimental Tasks	Interaction Type	Type of Feature Inputs	Machine Learning Algorithms	Dataset Management
H. Drimalla et al., 2020 [63]	ASD/TD Prediction	ASD: n= 37, TD: n = 43	Simulated Interaction Task	Human-Human Interaction	Facial Expression Features, Gaze Behaviour Features, Voice Characteristics	Random Forest	nested cross-validation (an outer leave-one-out cross-validation loop and an inner 3-fold cross-validation loop for hyperparameters tuning). **Threshold:** 0.5
K. S. Lohan et al., 2018 [64]	ASD/TD Prediction	ASD: n= 24, TD: n = 29	Avatar using "Blende" 2.6 for participants to learn RJA/IJA	Human-Agent Interaction	Pupil Diameter Data, AOI	Sequence to Sequence recurrent neural network (LSTM)	four-way cross-validation
D. Alie et al., 2011 [65]	ASD/TD Prediction	ASD: n= 6, TD: n = 26	Parents and Children Interaction	Human-Human Interaction	Gaze Pattern	VMM, HMM	leave-one-subject-out cross-validation
N. Qiu et al., 2020 [66]	ASD/TD Prediction	ASD: n= 40, TD: n = 43	The Still-Face Paradigm	Human-Human Interaction	Behavioural Characteristics	SVM, Naïve Bayes, Random Forest	-

(Continued)

Table 2.1. *(Continued)*

Authors	Research Focus	Sample Sizes	Experimental Tasks	Interaction Type	Type of Feature Inputs	Machine Learning Algorithms	Dataset Management
J. Zhang et al., 2022 [67]	ASD/TD Prediction	ASD: n= 143, (Aged 4-7 years), TD: n = 113, (Aged 4-7 years), TD adult: n = 43 (Aged 19-32 years)	Stimulus Video (Eye-Gaze-Shifting, Gesture Shifting)	Human-Agent Interaction	Valid Sampling Rate, Number of RJA Events	Logistic Regression	10-fold cross-validation
M. Varma et al., 2022 [68]	ASD/TD Prediction	ASD: n= 68, TD: n = 27	Social Gameplay Videos (Semi-Structured)	Human-Human Interaction	Gaze Fixation Patterns (Coordinates), Window And Shift Parameter Values Tuned.	LSTM	held-out test set. **Threshold:** Bonferroni-corrected significance threshold of 0.0031
M. Alcaniz et al., 2022 [69]	ASD/TD Prediction	ASD: n= 35, TD: n = 20	VR Static And Dynamic Social And Non-Social Stimuli	Human-Agent Interaction	Number of AOIs seen per scene on social and non-social items and humans	Naïve Bayes, XGBoost, K-NN, Random Forest, and SVM	5-fold cross-validation

(Continued)

Table 2.1. *(Continued)*

Authors	Research Focus	Sample Sizes	Experimental Tasks	Interaction Type	Type of Feature Inputs	Machine Learning Algorithms	Dataset Management
I. Akin-Bulbul et al., 2022 [70]	ASD/TD Prediction	ASD: n = 61, TD: n = 72	Passive Visual Stimuli Social Interaction and Toy Dataset; Passive Visual Stimuli Social Interaction and Animation Dataset	Human-Agent Interaction	Fixation Count, Fixation Duration, Dwell Time, Net Dwell Time, Diversion Duration	Decision Tree, Naïve Bayes, Random Forest, SVM	10-fold cross-validation, Feature selection (ReliefF, Information Gain, Wrapper)
A. Adham et al., 2022 [71]	ASD/TD Prediction	ASD: n= 54, TD: n = 32	A Battery of Eye Tracking Video Stimuli (Activity Monitoring, Social Referencing, Theory of Mind, Dyadic Bid Tasks)	Human-Agent Interaction	Eye Gaze Scan Path (Spatial, Temporal, Gaze Velocity)	DNN, CNN	train/test ratio (7:3) **Threshold:** velocity thresholds
Z. Zhao et al., 2021 [72]	ASD/TD Prediction	ASD: n= 19 from 20 recruited, TD: n = 20 from 23 recruited	Structured Conversation	Human-Human Interaction	Gaze Fixation Time (Three Features)	SVM, Linear Discriminant Analysis, Decision Tree, Random Forest	leave-one-out-cross-validation. **Threshold:** velocity thresholds
P. R. Krishnappababu et al., 2021 [73]	ASD/TD Prediction	ASD: n= 40, TD: n = 396	Recorded 6 Movies with Social and Non-Social Components	Human-Agent Interaction	Integrated Entropy, Positive Energy on Different Movies	-	leave-one-out-cross-validation
N. Kojovic et al., 2021 [74]	ASD/TD Prediction	ASD: n= 68+101, TD: n = 68	Social Interaction Videos (ADOS Assessment)	Human-Human Interaction	Skeletal Information, Video 5 Seconds	VGG16 CNN+ LSTM	80-20 train-test split

(Continued)

Table 2.1. *(Continued)*

Authors	Research Focus	Sample Sizes	Experimental Tasks	Interaction Type	Type of Feature Inputs	Overall Machine Learning Algorithms	Dataset Management
N. Zhang et al., 2022 [75]	ASD classification (3 levels)	ASD: n= 33	ADOS-2 Interview	Human-Human Interaction	Local Phase Quantization in Three Orthogonal Planes (LPQ-TOP)	Dictionary learning through sparse coding (K-SVD, k-singular value decomposition) + dimensionality reduction through factor analysis (Marginal Fisher Analysis)	10-fold cross-validation **Threshold:** 0.75
J. T. Megerian et al., 2022 [76]	Multiple level prediction (ASD/ not ASD/ Indetermined)	ASD: n= 122	Questionnaire + two social interaction videos	Human-Human Interaction	-	-	**Threshold:** PPV greater than 65% and NPV greater than 85%
A. Di Nuovo et al., 2018 [77]	Attention Estimation (Attention/ Distraction)	ASD: n= 6	VB-MAPP Imitation Tasks	Human-Robot Interaction	Face Detection, Facial Landmark	Naïve Classification, K-NN Classifier	10-fold cross-validation
C. Eunji et al., 2017 [78]	Eye Contact/ No Eye Contact	ASD: n= 50, TD: n = 50	Semi-Structured Play Interaction	Human-Human Interaction	Eye Contact Detection	PiCNN, AlexNet, PEEC, GazeLocking	5-fold cross-validation, 80% training, 20% testing for each disjoint train/test splits **Threshold:** 0.9

(Continued)

Table 2.1. *(Continued)*

Authors	Research Focus	Sample Sizes	Experimental Tasks	Interaction Type	Type of Feature Inputs	Overall Machine Learning Algorithms	Dataset Management
E. Chong et al., 2017 [79]	Detection of Gaze Shift for Joint Attention	ASD: n= 8, TD: n = 8	ESCS (Early Social Communication Scales)	Human-Human Interaction	Gaze Shift	SVM	20% of samples trained
G. Nie et al., 2018 [80]	Predict RJA Performance Level (Low/Medium/High)	ASD: n= 34	ESCS (Early Social Communication Scales) (RJA part)	Human-Robot Interaction	GMM Of Head Pose, Hard Histogram Of Head Pose, Soft Histogram Of Head Pose, Hard Histogram Of Head Pose Motion, Soft Histogram Of Head Pose Motion	Transductive Support Vector Machine (TSVM) with radial basis function kernel	-
H. Javed et al., 2020 [81]	Social Engagement Estimation	ASD: n= 5, TD: n = 13	Social Engagement Estimation (Social interaction included)	Human-Robot Interaction	Movement + Expression	CNN, SVM, Random Forest, Decision Tree, K-NN	10-fold cross-validation (train/test split of 0.8/0.2
E. Chong et al., 2020 [82]	Eye Contact/ No Eye Contact	ASD: n= 66, TD: n = 36	Social Interaction Play with examiner (ESCS and BOSCC)	Human-Human Interaction	Cropped Face Regions	Deep Learning (ResNet-50 backbone)	**Threshold:** 0.9
W. B. Liu et al., 2017 [83]	Responsiveness score prediction	ASD: n= 22, TD: n = 21	Response to Name	Human-Human Interaction	Latency, Weighted Duration, Latency + Weighted Duration	Decision Tree	-

(Continued)

Table 2.1. *(Continued)*

Authors	Research Focus	Sample Sizes	Experimental Tasks	Interaction Type	Type of Feature Inputs	Overall Machine Learning Algorithms	Dataset Management
G. Zhang et al., 2021 [84]	Mutual Gaze Score	ASD: n= 10	Randomized trial (Standard Therapy: table-top reading activity, play therapy: drumming activity, singing activity)	Human-Human Interaction	Prediction Scores and Ground Truth by Experts	Deep CNN (Three branches)	pretrained on AVA/UCO-LAEO, AFLW datasets **Threshold:** mutual gaze score (0.6)
C. Wu et al., 2021 [85]	Facial Behaviour Recognition (Smile, Look Face, Look Object, Vocalization)	ASD: n= 133 at risk (high/low)	adult-child play task	Human-Human Interaction	Image, Facial Features	ResNet-18	horizontal flip
J. Li et al., 2021 [86]	Emotion Recognition (Negative And Non-Negative, Neutral, And Positive)	ASD: n= 6	therapeutic games for children to play (ASD-affect dataset)	Human-Human Interaction	Facial And Speech Emotion	ResNet-18	data augmentation and 5-fold cross-validation **Threshold:** 0.5
D. Anagnostopoulou et al., 2021 [87]	Estimate Engagement Level (1-3)	ASD: n= 7, TD: n = 25	play with robots (show me the gesture, express the feeling, pantomime, guess the object and joint attention)	Human-Robot Interaction	Pose Features	AlexNet,2D CNN, LSTM-based RNN, 1D multi-channel CNN	train/validation, data augmentation (flip vertically with 0.3 probability, add gaussian noise with 0.5 probability)

IJA, initiation of joint attention; RJA, response to joint attention; AOI, areas of interests; SVM, support vector machine; LSTM, Long Short-Term Memory; VMM, Variable-order Markov Models; HMM, Hidden-order Markov Model; K-NN, K-nearest neighbour; PPV, positive predictive value; NPV, negative predictive value; AU, action units.

2.3.3 Risk Of Bias And Applicability

Quality assessment of the eligible studies was conducted using QUADAS-2 [59] (as shown in Figure 2.2 and Supplementary Table 6). Of the final 27 studies selected, eighteen were considered to have a low risk of bias in patient selection because of a clear description of patient enrolment, inclusion, and exclusion criteria. One study lacked a clear description of patient selection and was therefore categorised as unclear. Eight studies were classified as high risk of bias due to highly imbalanced participant numbers or because the included participants were classified as high-risk and had not received a formal diagnosis of ASD.

All 27 studies were categorised as low risk of bias for both index tests and flow and timing domains, as the selected papers all had at least one machine learning or deep learning model (index test), and the model training procedures were conducted after data collection and cleaning. Therefore, there were no strict requirements on the timing. Of these, twenty-two studies reported reference standards and therefore classified as low risk of bias. The remaining five studies were classified as unclear due to a lack of description of the exact reference standard applied at enrolment.

For applicability concerns, one study reported participants who were at low or high risk of ASD but were not confirmed at enrolment. Therefore, this study was rated as unclear regarding whether the recruited patients matched the review question. The remaining studies showed low risk in terms of patient selection. The index test domain was consistent with the review question; therefore, all 27 studies were rated low risk. There were 25 studies rated as low risk for concerns about reference standards for target conditions, while only two studies were classified as unclear because they did not provide sufficient information to confirm their applicable reference standards.

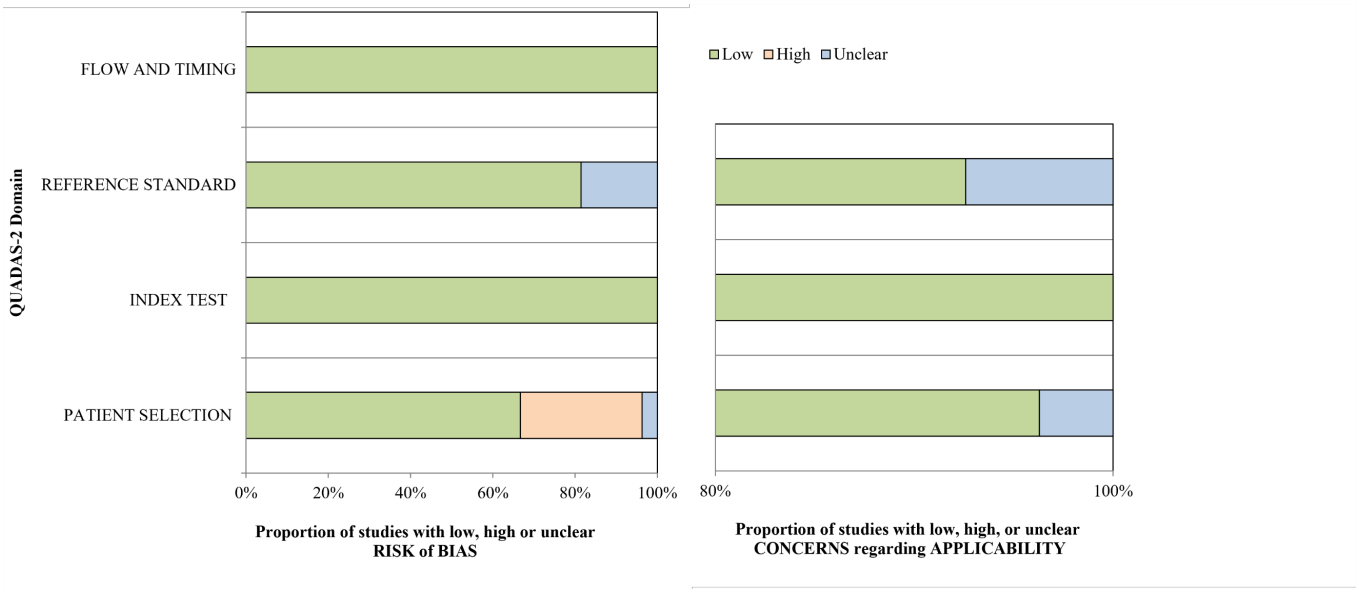


Figure 2.2. Quality assessment of the eligible studies by the QUADAS-2 checklist.

2.3.4 Systematic Review: Social Interaction Tasks And Artificial Intelligence Algorithms Applied To Autism Diagnosis

Figure 2.3 shows the types of classification algorithms and the types of input data modalities applied in ASD diagnosis studies and behaviour recognition studies. Within human-to-robot interaction studies, there was an investigation into facial and posture features for training machine learning models. For behaviour detection, deep neural networks, including ResNet and other CNN models, were predominantly used in human-to-human interaction studies. In studies reporting ASD classification, facial and gaze features were commonly trained on well-established machine learning models, with fewer studies exploring deep neural networks. This is consistent with studies involving interactions between humans and human-to-agent interaction. In addition, gaze pattern features were only trained on the LSTM deep learning algorithm in human-robot interaction settings.

Structured Interpersonal Social Interaction

Traditional diagnostic assessments require trained clinicians to interact with the participants as guided by structured plays or interviews, such as the ADOS-2. These assessment tools involved human-human interactions to assess the behavioural differences between people with ASD and TD. For instance, Tang et al. [61] used head movement, facial appearances and vocal features extracted from a two-minute "mother amused child without touch of body" session and a one-minute still-face paradigm to discriminate infants at high-risk of ASD from TD. The machine learning algorithms applied in the study are K-Nearest Neighbour (K-NN) and Support Vector Machine (SVM). Qiu et al. [66] used the same still-face paradigm data and applied children's behaviour characteristics for this classification task on Naïve Bayesian (NB), Random Forest (RF), and SVM algorithms. To retain a standardised diagnostic process for social interaction assessment, Kojovic et al. [74] video recorded each ADOS-2 assessment administrated in their study and applied a VGG-16-based convolutional neural network to extract pose features, and opted in the Long Short-Term Memory (LSTM) architecture for ASD classification. Similarly, Zhang et al. [75] analysed video recordings of participants aged 16-37 years receiving ADOS-2 interviews for ASD diagnosis. A discriminative few-shot learning approach was used to extract facial dynamic information (facial appearance, static expression, and eye movements) from interview scenes to derive predictions of ASD classification and match the classification into three categories (autism, autism spectrum and non-spectrum), which in line with the ADOS-2 assessment scheme.

Other studies have used the recorded ADOS-2 assessment tool [58, 88] to categorise participants into ASD or TD groups, then designed their unique tasks to assess participants' social behaviours observed during the assessment to explore the potential use of machine learning algorithms in discriminating ASD from TD. For example, Zhao et al. [72] carried out a structured face-to-face conversation task between an interviewer and child participants, and features on eye gaze fixation duration in each area of interest (AOI) and the duration of each conversation task were parsed for machine learning model training (algorithms included Linear Discriminant Analysis (LDA), Decision Tree (DT), RF and SVM) for ASD classification. Drimalla et al. [62, 63] introduced a simulated interaction task to extract ASD-related social biomarkers (facial expression, gaze, and voice) for ASD classification, and the performance of RF, SVM and CNN models were compared. Alie et al. [65] proposed a three-minute parent-child face-to-face interaction task. They applied Markov models to analyse the eye gaze

patterns of participating infants to explore the potential for identifying infants at risk of ASD at an early age.

Unstructured Interpersonal Social Interaction

Previous studies have also explored the potential of ASD diagnosis using video recordings collected in non-clinical and unstructured home play environments. Varma et al. [68] used automatically annotated eye gaze fixation and scan-path features of their recruited child participant, applying an LSTM-based neural network model to obtain preliminary results of 60% accuracy. Megerian et al. [76] reported a medical device currently under a multi-site cohort clinical trial to provide a preliminary ASD screening as a reference for healthcare providers. In their studies, caregiver and healthcare provider questionnaires and two home-taken videos were analysed, and participants were rated using a gradient-boosted decision tree model (autistic, non-autistic and indeterminate).

Human-Agent Interaction

An increasing number of research findings suggest that behavioural markers, such as eye gaze, facial expression, and body movements, can be used for diagnosing ASD. The inclusion of external devices for human-agent interaction as part of clinical assessments has gained attention [89, 90]. Zhang et al. [67] proposed a logistic regression machine learning model to identify ASD by quantifying participants' responses to joint attention to video stimuli containing eye-gaze and gesture-shifting scenes. Akin-Bulbul et al. [70] demonstrated that machine learning algorithms such as DT, NB, RF, and SVM could be trained on eye gaze features collected from social (conversation and play between actor and a toddler) and non-social (toys and animations) video stimuli to perform ASD screening. Atyabi et al. [71] proposed a mechanism that uses a collection of eye-gaze scan path, temporal information, and pupil velocity features trained on CNN and DNN models to classify ASD and TD. The video stimuli tasks conducted in their study contain Activity Monitoring, Social Referencing, Theory of Mind, and Dyadic Bid. Other movie stimuli tasks containing both social and non-social components were introduced by Krishnappababu et al. [73]. They suggested that atypical facial dynamics and facial expressions recorded when participants were watching the stimulus on a tablet could assist with ASD prediction using DT algorithms. The movie stimulus consists of the blowing bubbles, spinning top, rhymes and toys, make me laugh, mechanical puppy, and dog in grass right-right-left tasks. Alcaniz et al. [69] proposed that Virtual Reality (VR) containing static and dynamic social and non-social stimuli could be an assessment tool for ASD diagnosis. They investigated the prediction performance of gaze features recorded during VR play trained through conventional machine learning algorithms such as SVM, RF, NB, XGBoost and K-NN. Lohan et al. [64] suggested that children's eye gaze movements guided by avatars could be used for training responses to joint attention and identify ASD from the TD group through a trained LSTM model.

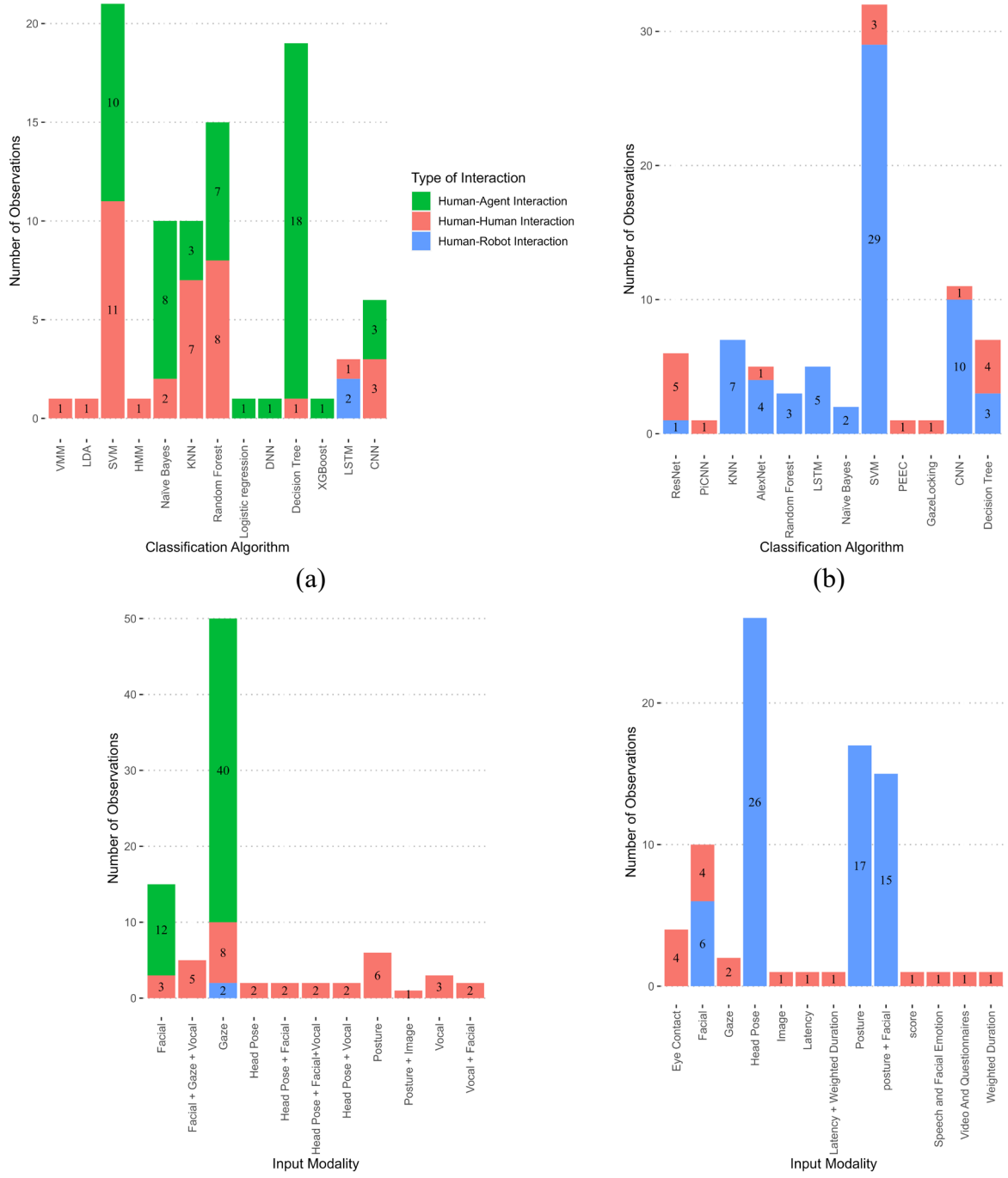


Figure 2.3. Bar chart illustrating the input modalities and the classification algorithms applied in autism diagnosis and social behaviour detection studies. For all plots, the results are displayed in descending order of accuracy. The subplots are categorised based on the type of social interactions: Human-to-agent interaction (green), Human-to-human interaction (red) and Human-to-robot interaction (blue). Subfigure a) presents the number of observations (outcomes) corresponding to different classification algorithms employed in ASD diagnosis. Subfigure b) presents the use of classification algorithm in behaviour detection studies. Subfigure c) presents the types of input modalities used for training AI models in ASD diagnosis studies. Subfigure d) presents the types of input modalities used in behaviour detection studies.

2.3.5 Quantitative Meta-Analysis Of Diagnostic Accuracy Tests For ASD

Of the 27 studies in this systematic review, 14 focused on using machine learning and deep learning algorithms to distinguish ASD from TD groups in social interaction assessments. Of these studies, only seven provided sufficient information (TP/TN/FP/FN values) to accumulate a total of 32 contingency table samples for synthesis of the main effect (Table 2.2). These samples were drawn from five different human-to-human interaction experiments (Still-Face paradigm, Simulated Interaction Task, Face-to-Face Interaction, Structured Face-to-Face Conversion, and the ADOS-2 Diagnostic Assessment). They contained 135 individuals diagnosed with ASD and 166 from the TD group. The remaining studies on ASD diagnosis were not included in this meta-analysis due to a lack of information on the two-by-two contingency table.

The results of each univariant meta-analysis (Figure 2.4) revealed a significant overall effect. Reported pooled sensitivity was 0.773 (95% CI [0.711 to 0.825]) with significant differences observed between experimental tasks ($Q(4) = 23.74$, $p<0.0001$), and pooled specificity of 0.837 (95% CI [0.788 to 0.876]) with observed significant between-group differences among different experimental tasks ($Q(4) = 20.25$, $p<0.001$). A significant overall heterogeneity in sensitivity ($I^2 = 78\%$, $\tau^2=0.6521$, $p<0.0001$) and specificity ($I^2 = 75\%$, $\tau^2=0.5946$, $p<0.0001$) were also observed. Furthermore, Figure 2.5. revealed a significant pooled effect on the diagnostic odds ratio (DOR) of 19.929, 95% CI [11.181 to 35.523], and significantly high heterogeneity ($I^2 = 83.6\%$, $\tau^2=2.230$, $p<0.0001$). The between-group differences among experimental tasks were significant ($Q(4) = 28.31$, $p<0.0001$). A bivariant SROC curve (Figure 2.6) demonstrated a significantly high pooled sensitivity and specificity of 0.776 (95% CI [0.712 to 0.829]) and 0.839 (95% CI [0.790 to 0.879]), respectively, with an AUC of 0.878.

Multi-variate Meta-Regression Analysis

The study-level covariant effects as a function of the experimental tasks were conducted to explore the potential sources of heterogeneity observed across the studies. Figure 2.7 revealed a significantly high pooled estimates of 13.92 (95% CI [4.97 to 38.98], $p<0.001$) and between-study variation ($\sigma^2 =0.974$, $Q(31) = 186.417$, $p<0.001$). It suggests that the between-study heterogeneities cannot be explained by differences in experimental tasks alone.

Table 2.2. Contingency table for the meta-analysis.

Study	Sample Size (ASD/TD)	Experimental Tasks	Classification Algorithm	Modality	TP	FP	TN	FN	Sensitivity	Specificity
C. Tang et al., 2020 [61]	40/43	Still-Face	SVM	Head Pose	36	30	13	4	0.9	0.302
				Facial	34	3	40	6	0.85	0.930
				Vocal	36	6	37	4	0.9	0.861
				Head Pose + Vocal	36	7	36	4	0.9	0.837
				Head Pose + Facial	34	2	41	6	0.85	0.954
				Vocal + Facial	38	1	42	2	0.95	0.977
				Head Pose+ Facial+ Vocal	38	1	42	2	0.95	0.977
			K-NN	Head Pose	17	15	28	23	0.425	0.651
				Facial	26	2	41	14	0.65	0.953
				Vocal	36	7	36	4	0.9	0.837
				Head Pose + Vocal	36	6	37	4	0.9	0.861
				Head Pose + Facial	26	2	41	14	0.65	0.954
				Vocal + Facial	37	2	41	3	0.925	0.954
				Head Pose + Facial+ Vocal	36	2	41	4	0.9	0.954

(Continued)

Table 2.2. *(Continued)*

Study	Sample Size (ASD/TD)	Experimental Tasks	Classification Algorithm	Modality	TP	FP	TN	FN	Sensitivity	Specificity
H. Drimalla et al., 2019 [62]	36/41	Simulated Interaction Task	Random Forest	Facial + Gaze + Vocal	24	7	34	12	0.667	0.829
			SVM		23	9	32	13	0.639	0.780
	36/40		CNN		9	14	26	27	0.25	0.650
					19	13	27	17	0.52	0.675
H. Drimalla et al., 2020 [63]	36/43	Simulated Interaction Task	Random Forest	Facial	22	12	31	14	0.611	0.721
				Gaze	19	17	26	17	0.528	0.605
				Vocal	26	10	33	10	0.722	0.767
				Facial + Gaze + Vocal	24	9	34	12	0.667	0.791
D. Alie et al., 2011 [65]	6/26	Parents and Children Interaction	VMM	Gaze	6	2	24	0	1	0.923
			HMM		2	2	24	4	0.333	0.923
N. Qiu et al., 2020 [66]	40/43	Still-Face	Naïve Bayes	Posture	30	7	36	10	0.75	0.837
					31	6	37	9	0.775	0.860
			Random Forest		31	7	36	9	0.775	0.837
					31	5	38	9	0.775	0.884
			SVM		35	10	33	5	0.875	0.767
					34	7	36	6	0.85	0.837
Z. Zhao et al., 2021 [72]	19/20	Conversation	SVM	Gaze	16	0	20	3	0.842	1
N. Kojovic et al., 2021 [74]	169/68	ADOS Assessment	CNN	Posture + Image	29	8	26	5	0.854	0.765

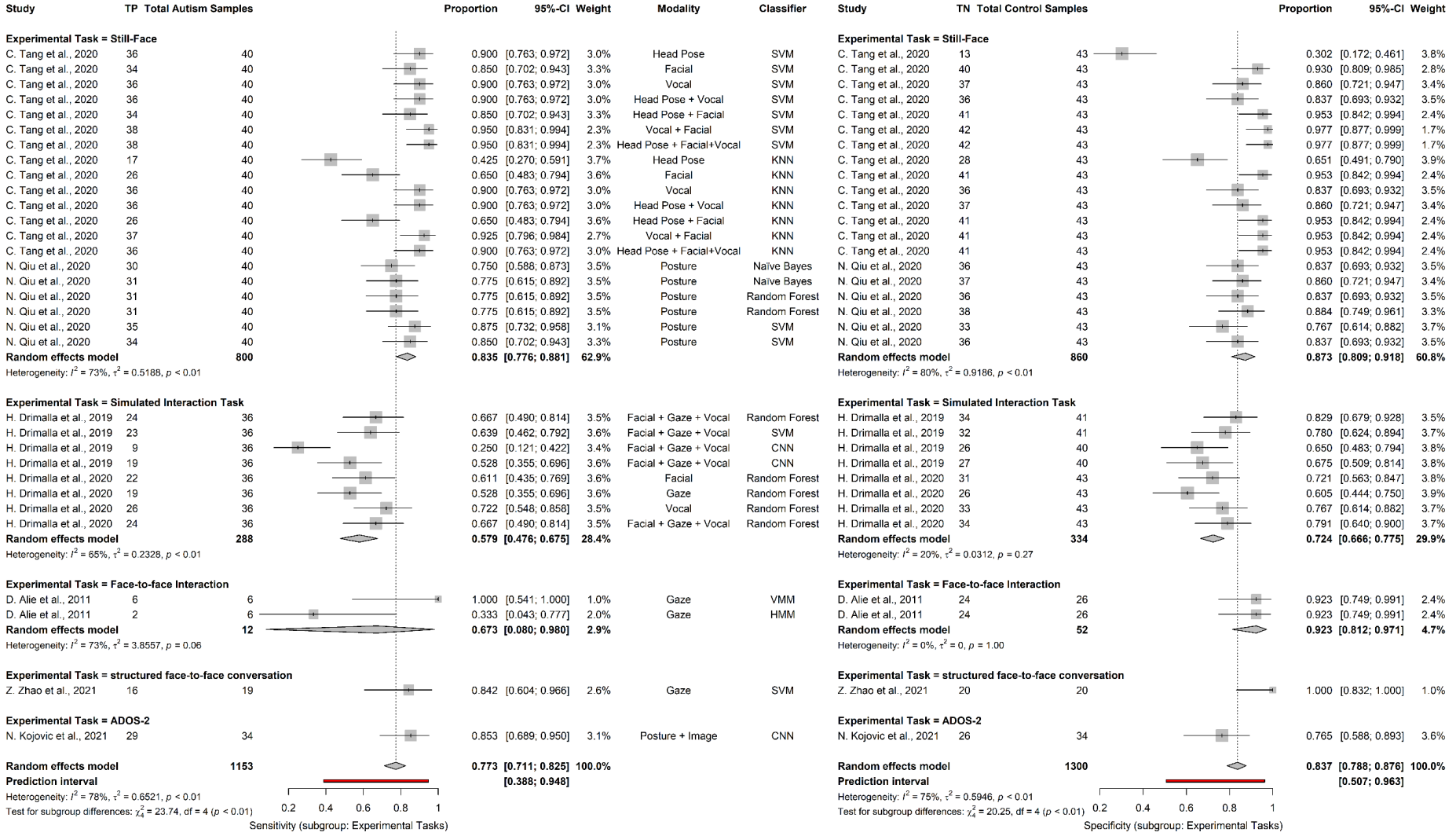


Figure 2.4. Forest plot of pooled sensitivity and specificity for an univariant meta-analysis.

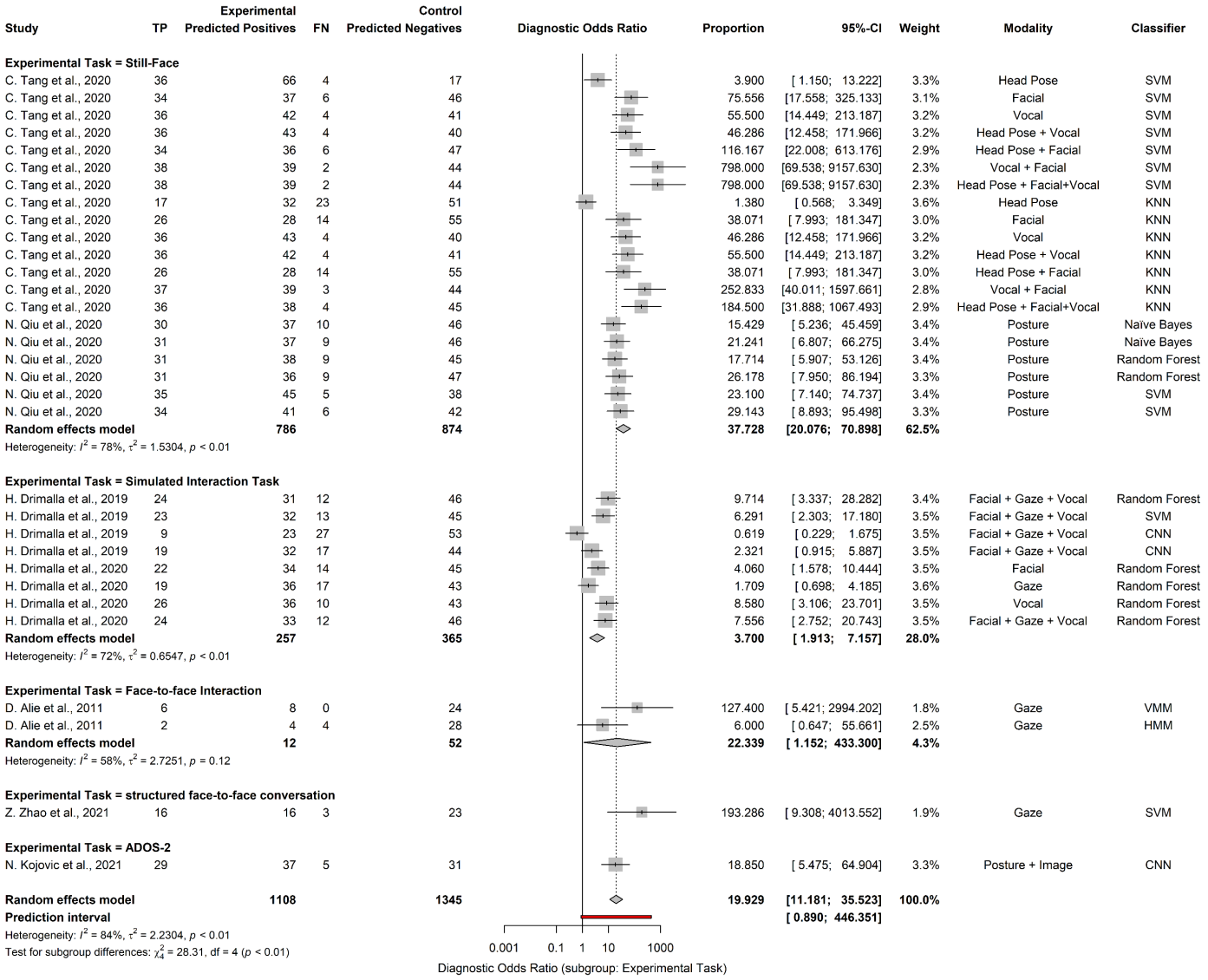


Figure 2.5. Forest plot of pooled diagnostic odds ratio for an univariant meta-analysis.

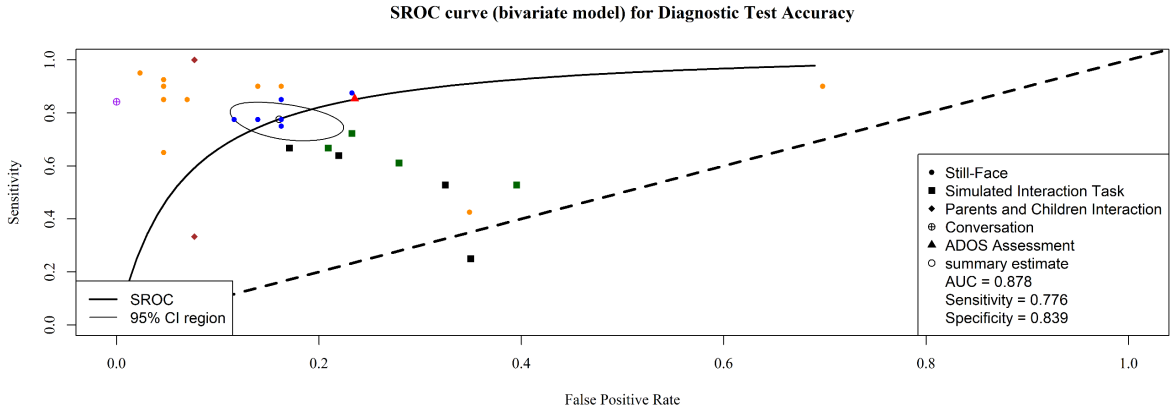


Figure 2.6. Summary Receiver Operating Characteristics Curve of the studies eligible for meta-analysis.

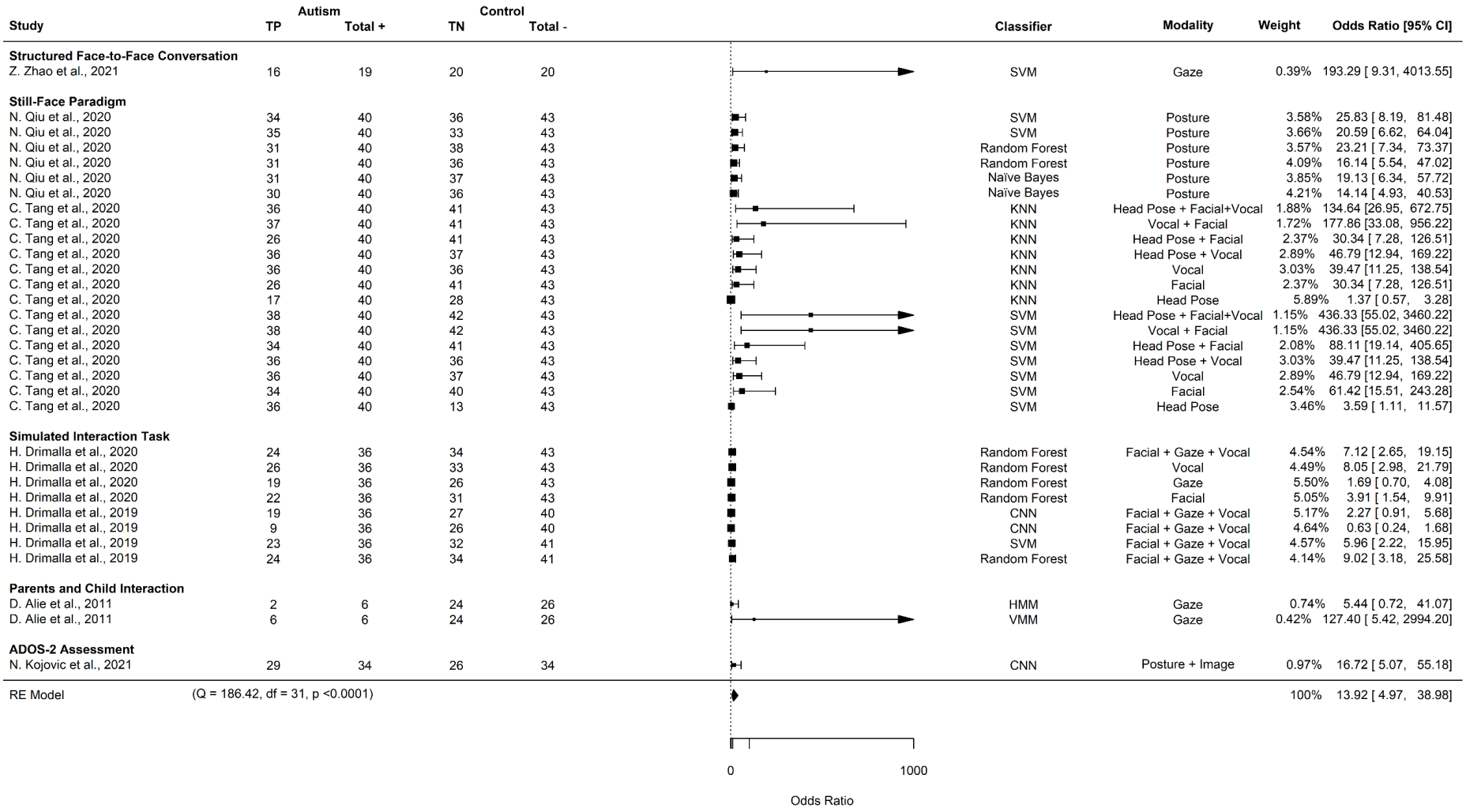


Figure 2.7. Forest plot of multi-variate meta-regression with the experimental task of each reported studies as a covariate.

Moderator Analysis

The between-study omnibus tests (k=32, n=7) of all moderators was statistically significant (p-value < 0.0001), which reveals that moderators including classification algorithms, input data modality and experimental tasks applied in ASD diagnostic studies all significantly impact the effect sizes observed in the seven included studies. For example, for the classification algorithm subgroup, the estimated variance components (σ^2= 1.399) were determined with tests of residual heterogeneity (QE(df = 25) = 127.563, p-value < 0.0001) and moderators (QM(df = 7) = 45.579, p-value < 0.0001). Similarly, for the modality subgroup, the estimated variance components (σ^2= 2.912) were obtained with test of residual heterogeneity (QE(df = 21) = 56.737, p-value < 0.0001), and moderators (QM(df = 11) = 98.519, p-value < 0.0001). The findings suggest that the observed heterogeneity between studies cannot be explained by data modalities and classification algorithms alone. However, when multiple moderators were considered (combining both the classification algorithms and modalities), most of the heterogeneity across studies was explained (with the estimated variance components σ^2 = 2.9110, test of residual heterogeneity (QE(df = 15) = 24.582, p-value < 0.0558), and test of moderators (QM(df = 17) = 117.783, p-value < 0.0001)). When considering the types of experimental tasks as a moderator, the estimated variances components (σ^2= 0) were observed with a test of residual heterogeneity (QE(df = 13) = 6.501, p-value =0.926, and QM(df = 19) = 667.161, p-value < 0.0001). Detailed moderator effects are shown in Supplementary Table 7-10.

Publication Bias

Funnel plot asymmetry was detected (Figure 2.8), indicating a possible small-study effect with a significantly larger intercept of the regression model ($\widehat{\beta_0}$= 6.353, t= 5.11, p<0.0001, multiplicative residual heterogeneity variance (τ^2= 9.092)). A trim and fill analysis (Figure 2.9) imputed 13 studies, and the bias-corrected estimates of the true effect is 1.776 (95% CI [1.106 to 2.447], z-value = 5.190, p-value <0.0001), quantifying heterogeneity (τ^2= 4.715 [3.191 to 8.557]; τ= 2.1715 [1.786 to 2.925], I^2 = 94.4% [93.3% to 95.4%]; H = 4.24 [3.86 to 4.65], Q(44) = 790.20, p-value <0.0001). This indicates a high small-study publication bias.

Sensitivity Analysis

Sensitivity analyses, shown in Figure 2.10, compared a correlated effects model with small-sample corrections. The robust variance estimation analysis revealed that the model assumptions of the main analysis were robust (I^2 = 80.04%, τ^2=1.82, estimates = 2.72, t= 5.56, p=0.00176, 95% CI [1.51 to 3.94]).

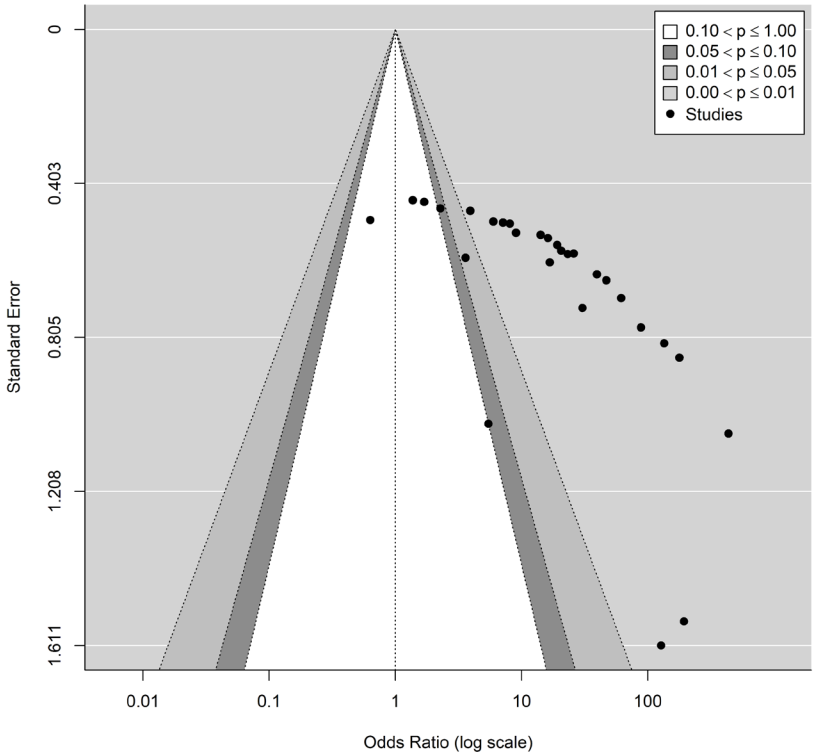


Figure 2.8. Funnel plot for publication bias.

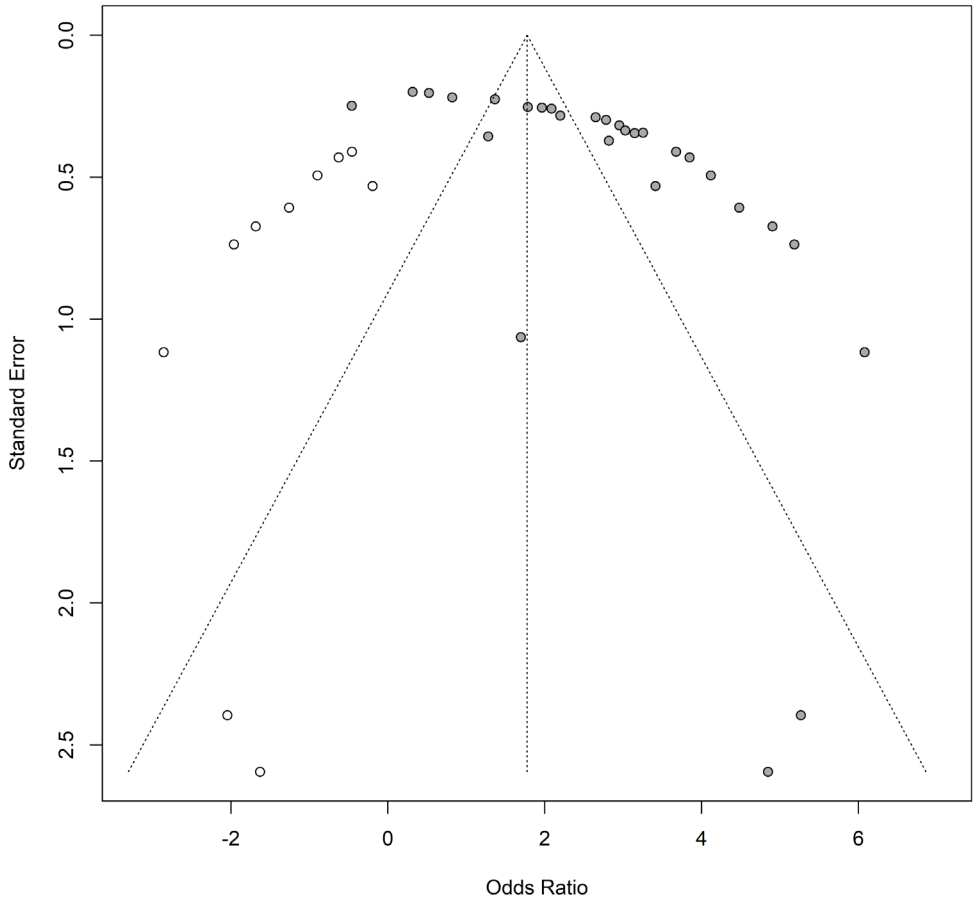


Figure 2.9. A funnel plot with trim and fill analysis on publication bias.

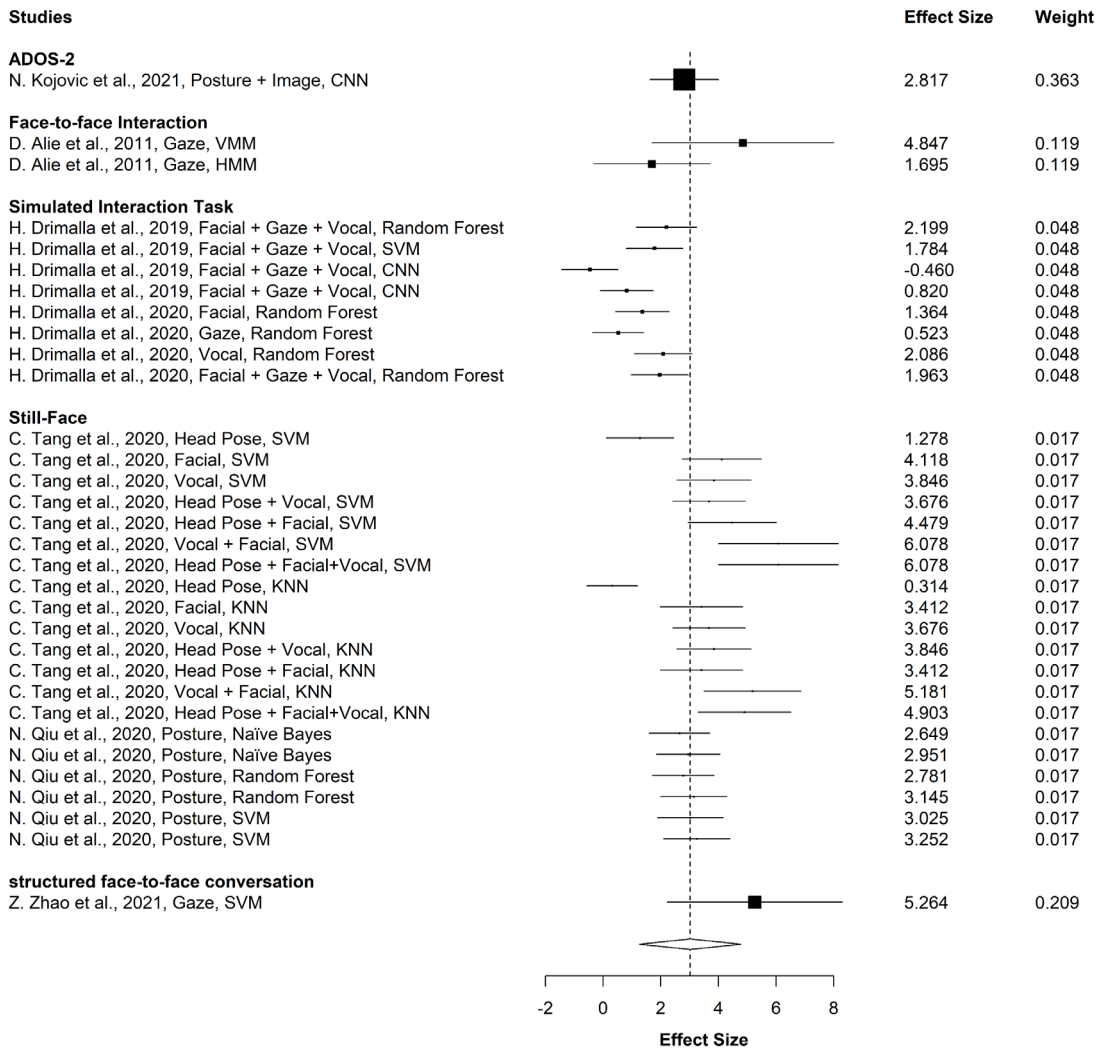


Figure 2.10. Forest plot for robust variance estimation on the reported outcomes.

2.4 Discussion

This systematic review and meta-analysis provides promising evidence to suggest that artificial intelligence algorithms containing digitalised facial phenotypic information discriminate between ASD and TD groups during social interaction assessments. The systematic review summarised existing AI applications in various types of interaction modes (human-human, human-social robot, and human-computer agent interaction) during structured social interaction assessments and therapy sessions. Findings showed that AI can identify atypical facial features and social behavioural markers associated with ASD. However, differences in experimental settings significantly impacted on the generalisability of AI algorithms, suggesting that the evidence for these methods remains at an early stage for real-world applications.

Furthermore, this review suggests that machine learning algorithms, such as SVM (Figure 2.3a), performed better in differentiating between ASD and TD in human-human and human-agent interactions relative to other algorithms. Facial and gaze features (Figure 2.3c) are the most dominant features selected for the model training. There have been no reported studies of human-agent interactions related to social behaviour detection and social responsiveness level predictions. Most studies explored the use of social robots as a treatment tool to improve social

skills of participants with ASD. In those studies, deep learning algorithms, such as ResNet and AlexNet-based models, and machine learning models, such as K-NN and Random Forest, performed better than SVM, although SVM was one of the most explored algorithms (Figure 2.3b). In human-human interaction studies, ResNet-based deep neural networks and Decision Tree models were the most investigated algorithms for behaviour detection, with ResNet performing better. In addition, as demonstrated in Figure 2.3d, eye contact scenes and facial features extracted during human-human interactions were widely studied for behaviour detection. Head pose and body movement features were commonly used for robot-assisted studies, and facial features (including facial landmarks and head pose) performed better.

The meta-analysis revealed estimates of AUC = 87.8%, sensitivity = 77.3%, specificity = 83.7%, DOR = 19.929. The study-level multi-variate meta-regression analysis showed a pooled estimate of 13.92 DOR, indicating adequate evidence for the use of AI models in ASD research. The moderator analysis suggested that heterogeneity observed across studies was influenced by a combination of 1) experimental tasks, 2) classification algorithms used for model training, and 3) different input modalities learned by each model. The moderator analysis (Supplementary Table 10) also showed a positive effect on diagnostic accuracy when considering facial, gaze, posture movement, and vocal features as AI model inputs, while there is no significant effect using only head pose features. In model selection, machine learning models such as random forest, SVM, and Naïve Bayes have shown significant positive effects, while HMM show a weaker negative effect, and K-NN models show a positive but insignificant outcome. In comparing social interaction tasks, the simulated interaction task showed a significant negative effect on ASD diagnosis. There was only relatively stable performance in the still-face paradigm. More complex social settings have not been found to achieve competitive diagnostic test accuracy (Figure 2.6).

Systematic Review On Social Behavioural Marker Recognition Studies

A systematic search for potential review articles on the application of AI algorithms in ASD research identified three review articles and one meta-analysis. These review articles covered brain-based markers [48], eye tracking in social visual attention tasks [14], emotion recognition algorithms [15], and other behavioural markers that differentiate between ASD and TD groups [13]. However, studies have yet to explore the effectiveness of AI in analysing natural social behaviours in social interaction contexts for clinical assessments and interventions. Multi-modal feature learning mechanisms have gained research attention to understand the underlying human perceptions and behaviours of cognitive processes. Among them, facial information plays a pivotal role in the clinical diagnosis of ASD [91]. Research has shown that visual attention and distraction behaviours could be automatically recognised using trained deep learning and machine learning models through facial emotion and movement patterns recorded in social contexts. For example, in a human-robot therapy session proposed by Nuovo et al. [77], the softbank robotics, NAO, uses Naïve Bayes and K-NN machine learning models to capture facial images of participants conducting the imitation tasks in the Verbal Behaviour Milestones Assessment and Placement Program (VB-MAPP), and to detect their attention and the participants' training performance. Chong et al. [78, 82] conducted a wide range of naturalistic social tasks to detect eye contact behaviours during adult-child interactions. In their studies, the adult assessor captures facial images of the child participant using wearable

cameras embedded in glasses. Deep learning models such as ResNet-50, AlexNet, Pose-implicit Convolutional Neural Networks (PiCNN), Pose-Dependent Egocentric Eye Contact (PEEC) detector and GazeLocking were applied for the detection of eye contact moments when the child participants were looking into the wearable glasses on adult. The experiments conducted in their studies included: 1) the Early Social Communication Scales (ESCS) for children under 60 months, where a semi-structured interaction task examined the non-verbal communication behaviours exhibited by participants when playing with toys; 2) the Marcus session, which simulated a therapy session in which the therapist allowed children to play freely with toys, assigned them to play with toys, and prompted children to initiate social interactions by withholding toys; 3) a structured Rapid-ABC task in which assessors stimulate responses by using small balls for turn taking, reading aloud and encouraging children to turn the pages and touching pictures in the book with their fingers, placing the book on their head as a surprise, and playing a gentle tickling game; and 4) administered the Brief Observation of Social Communication Change (BOSCC) task to provide naturalistic free play sessions, and snack time or conversation to encourage children to initiate joint attention. Chong et al. [79] also proposed an SVM classifier to predict gaze shifts that initiate joint attention behaviour through the obtained head pose and camera pose estimation features. As an important social biomarker, detecting interpersonal mutual gaze in ASD therapy training was the focus of Guo et al. [84]. They proposed a deep CNN model to score mutual gaze for each detected head pose between child-trainer pairs.

In addition to gaze behaviour detections, research also examined participants' engagement in receiving therapies in the context of human-robot interaction. Javed et al. [81] proposed a deep CNN model in comparison to conventional machine learning models (SVM, RF, DT, K-NN) to detect social engagement behaviours presented in both ASD and TD groups through the skeleton posture and facial expression features. The engagement behaviours analysed in their study include gaze, smile, triadic interaction, vocalisation, imitation, and self-initiated interaction. Anagnostopoulou et al. [87] proposed deep learning models (1D CNN, ResNet-50, LSTM-based RNN, 2D CNN, AlexNet) to estimate engagement levels based on posture key points detected in social play therapy video recordings.

Research also emphasises that multi-modal feature learning supports the recognition of complex social behaviours related to ASD during assessments and therapy sessions. For example, Liu et al. [83] introduced a "response-to-name" dataset in which clinicians rated child responses to a name call during toy play. The study showed that the TD group responded more strongly than the ASD group, and therefore, the responsiveness score in the ASD group was lower. Through a collection of face detection, head pose estimation, and response latency and weighted duration information, their DT models were able to predict social responsiveness scores from the recorded videos. G. Nie et al. [80] proposed that head pose features (such as hard and soft histogram of head pose data, head pose on different areas of interest extracted by Gaussian Mixture Model, and head pose motion) trained on transductive SVM models with radial basis function kernel enhance the prediction of response to joint attention performance in interpersonal interaction. They suggest that the ESCS test conducted during a human-robot interaction could help predict performance during human-to-human social contexts. Li et al. [86] suggested that facial emotion features extracted from the "ASD-Affect" dataset combined with speech emotion recognition improved the prediction of affect states (positive, negative, and neutral) of children with ASD. Wu et al. [85] proposed a ResNet-18-based deep learning

model to identify infants' behaviour associated with ASD diagnosis through adult-child interaction video recordings and facial features (such as face and eye landmarks, facial action units, head pose and gaze direction estimated by OpenFace2.0 library) extracted.

Systematic Review Of Feature Extraction Algorithms

Historically, researchers have considered atypical biological and behavioural features as markers used in ASD diagnosis. Likewise, extensive research on AI algorithm development has been dedicated to extracting unique feature patterns for each category from multi-modal datasets for classification. In the context of eye gaze related studies, algorithms were initially designed to extract participants' faces, extract facial landmarks, calculate facial expressions, and estimate gaze direction. As shown in Supplementary Table 5, prediction models for eye gaze analysis applied facial recognition and extraction algorithms, including Viola-Jones, MTCNN, Fast R-CNN and low-dimensional facial embedding algorithms. In addition, some studies suggested the use of off-the-shelf libraries, including Omron OKAO, Dlib, OpenFace2.0, OpenPose and Amazon Rekognition.

To further extract facial landmarks and head pose features, studies utilised algorithms such as MTCNN, HOGs, IntraFace and Perspective-n-Point Algorithm, Gaussian Mixture Model and hard or soft histogram, OpenFace, OpenPose, and local phase quantization on three orthogonal planes. External devices, such as the Tobii eye tracker device, support the extraction of gaze-shifting patterns and provide participants' gaze fixation information on video stimuli. Some studies have also used OpenFace2.0 for eye gaze estimation, while others preferred to use pupil diameter changes and gaze fixation features on predefined areas of interest for gaze estimation. A commonly used software tool when using body movement data in ASD research is OpenPose, and Laban Movement Analysis has been introduced in other studies. Studies exploring the use of vocal features for behavioural marker detection, including openSMILE, librosa and proscodi libraries and Kaldi libraries, these features were individually input or fused as a combined input to the developed machine learning and deep learning algorithms purposed for the behaviour recognition or ASD classification tasks.

Limitations And Future Recommendations

Although the meta-analysis showed promising results, further studies are needed to validate these findings due to the limited number of eligible studies and reported participants included in the analysis. For instance, only the studies conducted during still-face paradigm and conversation tasks were found within and above the summary estimates. Further, the parent and children interaction intervention study has noticeably imbalanced sample sizes (6 ASD samples with 26 TD samples), suggesting caution is needed in the interpretation of the effectiveness of AI models in ASD diagnostics.

In addition, not all diagnostic test studies were included in the analysis due to the lack of critical information, such as sensitivity, specificity, and contingency table required to conduct meta-analysis. This is a common phenomenon in machine learning studies, as often the collected data needs to be split into training, testing (and validation) sets before the models can be trained. Not all studies are able to report contingency tables in terms of number of participants. Therefore, some studies on human-to-robot and human-to-computer agent types of social interaction tasks were not included due to a lack of information. An email was sent to the study authors for this information, with only two provided relevant data for the meta-

analysis. No responses were received for other studies. Secondly, few studies applied cross-validation and data augmentation methods to reinforce the generalisability of trained models and reduce data imbalances, which should be widely used in future studies to avoid exaggerating the performance of the models. With only seven studies eligible for the meta-analysis, in which five different types of experimental tasks were introduced, the heterogeneity effect caused by the experimental tasks may not be representative.

Further, this review also collected diagnostic criteria, age and IQ information during data extraction. However, due to the limited information provided in the included studies, there was not a sufficient sample to investigate their moderator effect. Therefore, it is recommended that future studies report this information to inform other possible sources of heterogeneity further.

Moreover, the ethical utilisation of assistive technology (such as robot-assisted intervention) incorporating AI in ASD research and translating AI in clinical practices for patient care (such as early diagnosis and intervention) requires further examination of the validity and generalisability of the developed algorithms across multiple sites and populations. As summarised in Table 2.1, there is no consensus on data management plans and the acceptable performance for both the ASD diagnosis studies and the social biomarker detection studies because of the diversity of experimental settings covered in this review, and there is a lack of comparative validation results with domain experts in those tasks. For AI digital tools to be used in clinical services, it requires domain experts to understand the underlying mechanisms of AI, establish acceptable accuracy for clinical validation, analyse patient acceptance and trust, and ensure robust data privacy and adherence to ethical guidelines and regulatory processes.

2.5 Conclusion

In conclusion, this systematic review and meta-analysis provides an overview of current clinical applications of AI algorithms to ASD diagnosis and social marker detection in social interaction settings. A high AUC score obtained in the meta-analysis indicates that AI algorithms developed on facial features have the potential to assist in diagnosis. However, future studies should consider analysing these biomarkers in unrestricted social environments, such as conducting free play type of games both in clinics and at home. Given the complexity of social interaction assessment, deep learning models designed for action recognition may support detecting ASD-related social behaviours. Further research and validation will help advance the use of these algorithms in clinical practice, facilitating the development and socialisation of individuals with ASD.

2.6 Summary

The systematic review and meta-analysis conducted in this chapter provide a detailed analysis of the use of digital information, including facial features, for ASD classification or studying related social behavioural markers in social interaction tasks. It was found that existing studies have made preliminary use of AI algorithms in structured social interaction settings. However, they have not yet taken into account naturalistic social responses in unrestricted free play environments. Therefore, this thesis will focus on using the advantages of computer vision for action recognition to analyse social behavioural markers and physiological biomarkers that have clinical relevance, and that can be objectively quantified and used for ASD classification.

Chapter 3: A Multi-Modal Behaviour Recognition Framework For Eye Contact Detection In Free Play

The systematic review and meta-analysis discussed in Chapter 2 suggested that facial and gaze features are the primary sources of information for ASD classification and social behaviour detection using ML and DL algorithms during human-human social interaction. The experimental designs involved are relatively homogenous, such as the still-face paradigm and the computer-based simulated interaction tasks conducted in structured social contexts. The review also suggested that DL algorithms perform better than ML in detecting social behavioural responses through facial modalities. Due to the restriction of body movements, these experiments did not analyse heterogeneous natural social responses of the participants in unrestricted social situations. This chapter addresses the first research question to examine whether DL models can identify critical social skills exhibited by child participants in a natural social interaction setting.

Current ASD assessments consider atypical eye contact as a critical behavioural marker. The behavioural recognition framework proposed in this chapter detects direct eye contact episodes observed during free play in a social interaction assessment. In addition, this chapter also investigated the usability and interpretability of DL frameworks for multi-view, multi-modal feature learning to detect social behaviours from video-based input modalities collected in complex social environments.

Source Code In Code Availability:

The codes used to generate the reported results can be accessed via this link https://github.com/CarterSunUSYD/EyeContact_Recognition_Freeplay.

Acknowledgements

This study was supported by the Brain and Mind Centre Child Neurodevelopment and Mental Health team, the National Health and Medical Research Council (Project grants 1043664 and 1125449) and the Bupa Health Foundation independent research grants program.

Competing Interests

This paper has data that is the subject of an ongoing patent filing.

3.1 Introduction

Eye contact is broadly acknowledged as a prominent indicator of social competence when evaluating human social interactions [92]. It is believed to be an essential non-verbal communication behaviour in social interaction that assists with understanding other people's social attention [93], emotional changes [94], relationship intimacy [95, 96], and trust [96, 97]. In addition to impacting relationship quality, other research has pointed to the role eye gaze plays in clinical disorders and mental well-being. It is believed to contribute to early social development [98], with evidence that reduced eye contact early in life is often associated with brain and cognitive divergence [99, 100]. Avoidance or reduced eye contact is associated with heightened anxiety and other clinical mental health issues [101]. As a result, reduced eye contact is sometimes linked to social withdrawal [102-104] and serves as a potential biomarker for diagnosing conditions such as social anxiety disorder [105] and depression [106]. Moreover, atypical eye contact in social interactions forms part of the diagnostic criteria for ASD [107] in the DSM-5 [108]. However, challenges persist in understanding the underlying mechanisms of eye contact manifestations in social interaction due to the heterogeneous nature of experimental settings, subjective human bias in assessments and therapeutic sessions, and the lack of quantitative analytical mechanisms that can be universally applied to capture multi-individuals eye contact events in social interaction. In clinical practices, objectively assessing the impact of reduced eye contact on the quality of social interactions provides an opportunity to investigate its utility as a clinical biomarker.

The quality of social interactions between children and their caregivers is often studied in free play environments. Free play in early childhood contributes to the development of individual adaptivity, social skills, mental health, and increased creative thinking through self-directed activities that enhance cognitive and brain development [109-114]. Free play allows the assessors to observe participants' natural interactions with other people and the awareness of objects in their surroundings in an environment that does not restrict their movements. It is also used as a standardised social interaction task to assess clinical conditions [115]. Within these more naturalistic play tasks, measuring eye contact in social interactions has traditionally involved an observer using video-recorded play sessions to count the frequency and duration of each eye contact episode [96, 116]. Such approaches are often very time-consuming and subject to human error in coding [117, 118], making these resource-intensive research measures impractical for routine clinical use.

Several studies have suggested the use of simulated interactive and social communication environments as a new form of assessment to reduce workload in observational coding [119, 120]. In these settings, eye trackers are required to record face and eye movements during the experiment [121, 122]. Eye contact analysis using these approaches is performed by counting the frequency and duration of participants' gazes on a screen presenting a video stimulus. Other studies suggested developing robot-assisted treatment mechanisms [123] that allow programmable or remotely controllable robots to interact socially with participants and prompt accordingly based on the participants' reactions [124]. However, these methods do not allow a direct analysis of natural human social interaction. In addition, these methods require observing the frontal face of the participant to detect relevant social biomarkers. Such approaches also have additional requirements for calibration that can sometimes make it challenging to collect data from different samples (e.g., young children).

In recent years, wearable technology has gained popularity in clinical research to enable recording eye contact behaviour in human-to-human social interaction [125, 126]. Such as the use of an egocentric camera [82] embedded with wearable glasses to record the child participant's gaze direction. These studies defined eye contact behaviour when the participant's

gaze was directed towards the assessor's eye regions where the embedded cameras are located and considered no eye contact when the eyes moved away. This approach satisfies the need to record information about the participant's frontal aspect with a close camera view in face-to-face interaction [127]. The limitation of this wearable technology is that the glasses move with the wearer, which can make calibration of the participant's position in real-time complex. In addition, it adds additional devices to the face, potentially obstructing facial information. It also may be challenging to use such devices when the person wearing them (young children, people with sensory sensitivities of glasses).

This chapter introduces a synergetic multi-modal, multi-view deep learning framework for recognising eye contact behaviours among multiple individuals in a free play clinical setting. The aim is to provide objective and quantifiable discriminants for assessing social competence. In real-world testing, participant movement is often restricted due to the challenges of predicting heterogenous movements and the risk of recording occluded video recordings during natural movements. Digitised behaviour patterns extracted through deep learning algorithms have the potential to be a new paradigm for objectively analysing behavioural markers in psychological assessments.

Firstly, this chapter tests the effectiveness of training spatio-temporal features of expert-annotated eye contact segments using DL algorithms designed for behaviour recognition [18]. It examines whether features from multiple camera views could help identify highly heterogeneous multi-individuals' eye contact behaviours. Secondly, this chapter compares the sensitivity of the multi-view and multi-modal fusion mechanisms with that of the single-view, single-modal behaviour recognition. The comparison aims to highlight challenges when video camera views are obstructed due to the free movement of participants or the presence of multiple individuals in the assessment processes. Thirdly, this chapter tests the generalisability of each data modality in recognising eye contact behaviour in single or multiple-angle videos from a free play task. It assesses whether the framework could accurately identify eye contact episodes in video recordings of children with ASD and TD during social interaction assessments. Lastly, this chapter presents a visual evaluation of the trained model's decision basis for eye contact detection, enhancing the transparency and interpretability of the model. By analysing the recognition sensitivity of the trained model in situations with or without eye contact behaviour, this research demonstrates how an objective behaviour recognition framework using digital representation features could assist future clinical assessments for screening, diagnosing, and intervening in psychological disorders.

3.2 Material And Methods

3.2.1 Participants

Participants were recruited from the Clinic for Autism and Neurodevelopmental (CAN) Research at the Brain and Mind Centre, the University of Sydney as part of a larger clinical trial. ASD diagnoses were confirmed using the DSM-5 criteria [108] following administration of the Autism Diagnostic Observation Schedule 2nd edition (ADOS-2) [128]. Assessments (ADOS-2) were administered by research reliable assessors and independent interviewers. TD children were recruited from the broad community via newspapers and flyers and assessed over the phone during an interview including the structured clinical interview for DSM-5. No participants were receiving psychotropic medication.

The clinical trial was approved by the University of Sydney Human Research Ethics Committee (2013/502) and the Sydney Local Health District Human Research Ethics Committee (HREC/18/RPAH/157). Written consent was obtained from all participants' caregivers prior to their inclusion in the trial.

3.2.2 Procedures

Each child participant completed a 60-minute semi-structured social interaction session in which all conversations and behavioural actions of participants, the assessors, and their caregivers were recorded by four HD cameras (25Hz) mounted on the walls, see Figure 3.1 (Step A) for data collection. At the same time, audio signals of the session were collected by a 4K microphone attached to the ceiling at a height difficult for the child participant to reach. ECG signals were collected on the child participant and assessor simultaneously with the video and audio signals in the Observer XT software to avoid signal desynchronisation problems. The contribution of this research was to recognise direct eye contact behaviours through the multi-view camera recordings. Audio and physiological signals were not analysed.

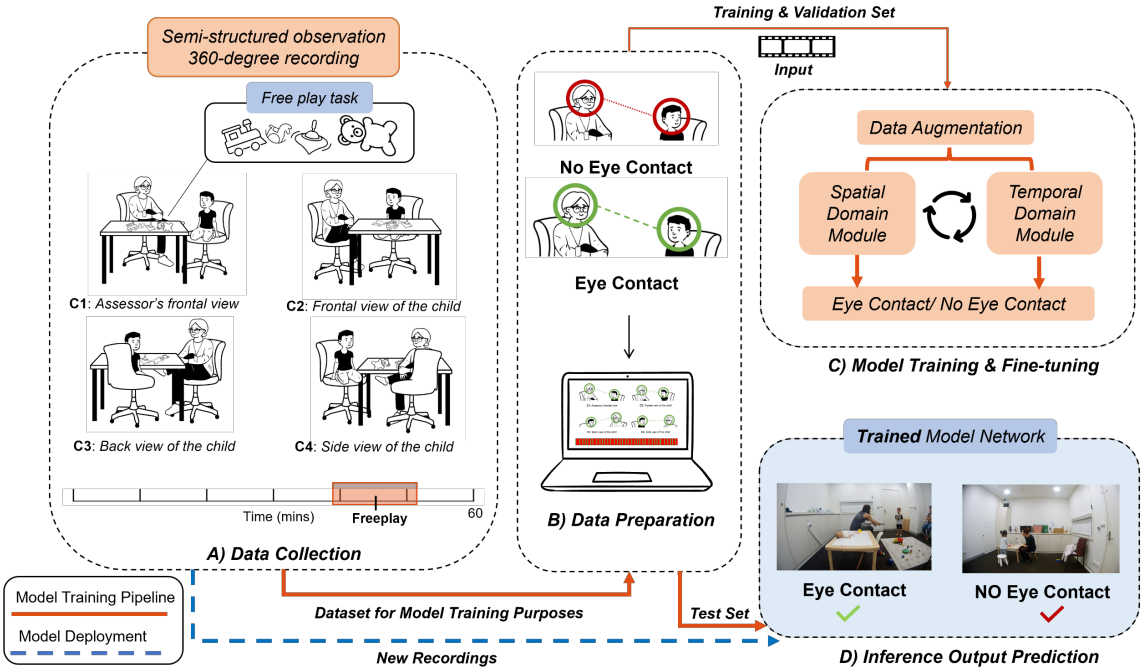


Figure 3.1. Overview of social interaction assessments and model development for eye contact detection during free play. Step A shows the assessment setting and extraction of free play task used in this research. Step B illustrates human data annotation for model training and fine-tuning (Step C). The eye contact behaviours highlighted in green, and no eye contact in red. The trained model was tested on both the test set and newly recorded videos without human annotation.

The Free Play Task

This research focuses on the free play task derived from a baseline social interaction session to assess the social skills of children interacting with the assessor in a simulated real-life interaction scenario. The free play task allows the participants to play in a natural social interaction environment where there are no restrictions on their movement in the trial room. During the free play task, the assessor in the trial room would provide the child participant with toys on a table or, if participants preferred to play on the floor, the toys would be provided on the blanket next to where they were sitting in the trial room (see Figure 3.1 Step D). Children were instructed to play on their own for at least 3 minutes. The assessor was allowed to join play spontaneously by interacting with the child or directing the child's attention to various toys. If the child could not focus on the toys and play, parents could help get the child interested in toys. The free play task was assigned 5 minutes, and another observer in the next room

observing the social interaction session through video recordings would signal the assessor in the trial room by knocking on the door when there was enough time in the task to end the free play. The full 360-degree camera recording setup obtains as much information as possible about the child participant at different moments.

3.2.3 Dataset Preparation

Behaviour Annotation

The free play task in the social interaction session began when the assessor took the relevant toys out of the storage box and placed them on the table or the floor next to the child participant. After each session, a researcher recorded the start time of each task from each participant's video recording. Based on the recorded start time stamp, the FFmpeg (version 4.2.4) library [129] segmented a standardised free play task from each 60-minute social interaction session. Two independent coders with domain knowledge were trained with the lead clinical researchers and used the MATLAB (version R2021a) Ground Truth Labeler application to annotate each eye contact event within the segmented videos (Figure 3.1 STEP B). Based on the current ASD assessment criteria, an eye contact event was defined as the child participant directing visual attention to the eyes of the co-located assessor while the assessor gazed into the child's eyes. The coders first uploaded the four-view video recordings into the Labeler and reviewed the recorded free play videos. Each eye contact event's start and end points were labelled. The free play task ends at the start of the still-face paradigm and reunion task. If the still-face paradigm occurred within the segmented videos, coders also marked the start time of its session and removed the annotations for this period.

The inter-rater reliability of the two trained coders was assessed after data annotation using Cohen's Kappa values. All 33 free play sessions for the TD group and 16 sessions for the ASD group were coded by both coders. The results showed moderate-to-high inter-rater reliability for annotations of eye contact. This included the Kappa value for the TD group (mean 0.72, standard deviation 0.25) and the Kappa value for the ASD group (mean 0.84, standard deviation 0.15). All annotated videos were cross-checked by the lead clinical researchers in the team to resolve discrepancies in the coded behaviours.

Data Extraction

The free play segments were further trimmed using the FFmpeg library based on the two categories of annotated behaviours: "Eye Contact" and "Other", where "Other" refers to frames without any social interaction labels. The dataset was then prepared using a 5-fold stratified cross-validation and random shuffling. This examined a predetermined percentage of participant data (60%) in each fold to train the model, then used the rest to validate and test the model (20% each). These percentages captured a unique set of participant data across the ASD and TD groups within each fold separately.

To prepare for frame-level model training, the Python library DenseFlow [130] was used to decode each trimmed clip into RGB frames (representing the true colour of the image) and optical flow [131, 132] images in the x and y-axis directions (presenting the motion pattern between two consecutive frames in each axis direction). The dataset consists of 1,726 clips of eye contact events (42,110 RGB frames with 40,384 optical flow images in each axis direction) and 2,435 clips of "Other" events (3,849,732 RGB frames with 3,847,297 optical flow images in each axis direction).

3.2.4 Proposed Deep Learning Framework For Eye Contact Detection

To identify eye contact behaviour between the child participant and the assessor in a natural interaction setting during the free play task, multi-view video recordings allow for a complete and maximum capture of eye gaze direction. The coherence and complexity of human action have led research efforts in action recognition to use multi-modal, continuous data to detect target behaviours through deep learning algorithms [18]. Classical deep learning excels at learning continuous behaviour processes in both the temporal domain (stacked sequence of images) and spatial domain (2-dimensional single image features, or 2-dimensional image with depth features (formed 3D features [133])), such as two-stream convolutional networks [31].

The Multi-Modal Social Behaviour Recognition (MSBR) framework, outlined in Figure 3.2, is proposed in this research for detecting eye contact events between children with or without ASD diagnosis aged 3-12 years and the assessor during a free play task from multiple camera views. The MSBR framework is divided into two components: a spatial domain branch (highlighted in blue) employing multiple single-frame RGB images and a temporal domain branch (highlighted in grey) with segmented clips of consecutive optical flow images. It is based on the structure of the Temporal Segment Network [32, 134] using ResNet-50 [135] as the backbone to train the model branches to avoid sample imbalance by randomly intercepting a uniform length sequence from long videos. ResNet-50 was chosen because it outperforms many other backbones in action recognition [136].

Data Pre-Processing

The MSBR framework performs integrated prediction of the input videos based on the sequence of segments decoded into two-dimensional images. In the input data pre-processing phase (Figure 3.2), each video clip is decoded into RGB frames and optical flow images with its labels (action class and camera angle) and shuffled. The random shuffling allows model generality and reduces overfitting caused by the model training irrelevant features. A data augmentation strategy was introduced to give the model a larger and more diverse sample size for training and to prevent severe overfitting due to the lack of data samples.

Data augmentation techniques, such as MultiScaleCrop (with the scales 1, 0.875, 0.750, 0.660) on the RGB images channel (the spatial domain CNN branch) and RandomResizedCrop on the optical flow images channel (the temporal domain CNN branch) were adopted. Then, all these images were resized to a uniform size, 224 x 224, to meet the requirements of the input channels of the ResNet-50 backbone architecture. These images were further horizontally flipped with a 50% ratio to increase the data diversity, and batch normalisation was conducted to transform image features to a similar scale to help the model converge faster. These data augmentation techniques were not applied to the validation and testing datasets, except the resizing and normalisation steps, to reduce the computational cost when processing newly collected videos and to reflect whether the model is generalisable.

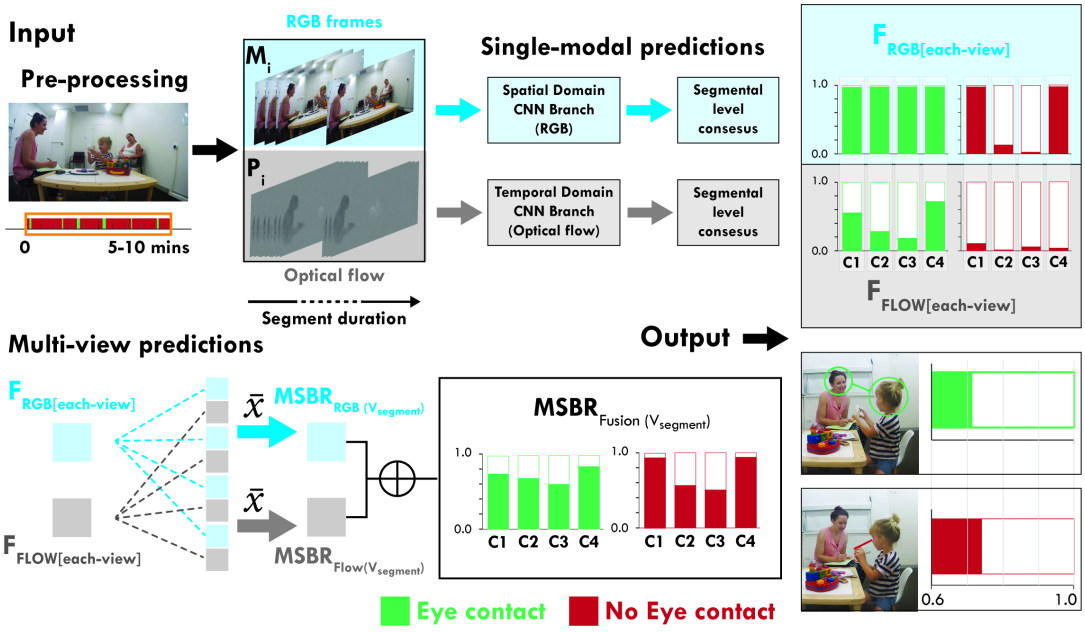


Figure 3.2. The multi-modal social behaviour recognition framework flowchart. The input video from each camera angle is segmented into RGB frames (Mi) and grayscale optical flow image stacks (Pi). Then, it is fed into spatial (blue) and temporal (grey) channels, respectively, for training. The segment-level consensus scores from each single-modal of each input video clip are first calculated (bar graphs, F_{Flow}[each-view] and F_{RGB}[each-view]). Multi-modal predictions for all views of the same period are then aggregated separately through each channel. Multi-view late fusion ($MSBR_{Fusion}(V_{segment})$) is then performed, and the predictions from both channels (spatial and temporal) are averaged to produce a final prediction score on the presence or absence of eye contact behaviour.

Single-View And Single-Modality Feature Learning

In the single-modal predictions phase (Figure 3.2), the framework randomly selects equally spaced RGB images (applied to the spatial domain CNN branch) and consecutive optical flow images (applied to the temporal domain CNN branch) from each single-view video segment and loads them into the respective branch. The spatial domain CNN branch performs frame-level prediction for each of the selected $\boldsymbol{n}$ RGB frames ($\boldsymbol{M_i}$). Then, an average consensus is reached for all frame-level prediction scores as the prediction result of the input video (F_{RGB}[each-view]). A temporal domain CNN branch is performed to train the optical flow feature patterns. Unlike the spatial domain, the input of the temporal domain CNN requires consecutive optical flow images ($\boldsymbol{P_i}$) of length $\boldsymbol{l}$, as a segment with equal interval. An average consensus is reached for all segment-level prediction scores as the prediction result of the single-view input video (F_{Flow}[each-view]).

Multi-View Prediction

The prediction of behavioural events from multiple camera views is achieved by aggregating segment-level predictions from individual camera views. The input video segments are denoted by the function $\boldsymbol{H_c}$, where "$\boldsymbol{c}$" represents each camera angle. Spatial domain ($\boldsymbol{MSBR_{RGB}(V_{segment})}$) and temporal domain ($\boldsymbol{MSBR_{Flow}(V_{segment})}$) predictions from each camera view are averaged to obtain late-fusion across all camera views, $\boldsymbol{MSBR_{Fusion}(V_{segment})}$, to calculate the final prediction scores for each action class (where class 1 refers to eye contact behaviours presented between the participant and the assessor, while class 0 refers to no eye contact behaviours presented).

$$MSBR_{RGB}(V_{segment}) = \frac{1}{4}\sum_{c=1}^{4} H_c\left(\frac{1}{n}\sum_{i=1}^{n}(F(M_i;W))\right) \quad (3.1)$$

$$MSBR_{Flow}(V_{segment}) = \frac{1}{4}\sum_{c=1}^{4} H_c\left(\frac{1}{u}\sum_{i=1}^{u}\left(F(P_{l,i};W)\right)\right) \quad (3.2)$$

$$MSBR_{Fusion}(V_{segment}) = \left(MSBR_{RGB}(V_{segment}) \oplus MSBR_{Flow}(V_{segment})\right) \quad (3.3)$$

3.2.5 Implementation Details

The proposed method was implemented in PyTorch and ran on two 11GB NVIDIA GeForce RTX 2080Ti GPUs. In our experiments, all hyper-parameters tuning were decided using 5-fold cross-validation based on the performance of the validation sets. The best results were obtained when the learning rate was set to $3x10^{-5}$, the batch size was set to 10, the dropout ratio was set to 0.1, and the spatial domain CNN branch was trained by extracting 10 RGB images per segment with an equal interval of 1 frame. For the training of the time-domain CNN branch, the best performance was obtained when the learning rate was set to $8x10^{-6}$, the batch size was set to 32, the dropout rate was set to 0.6, and the input video was segmented in three clips, each clip containing four consecutive optical flow frames with an equal interval of 1 frame. Our framework was initialised using the approach presented in Wang, L. et al. [32, 134], and adaptive-moment-estimation (Adam [137]) was used for network optimization with CosineAnnealing learning rate scheduler [138] due to its ability to fast coverage and less parameter-tuning. A standard back-propagation process was used to calculate the gradient of loss and to update model parameters (weight and biases) of the deep learning architecture [32, 134]. The training was set to stop at 300 epochs and used the standard categorical cross-entropy loss function.

3.2.6 Evaluation Of Eye Contact Recognition

This study evaluated the performance of the proposed MSBR framework in identifying eye contact events occurring in free play. Performance evaluation metrics for individual segments on the 5-fold training and validation datasets were calculated, including top-1 accuracy, precision, recall, and f1 scores. From there, the best social behaviour recognition model was selected in both branches of each fold. Statistical significances between single-view and multi-view prediction were compared using paired samples t-tests. The model training and fine-tuning processes are illustrated in Figure 3.1(STEP C). Given the nature of the imbalanced sample proportion, where most of the time the child participants did not have eye contact with their social partner during the free play task (defined as a class: Other), the selection of precision and recall values are better metrics to test the actual performance of the targeted "Eye Contact" class, compared with the receiver operating characteristics (ROC) curve. This is because the ROC curve could be overly optimistic about the performance of models trained with an imbalanced sample ratio. The selected models in each fold are then further tested on their corresponding test datasets to evaluate the generalisation of the models, as illustrated in Figure 3.1 (STEP D). The prediction results of single-view and multi-view of the same time-stamped segments were also analysed in three aspects: spatial domain, temporal domain, and fused two-domain CNN branch prediction scores, respectively. F1-score represents the harmonic mean of precision and recall, which counts both the "false positives" (the segment did not show eye contact, but the model predicted it as eye contact) and "false negatives" (the model predicted the input segment as "Other", but the segment did show eye contact), which was used to analyse statistical differences between settings.

Furthermore, to aid interpretability, we adopt the Gradient-weighted Class Activation Mapping (Grad-CAM [139]) visual heatmap technique, which localises the regions where the

MSBR framework makes decisions in identifying instances where eye contact events occurred and those where no eye contact was made ("Other" class). Grad-CAM presents a visual explanation of the trained model through gradient-based localisation. In this study, the Grad-CAM algorithm was applied at the three activation layers of the trained model (layer1.0, after the first convolutional block after input; layer4.1 and layer 4.2 before the classification layer) to explore the feature regions of interest learned by the ResNet-50 backbone used to make the final decision.

3.3 Results

The direct eye contact behaviours embodied in 136 free play tasks (totalling 733.140 minutes) from the social interaction assessments were annotated and analysed, with annotated video recordings from 103 children diagnosed with ASD and 33 TD children aged 3-12 years. All tasks were recorded with four high-definition cameras mounted on the walls. Of these annotated videos, eye contact events occurred in 77 of the ASD group and 25 of the TD group. The total duration of eye contact events was 567.11 seconds (totalling 424.750 seconds in the ASD group and 142.360 seconds0 in the TD group), with an average duration of 4.170 seconds (SD=6.800) and an average frequency of 3.940 times (SD=5.270) across all included participants. The average duration of free play was 5.390 minutes (SD= 2.200).

3.3.1 Demographics

Descriptive statistics were calculated for age, gender and Nonverbal Intelligence Quotient (NVIQ) for the sample (see Table 3.1). Independent samples t-tests showed no statistically significant differences in age ($t(134)=1.245$, $p=0.215>0.050$). However, statistically significant differences were found in NVIQ ($t(70.895) = 7.088$, $p <0.001$, two-sided), with 25 missing data points. There was a statistically higher proportion of males to females, which is expected given it was drawn from the ASD population that has a higher prevalence for males (χ^2 (1, N=136) = 4.976, $p=0.026<0.050$).

Table 3.1. Demographics and descriptive statistics.

	ASD	***TD***
Demographics	**Mean (SD)**	**Mean (SD)**
N	103	33
Gender (M/F)	87/16	22/11
Age (years)	6.960 (2.720)	7.63 (2.670)
NVIQ	89.180 (24.740)	116.04 (13.260)
Data Set (Seconds)		
Eye Contact Duration	4.120 (7.010)	4.31 (6.210)
The Free Play Task Duration	344.550 (142.660)	257.570 (51.430)

NVIQ = Nonverbal Developmental Quotient/Nonverbal Intelligence Quotient

3.3.2 Training Convergence

The MSBR model was evaluated on the training and validation datasets to measure the loss and Top-1 accuracy for each modality in single-view video inputs. In addition, the mean class accuracy, F1 score, precision and recall scores were calculated for the validation dataset. As illustrated in Figure 3.3, the training loss on the RGB channel converged faster in comparison to the optical flow channel. Similarly, the Top-1 accuracy on the training dataset (Figure 3.3 a and e) was higher on the RGB channel. The validation dataset also revealed that the mean class accuracy, F1 score, precision, and recall scores performed better on the RGB channel.

3.3.3 MSBR Performance In Eye Contact Recognition

Table 3.2 summarises the average prediction performance of the MSBR framework for behaviour recognition by single-view and multi-view video segments. The distribution of the number of single view segments allocated to each of the five folds in the train/test/validation dataset is shown in Supplementary Table 11.

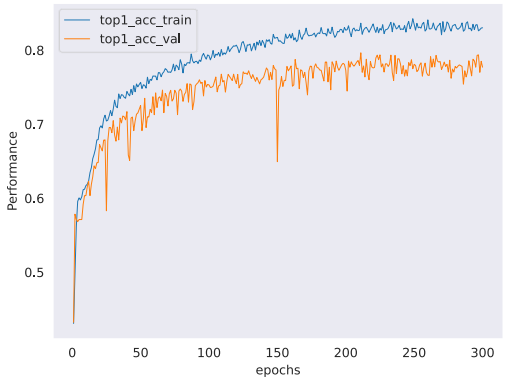
The overall performance of the model was evaluated by calculating the Top-1 accuracy of each single-view video segment for different modalities (spatial domain, temporal domain, and fusion of the two domains) separately on the test set. Precision, Recall and F1 scores were also reported (in Table 3.2) along with the performance of each fold (See Supplementary Table 12-14). In addition, the prediction performance of multi-view video segments integrated from the video recordings with the same timestamp was compared (see Supplementary Table 15-17). The precision-recall curve shown in Figure 3.4 from the first split of the test sets also indicated that multi-view prediction generally achieved higher performance than that predicted by single-view video segments. The spatial domain (RGB images as input) performed better on average compared to the temporal domain using optical images as the input data source. The fusion of single-view and multi-view predictions in the spatial and temporal domains (ratio of spatial to temporal domains of 1) achieved competitive or better results in both classes ("eye contact" and "other"), as shown in Table 3.2. The statistically significant differences in F1 scores of the 5-fold stratified cross-validation demonstrate that multi-view predictions perform better than single-view across all modalities (Table 3.3).

3.3.4 Performance Between Diagnostic Groups

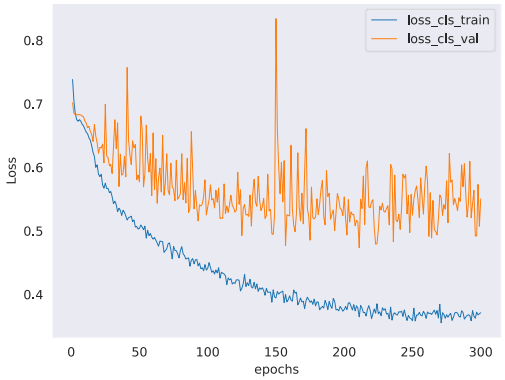
Figure 3.5 highlights the classification performance of the MSBR framework on the first test dataset when classifying eye contact and no eye contact ("other") for the respective F1 scores of the ASD and TD groups. For the eye contact dataset, the F1-score is generally comparable across all modalities on both diagnostic groups, similar to the "other" group, with the exception that optical flow was over 5% higher in the TD group.

Table 3.2. The average prediction performance of eye contact (other) behaviours on the test sets.

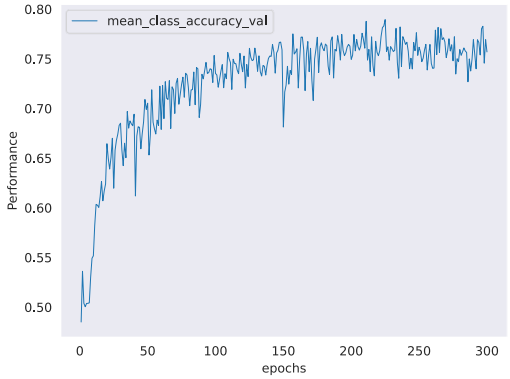
			Evaluation Metrics		
Camera Angle	**Modality**	**Top-1 Accuracy**	**Precision**	**Recall**	**F1**
Single-View	Spatial Domain	0.850	0.803 (0.886)	0.841 (0.856)	0.820 (0.870)
	Temporal Domain	0.798	0.799 (0.805)	0.690 (0.872)	0.734 (0.835)
	Fusion	**0.882**	**0.863 (0.896)**	**0.848 (0.906)**	**0.854 (0.900)**
Multiple-View	Spatial Domain	0.924	0.886 (0.953)	0.936 (0.916)	0.912 (0.932)
	Temporal Domain	0.851	0.890 (0.840)	0.734 (0.931)	0.796 (0.881)
	Fusion	**0.935**	**0.925 (0.942)**	**0.915 (0.948)**	**0.920 (0.945)**



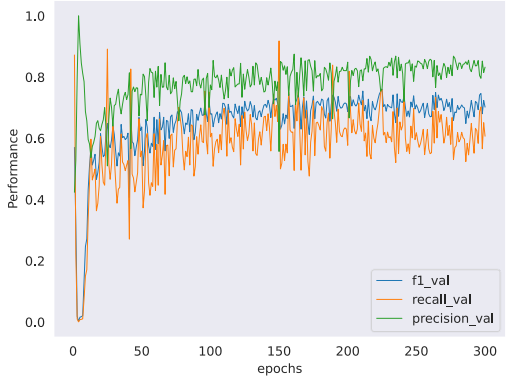
(a) Top-1 accuracy on the training and validation of the optical flow datasets



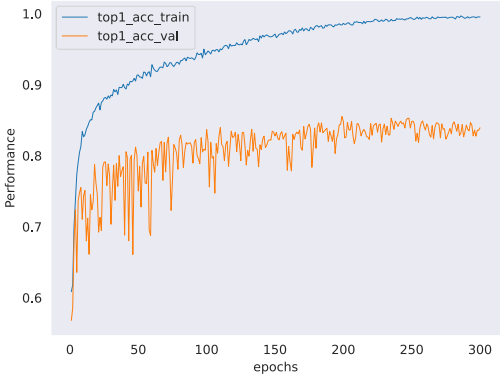
(b) Training and validation loss on the optical flow datasets



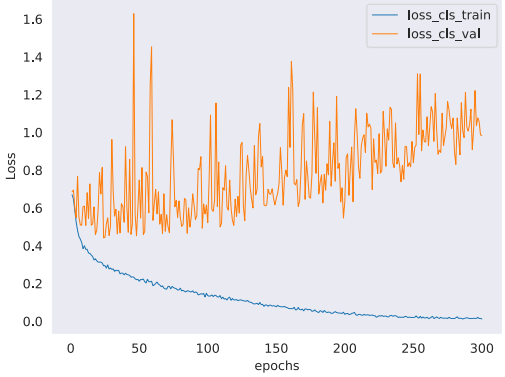
(c) Mean class accuracy on the validation optical flow datasets



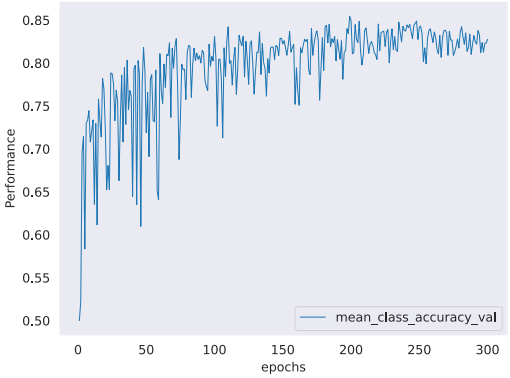
(d) F1, precision and recall scores on the validation optical flow datasets



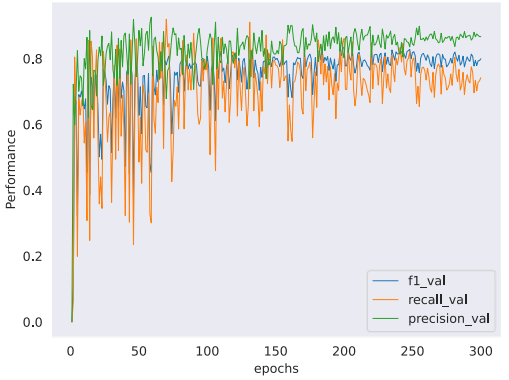
(e) Top-1 accuracy on the training and validation of the RGB datasets



(f) Training and validation loss on the RGB datasets



(g) Mean class accuracy on the validation RGB datasets



(h) F1, precision and recall scores on the validation RGB datasets

Figure 3.3. Training and validation performance for each modality (Optical Flow, RGB).

Table 3.3. Paired samples t-tests of the five-fold cross-validation results on the F1 scores.

	Eye Contact		Other	
Single View – Multiple View	**t-test**	**p-value**	**t-test**	**p-value**
Spatial Domain	-12.91	<0.001	-20.79	<0.001
Temporal Domain	-3.04	0.038	-4.04	0.016
Fusion	-14.94	<0.001	-12.19	<0.001

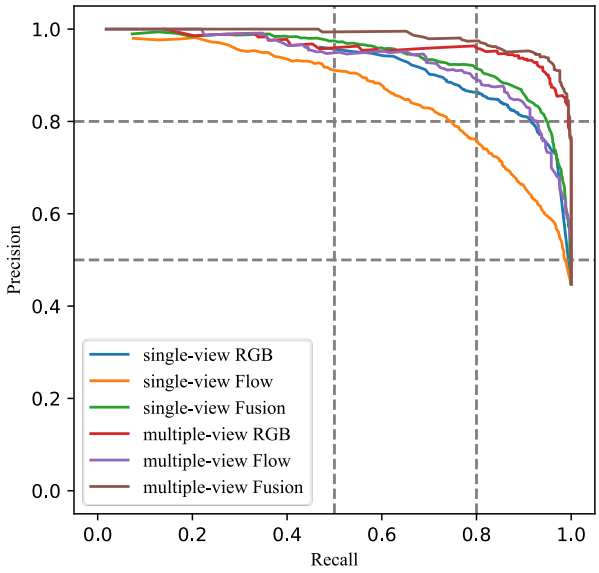


Figure 3.4. The precision-recall curve of eye contact behaviour recognition on the first-fold of the test set.

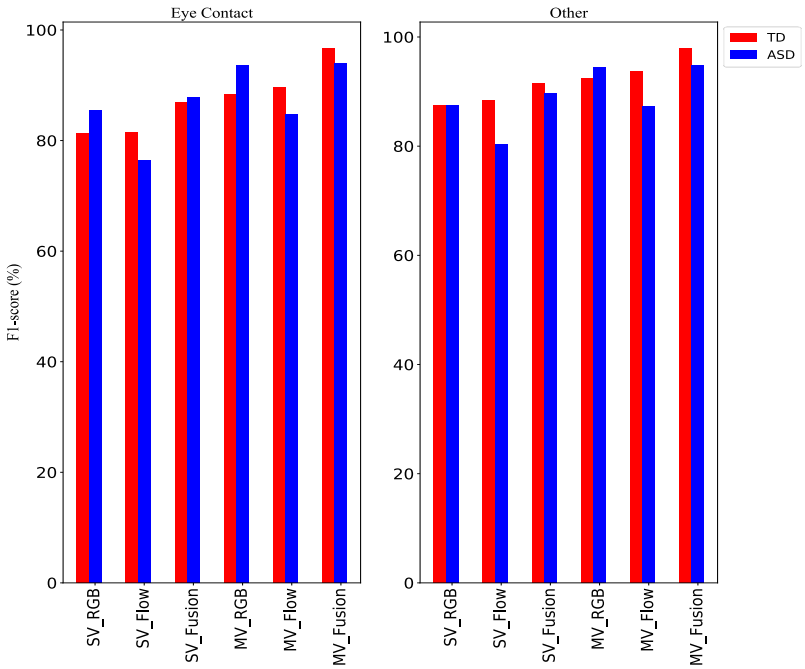
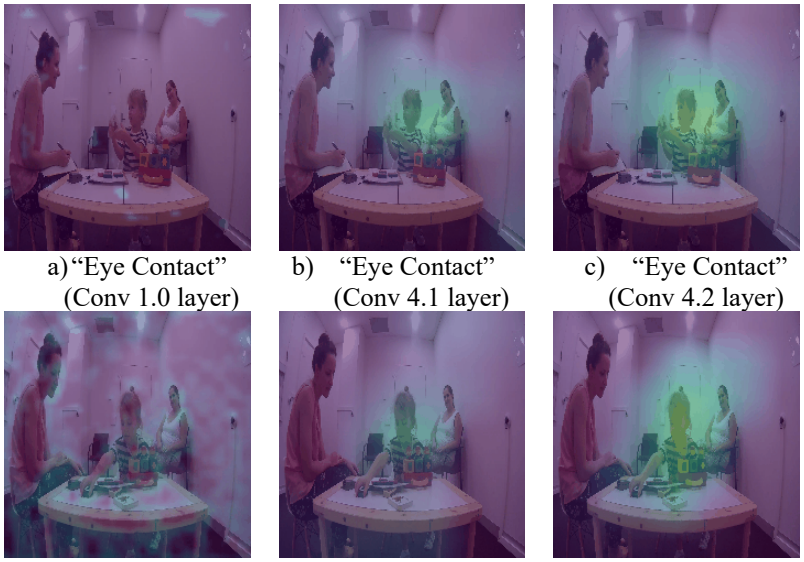


Figure 3.5. F1 score bar graph comparing the trained model's performance for eye contact (and other) behaviour detection on the first-fold test set, where SV on the x-axis stands for single view and MV for multiple views.

a) b)

Figure 3.6. Sample images of optical flow from eye contact in free play. a) Clear and b) unclear movement patterns.



a) "Eye Contact" (Conv 1.0 layer) b) "Eye Contact" (Conv 4.1 layer) c) "Eye Contact" (Conv 4.2 layer)

d) "Other" (Conv 1.0 layer) e) "Other" (Conv 4.1 layer) f) "Other" (Conv 4.2 layer)

Figure 3.7. Grad-CAM heatmap visualisation of the classification of "eye contact" and "other" behaviours in free play by the MSBR framework on the input-side and output-side convolutional layers of the ResNet-50 architecture. The yellow and green shaded areas emphasise the model's focus on classification, while the dark red shaded areas do not present these behaviours.

3.4 Discussion

This study demonstrates the capability of deep learning algorithms to objectively identify eye contact behaviours between multiple individuals in a free play environment through panoramic video recordings captured on fixed walls. The proposed multi-view, multi-modal feature learning mechanism holds great potential for quantitatively parsing social skill assessments. Given the stochastic nature of interactive behaviour in free play challenges in accurately quantifying social behaviours through video recordings, multiple-view video recordings provide more comprehensive visual information, offering an overall understanding of social dyads. However, due to the unpredictable and variable nature of social interaction behaviours among multiple individuals, the subjective interpretation of social behaviours by assessors, the inability to record social responses of multiple individuals in a single camera view, and the expensive labour costs associated with annotating video data contribute to the complexity of real-world testing.

To address these challenges, this study employed multi-view recordings to record social interaction behaviours, applying a multi-modal learning approach, and introduced a deep learning framework that could significantly enhance the model's predictive performance in eye

contact detection. These results show promising preliminary evidence for a feasible solution in recognising multiple and diverse interactive behaviours in complex environments. These findings raise critical questions for future clinical assessment and treatment, encouraging the consideration of reducing the impact on natural human responses while objectively and comprehensively analysing social behaviours.

The MSBR framework presented in this study facilitates the application of multi-modal deep learning models in recognising non-verbal social behaviours in naturalistic interactions. Benefiting from the recording devices fixed on the surrounding walls, the interacting parties can engage in free play without the need for wearing external devices, minimising the distraction of participants caused by the "newly introduced" devices, such as eye-tracking devices or robots and improving the ecological validity of the study. Our findings suggest that comprehensive identification of co-located participants' eye contact events from multiple camera angles is feasible in a free play environment that does not restrict body movements during play, providing opportunities to evaluate social competence in real-world testing.

Furthermore, this is the first study to apply action recognition algorithms to identify social skills, such as eye contact events, without constraining the free play of multiple participants, including child participants and assessors. Notably, this approach eliminates the need for additional devices during clinical assessments. Previous studies implementing free play tasks focused on assessing infants' face-looking and mutual gaze skills [140]. These studies often used setups involving head-mounted eye trackers on both infants and assessors, and camera recordings capturing whole-body motor behaviours of each participant in the scene, and the social behaviours were coded on a frame-by-frame basis for assessment. In contrast, Tang et al. [61] proposed a behaviour screening method utilising a support vector machine classifier to evaluate social stress responses and attention to the caregiver by multi-cue features in a still-face paradigm from multi-view camera recordings. However, their study was limited to observing the responses of the infant participants at rest and did not analyse behaviour during free movement. The deep neural network proposed by Chong et al. [82] was the most recent study to validate its accuracy of eye contact recognition during free play. However, this approach was limited by the assessor's unilateral use of egocentric glasses, limiting the recognition of eye contact between interacting parties. In this study, the MSBR framework identifies eye contact behaviours within a completely unrestricted free play environment. It allows all participants (child participants, assessors, and caregivers) to enter the trial room. This distinctive approach addresses limitations observed in prior studies and contributes to a more realistic understanding of social competence in simulated natural social interaction scenarios.

Moreover, the precision-recall curves illustrated in Figure 3.4 demonstrate the promising ability of deep learning models trained on sparse samplings of RGB images and stacks of optical flow images to recognise eye contact events. In both modalities, the models achieved precision and recall trade-off performance exceeding 75% when utilising a single-view video recording. This can be attributed to the digital features learned by the models, effectively minimising environmental features specific to the video capturing angles (as illustrated in Figure 3.7) and enabling the recognition of eye contact behaviours of multiple individuals at each camera view. Furthermore, in comparison to previous approaches [141], which trained only a limited number of features and required presenting frontal facial information to recognise eye gaze movements, this study enhances eye contact behaviour recognition from the lateral views captured from other camera angles.

In addition, the importance of achieving an average consensus across modalities from multiple angles is demonstrated in Figure 3.4. A multi-view and multi-modal consensus training strategy is considered necessary to mitigate the possibility of incorrect predictions

resulting from the recognition of invalid facial information by any single camera view. For instance, situations where facial information in a conversation is obscured, making it challenging for people to intuitively determine whether there was an eye contact event happened during social interaction. Implementing synergistic multi-view prediction improved the overall performance by approximately 5%. To ensure model generalisability and prevent participant video overlap, which may introduce individual differences in learning, this study scaled the distribution of videos from both diagnostic groups during the data allocation process before model training.

Model Acceptance

The model performance was not found to be heavily affected by noise from the uncertain number of people in the room during social interactions. Uncertainty about the number of people in the room arose from the fact that caregivers were allowed to sit in the room and observe the assessment. They were asked to stay out of the interaction in case they affected the child's independent responses, thus affecting the smooth running of the session or the reliability of the child's behavioural data. As shown in Figure 3.7, the features learned by the model on the input layer (the convolutional layer 1.0) focus on the rich textural detail, such as the shape of each human body, the structure of the trial room (Figure 3.7a and d); while the attention in the last two layers of the ResNet-50 deep architecture (convolutional layers 4.1 and 4.2) gradually narrowed to present only the areas where most behavioural actions were performed, specifically focusing on the participants' faces and surrounding areas (Figure 3.7b-c, e-f). This is consistent with the regions clinicians focus on in situations where they perceptually distinguish between the presence and absence of eye contact for clinical diagnosis.

Limitations

However, Figure 3.5 demonstrates that the spatial-domain models using RGB images generally outperform the temporal-domain models trained with optical flow images. This finding contradicts previous action recognition articles [18, 134]. This may be due to the transient nature of eye contact behaviours, which is characterised by small movement amplitudes and trajectories, making it challenging to leverage the time-domain information provided by optical flow images in multi-modal learning for short-duration social behaviour recognition (Figure 3.6b). The use of multiple discrete RGB images cannot compensate for the lack of temporal information due to hardware resource consumption (e.g., higher GPU memory usage and more time costs in data processing and training). Nevertheless, leveraging multi-modal feature learning can enhance inter-modality information complementarity, thereby improving the overall prediction performance of the model.

This study adopts a two-stream convolution model to recognise behaviour from image features' spatial and temporal aspects [31]. It further improves the framework of sparsely sampled videos for behaviour recognition [32]. Hence, the MSBR framework proposed in this paper is more applicable to identifying social behaviours in clinical assessment with multiple cameras and multiple individuals interacting socially in a natural environment. However, due to less information presented in optical flow images, future studies should incorporate other modalities of temporal feature training that are not limited to video-based information, such as incorporating gestures and audio-visual signals. In addition, due to the high demand for data annotation and cross-checking before commencing deep learning model training. This approach requires extensive labour costs and potentially introduces subjective biases. Future studies should also develop action localisation algorithms to identify behaviours in unlabelled video recordings to help locate target social behaviours, use pre-trained models and techniques

to select key features to reduce annotation requirements and incorporate gold standard clinical assessments to aid in the quantitative analysis of social competence.

3.5 Summary

In summary, the multi-view, multi-modal feature learning mechanism utilised in this chapter allows the deep learning framework to recognise digital phenotypes of eye contact events between co-located individuals interacting in a free play social environment. The model was found to primarily rely on the facial regions of the target population to detect this behaviour, which is consistent with the subjective perception of human sense and matches with experts' knowledge during clinical assessments. However, this framework design was centred on the recognition of multi-view social behaviour from video recordings to aid in social competence analysis in clinical settings, and it was not designed to discuss the potential of digitalised facial information detected in video records as a biomarker for neuropsychiatric conditions detection. Therefore, the next chapter will explore whether facial representations of children extracted during social interaction in free play can distinguish between diagnostic groups.

Chapter 4: Facial Feature In Free Play: An Application of Integrating Statistics And Data-Centric Machine Learning For Autism Spectrum Disorder Classification

Clinical screening for developmental disorders, including ASD, typically begins with an initial evaluation of social competence, followed by a formal diagnosis based on the standard reference manuals. The previous empirical chapter addressed the first step, focusing on the use of deep learning frameworks with a multi-modal, multi-view feature learning mechanism, allowing an objective detection of social behaviours exhibited by co-located human-human interactions in free play. The current chapter focuses on the second research question, exploring potential facial features from social interaction episodes as discriminators for ASD classifications.

According to the systematic review and meta-analysis previously discussed in Chapter 2, it was determined that for ASD classification, machine learning is usually preferred when the features contain well-established domain knowledge, exhibit potentially strong correlations, and are particularly widely applied when only a limited sample size is available. In contrast, deep learning is computationally expensive and excels in handling complex and irregular features with less domain knowledge. To examine the potential of facial features extracted during social interactions as a biomarker in classifying individuals with ASD, this chapter aims to explore highly correlated multi-view facial features for classification purposes through statistical mechanism analysis and data-centric model training approaches. Clinical diagnoses rely heavily on patient's facial information to interpret cognitive development and perceptions. With the development of machine learning algorithms, these facial features can now be extracted by systematic image feature engineering algorithms. In addition, the extraction of those digital features is widely recognised by domain experts, as it aligns with current clinical assessment standards. Therefore, machine learning algorithms are suitable for developing such classification algorithms.

Analysing joint attention episodes is critical for assessing social competence and diagnosing ASD. However, current studies in real-world testing often assume that capturing frontal facial information is necessary for assessments, which limits the hardware utilisation to focus on only one social party. This chapter analyses potential discriminative features extracted from all facial orientations during joint attention episodes in free play. The objective is to eliminate the need for wearable devices and discover the potential efficacy of lateral-view and occluded facial features in ASD classifications.

4.1 Introduction

Effective social communication involves expressing mutual interests through both verbal and nonverbal means. Proficiency in understanding social cues and engaging in joint attention are pivotal indicators for assessing social competence [142]. Joint attention is a set of behaviours where individuals coordinate their attention with others towards a shared object or event, using mechanisms such as eye contact, pointing, touching and vocal cues [143]. It is presumed to play a significant role in fostering social cognition development [144, 145], language acquisition [146-148], emotion regulation and socialisation [149].

Deficits in joint attention are often associated with social developmental delays and disorders such as ASD [150-153]. These deficiencies manifest early in childhood, often leading to challenges in language and social development [154, 155]. Research indicates that children with ASD may exhibit atypical gaze patterns during joint attention episodes, potentially due to heightened social anxiety and stress when looking into other's eye regions [84, 151]. This atypical use of eye gaze may reflect underlying differences in social information processing in ASD [101, 156], influencing social judgement based on facial features. It was evidenced that individuals with ASD tend to focus more on the mouth area rather than on the eye regions for facial emotion recognition [157]. Furthermore, rich information from facial expressions [158], including facial muscle movements, eye gaze direction, and lip movement patterns, could interpret different emotions and social attention cues, which are essential for establishing joint attention, shared attentional focus, and effective social communication [159].

To further understand the mechanisms of initiating and responding to joint attention events in the ASD population, recent studies have proposed quantitative methods to analyse gaze and head posture patterns [160]. An eye-tracking study suggested that atypical behaviour was observed in initiating joint attention (IJA), but no significant differences were found in response to joint attention (RJA) for infants [161]. Clifford and Dissanayake [162] found abnormal behaviours in interpersonal interactions recorded in home videos, and deficits in joint attention were evidenced at the early developmental stage. The study recommended that future early intervention studies consider these behavioural deficits an important assessment component. Isabella et al. [163] suggested that low-risk ASD infants were more likely to respond to complex joint attention cues relative to TD and high-risk ASD infants. The study analysed recorded videos of infants' RJA cues initiated by adults in the form of eye gaze alternation, head turns, finger pointing and vocal guidance during free play.

With the emergence of computer vision algorithms developed for facial information and posture movement tracking, ASD diagnosis and therapy have gradually transformed from manual coding to automation in AI. Several studies have used digital features extracted using AI algorithms to analyse joint attention in controlled assessment settings. Ko et al. [164] proposed a deep neural network to detect social behaviours in joint attention tasks, including IJA and low and high-level RJA. Chong et al. [79] proposed that gaze-shifting features in interpersonal interactions could be analysed to recognise IJA using conventional machine learning algorithms such as SVM. Zhang et al. [67] proposed a logistic regression model to analyse gaze- and gesture-shifting patterns recorded when participants followed the pointing stimulus videos to quantify RJA and further differentiate between ASD and TD groups. Nie et al. [80] utilised head pose and head pose motion features collected during a human-robot interaction setting to predict the performance of RJA in interpersonal therapies by Transductive SVM models. Most of these studies were conducted to analyse joint attention during a controlled environment, and less research has considered the use of facial information detected during joint attention events that occurred under the observation of naturalistic responses in unrestricted free play settings.

This chapter examined the facial responses of child participants when participating in joint attention with social partners in an unrestricted free play environment aimed at identifying digital facial features that may contribute to ASD diagnosis. The study begins with facial recognition and landmark detection by examining facial images of participants (child, assessor and two caregivers) in videos taken from multiple camera views. Then, facial behavioural markers such as gaze shifting, head pose movements, and facial action units were extracted by the OpenFace2.0 toolkit [21]. In addition, the facial features of the assessor and caregivers were excluded using the cosine similarity algorithm, and children's facial images with detection confidence levels below 80% were filtered. The head movement features (displacement, velocity, and acceleration) were then computed from the remaining children's facial features captured during joint attention episodes to examine their statistical effectiveness in distinguishing the ASD and TD groups. The application of SVM models and feature reduction strategies on these features was also investigated to determine their feasibility for ASD prediction. In addition, facial features were categorised into frontal, lateral and occlusion views to analyse the potential impact of occluded views on the classification task.

4.2 Methods

4.2.1 Participants

There were 73 children diagnosed with ASD and 33 age-matched TD children recruited from the Clinic for Autism and Neurodevelopmental (CAN) Research at the Brain and Mind Centre, the University of Sydney. ASD diagnoses were confirmed using the DSM-5 criteria, and the licensed assessors administered the Autism Diagnostic Observation Schedule 2nd edition (ADOS-2) at the clinic. TD group was recruited from the broad community through newspapers and flyers. Initial screening with a telephone interview, including the DSM-5 structured clinical interview [165], was conducted to determine their eligibility to participate in the study. The study was approved by the University of Sydney Human Research Ethics Committee (2013/341). Written consent was obtained from all participants' caregivers before inclusion in the study.

4.2.2 Procedures

All eligible participants completed a social interaction session based on the protocol to assess their social communication and reciprocity skills. This assessment did not consider the participants' level of verbal skills, and all participants completed the same tasks. The focus of this study was to determine participants' naturalistic social response during unrestricted free play with other social partners, such as the assessor and caregivers. The free play task in the social interaction session began when the assessor pulled the relevant toys out of the box. The toys were placed on the table (as shown in Figure 4.1) if the child could remain seated in a chair. Otherwise, the toys were placed on a blanket on the floor if the children could not tolerate sitting. Children were asked to play alone for at least 3 minutes. If the child could not focus on the toys and play, the assessor could stimulate their attention to participate in play by interacting with them or directing their attention to other toys. Caregivers could help the child become interested in the toys if instructed by the assessor. The total duration of the free play task was 5 minutes. All conversations and behaviours presented by both the child participants and assessors, as well as their caregivers, were recorded by four HD cameras (25Hz) mounted on the surrounding walls, obtaining a full 360-degree view of the activities presented in the trial room.

Figure 4.1. The multiple camera system setups for the social interaction assessment sessions. a) Camera 1, primarily the assessor's frontal view; b) Camera 2, mainly the child participant's frontal view; c) Camera 3, back view of the child and lateral view of the assessor; and d) lateral view of the child and back view of the assessor.

4.2.3 Behavioural Data Coding

As an important aspect of social interaction and communication, joint attention behaviours between the child participants and the assessor presented during the free play task were manually coded by two independent coders using the MATLAB (version R2021a) Ground Truth Labeler application. Joint attention is a type of social skill that describes the ability to follow a social partner's gaze or their pointing direction to objects of mutual interest and the ability to share interests with other social partners. It is common that both social parties have direct eye contact first, and one directs the other's attention to an object, whether by eye gaze or pointing with a finger [142, 151]. Joint attention in a social setting must begin with direct eye contact. A joint attention event may be considered complete if one of the participants focuses on the pointed object for an extended period without making eye contact with the other again or has direct eye contact with each other after following the social prompts.

The FFmpeg (version 4.2.4) library [129] was used to segment a standardised 5-minute free play task from each 45-minute social interaction session on the recorded videos. Two independent coders annotated the beginning and end points of each behavioural event within the segmented free play videos. The free play task ends at the beginning of the still-face paradigm and reunion task. If the still-face paradigm occurred within the segmented videos, coders also marked the start time of its session and removed the annotations for this period.

Cohen's Kappa values were used to assess the inter-rater reliability of the two coders. All 33 free play sessions from the TD group and 16 sessions from the ASD group were coded by both coders. The results showed moderate-to-high inter-rater reliability in both groups. This

included kappa values for labelling joint attention behaviours in the TD group (mean 0.87, standard deviation 0.18) and in the ASD group (mean 0.87, standard deviation 0.19). The lead clinical researcher on the team further checked the reliability of the annotation of the coders on the remaining non-overlapping video segments.

4.2.4 Feature Extraction And Data Preparation

This study uses the OpenFace2.0 open-source toolkit to extract the facial features of each person appearing in the multi-view video segments recorded during free play. The OpenFace2.0 toolkit has been trained on many open-access datasets in facial landmark detection, action unit prediction relevant to emotion recognition, head pose, and eye gaze tracking to comprehensively extract human facial behavioural features. The toolkit performs frame-by-frame image feature extraction on the video recordings. Hand-crafted movement of the gaze shifts (left-right, up-down) and head pose features (pitch, yaw, roll) were further calculated from the coordinates provided by the OpenFace2.0 toolkit. Angular displacements, velocities and accelerations were calculated for these gazes and head pose features on a frame-by-frame basis.

To examine the differences in extracted features between the ASD and TD groups and to compute event-based features, hand-annotated behavioural labels for each joint attention event were pooled into a customised Python script. In addition, multi-view videos recorded for the same participant session were merged and labelled with their corresponding camera angles. The rest of the data recorded in free play were filtered out because they were not related to behavioural events. Furthermore, since the current publicly available training datasets are limited in containing multiple people moving freely in the same scene, current face detection algorithms, including OpenFace2.0, cannot guarantee to capture the facial information of all participants throughout the recordings. In particular, when people move, some facial information may be lost due to head rotation, and the algorithms cannot always assign facial labels uniformly throughout. Therefore, the data with less than an 80% confidence level were also filtered out.

In addition, the use of a multi-view camera setup increased the chances of capturing the full-frontal view of all participants in free play. The optimal performance of current facial recognition algorithms in facial landmark detection tasks is when the full-frontal view of human faces is reflected in the image. The camera angle relative to each person's position in the view or when an object obscures the facial areas may affect the performance in face detection. Therefore, based on the aligned face images extracted by OpenFace2.0, additional manual annotation was performed to label the head positions into three categories: frontal, lateral, and occlusion. Before this manual annotation, cosine similarity was applied to reassign each aligned face image with a maximum of four participants (the child participant, their caregivers (two people maximum, sitting in the chairs next to the trial entrances), and the assessor) using the Histogram of Oriented Gradients (HOG) features generated by OpenFace2.0. However, due to the head orientation issues, the cosine similarity approach was only able to assign a face label for images containing similar facial texture details. It cannot perform recognition of the role of the participants in the video. The caregivers' facial data were filtered out during the labelling process, and the assessors' and the child's facial data were labelled simultaneously. After the additional manual labelling of the head orientation (frontal, lateral, occlusion) and each participant role (child or assessor), the multi-view datasets were further differentiated according to each social event, and the assessors' data were excluded due to the research focus in this study was on the children's responses. Furthermore, only the highest confidence face recognition result from repeated timestamps across multiple camera angles was retained for the multi-view dataset. This helps ensure accurate labelling of facial

landmarks, leading to more precise predictions of action units and gaze direction captured in images.

Moreover, movement features for each episode of a joint attention event were calculated by averaging the values across all frames. The initial state (at time t=0) of hand-crafted gaze and head pose features with angular displacement, velocity, and acceleration equal to zero were excluded for calculation. A detailed description of the extracted features is given in Supplementary Table 18.

4.2.5 Statistical Analysis

Statistical analyses of the hand-crafted and OpenFace2.0 generated features and demographic characteristics were performed on SPSS (version 28). A two-tailed independent samples t-test or Mann-Whitney U test was applied to analyse the mean (or mean rank) of each feature to compare the differences in joint attention behaviour between ASD and TD groups. In our multi-view setup, group differences in facial features may also differ due to the different facial orientations and camera positions of each child detected in the recordings. Therefore, statistical analyses were repeated for each facial orientation and defined as statistically significant when p-values were smaller than 0.05 (two-tailed).

The normality of each feature was assessed using the one-sample Kolmogorov-Smirnov (KS) test. A parametric independent samples t-test was applied when the features were normally distributed, and the p-value of the KS test was greater than 0.05. Otherwise, a Mann-Whitney U-test was applied if the data distribution was skewed. A log-base-10 transformation was not applied in this study because behavioural features have values less than or equal to zero.

For the demographic characteristics analysis, independent sample t-tests were performed for age and nonverbal developmental quotient (NVDQ). The Chi-square tests were conducted for between-group gender differences analysis.

4.2.6 Multi-View Facial Feature Combination And Model Training

This study subdivided facial features recorded from multiple camera angles into three domains (gaze, pose, and action units with facial orientation labels) and explored the feasibility of ML algorithms in differentiating children's facial information captured during spontaneous responses to joint attention with other social partners in free play. The ML algorithms were trained based on SVM models combined with multiple kernel methods. The impact of the comprehensiveness of facial information on the recognition of group differences in joint attention behaviours was also analysed. In addition, the impact of different combinations of feature information used for model development was analysed using a multi-modal feature learning approach and the mean cross-validation accuracy of each model was compared.

SVM is an ML algorithm commonly used with linear and non-linear classification tasks. In non-linear classification, it maps linearly non-separable features to high-dimension space, which helps to use linearly separable hyperplanes with a maximum margin for classification tasks [166]. In this study, we hypothesised that facial features extracted from multiple facial orientations of ASD and TD individuals could be classified using an SVM combined with a kernel approach. The conventional linear, radial basis function (RBF) and sigmoid kernels were applied to SVM for comparison. Feature dimensionality reduction techniques, such as principal component analysis (PCA) and least absolute shrinkage and selection operator (Lasso) regression, were used to reduce redundant facial features and select highly relevant features for classification. In this regard, the PCA algorithm projects all input features onto new principal components with maximum variance characteristics to perform classification on a lower-dimensional plane. On the other hand, Lasso regression produces sparse models by applying

L1 regularisation to remove irrelevant or redundant features, retaining only the features with non-zero coefficients. Lasso assumes a linear relationship between the input variables and the classification output, so the features selected by the lasso regression are more interpretable and have a linear relationship between the features.

4.2.7 Machine Learning Model Validation

A 5-fold cross-validation strategy was used to compute the accuracy, precision, recall, and F1 scores to evaluate the performance of applying different classification methods on the multi-view facial features. Specifically, separate classification models were trained for each facial orientation (frontal, lateral, occlusion, and all views) from the three facial feature domains and combinations. The facial features were first grouped into a training set and a test set in a ratio of 8 to 2. The training set was then assigned to equal proportions of five datasets. During model training, a subset of the training set was selected as the validation set for hyperparameter tuning, and the other subsets were used to train the models. This process recurred five times without repeating the same validation dataset.

A grid search strategy was used on the training samples to train models by initialising different combinations of hyperparameters to find the optimal hyperparameter values for the models' weights. The StandardScaler algorithm from the sklearn Python library was used to normalise the data before commencing model training so that the models could converge faster and avoid the convergence difficulties caused by large numeric numbers. The regularisation parameter C controls the hardness of the hyperplane margin, with a small C leading to a soft margin and a large C leading to a hard margin. This study uses the regularisation parameter in all kernels from 0 to 1 with a step size of 0.1. The kernel coefficient parameter, gamma, defines the size of the kernel, which is applied to the RBF and sigmoid function kernels and affects the shape of the decision boundary. The smaller the gamma value, the further the decision boundary is from the data points. These parameters are set in the range of 0 to 1 with a step size of 0.1. The grid search strategy iterates different combinations of hyperparameter values on the model weights through 5-fold cross-validation. It reports the best hyperparameter values to test model performance on the test set.

4.3 Results

4.3.1 Demographics

Of the 106 child participants aged 3-12 years who participated in this study, 24 children diagnosed with ASD and 22 from the TD group showed joint attention with the assessor during free play, with 104 joint attention episodes totalling 487.6 seconds in length. The average duration of joint attention between child participants and the assessor was 10.6 seconds (SD: 13.15), and the average duration of free play session was 262.49 seconds (SD: 50.75). Descriptive statistics were calculated for the age, gender, and Nonverbal Intelligence Quotient (NVIQ) of the sample (see Table 4.1). Independent-sample t-tests showed no statistically significant differences in age ($t(44) = -1.292$, p-value $= 0.203 > 0.05$ two-tailed). However, statistically significant differences were found in NVIQ ($t(35) = 3.923$, p-value <0.001, two-sided), with nine missing data points. There was no statistically significant gender difference (χ^2 (1, N=46) $= 3.486$, p-value $= 0.062 > 0.05$).

Table 4.1. Demographics and descriptive statistics.

	ASD	*TD*
Demographics	**Mean (SD)**	**Mean (SD)**
N	24	22
Gender (Male/Female)	23/1	17/5
Age (years)	8.16 (2.06)	7.25 (2.74)
NVIQ	96.67 (15.50)	115.19 (12.33)
Data Set (Seconds)		
Average Free Play Task Duration	344.55 (142.66)	257.57 (51.43)
Average Joint Attention Behaviour Duration	12.43 (15.01)	8.60 (10.76)
Total Joint Attention Duration	298.36	189.24

NVIQ = Nonverbal Developmental Quotient/Nonverbal Intelligence Quotient

4.3.2 Statistical Analysis Of Facial Information

A statistical comparison of the extracted facial features (presented in Supplementary Table 21), including eight gaze movements, 35 action units, and 11 head pose movements, showed that there were statistically significant differences between groups in facial muscles and postural movements that could be used to classify individuals with ASD (as shown in Table 4.2). By comparing the facial features presented by child participants in all camera views, the facial action unit AU15 (lip corner depressor) was statistically significantly higher in the TD group than in the ASD group. In addition, there were significant differences in head rotation features. For example, velocity and acceleration of head pose movement in the roll plane (rotation of the head from side to side in the horizontal direction), acceleration of head poses in the yaw plane (left-right head rotation), and angular displacement of head pose in the pitch plane (head up or down) were statistically significantly higher in the ASD population. In the frontal view dataset, AU01 (inner brow raiser) and AU17 (chin raiser) were significantly higher in the TD group than in the ASD group.

Table 4.2. Between group differences in facial features.

(A) Independent Samples t-tests

	Significant Feature	*Mean diff (SE)*	*t-test (df)*	*p-value*	*Hedges' g [95% CI]*
All views	AU15_c	0.187 (0.0729)	t(44) = 2.570	0.014	0.746 [0.152 to 1.331]
	Yaw_Ry_acceleration	2.954 (1.435)	t(44)= 2.058	0.046	0.597 [0.012 to 1.176]

(B) Non-parametric Mann-Whitney U Tests

	Significant Feature	*U*	*p-value*	*Significant group*
All views	Pitch_Rx_angular_displacement	172	0.043	ASD
	Roll_Rz_velocity	167	0.033	ASD
	Roll_Rz_acceleration	164	0.028	ASD
Frontal view	AU01_r	142.5	0.050	TD
	AU17_r	143.5	0.050	TD

4.3.2 Machine Learning Model Performance

Table 4.3 showed that the SVM model using the RBF kernel performs better, especially in unimodality training (gaze, pose, and action units with camera angle information). The mean cross-validation accuracy performances were further improved using feature dimensionality reduction algorithms, such as PCA and Lasso regression. It also demonstrated that the classification performance of action unit features and camera angle information outperformed that of using gaze or pose features. Using a combination of feature inputs (gaze and action unit, pose and action unit) was better than a single modality (gaze or pose features only). Using all three modalities as input information (gaze, pose and action unit features), the models achieved competitive or higher performance than other feature combinations (as shown in Figure 4.2 and Table 4.3).

4.3.4 Feature Dimensionality Reduction And Selection

Figure 4.2 illustrated that when performing data dimensionality reduction on the action unit and pose features, the Lasso regression method performed better than PCA. In contrast, when selecting an SVM classifier with an RBF kernel, using PCA for dimensionality reduction achieved the same or better performance than using Lasso regression (see Table 4.3 for the mean cross-validation accuracy performance of each model).

Table 4.3. Comparison of the mean cross-validation accuracy of machine learning models trained on different feature modalities across all camera views during joint interaction episodes.

JA (All views) (Mean (SD))	**Gaze**	**Pose**	**Action Units+Camera Angle**	**Gaze+AU+ Camera Angle**	**Pose+AU+Camera Angle**	**All features**
linear SVM	0.593 (3.199e-4)	0.593 (3.199e-4)	0.761 (9.585e-3)	0.762 (7.316e-3)	0.764 (6.681e-3)	0.777 (6.068e-3)
RBF SVM	0.594 (5.807e-4)	0.607 (5.779e-3)	0.975 (5.096e-3)	0.840 (1.458e-2)	0.759 (1.244e-2)	0.743 (9.786e-3)
Sigmoid SVM	0.594 (3.199e-4)	0.593 (3.199e-4)	0.706 (1.435e-2)	0.593 (3.199e-4)	0.593 (3.199e-4)	0.593 (3.199e-4)
PCA+ Linear SVM	0.593 (3.199e-4)	0.593 (3.199e-4)	0.738 (6.313e-3)	0.736 (5.198e-3)	0.755 (1.457e-2)	0.754 (1.049e-2)
PCA + RBF SVM	0.723 (8.276e-3)	0.607 (5.779e-3)	0.966 (6.137e-3)	0.962 (6.230e-3)	0.962 (4.493e-3)	0.961 (6.247e-3)
PCA+ Sigmoid SVM	0.593 (3.199e-4)	0.593 (3.199e-4)	0.723 (7.867e-3)	0.724 (7.574e-3)	0.735 (1.104e-2)	0.733 (1.332e-2)
Lasso Regression +Linear SVM	0.593 (3.199e-4)	0.593 (3.199e-4)	0.761 (9.485e-3)	0.761 (8.124e-3)	0.763 (7.383e-3)	0.777 (5.662e-3)
Lasso Regression + RBF SVM	0.613 (2.456e-3)	0.619 (6.510e-3)	0.975 (5.096e-3)	0.896 (1.024e-2)	0.832 (1.423e-2)	0.743 (8.551e-3)
Lasso Regression + Sigmoid SVM	0.594 (4.196e-3)	0.594 (1.379e-3)	0.706 (1.435e-2)	0.593 (3.199e-4)	0.593 (3.199e-4)	0.593 (3.199e-4)

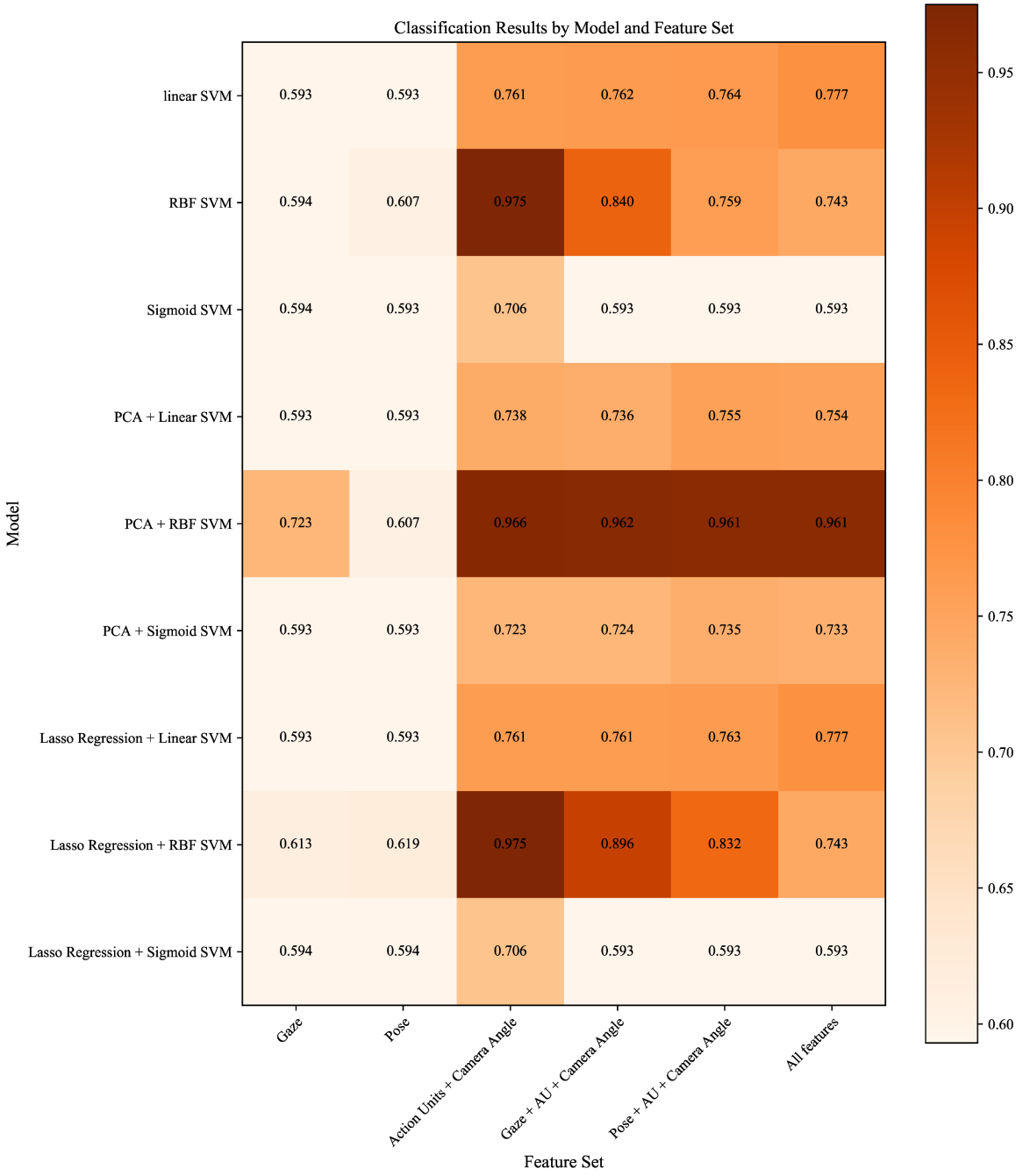


Figure 4.2. Heatmap for the mean cross-validation accuracy performances using different input feature combinations.

4.3.5 Effect Of Face Orientation On Model Prediction

In the comparison between Table 4.3 and Supplementary Table 19, which presents the predictive results of ASD using facial features across all modalities, the ML model trained underwent feature dimensionality reduction through Lasso regression and demonstrated superior performance when utilising single facial orientation feature patterns compared to training with all facial orientations. Of the ASD prediction results using all three facial feature modalities, models trained with Lasso regression with feature reduction showed that single facial orientation features yield better performance than training all facial orientations.

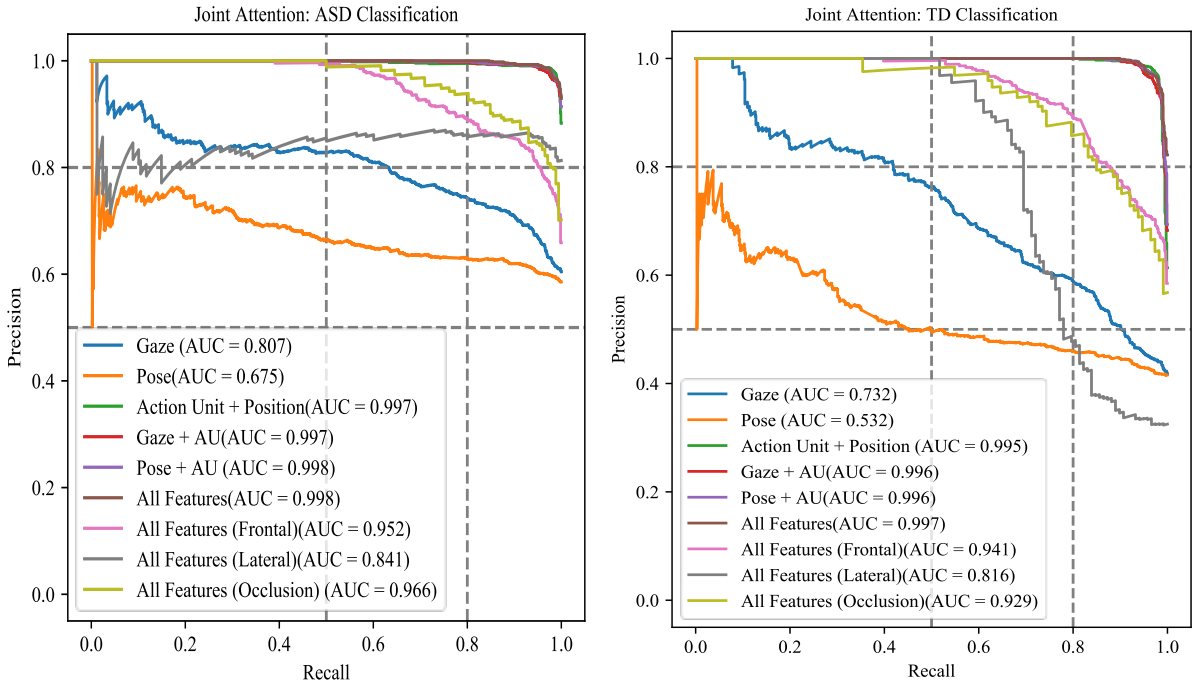


Figure 4.3. Model performance comparisons for ASD and TD classification using different feature inputs extracted from joint attention events in free play. Subfigure a) shows the precision-recall curve in ASD classification, and b) shows the precision-recall curve in TD classification.

Supplementary Table 19 also showed a linear relationship among the facial features after dimensionality reduction using Lasso regression. For instance, the linear SVM model trained with lateral view features achieved an accuracy of 88.3%. Similarly, the accuracy reached 88.9% when partially occluded facial features were used. In addition, frontal view facial features were found more suitable for the RBF kernel-based SVM model, which achieved an accuracy of 86%, surpassing the linear SVM model by 5.3%. The results also indicated that using sigmoid kernel-based SVM models did not show improvement following the Lasso feature reduction. Furthermore, Supplementary Table 20 provided a summary of the features obtained after conducting Lasso regression on each modality of the facial information.

4.3.6 Evaluating Classification Performance Of Machine Learning Classifiers With Different Feature Inputs

Figure 4.3 demonstrated that the prediction accuracy is more dominant when using a feature set that contains action unit features than when using gaze or pose features alone. The best classification accuracy was obtained by including all three types of extracted features. In addition, the SVM classifier with an RBF kernel performed well after PCA dimensionality reduction. This approach yielded ASD classification of AUC = 0.998 and TD of AUC = 0.997.

4.4 Discussion

This study innovatively proposed a quantitative method to address the challenges of health professionals in objectively assessing children's facial responses in joint attention with other social partners in a free play environment. It combined digitalised facial feature extraction algorithms and multi-view, multi-modal facial information fusion methods to establish a new automated, objective mechanism for ASD classification. In addition, we break through the existing research bottleneck of obtaining only frontal facial information for behaviour

assessments and ASD prediction. The SVM-based algorithm proposed in this study was found suitable for quantitative analysis of facial biomarkers from lateral and occluded facial views to assist in ASD classification from irregular head movement features in free play.

This study emphasises the potential of digital representations of facial features derived from joint attention episodes in the free play task of social interaction sessions to discriminate between children with and without ASD accurately. The experimental tasks were recorded using four fixed cameras on the walls to ensure the successful capture of facial information from multiple angles during play, which helped eliminate any potential failure in facial feature extraction due to movement and head rotations. The statistical analysis in Table 4.2 revealed significant differences in posture and facial muscle movements (action units) between children's joint attention behaviours in free play. In addition, the linear SVM models trained using all feature modalities (gaze movement, action units, and posture movement) after Lasso feature reduction achieved over 77.7% mean cross-validation accuracy in discriminating ASD from TD groups. Furthermore, Supplementary Table 20 showed that when action unit features were included in the model input feature vector, the majority of action unit features contributed to the predictive model for distinguishing between ASD and TD groups, but the AU28 (Lip Suck) feature may not have a significant impact. Moreover, utilising feature patterns across all camera views, it was found that features such as rotation velocity of the head from left to right, up and down, horizontal directions, and angular acceleration of the head rotation from left to right were associated with ASD classification.

Given the complexity of ASD, the findings suggest that action units are more discriminative compared to other facial information, possibly because action units can be digitalised using AI models that quantify muscle movement and are considered important biomarkers for assessing nonverbal communication and emotional state. Clinical experiences [167] have evidenced that people with ASD show fewer or more exaggerated facial movements due to a lack of social perception.

Furthermore, a substantial influence of facial orientation on the machine learning model performances was observed. Different from considering features from all views, individual facial orientation data revealed that gaze and head pose rotation angular accelerations (left-right shifting) were identified as two features that were deemed irrelevant for predicting ASD. In the frontal and occluded views, the head pose rotation angular velocity and acceleration for left-to-right rotation were also considered irrelevant. For frontal view only features, it was determined that gaze displacement and velocities associated with looking left-right were irrelevant for the classification task. Similarly, for occlusion view only features, the head pose rotation from left to right, angular acceleration of the head pose rotation from side to side in the horizontal direction, and angular displacement of the head pose rotation from left to right were deemed irrelevant. In lateral view features, gaze angular acceleration in up-down direction was considered irrelevant information.

In addition, SVM models could classify groups even when some facial parts were obscured by external objects, so long as the features extracted by OpenFace2.0 still yielded over 80% certainty. As it summarised in Supplementary Table 19, using features extracted from the lateral view, the linear SVM model trained with Lasso feature reduction achieved 88.3% accuracy, while the occlusion view achieved over 88.9%. The use of frontal view facial information achieved 86% accuracy with an SVM model containing an RBF kernel. In contrast, in some cases, feature dimension reduction using PCA performed better than Lasso regression. For example, utilising frontal view features, model prediction performance achieved an AUC of over 94%. The occluded features achieved over 92% AUC for the TD group prediction and over 96% AUC for the ASD group classification. A possible reason could be that PCA projects high-dimensional features onto new principal components using all the extracted features. On the other hand, Lasso regression reduces the dimensionality of the original dataset by

eliminating features that may result in the omission of non-linear feature patterns. These findings emphasise that collecting facial features from the frontal view is essential, as it represents the full facial perception patterns during social interaction. However, other views could also be used to capture digital feature patterns for ASD classification using ML algorithms.

This study provided a comprehensive evaluation of the effectiveness of using facial information as predictive factors to differentiate between ASD and TD individuals by reporting both the statistical analysis and machine learning performance evaluations. To the best of our knowledge, this is the first study that analysed children's facial information extracted during free play that overcame the challenges faced in traditional studies using eye tracking devices to analyse children's social responses to video stimuli, such as the need to remain seated and the recording devices must capture the frontal view.

Limitations And Future Research Directions

While this study provides valuable insights into the use of facial information in ASD diagnosis, several limitations exist. First, this study only considered children's responses without analysing the reactions of other social partners. This limits our comprehensive understanding of the social contexts and underlying reasons for the facial responses exhibited by children. Therefore, considering the potential impact of other social partners' reactions is crucial for gaining further insights into the psychological processes involved in joint attention in children. Furthermore, a more direct analysis of joint attention behaviour may be appropriate through other social tasks rather than free play. As discussed by Bruinsma et al. [151], TD did not show more initiation of joint attention bids in free play than in the ASD group. Thus, the small sample size and the small heterogeneity learned by the model may introduce bias to model training and lead to overly desirable results. Future studies should also consider exploring social tasks based on facial feature analysis to investigate its potential use in ASD diagnosis. Furthermore, given the complexities of ASD in clinical assessment, the feature extracted should be associated with emotional expressions and integrated with behavioural observation coding with multi-modal information to understand social competencies and communication difficulties. Also, advanced deep learning algorithms in facial recognition from multiple head orientations should be trained to support facial feature extraction from multiple camera views due to natural responses to social interaction leading to unpredictable head movement.

4.5 Summary

This chapter provided an in-depth analysis of children's facial features during joint attention behaviour presented in free play. By integrating statistical and ML analysis, the study proposed clinically relevant facial features that can be used for ASD assessments. The finding was consistent with the prior knowledge from existing literature about analysing facial feature changes in social interactions. Moreover, this study reinforces the significance of digital information in classifying and recognising ASD-related behavioural markers in unrestricted natural interaction settings. It highlights the effectiveness of artificial intelligence algorithms in mining feature patterns from multiple facial orientations that are challenging to quantify through observational approaches.

Chapter 5: Automated Electrocardiogram Signal Processing And Heart Rate Variability Calculation In The Clinical Assessments of Social Competencies in Autism Spectrum Disorder: Statistical Validation And Clinical Application

Part of the contents presented in this chapter is published in the EMBC conference 2023.

The preceding sections introduced a deep learning framework that recognises social behaviours between multiple individuals in free play environments with multi-view, multi-modal information. These chapters also revealed that human facial responses in natural social interaction environments play a decisive role in ASD classification using machine learning algorithms.

In addition to behavioural markers, physiological signals are widely used in hospital settings as an essential biomarker reflecting the cardiovascular and physiological health of the human body. In particular, heart rate variability (HRV) metrics derived from electrocardiogram (ECG) have been signified to be associated with various psychological conditions and psycho-emotional changes in people during social interactions. Thus, it is considered a reference for assessing social competencies and attention. Heterogeneity in data labelling is introduced with the large volumes of ECG data collected in the clinical setting, which vary in data quality (differences in sampling rate, motion artefacts and environmental noise) and device used. In addition, data collection often involves significant highly skilled labour costs, which could introduce human error and subjective bias, increasing the risk of recording incorrect clinical information.

Limited by the current prevalence of manual signal processing and R-peak detection in clinical assessments, most of the literature has reported only short-term baseline data, while few clinical trials have reported long-term recordings in social interaction settings. Therefore, this chapter addresses the third research question, aiming to investigate whether there are statistical and clinical differences in HRV derived from ECG signals collected during natural social interaction tasks. First, this chapter presents an automated ECG signal cleaning and HRV calculation toolbox (a modified version of NeuroKit2), and it was validated against human annotated signals (such as commercialised manual signal cleaning toolboxes) on publicly available datasets (collected with both resting-state and exercises) and on our clinical baseline dataset. Then, the clinical utility of the tool in social interaction assessment tasks is assessed by comparing the changes in HRV between baseline tasks (resting state and watching a movie) and free play tasks (unrestricted movement and playing toys).

5.1 Introduction

Physiological monitoring has long been an important part of healthcare services and personal training, providing crucial insights into physical conditions. These physiological signals can be collected through the electrical activities of the human body. For instance, an electrocardiogram (ECG) reflects the cardiovascular condition of the heart. The heart rate variability (HRV) metrics obtained from the fluctuations in inter-beat-interval (IBI) durations of the ECG signals [168, 169] reflect the autonomic nerves' working condition. Reduced HRV indicates a higher risk of mental disorders and cardiovascular diseases [170]. Over the past decades, non-invasive ECG devices have been revolutionised as a valuable source of information for recording cardiac variability, transitioning from traditional bedside 12-lead dry electrode ECGs to wearable 2-lead ECGs, and photoplethysmography (PPG) using optical sensors. Technological innovations seek to produce portable, easy-to-use, and low-cost products for recording ECG signals. However, long-standing challenges persist about the inconsistency of data processing algorithms. These inconsistencies are partially due to differences in the nature of the signals and the design of the signal acquisition systems [171]. For example, while ECG provides a direct measure of the electrical activity of the heart, PPG measures the mechanical response at some distance from the heart, introducing a lag from the ECG beats that varies with time, health conditions, and individual differences (pulse transit time) [172]. Addressing these challenges is critical to ensure the reliability and efficacy of reporting physiological health through monitoring systems in real-world applications.

There is currently no consensus on the optimal sampling frequency for digital ECG data recording due to the different product standards between countries and the development of emerging algorithms in physiological signal processing. At lower sampling frequencies, typically ranging from 50 to 100Hz, signal interpolation [173] is required as the low-frequency signals are more sensitive to motion artifacts. In contrast, higher frequency signals (e.g., above 500Hz) usually require better computational resources to evaluate the electrical activity of the heart. Despite the lack of consensus on sampling frequencies, a unified quantitative algorithm on HRV metrics calculation has reached a consensus, that is to accurately locate the QRS complexes [5] from ECG signals. This is because the QRS complex indicates ventricular depolarizations characterised by high amplitude electrical potential changes. It involves calculating the change in duration between adjacent R-peaks, which gives an indication of HRV in the time domain. However, the differences in sampling frequencies may have an unpredictable impact on the R-peak detection process, resulting in inconsistent HRV calculations [174]. Therefore, developing a standardised digital signal processing tool is considered critical.

Traditional approaches in manual ECG signal quality and R-peak detection inspection are impractical for long-term datasets (for instance, several hours to days). Therefore, many studies have analysed only short-term data (e.g., about 5 minutes) for HRV calculations [175]. The most commonly used signals to validate the reliability of HRV calculations are resting-state ECGs, while fewer studies have used ECG data collected under stress to validate algorithms [176]. In studies diagnosing social competence and quality of life in individuals with mental disorders, the response to individuals under social stress is a critical component of such assessments [42, 177]. Given the complexity of data cleaning, emerging software for automated R-peak detection and HRV calculation is urgently needed for ECGs collected in social interaction assessments.

Several automatic physiological signal processing toolboxes have been proposed in the literature [178, 179]. However, these have not been widely applied to psychological studies for the diagnosis of mental disorders. This chapter presented a modified NeuroKit2 toolbox [179] for ECG signal processing and HRV calculation. The toolbox was validated to evaluate its performance in processing a public dataset (the MIT-QT databases) collected from healthy adults. It was then validated through our baseline ECG signals collected for social interaction

assessments on children with and without ASD. The QT database [180] consists of expert-annotated ECG signals for the measure of QT interval. The annotation includes the onset and offset of QRS complexes and T-waves, which indicates cardiac repolarisation of the heart. The HRV calculation differences in time-domain metrics between our modified NeuroKit2 toolbox and its default setup in relation to the human annotations were compared using the QT database. The ECG signals collected during the baseline social interaction task were annotated by clinicians through the commercialised MindWare software, and HRV metrics were compared with the modified NeuroKit2 toolbox.

The toolbox was applied to child participants who completed a free play task in the social interaction assessment after statistical validation of the HRV metrics calculation on ECG signals collected at rest and during exercise. Statistical and clinical differences between algorithms were assessed using the Two One-Sided t-tests (TOST) [181, 182]. Furthermore, the clinical significances between ASD and TD populations on the social interaction assessment tasks were analysed using independent sample t-tests or non-parametric tests. In addition, changes in HRV between the baseline and the free play tasks of the assessments in both clinical groups were further examined using paired-samples t-tests and general linear models for repeated measures.

5.2 Materials And Methods

5.2.1 Datasets Overview

The QT Database

Three datasets were used in this chapter to validate the NeuroKit2-based automatic HRV calculation toolbox: the MIT-BIH Normal Sinus Rhythm dataset (NSR) [183] and the MIT-BIH ST Change dataset (ST change) [184] provided by the QT database and our clinical baseline dataset. The NSR dataset consists of 10 records in total 2 hours and 30 minutes long, with 300 annotated beats. The ST change dataset contains six recordings in total of 1 hour and 30 minutes long with 206 annotated beats. These two datasets represent a group of healthy adult participants who show normal sinus rhythm when remaining seated and show ST-segment elevation or depression during stress tests (ST change). The frequency published in those QT databases was sampled or converted at 250Hz.

Social Interaction Assessment Datasets

To further validate the toolbox for calculating HRV metrics in a child population and to explore its potential for clinical application, this study included 106 participants who completed a baseline social interaction assessment. The 106 child participants (ASD = 64, and TD = 42), aged 3-12 years, included in this study were recruited from the Clinic for Autism and Neurodevelopmental (CAN) Research at the Brain and Mind Centre, the University of Sydney. Participants in the ASD group were confirmed using the DSM-5 criteria [165], followed by the Autism Diagnostic Observation Schedule 2nd edition (ADOS-2) assessment [128], administered by research reliable assessors. The TD children were recruited from the broad community via newspapers and flyers. A telephone interview was administered to assess by an independent interviewer. Eligible participants were required to be free of any cardiac condition and use cardiovascular medication, and no participants were receiving psychotropic medication. The study was approved by the University of Sydney Human Research Ethics Committee (2013/502) and the Sydney Local Health District Human Research Ethics Committee (HREC/18/RPAH/157). Written consent was obtained from all participants' caregivers prior to their inclusion in the study.

During the social interaction assessments, the cardiac activities of the child participants were recorded using a portable ECG Holter device fitted with standard 2-lead electrodes (Ag/AgCI) on the child's chest, abdomen and back [185]. The Holter device sampling rate was at 500Hz

with a powerline noise of 50Hz. The ECG signals were saved through a MindWare Mobile Recorder and transferred to a lab computer with a wireless connection.

The first five-minute baseline task assessed the validity of applying the modified toolbox in processing ECG signals and calculating HRV metrics from the child population. During these five minutes, participants were asked to watch a pre-recorded video, remain calm, and reduce their movements. A total of 8 hours and 50 minutes of ECG signals collected during the baseline tasks were annotated by clinicians using MindWare software for validation. After validation of the NeuroKit2 toolbox , further analysis was conducted on participants who completed the free play task of the baseline social interaction assessment. Group differences in HRV between ASD and TD individuals during these sessions were compared.

5.2.2 ECG Signal Pre-Processing

Before extracting HRV metrics, ECG signal pre-processing is needed first to reduce the baseline wander and muscle noises caused by body movement and powerline interference that occurs in the power supply (50 or 60Hz) [186]. The 50Hz power supply is used in Europe, Asia, Australia and New Zealand, whereas 60Hz is used in North and South America. The NeuroKit2 toolbox, with its default settings, removes the baseline wandering in the raw signal below 0.5Hz using a 5^{th} order high-pass Butterworth filter. A powerline smoothing algorithm is used with a moving average kernel. The window size equals one period of the powerline frequency, with a default setting of 50Hz. To obtain nearly consistent ECG data cleaning processes with the MindWare software used in our previous study [187], we created a custom Python script based on the original NeuroKit2 package. A 9^{th} order band-pass Butterworth filter between 0.25 and 40Hz was applied to remove the baseline and muscle noise, as in

$$H(j\omega)=\frac{1}{1+\left(\frac{\omega}{\omega_0}\right)^{2n}} \tag{5.1}$$

Where ω in equation (5.1) is the radian frequency, ω_0 represents the band-pass frequency and the order n.

A notch filter was then applied to remove powerline noise (in s-domain), as in

$$H(s)=\frac{s^2+\omega_0^2}{s^2+\frac{\omega_0}{Q}s+\omega_0^2} \tag{5.2}$$

The band-stop frequency was set to 50Hz on our baseline dataset or 60Hz after visual inspection using a power spectrum density plot on the MIT-BIH datasets. Where ω_0 in equation (5.2) is the band-stop frequency, and the quality factor Q is set to 30.

5.2.3 Data Annotation and R-Peak Detection

R-Peaks Detection On NeuroKit2 Toolbox

After ECG signal pre-processing, the NeuroKit2 toolbox extracts QRS complexes using the steepness of the absolute gradient of the ECG signal, and the position of the local maxima in the QRS complex was labelled as the R-peaks (as shown in Figure 5.2(a)), which is consistent between our methods and the default settings in NeuroKit2.

Experts Annotation On R-Peaks

The expert annotated R-peak locations in the MIT-BIH datasets were extracted using the WFDB library. In contrast, the clinicians used the automatic peak detection algorithm in MindWare and visually inspected all detected R-peaks for artefact removal and corrected for ectopic beats that were less than 10% of the max R-peaks annotated of our baseline dataset. In MindWare, the artifacts and range settings were set to the minimum heart rate of 40BPM (beats per minute) and maximum of 200BPM. The IBI min/max and MAD/MED checks [188] were

turned on in the MindWare software. The peak detection was set as its default value with a sensitivity of 1.

5.2.4 Heart Rate Variability Feature Extraction

Time-Domain Feature Extraction

The Inter-beat-interval (IBI) series representing the mean of the time interval between two heartbeats was calculated from the detected R-peaks. A pool of time-domain components of HRV was derived from the IBI series according to the guidelines suggested by Quintana et al. [189]. Time domain features were analysed to reflect HRV over consecutive R-R intervals. In this study, we report SDNN (standard deviation of R-R intervals in milliseconds), RMSSD (root mean square of successive R-R interval differences), NN50 (the number of pairs of successive R-R intervals that differ by more than 50ms) and pNN50 (the proportion of successive R-R intervals that are larger than 50ms) metrics on all three datasets. In addition, Mean Heart Rate (the average heart rate in beats per minute, MHR) values were calculated and compared only on our baseline dataset due to the lack of information in the MIT-BIH datasets. The MeanNN (mean of R-R intervals in milliseconds) values were compared on the QT database and not reported on the child baseline data because the MindWare software does not provide this metric.

Frequency-Domain Feature Extraction

Due to the lack of ground truth information about the frequency-domain features in the MIT-BIH datasets, this study only analysed the frequency-domain features on the baseline dataset, which includes low-frequency (LF) and high-frequency (HF) components, and respiratory sinus arrhythmia (RSA). The RSA values were estimated by taking the natural logarithm of the HF components (LnHF). According to the guidelines in Quintana et al. [189], the frequency-domain analysis of ECG recordings in children needs to be set between 0.04 and 0.24 Hz for the low-frequency band and between 0.24 and 1.04 Hz for the high-frequency band. In MindWare, the windowing function was set to Hamming window, and the entire five-minute window size was used for the Fast Fourier Transform (FFT). However, in our modified NeuroKit2 toolbox, the FFT spectrum using the Welch periodogram [37] was used. The Welch method calculates the power spectrum density using a sequence of overlapping segment intervals rather than taking the entire data into one window.

5.2.5 Statistical Analysis

Validation Of The Modified NeuroKit2 Toolbox

HRV metrics calculations from the hand-annotated data and the NeuroKit2 toolbox were compared using SPSS (version 28). Descriptive statistics and One-Sample Kolmogorov-Smirnov (KS) Tests were first conducted to assess data normality. The data were found not normally distributed using a two-tailed t-test (p-value < 0.05). TOST [181, 182] were conducted to test equivalences between the two signal processing methods implemented on the NeuroKit2 toolbox (NeuroKit2 default settings and modified settings conforming to the manual labelling tool, MindWare) and the labels provided by the three datasets. The null hypothesis was that the mean differences between the two signal processing approaches were not equivalent and that the observed true differences were outside the equivalence range. A sensitivity analysis was conducted using G*Power (version 3.1) to determine the smallest effect size of interest (SESOI) [182] for the equivalence bounds. The alpha level was set to 0.05. The "TOSTtwo" function in R (version 4.2.2) was applied to conduct the TOST test.

Differences Between The Clinical Groups In The Social Interaction Assessment

The statistical differences of the HRV collected during the baseline and free play tasks in the social interaction assessment between the ASD and TD children were compared using

independent samples t-tests and non-parametric Mann-Whitney U tests. A logarithmic base-10 transformation, Log10 (x + 1), where x represents each HRV metric, was applied due to violations of normality assumptions with One-Sample KS tests (p-values <0.001).

In addition, the changes in HRV between the two tasks were assessed using paired samples t-tests and non-parametric Wilcoxon signed ranks tests. The General Linear Model (GLM) repeated measure analyses were conducted to investigate the within-subjects and between-subjects effects on time, group (ASD or TD), and time and group interactions. Following the central limit theory, the changes in HRV metrics between tasks were analysed in a natural format; logarithm transformations were not applied.

5.3 Results

5.3.1 ECG Signal Pre-Processing And R-Peak Detection

All ECG recordings included in this study were filtered through two signal processing methods separately (that is, the NeuroKit2 default method and our modified NeuroKit2 signal processing approach similar to the MindWare software). As the two MIT-BIH datasets were collected from both the U.S. (the normal sinus rhythm group) and European research sites (the ST change subsets), powerline noise may differ between sites. To apply a correct powerline noise filtering algorithm using our modified NeuroKit2 toolbox, the power spectrum density of the original ECG signal was first visualised. Figure 5.1(a) shows an example where a 60Hz powerline reference was recorded in the original ECG signal. In our modified NeuroKit2 toolbox, a powerline interference at 60Hz and motion artifacts were filtered, resulting in a signal with less noise (Figure 5.1(b)). The NeuroKit2's default signal processing approach applied powerline smoothing at 50Hz.

5.3.2 Heart Rate Variability Metrics Validation

Time-Domain Analyses On The MIT NSR And ST Change Datasets

A total of 506 annotated R-peaks from the 16 selected ECG recordings provided by the QT database were used to calculate time domain HRV metrics and considered as the ground truth labels. Corresponding time-domain HRV metrics were calculated from the two signal processing methods applying the NeuroKit2 toolbox. The normality of these calculated metrics was tested using the One-Sample KS Test. The SDNN and RMSSD values from the NSR dataset and the SDNN value from the ST change dataset were significant (p-value < 0.05), indicating that the assumption of a normal distribution of the data was violated. The TOST tests revealed that all time-domain HRV metrics were statistically significant (p-value < 0.05), suggesting equivalences between the time-domain HRV metrics of the QT database (NSR and ST change datasets) and the two automated signal processing approaches using NeuroKit2, as shown in Table 5.1-5.4.

Time- And Frequency-Domain Analyses On Our Baseline Dataset

The normality of our baseline dataset showed that only the mean heart rate and RSA values were non-significant (p-value = 0.2); all other metrics compared were found significant (p-value < 0.05), indicating that the assumption of a normal distribution of the data was violated. The TOST tests were also evaluated using our baseline dataset, where the HRV metrics were calculated by the MindWare software and annotated by humans, serving as the ground truth labels. As it demonstrated in Table 5.5, time-domain metrics such as mean heart rate, pNN50, NN50, RMSSD and SDNN indicated statistically significant results for the equivalence tests (p-value < 0.05). However, among the frequency-domain HRV metrics, only the low-frequency component indicated statistically significant results (p-value <0.05). The high-frequency components demonstrated statistically non-significant results, for example, high-frequency power (t (210) = 1.609, p-value = 0.0546) and RSA (t (210) = 1.080, p-value = 0.141).

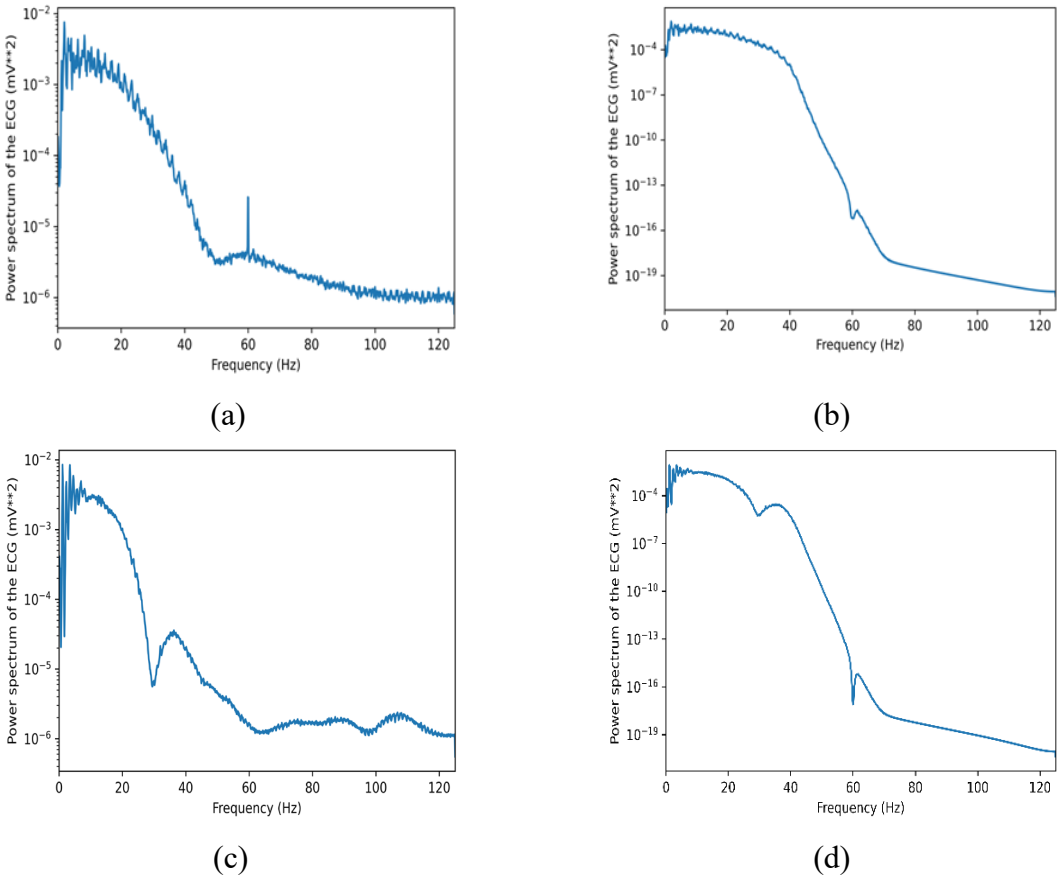


Figure 5.1. Power spectrum density plots. (a) Before signal processing (powerline noise at 60Hz); (b) after signal processing using our modified NeuroKit2 toolbox; (c) before signal processing (no powerline noise evident), and (d) showing an over-processed signal.

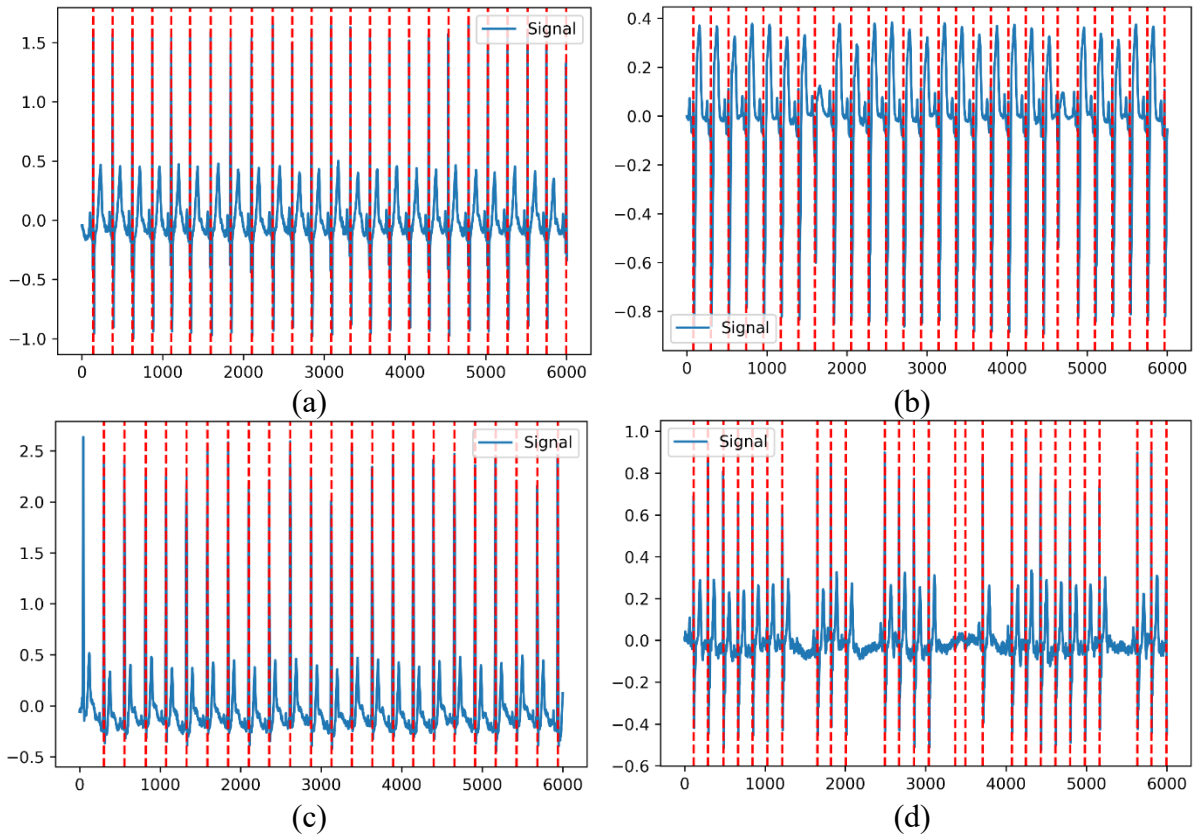


Figure 5.2. ECG signal and R-peak location detection using our modified NeuroKit2 toolbox. (a) All R-peaks were correctly identified on a non-inverted ECG signal; (b) all R-peaks were correctly identified on an inverted ECG signal; (c) the first R-peak was not detected due to incomplete QRS complex observed, and (d) additional R-peaks detected due to noisy signals.

Table 5.1. Two one-sided tests for equivalence between time-domain heart rate variability metrics were calculated by the QT databases and the NeuroKit2 toolbox with default settings in the Normal Sinus Rhythm group.

Normal Sinus Rhythm	**QT – NeuroKit2 Default**			
	Mean Diff (SE)	***p-value (lower/upper bound)***	***TOST [90% CI]***	***Equivalence Bounds***
MeanNN	0.072 (0.096)	0.010/ 0.011	[-91.516, 91.660]	[-133.531, 133.531]
SDNN	-0.031 (0.330)	0.011/ 0.010	[-33.941, 33.879]	[-49.439, 49.439]
RMSSD	0.227 (0.606)	0.010/ 0.011	[-34.380, 34.834]	[-50.456, 50.456]
NN50	4.700 (1.545)	0.009/ 0.012	[-148.896, 158.296]	[-223.935, 223.935]
pNN50	0.453 (0.139)	0.010/ 0.011	[-18.092, 18.998]	[-27.038, 27.038]

Table 5.2. Two one-sided tests for equivalence between time-domain heart rate variability metrics calculated by the QT databases and the NeuroKit2 toolbox with our modified settings in the Normal Sinus Rhythm group.

Normal Sinus Rhythm	**QT – NeuroKit2 Ours**			
	Mean Diff (SE)	***p-value (lower/upper bound)***	***TOST [90% CI]***	***Equivalence Bounds***
MeanNN	-0.020 (0.018)	0.011/ 0.011	[-91.595, 91.554]	[-133.511, 133.511]
SDNN	0.336 (0.084)	0.010/ 0.011	[-33.597, 34.270]	[-49.474, 49.474]
RMSSD	0.944 (0.310)	0.010/ 0.012	[-33.755, 35.644]	[-50.590, 50.590]
NN50	5.500 (2.669)	0.009/ 0.012	[-148.755, 159.755]	[-224.896, 224.896]
pNN50	0.529 (0.257)	0.009/ 0.012	[-18.097, 19.156]	[-27.157, 27.157]

Table 5.3. Two one-sided tests for equivalence between time-domain heart rate variability metrics calculated by the QT databases and the NeuroKit2 toolbox with default settings in the ST Change group.

ST Change	**QT – NeuroKit2 Default**			
	Mean Diff (SE)	***p-value (lower/upper bound)***	***TOST [90% CI]***	***Equivalence Bounds***
MeanNN	-0.058 (0.132)	0.010/ 0.010	[-222.298, 222.182]	[-335.539, 335.539]
SDNN	-0.645 (0.489)	0.011/ 0.010	[-45.814, 44.523]	[-68.196, 68.196]
RMSSD	-6.310 (2.233)	0.019/ 0.006	[-39.159, 26.540]	[-49.597, 49.597]
NN50	1.833 (4.028)	0.010/ 0.011	[-90.329, 93.996]	[-139.148, 139.148]
pNN50	0.264 (0.366)	0.010/ 0.011	[-9.800, 10.327]	[-15.194, 15.194]

Table 5.4. Two one-sided tests for equivalence between time-domain heart rate variability metrics calculated by the QT databases and the NeuroKit2 toolbox with our modified settings in the ST Change group.

ST Change	**QT – NeuroKit2 Ours**			
	Mean Diff (SE)	***p-value (lower/upper bound)***	***TOST [90% CI]***	***Equivalence Bounds***
MeanNN	-0.148 (0.201)	0.011/ 0.010	[-222.381, 222.085]	[-335.529, 335.529]
SDNN	-0.659 (0.522)	0.011/ 0.010	[-45.857, 44.538]	[-68.239, 68.239]
RMSSD	-6.538 (2.387)	0.019/ 0.006	[-39.794, 26.718]	[-50.210, 50.210]
NN50	1.167 (3.643)	0.010/ 0.011	[-91.975, 94.308]	[-140.626, 140.626]
pNN50	0.167 (0.331)	0.010/ 0.011	[-10.019, 10.352]	[-15.377, 15.377]

Table 5.5. Two one-sided tests for equivalence between time- and frequency-domain heart rate variability metrics calculated by mindware software and the NeuroKit2 toolbox with our modified settings in our child baseline task.

Baseline Task	MindWare – NeuroKit2 Ours			
	Mean Diff (SE)	***p-value (lower/upper bound)***	***TOST [90% CI]***	***Equivalence Bounds***
Mean Heart Rate	0.0199 (0.0332)	0.010/0.010	[-3.236, 3.197]	[-4.558, 4.558]
SDNN	0.0142 ($7.77e^{-3}$)	0.019/0.005	[-7.23, 5.238]	[-8.835, 8.835]
RMSSD	0.0206 ($7.78e^{-3}$)	0.023/0.004	[-11.030, 7.275]	[-12.971, 12.971]
NN50	0.0138 (0.0107)	0.010/0.011	[-16.859, 17.236]	[-24.159, 24.159]
pNN50	$9.20e^{-4}$ ($6.95e^{-3}$)	0.007/0.014	[-4.053, 4.790]	[-6.266, 6.226]
HF	0.0972 (0.0190)	0.055/ 0.001	[-721.321, 278.149]	[-708.201, 708.201]
LF	0.0238 (0.0159)	0.038/0.002	[-499.274, 247.570]	[-529.197, 529.197]
RSA (LnHF)	0.0145 ($2.79e^{-3}$)	0.141/ 0.0002	[-0.520, 0.0700]	[-0.418, 0.418]

5.3.3 Between Group Differences In HRV Metrics

The independent samples t-tests and non-parametric Mann-Whitney U tests presented in Table 5.6 indicated statistically significant differences between the ASD and the TD groups in the time-domain HRV metrics (NN50, pNN50) and frequency-domain HRV metrics (HF, LnHF) during the 5-minute baseline task while watching pre-recorded videos. These results revealed significantly lower HRV metrics in the ASD child population. No statistically significant mean or median differences were observed during the free play session, as it shown in Table 5.7.

HRV Changes Between Social Interaction Tasks

Table 5.8 presented the paired samples t-tests, and non-parametric Wilcoxon signed ranks tests conducted to assess the differences in HRV metrics between the baseline and free play sessions of the TD group. It revealed that the LF component showed no statistical significance (p-value > 0.05). By contrast, the mean heart rate of the TD children in free play was statistically significantly higher compared to the baseline task (p-value < 0.001). All other heart rate variables, including SDNN, RMSSD, NN50, pNN50, HF, and LnHF, showed statistically significant results (p-value < 0.05) and exhibited lower mean differences in free play.

Table 5.9 showed that children with ASD had a higher mean heart rate during free play (p< 0.021). The time-domain HRV variable, NN50, was statistically significant (p<0.05), with a higher mean difference in the baseline task. All other HRV metrics were not statistically significant (p-value > 0.05).

Table 5.6. Independent samples t-tests and Mann-Whitney U tests for between group differences in the baseline task.

Variable	NeuroKit2 Ours				NeuroKit2 Ours		
	t-test for Equality of Means				Mann-Whitney U test		
	Mean Diff (SE)	*t-test results*	*p-value*	*95% CI of the Differences*	*U- Test Statistic*	*Z Value*	*p-value*
Mean Heart Rate	-5.125 (2.770)	-1.850	0.067	[-10.618, 0.369]	1062	-1.822	0.069
Log10(SDNN +1)	0.0221 (0.0385)	0.574	0.567	[-0.0543, 0.0985]	1315	-0.187	0.851
Log10(RMSSD+1)	0.102 (0.0550)	1.851	0.067	[-0.00728, 0.211]	1134	-1.356	0.175
Log10(NN50 +1)	0.212 (0.121)	1.745	0.084	[-0.289, 0.453]	1036	-1.990	0.047
Log10(pNN50 +1)	0.200 (0.104)	1.917	0.058	[-0.00690, 0.407]	1041	-1.957	0.050
Log10(HF +1)	0.268 (0.107)	2.508	0.014	[0.0562, 0.481]	1017	-2.112	0.035
Log10(LF +1)	0.0339 (0.0866)	0.391	0.696	[-0.138, 0.206]	1266	-0.504	0.614
Log10(LnHF +1)	0.0373 (0.0154)	2.242	0.017	[0.00678, 0.0678]	1017	-2.112	0.035

Table 5.7. Independent samples t-tests and Mann-Whitney U tests for between group differences in the free play task.

Variable	NeuroKit2 Ours				NeuroKit2 Ours		
	t-test for Equality of Means				Mann-Whitney U test		
	Mean Diff (SE)	*t-test results*	*p-value*	*95% CI of the Differences*	*U- Test Statistic*	*Z Value*	*p-value*
Mean Heart Rate	-2.368 (3.499)	-0.677	0.50	[-9.323, 4.587]	760	-1.180	0.238
Log10(SDNN +1)	-0.0321 (0.0584)	-0.549	0.584	[-0.148, 0.0840]	870	-0.226	0.822
Log10(RMSSD+1)	0.0118 (0.781)	0.151	0.880	[-0.143, 0.167]	830	-0.572	0.567
Log10(NN50 +1)	0.210 (0.118)	1.779	0.079	[-0.0247, 0.445]	713.5	-1.583	0.113
Log10(pNN50 +1)	0.162 (0.104)	1.568	0.120	[-0.0435, 0.368]	729	-1.449	0.147
Log10(HF +1)	0.0598 (0.122)	0.490	0.625	[-0.183, 0.302]	815	-0.703	0.482
Log10(LF +1)	-0.0370 (0.142)	-0.261	0.795	[-0.319, 0.245]	833	-0.546	0.585
Log10(LnHF +1)	0.139 (0.282)	0.494	0.622	[-0.421, 0.699]	815	-0.703	0.482

Table 5.8. Within-group differences in HRV metrics of the TD group (Paired Samples t-tests, Wilcoxon Signed Ranks Test).

Variable (Freeplay – Baseline)	Paired Samples t-tests				Wilcoxon Signed Ranks Test	
	Mean Diff (SE)	*t-test results*	*p-value*	*95% CI of the Differences*	*Z Value*	*p-value*
Mean Heart Rate	5.148 (0.578)	8.906	<0.001	[3.969, 6.327]	-4.731	<0.001
SDNN	-8.747 (3.098)	-2.824	0.008	[-15.064, -2.429]	-2.599	0.009
RMSSD	-14.584 (4.426)	-3.295	0.002	[-23.612, -5.556]	-3.254	0.001
NN50	-37.906 (7.471)	-5.074	<0.001	[-53.143, -22.670]	-3.843	<0.001
pNN50	-10.196 (1.902)	-5.360	<0.001	[-14.075, -6.317]	-4.076	<0.001
HF	-1060.424 (347.088)	347.088	0.005	[-1768.310, -352.530]	-3.964	<0.001
LF	-276.535 (238.998)	238.998	0.256	[-763.974, 210.903]	-0.580	0.562
RSA (LnHF)	-0.559 (0.139)	-4.021	<0.001	[-0.842, -0.275]	-3.366	<0.001

Table 5.9. Within-group differences in HRV metrics of the ASD group (Paired Samples t-tests, Wilcoxon Signed Ranks Test).

Variable (Free play – Baseline)	Paired Samples t-tests				Wilcoxon Signed Ranks Test	
	Mean Diff (SE)	*t-test results*	*p-value*	*95% CI of the Differences*	*Z Value*	*p-value*
Mean Heart Rate	1.149 (2.054)	0.559	0.578	[-2.972, 5.269]	-2.303	0.021
SDNN	77.400 (79.890)	0.969	0.337	[-82.837, 237.639]	-0.857	0.392
RMSSD	120.060 (120.705)	0.995	0.324	[-122.044, 362.164]	-1.132	0.258
NN50	-17.352 (5.557)	-3.123	0.003	[-28.498, -6.206]	-2.082	0.037
pNN50	-3.224 (2.176)	-1.481	0.144	[-7.589, 1.141]	-1.876	0.061
HF	448.066 (441.598)	1.015	0.315	[-437.667, 1333.799]	-1.477	0.140
LF	728804.440 (728105.764)	1.001	0.321	[-731590.777, 2189199.662]	-1.227	0.220
RSA (LnHF)	-0.0482 (0.152)	-3.160	0.753	[-0.354, 0.258]	-0.753	0.451

Between Group Differences

The between-group differences in the changes observed in HRV metrics between the free play and baseline sessions were summarised in Table 5.10. It revealed that changes in the mean heart rate were statistically significant ($p = 0.009$), and the mean difference was larger in the TD group. The changes in HRV metrics such as RMSSD, NN50, pNN50, HF and LnHF were statistically significant ($p < 0.05$), and the mean differences of the ASD group were larger.

Repeated Measures GLM

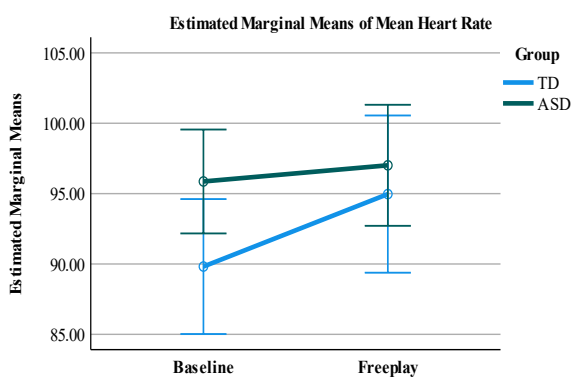
The GLM repeated measures analysis (as shown in Table 5.11 and Figure 5.3) revealed statistically significant time effects on mean heart rate between the baseline and free play tasks ($F(1, 84) = 5.393$, p-value = 0.023). Time-domain HRV metrics were also shown with statistical significance. For instance, NN50 demonstrated significant effects of time ($F(1, 84) = 35.854$, p-value <0.001), time-group interaction ($F(1, 84) = 4.961$, p-value =0.029), and between subjects group ($F(1, 84) = 4.988$, p-value <0.028). Similarly, significant time effects was showed on pNN50 ($F(1, 84) = 17.73$, p-value <0.001), time-group interaction ($F(1, 84) = 4.785$, p-value =0.031). In addition, significant effects were shown on high-frequency components. HF showed significant time-group effects ($F(1, 84) = 5.668$, p-value =0.020). LnHF showed significant time effects ($F(1, 84) = 7.256$, p-value =0.009), and time-group interaction effects ($F(1, 84) = 5.134$, p-value =0.026).

Table 5.10. Between-group differences of the changes in HRV metrics in the free play and baseline tasks (Independent Samples t-tests and Mann-Whitney U tests).

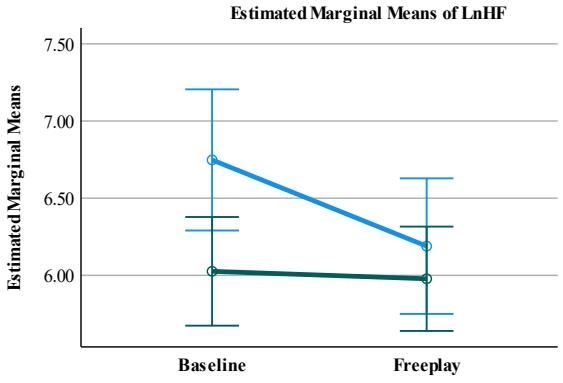
Changes between Free play and Baseline	**Independent Samples t-tests**					***Mann-Whitney U Test***		
	Mean Diff (SE)	***t-test results***	***p-value***	***95% CI of the Differences***	***Hedges'g Effect Sizes [95% CI]***	***U Test***	***Z Value***	***p-value***
Mean Heart Rate	4.0 (2.712)	1.475	0.144	[-1.393, 9.392]	0.326 [-0.111, 0.761]	573	-2.600	0.009
SDNN	-86.148 (104.058)	-0.828	0.410	[-293.078, 120.783]	-0.183 [-0.617, 0.252]	689	-1.563	0.118
RMSSD	-134.644 (157.217)	-0.856	0.394	[-447.287, 178.0]	-0.189 [-0.623, 0.245]	639	-2.01	0.044
NN50	-20.554 (9.228)	-2.227	0.029	[-38.906, -2.203]	-0.492 [-0.931, -0.051]	605.5	-2.310	0.021
pNN50	-6.971 (3.187)	-2.187	0.031	[-13.310, -0.634]	-0.484 [-0.922, -0.043]	571	-2.618	0.009
HF	-1508.491 (633.623)	-2.381	0.020	[-2768.520, -248.461]	-0.526 [-0.965, -0.084]	562	-2.698	0.007
LF	-729080.978 (948128.072)	-0.769	0.444	[-2614537.630, 1156375.672]	-0.170 [-0.604, 0.265]	715	-1.331	0.183
RSA (LnHF)	-0.510 (0.225)	-2.475	0.015	[-0.958, -0.062]	-0.501 [-0.939, -0.06]	662	-1.805	0.071

Table 5.11. General linear model for repeated measures for within-subjects and between-subjects effects.

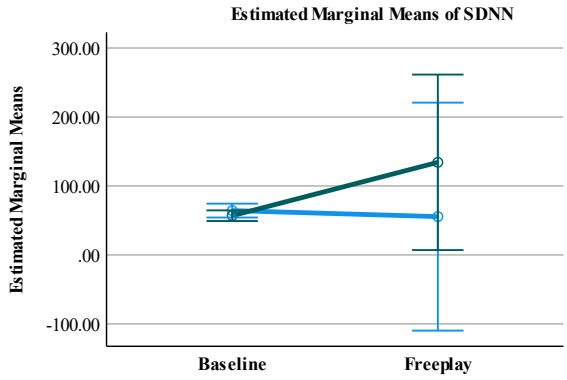
Variable	Experimental Task	HRV Metrics (Mean ± SD)		Effects (p-values)		
		TD (n = 32)	ASD (n = 54)	Time	Time-Group Interaction	Group
MHR	Baseline	89.818 ± 11.417	95.863 ± 14.787	0.023	0.144	0.183
	Free play	94.967 ± 11.329	97.012 ± 18.033			
SDNN	Baseline	64.274 ± 35.860	56.858 ± 23.266	0.511	0.410	0.503
	Free play	55.527 ± 26.435	134.259 ± 591.535			
RMSSD	Baseline	65.282 ± 60.660	43.285 ± 26.327	0.504	0.394	0.571
	Free play	50.698 ± 48.289	163.345 ± 890.677			
NN50	Baseline	116.09 ± 85.849	73.720 ± 68.143	<0.001	0.029	0.028
	Free play	78.190 ± 65.492	56.370 ± 55.178			
pNN50	Baseline	28.173 ± 22.500	17.373 ± 17.493	<0.001	0.031	0.057
	Free play	17.977 ± 17.236	14.149 ± 17.355			
HF	Baseline	2161.595 ± 3922.977	805.704 ± 998.467	0.337	0.020	0.266
	Free play	1101.171 ± 2487.044	1253.770 ± 3345.170			
LF	Baseline	1828.210 ± 2190.400	1401.486 ± 1518.665	0.444	0.444	0.445
	Free play	1551.674 ± 1381.928	730205.928 ± 5351300.596			
LnHF	Baseline	6.748 ± 1.394	6.189 ± 1.099	0.009	0.026	0.078
	Free play	6.026 ± 1.243	5.978 ± 1.330			



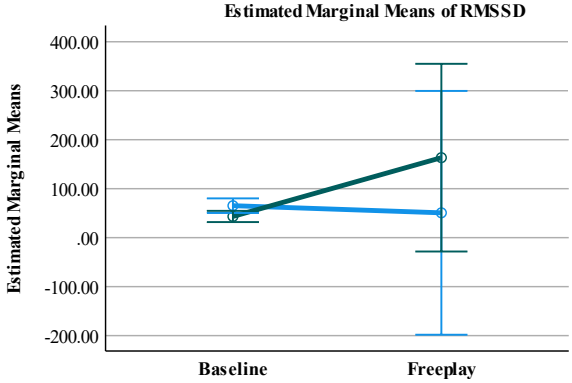
a) mean heart rate



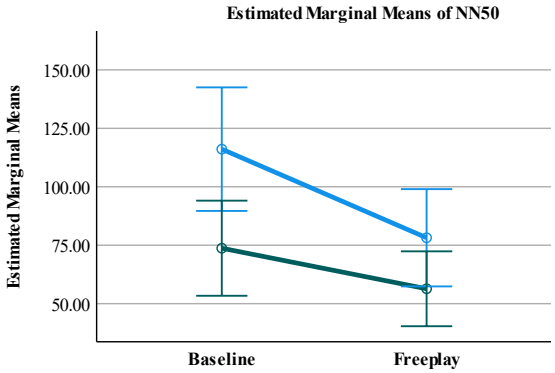
b) LnHF



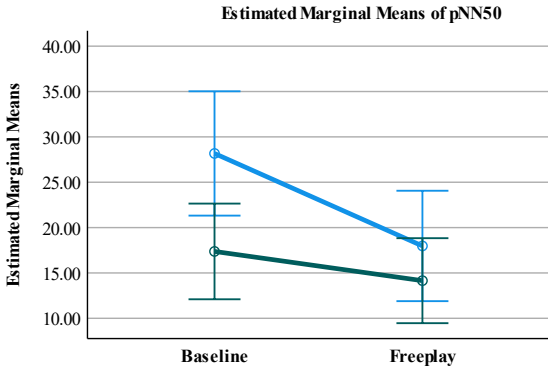
c) SDNN



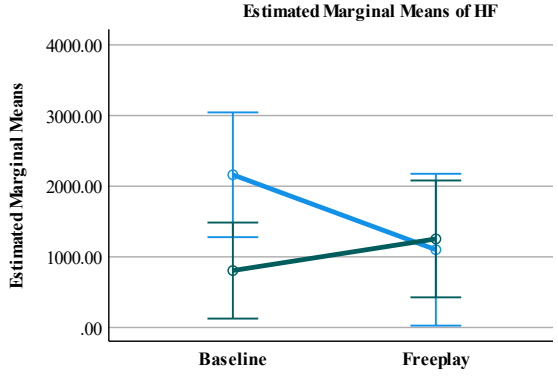
d) RMSSD



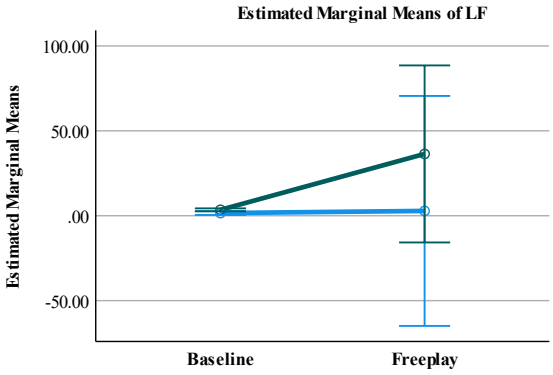
e) NN50



f) pNN50



g) HF



h) LF

Figure 5.3. Profile plots of the time-group interaction effects for the HRV metrics.

5.4 Discussion

This study first validated a customised NeuroKit2 toolbox developed based on the widely utilised commercial product "MindWare" for ECG signal processing and HRV metrics calculation. The validation was conducted in healthy adults from the publicly available QT database and our child clinical populations (with and without ASD) under resting and stress conditions. The proposed toolbox demonstrated statistical and practical equivalences in HRV metrics calculations compared to human-annotated ECG signals in both conditions. The significant changes in physiological responses over different social tasks (baseline and free play) suggest that the toolbox has a great potential for objectively processing physiological signals and deriving quantitative biomarkers for real-world clinical assessments.

The NeuroKit2 toolbox presented in this chapter is tailored to the band configuration guidelines [189] and MindWare software. These modifications were designed to achieve fully automated ECG signal cleaning and HRV analysis, reduce the labour cost of manual correction, and were not significantly different from the MindWare software. NeuroKit2 demonstrated superior accuracy in identifying the R-peak positions of most ECG signals collected during resting-state and under stress (e.g., exercises, free play) compared to the MindWare software. The MindWare software assumes a consistent IBI interval, leading to inaccurate recognition of R-peaks when there are sudden increases or decreases in heart rate, where R-peak location varies. However, discrepancies exist even though all the time-domain HRV metrics were statistically and practically equivalent in the QT database and our clinical populations. The large mean differences for the RMSSD and NN50 features observed in the QT databases may be due to the limitations of the small sample size. These differences were reduced when applying the toolbox to process our clinical datasets with a larger sample size. These differences may also stem from the fact that the ECG acquisition devices integrate digital filters before recording the data, and, therefore, no significant powerline noise was recorded in the ECG signals. The signals may have been over-filtered after the application of a 50Hz/60Hz notch filter (Figure 5.1(c) and Figure 5.1(d)). Further investigation would be needed to determine other potential causes for the observed differences in time-domain HRV measures between the two algorithms.

In addition, the statistical equivalences observed in low-frequency components further emphasise that R-peak detection using the modified NeuroKit2 toolbox is reliable and has no statistically significant difference from the MindWare software for visual inspection and correction. However, the high-frequency components (HF and LnHF) of our baseline children data showed significant upper-bound and insignificant lower-bound in the equivalence tests, indicating that the results were not statistically equivalent and, at the same time, were not statistically different. This indicates minor discrepancies between automated and human annotations during R-peak detection. Other sources of discrepancies could be the use of different window sizes and frequency transform algorithms (Welch and FFT) in handling high-frequency calculations. More specifically, the MindWare software used the entire sample signal as the window size for the FFT transformation. However, it has been suggested that applying the Welch method prior to FFT reduces the spectral leakage caused by non-stationary signals [37], which is common in natural movements. Therefore, this study applied the Welch method with Hamming windowing to the denoised ECG signal for the frequency domain analysis.

To the best of our knowledge, this is the first study to evaluate automated ECG signal processing and HRV calculations using the modified NeruoKit2 toolbox in a group of children with and without ASD who completed a social interaction assessment. Our findings suggest that time-domain HRV metrics, NN50 and pNN50, and frequency-domain HRV metrics, HF and LnHF (RSA) derived from the baseline ECG signals were significantly lower than in the TD group, which is consistent with the existing literature [190]. However, no significant differences in HRV metrics were observed between the two clinical populations in ECG signals collected in the free play task. Moreover, changes in physiological responses over the social interaction tasks suggest large changes in the mean heart rate in the TD group and RMSSD, NN50, and HF

metrics in the ASD group. There was statistical evidence of within- and between-group differences across tasks, with within-subjects time effects, time-group interaction effects, and between-subjects group effects. It highlights the potential of using HRV metrics as biomarkers in analysing social skills and social competencies in clinical assessments.

Limitations

NeuroKit2 has a limitation that if the first QRS complex in the ECG recording occurs at the very beginning, the R-peak detection algorithm may miss identifying the first R-peak due to the lack of reference information on a complete cardiac cycle (Figure 5.2(c)). In addition, NeuroKit2 may incorrectly determine the location of the R-peak when its amplitude is similar to the surrounding noise (as shown in Figure 5.2(d)). This is also a challenge in manual annotation, as the irregular, noisy signals could cover the regular pattern of a QRS complex. The advantage of applying NeuroKit2 towards ECG signal cleaning and HRV calculation in clinical practices is that, unlike MindWare software, it does not require manual correction of inverted signals. As demonstrated in Figure 5.2(b), NeuroKit2 remains effective even for inverted signals. It is commendable for its application in clinical long-term ECG recording and analysis.

However, due to the inconsistency in the number of R-peaks detected by manual annotation and automated methods, the mean difference reflected in the HRV metrics due to imperfect R-peak detection cannot be quantified using the Bland-Altman agreement [191]. Therefore, manual correction after automated labelling is recommended for a consensus check. Advanced machine learning algorithms validated across diverse ECG signals in R-peak detection could be applied after addressing model interpretability, transparency, and ethical consideration issues.

5.5 Conclusion

In summary, our modified NeuroKit2 toolbox effectively detected R-peaks in ECG signals and calculated HRV metrics for adult and child participants at rest and during exercises. The differences in HRV metrics calculation between the MindWare manual correction approach and our automated approach are not clinically significant, and significant changes between social interaction tasks were observed in this study, indicating that changes in HRV could be a physiological marker for ASD diagnosis.

5.6 Summary

This chapter presented an improved pipeline for automated ECG signal processing and HRV metrics calculation using the NeuroKit2 toolbox. The modified algorithms are similar to the MindWare HRV analysis tool used for manual cleaning and inspection of R-peak detection in social interaction studies. This chapter validated the toolbox against manual inspection when ECG signals contained motion artifacts and irregular ECG patterns and concluded statistical equivalence and clinically non-significant differences. This suggests that the toolbox can be applied to prolonged ECG recordings in social interaction to diagnose ASD. The practical significance and impact of the differences in the R-peak locations in clinical medicine should be explored in future studies.

Chapter 6: Conclusion

6.1 Summary

This chapter articulates the main outcomes of the thesis in applying digital representations to quantify atypical eye contact behaviours, facial phenotypes, and physiological biomarkers in children with ASD. It reveals the potential significance of digital phenotypes as a driving force in AI and algorithms development for assisting social competencies assessments and ASD diagnosis. The multi-view, multi-modal feature learning mechanism proposed in this thesis provided empirical explorations for assessing social developmental delays and their potential for translation into clinical practice. This chapter further discusses the challenges related to the use of AI in clinical applications and the prospects for future development of multi-modal research in ASD diagnosis.

6.2 Thesis Outcome

The focus of this thesis is to adopt the concept of multi-modal feature learning to construct novel mechanisms for parsing the natural reactions of children with ASD in the context of social interactions through digitalised features. Despite rapid development of multi-modal AI algorithms, pattern recognition and human action recognition is still in its infancy when it comes to applications to assist ASD diagnosis in social interaction settings. The systematic review and meta-analysis (Chapter 2) revealed that AI models could provide positive diagnostic accuracy for ASD diagnosis in social interaction tasks by aggregating multiple facial modalities (facial action units, gaze and posture movements, and vocal features). Specifically, atypical eye gaze and head pose movement features are considered critical information for training efficient machine learning and deep learning algorithms for classifying ASD from TD groups in human-human and human-computer agent interaction settings. Based on the existing literature, machine learning algorithms, including SVM and Random Forest, have been most widely explored in ASD research, with SVM outperforming other algorithms. In addition, posture features have been widely applied in human-robot interaction tasks to perform behavioural detection tasks (eye contact, facial emotion expression, and head movement) for therapeutic purposes. Human-robot interaction for diagnosis purposes is less explored. Furthermore, there are fewer studies using AI in human-human social interaction to assess social competencies.

There are two major challenges that exist in quantifying these biomarkers in clinical diagnosis. First, in clinical assessments, the target social behaviours are only quantified by post-session video recording. Assessment scoring schemes in the form of frequency and duration, and subjective professional experiences in diagnosis are used to determine atypical behaviours in ASD. Objective measurements of the group differences cannot be achieved without extensive and detailed video labelling. Second, research on facial feature analysis in unrestricted social interactions is limited. Current studies of eye gaze movements require the participants to limit their body movements to collect valid eye-tracking data. For studies using wearable sensors, it is difficult for the sensors to record the behavioural activities of multiple participants in a single device. Moreover, it is not conducive to a comprehensive study of the differences in interactions between multiple people. Therefore, based on the review conducted in Chapter 2, AI models trained in the still-face paradigm are deemed to be the most stable and accurate in diagnosing ASD. The use of AI for ASD diagnosis in unrestricted naturalistic interaction tasks is yet to be validated in translation into clinical diagnostic measures.

This thesis addresses these challenges from three main aspects. 1) A multi-modal social behaviour recognition (MSBR) framework is proposed to detect eye contact behaviours between child participants and assessors during free play. 2) Facial behaviour patterns extracted from child participants engaging in joint attention events with assessors during free play are further investigated to determine the effectiveness of using machine learning classifiers (as suggested by the systematic review) to distinguish groups. The statistical significance of these facial patterns is further investigated to explore the underlying mechanisms in distinguishing groups. 3) Adapted an automatic ECG signal processing and HRV metrics calculation toolbox to support the analysis of the changes in HRV across social tasks and to explore its performances in long-term data cleaning and statistical differences across social assessment tasks within and across individuals.

Inspired by the recent literature on action recognition studies [16-18, 31, 32, 192], the MSBR framework extends the application of action recognition algorithms to detect clinically relevant social behaviours in a multi-camera recording setting, for example, detecting direct eye contact episodes presented between a child participant and an assessor in free play. The assessment environment is designed with some familiarity for the participants, helping them to focus on the assessments. Therefore, the participants' caregivers are allowed to sit in the trial room but are asked to avoid interfering with the children's attention to the assessment. Chapter 3 presents the implementation details of the framework, visualises the causality of model prediction using the Grad-CAM algorithm [139], and, given the nature of imbalance in the eye contact samples, Chapter 3 utilises precision-recall curves to compare the true model performance for recognition of both eye contact and no eye contact behaviours. In general, single-view predictions using the MSBR framework could achieve over 88% accuracy, and multi-view consensus improves its performance to 93%. It is also observed that model performance is equally accurate for eye contact and no eye contact segments. The Grad-CAM algorithm reveals that the framework predicts eye contact behaviour primarily from the child participant's facial movements and attention shifting, consistent with the assessor's professional experiences.

While existing research has identified facial behaviour patterns as an essential phenotype for people to express emotions and initiate social concerns, distinctive eye gaze shifts and emotion changes during social interactions have been used as one of the novel diagnostic indicators of ASD [13, 14, 47]. However, few studies have analysed the most discriminative facial features by integrating statistical findings with machine learning applications. There is a lack of clinical evaluation schemes for evaluating these facial patterns when frontal views are occluded. Hence, Chapter 4 conducts a statistical analysis of the facial features exhibited by participants engaging in joint attention with others in free play, and with an exploration of facial features that could perform ASD and TD classification using SVM-based machine learning models. In this way, the data-centric machine learning models are verified to find potential indicators that possess statistical significance to assist ASD diagnosis. These analyses were conducted through the instantaneous movement features (displacement, velocity, and acceleration) of head pose, eye gaze, and the changes of action units. The results show that some head movement and action units features are statistically significant between the two diagnostic groups (ASD and TD). Feature selection algorithms further suggest that all action unit features are considered critical in classification. This is consistent across all head orientations (frontal, lateral, and occlusion views). With additional gaze and head movement features utilised in model training, model performance could be improved. Thus, SVM models containing facial action unit features, to

some extent, could be used to classify children with ASD in an interactive setting, which compensates for the assumption that information containing frontal facial features must be collected for classification in previous studies.

In addition to behavioural markers, physiological biomarkers are found to be associated with ASD diagnostic scores, as they reflect an individual's emotional and psychological activity. To understand psychological signal changes during naturalistic social interactions, EEG signal collection is not suitable because of the high intensity of movements during play and the discomfort caused by wearing EEG caps. However, wireless ECG signal acquisition is a mature technology widely used in various hospital environments. Therefore, Chapter 5 of this thesis focuses on analysing the changes in heart rate and HRV metrics across social interaction tasks and explores their statistical and clinical significance between diagnostic groups. A validation study was first conducted on a public ECG dataset of healthy adults at rest and under stress tests, then on our clinical participants engaged in a baseline task assessing social interaction.

Given that no statistical difference is observed between HRV metrics calculated using our modified NeuroKit2 toolbox and the hand-corrected approach by experts, the toolbox is considered reliable for analysing ECG signals collected in our clinical setting. More details of our modifications to the NeuroKit2 toolbox were discussed in Chapter 5. Moreover, there are within-subjects time effects and between-group time-clinical group interaction effects on the changes of HRV between the baseline and free play tasks, which suggests that change in HRV is a potential biomarker for ASD diagnosis.

This thesis argues that developing multi-modal, multi-view feature learning mechanisms helps understand the natural behavioural phenotypes among multiple individuals in a free play environment. The review chapter suggests that AI algorithms developed based on digitalised facial and posture features have the potential to assess social competencies and contribute to ASD diagnosis and therapy in socially interactive environments. However, due to the heterogenous experimental settings reported in current literature and the lack of exploration of the performance of developed systems in unstructured plays, there is no consensus on acceptable performance thresholds for these developed systems to be transitioned into clinical services.

The MSBR framework presented in Chapter 3 was developed to address these challenges by analysing unstructured free play tasks. It has been shown that by integrating prediction results from RBG images and optical flow images across all camera views, eye contact behaviours between multiple individuals can be recognised, resulting in a top-1 accuracy of 93.5%. This suggests that clinical assessments regarding social competence can integrate AI models to analyse video recordings automatically. Furthermore, as suggested in Chapter 4, by integrating digitalised facial features captured from multiple camera angles, these phenotypic markers can be objectively quantified to assist in diagnosing ASD, even in the absence of frontal facial views. This finding changes the paradigm and allows for a more holistic view of facial information to be analysed to support assessments.

In addition, prolonged physiological ECG signals and their derived HRV metrics reflect physiological status changes in different social environments. These signal analyses can be performed automatically with software toolboxes that are statistically and practically equivalent to domain experts. Chapter 5 also reveals that NN50 and HF metrics have the potential to differentiate between healthy controls and the ASD groups. Integrating our

improved NeuroKit2 toolbox into clinical services helps accelerate physiological biomarker analysis.

6.3 Limitations

Meta-analysis

The digitalised feature modalities and their use in machine learning and deep learning algorithms developed on human-computer agent and human-robot interactions were excluded from the meta-analysis due to the lack of a contingency table and the required sensitivity and specificity data to conduct the analysis. Therefore, the overall effects of AI in ASD diagnosis in social interaction contexts were analysed only in those human-human interaction studies that reported these data. In addition, information on demographics, such as age, IQ, gender, as well as diagnostic tools used for ASD screening or diagnosis, were extracted but not analysed as a covariate to explore the potential sources of heterogeneities observed in the meta-analysis because these details are rarely reported in the eligible studies. Furthermore, due to the limited number of studies suitable for the meta-analysis and the diverse social interaction assessment tasks reported in each study, the evidenced positive diagnostic odds ratio needs to be explained with caution.

Population and Sample Size

The sample size reported in each of the empirical studies (chapters 3-5) was small compared to the total population involved in the social interaction assessments due to the fact that social behaviours such as eye contact and joint attention were not frequently presented in our child population. In particular, because joint attention occurred so infrequently in children during free play, the multi-modal social behaviour recognition framework was only investigated for its performance in detecting eye contact events, and its usability in joint attention detection was not further investigated.

Furthermore, observable facial features captured from different camera angles vary due to unrestricted movements in free play; therefore, in Chapter 4, only machine learning algorithms were considered to explore the potential discriminative facial markers from video recordings. In addition, the number of physiological signals collected in the free play task were reduced in comparison to the baseline task due to children not being able to tolerate wearing the external ECG electrodes and their inability to focus on the assessment task at times.

Demographic Information

This thesis also did not take into account possible behavioural differences between age groups and IQs, ASD severity, as well as their ability to communicate, and comorbidity conditions (if any). Refinement in model training on those features may help to better understand and quantify social competencies and translate the refined models to clinical practices.

6.4 Future Research Directions

This thesis has investigated mechanisms to quantify social behaviours and revealed behavioural and physiological features dominant ASD classifications both statistically and using machine learning algorithms in free play. However, the validity of the multi-modal,

multi-view feature learning framework applied to assist with ASD assessments remains a key research question for future works. For instance, can the framework be applied to infant populations for early diagnosis and intervention? Is it necessary to have multiple cameras to record the social scenes, and can these systems be deployed in tele-consultation for their use in remote areas and other clinical environments? To answer those questions, future research should transfer the models trained in our setting and refine them to suit other social environments.

In addition, this thesis has not explored ways to quantify social synchrony between participants and their surrounding social partners, including both behavioural and physiological markers. This thesis has only examined features in video or image patterns that can be used for behavioural recognition and ASD classification. However, research on analysing acoustic features such as pitch, echolalia, and melodic contours, particularly in understanding verbal impairments in individuals with ASD, has not been addressed. In addition, gait analysis could help provide a detailed understanding of atypical body movement during naturalistic social interaction.

Furthermore, the heterogeneity in behavioural phenotypes among people has challenged the current ASD diagnostic procedures, as these phenotypes are difficult to objectively quantify through observations. Therefore, how well would the MSBR framework proposed in this thesis be generalisable in quantifying other social behavioural markers in social settings with unrestricted movements?

In translation AI models capable of clinical usage, future studies should develop automated video analysis toolboxes that involve semi-supervised key frames localisation algorithms with fewer requirements on data annotation, where clinicians can cross-check the reliability of the model in handling complex social scenes. The systems must also comply with ethical standards of AI in clinical practices, be generalisable for different patient demographics and clinical conditions, and have acceptable clinically proven performance. AI in medical research also needs to seek patient acceptance and trust. The underlying mechanism of AI needs to be fully understood by clinicians, and the toolboxes need to maintain data privacy and personal safety in using those tools.

6.5 Conclusion

In summary, this thesis focused on the implementation of multi-modal digital feature learning strategies into the study of differences in social behaviour and physiological changes associated with ASD diagnosis, providing a new research mechanism to quantify and analyse discriminant features associated with ASD diagnosis objectively. At the same time, multi-view fixed recording reduces the chance of missing valuable information and compensates for previous studies suggesting that frontal-view features are assumed. This thesis contributed to the literature by providing discriminatory criteria for accurate diagnosis of ASD, reducing labour costs for highly skilled specialists to label behaviours manually and analyse recorded data.

References

[1] J. Zeidan *et al.*, "Global prevalence of autism: A systematic review update," *Autism Res,* vol. 15, no. 5, pp. 778-790, May 2022, doi: 10.1002/aur.2696.

[2] P. Mundy, M. Sigman, J. Ungerer, and T. Sherman, "Defining the social deficits of autism: the contribution of non-verbal communication measures," *J Child Psychol Psychiatry,* vol. 27, no. 5, pp. 657-69, Sep 1986, doi: 10.1111/j.1469-7610.1986.tb00190.x.

[3] G. L. Simmons, S. Ioannou, J. V. Smith, B. A. Corbett, M. D. Lerner, and S. W. White, "Utility of an Observational Social Skill Assessment as a Measure of Social Cognition in Autism," *Autism Res,* vol. 14, no. 4, pp. 709-719, Apr 2021, doi: 10.1002/aur.2404.

[4] A. Mannion and G. Leader, "Attention-deficit/hyperactivity disorder (AD/HD) in autism spectrum disorder," *Research in Autism Spectrum Disorders,* vol. 8, no. 4, pp. 432-439, 2014.

[5] V. Khachadourian *et al.*, "Comorbidities in autism spectrum disorder and their etiologies," *Translational Psychiatry,* vol. 13, no. 1, p. 71, 2023.

[6] K. A. Boulton, A. J. Guastella, M. A. Hodge, E. A. Demetriou, N. Ong, and N. Silove, "Mental health concerns in children with neurodevelopmental conditions attending a developmental assessment service," *J Affect Disord,* vol. 335, pp. 264-272, Aug 15 2023, doi: 10.1016/j.jad.2023.04.098.

[7] K. A. Boulton, M. A. Hodge, A. Jewell, N. Ong, N. Silove, and A. J. Guastella, "Diagnostic delay in children with neurodevelopmental conditions attending a publicly funded developmental assessment service: findings from the Sydney Child Neurodevelopment Research Registry," *BMJ Open,* vol. 13, no. 2, p. e069500, Feb 1 2023, doi: 10.1136/bmjopen-2022-069500.

[8] E. Moricke, J. K. Buitelaar, and N. N. J. Rommelse, "Do We Need Multiple Informants When Assessing Autistic Traits? The Degree of Report Bias on Offspring, Self, and Spouse Ratings," *J Autism Dev Disord,* vol. 46, no. 1, pp. 164-175, Jan 2016, doi: 10.1007/s10803-015-2562-y.

[9] A. Masi, A. Lampit, N. Glozier, I. B. Hickie, and A. J. Guastella, "Predictors of placebo response in pharmacological and dietary supplement treatment trials in pediatric autism spectrum disorder: a meta-analysis," *Transl Psychiatry,* vol. 5, no. 9, p. e640, Sep 22 2015, doi: 10.1038/tp.2015.143.

[10] N. K. Kaufman, "Rethinking “gold standards” and “best practices” in the assessment of autism," *Applied Neuropsychology: Child,* vol. 11, no. 3, pp. 529-540, 2022.

[11] S. S. Joudar *et al.*, "Artificial intelligence-based approaches for improving the diagnosis, triage, and prioritization of autism spectrum disorder: a systematic review of current trends and open issues," *Artificial Intelligence Review,* pp. 1-65, 2023.

[12] R. L. Koegel and L. K. Koegel, *Pivotal response treatments for autism: Communication, social, & academic development* (Pivotal response treatments for autism: Communication, social, & academic development.). Baltimore, MD, US: Paul H. Brookes Publishing Co., 2006, pp. xiv, 296-xiv, 296.

[13] R. A. J. de Belen, T. Bednarz, A. Sowmya, and D. Del Favero, "Computer vision in autism spectrum disorder research: a systematic review of published studies from 2009 to 2019," *Translational Psychiatry,* vol. 10, no. 1, p. 333, 2020, doi: 10.1038/s41398-020-01015-w.

[14] M. E. Minissi, I. A. Chicchi Giglioli, F. Mantovani, and M. Alcaniz Raya, "Assessment of the Autism Spectrum Disorder Based on Machine Learning and Social Visual Attention: A Systematic Review," *J Autism Dev Disord,* vol. 52, no. 5, pp. 2187-2202, May 2022, doi: 10.1007/s10803-021-05106-5.

[15] A. Landowska *et al.*, "Automatic Emotion Recognition in Children with Autism: A Systematic Literature Review," *Sensors (Basel),* vol. 22, no. 4, Feb 20 2022, doi: 10.3390/s22041649.

[16] L. Wang, W. Hu, and T. Tan, "Recent developments in human motion analysis," *Pattern Recognit.,* vol. 36, no. 3, pp. 585-601, 2003, doi: 10.1016/S0031-3203(02)00100-0.

[17] D. Wu, N. Sharma, and M. Blumenstein, "Recent advances in video-based human action recognition using deep learning: A review," in *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017: IEEE, pp. 2865-2872, doi: 10.1109/IJCNN.2017.7966210.

[18] Y. Kong and Y. Fu, "Human Action Recognition and Prediction: A Survey," *International Journal of Computer Vision,* vol. 130, no. 5, pp. 1366-1401, 2022, doi: 10.1007/s11263-022-01594-9.

[19] L. Liu *et al.*, "Deep Learning for Generic Object Detection: A Survey," *International Journal of Computer Vision,* vol. 128, no. 2, pp. 261-318, 2019, doi: 10.1007/s11263-019-01247-4.

[20] E. Hjelmås and B. K. Low, "Face detection: A survey," *Computer vision and image understanding,* vol. 83, no. 3, pp. 236-274, 2001, doi: 10.1006/cviu.2001.0921.

[21] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018: IEEE, pp. 59-66, doi: 10.1109/FG.2018.00019.

[22] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 532-539.

[23] M. Cai, M. Li, Z. Xiong, P. Zhao, E. Li, and J. Tang, "An Advanced Deep Learning Framework for Video-Based Diagnosis of ASD," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, (Lecture Notes in Computer Science, 2022, ch. Chapter 42, pp. 434-444.

[24] H. Drimalla, I. Baskow, B. Behnia, S. Roepke, and I. Dziobek, "Imitation and recognition of facial emotions in autism: a computer vision approach," *Molecular Autism,* vol. 12, no. 1, p. 27, 2021, doi: 10.1186/s13229-021-00430-0.

[25] N. V. Valtakari *et al.*, "A field test of computer-vision-based gaze estimation in psychology," *Behav Res Methods,* Apr 26 2023, doi: 10.3758/s13428-023-02125-1.

[26] Z. Zhao *et al.*, "Atypical Head Movement during Face-to-Face Interaction in Children with Autism Spectrum Disorder," *Autism Res,* vol. 14, no. 6, pp. 1197-1208, Jun 2021, doi: 10.1002/aur.2478.

[27] G. Dawson *et al.*, "Atypical postural control can be detected via computer vision analysis in toddlers with autism spectrum disorder," *Scientific reports,* vol. 8, no. 1, p. 17008, 2018, doi: 10.1038/s41598-018-35215-8.

[28] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science,* vol. 349, no. 6245, pp. 255-260, 2015, doi: 10.1126/science.aaa8415.

[29] Q. Wei, H. Cao, Y. Shi, X. Xu, and T. Li, "Machine learning based on eye-tracking data to identify Autism Spectrum Disorder: A systematic review and meta-analysis," (in English), *Journal of Biomedical Informatics,* vol. 137, p. 104254, 2023, doi: 10.1016/j.jbi.2022.104254.

[30] K. K. Hyde *et al.*, "Applications of Supervised Machine Learning in Autism Spectrum Disorder Research: a Review," *Review Journal of Autism and Developmental Disorders,* vol. 6, no. 2, pp. 128-146, 2019, doi: 10.1007/s40489-019-00158-x.

[31] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," presented at the NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems, December, 2014. [Online]. Available: https://dl.acm.org/doi/10.5555/2968826.2968890.

[32] L. Wang *et al.*, "Temporal segment networks for action recognition in videos," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 41, no. 11, pp. 2740-2755, 2018, doi: 10.1109/TPAMI.2018.2868668.

[33] M. Kohli, A. K. Kar, V. G. Prakash, and A. P. Prathosh, "Deep Learning-Based Human Action Recognition Framework to Assess Children on the Risk of Autism or Developmental Delays," in *Neural Information Processing*, (Communications in Computer and Information Science, 2023, ch. Chapter 38, pp. 459-470.

[34] M. Khodatars *et al.*, "Deep learning for neuroimaging-based diagnosis and rehabilitation of autism spectrum disorder: a review," *Computers in Biology and Medicine,* vol. 139, p. 104949, 2021, doi: 10.1016/j.compbiomed.2021.104949.

[35] W. Bosl, A. Tierney, H. Tager-Flusberg, and C. Nelson, "EEG complexity as a biomarker for autism spectrum disorder risk," *BMC medicine,* vol. 9, no. 1, pp. 1-16, 2011, doi: 10.1186/1741-7015-9-18.

[36] F. H. Duffy and H. Als, "A stable pattern of EEG spectral coherence distinguishes children with autism from neuro-typical controls - a large case control study," *BMC Medicine,* vol. 10, p. 18, 2012, doi: 10.1186/1741-7015-10-64.

[37] M. V. Kamath, M. Watanabe, and A. Upton, "Heart rate variability (HRV) signal analysis: clinical applications," 2012.

[38] I. Bujnakova *et al.*, "Autism spectrum disorder is associated with autonomic underarousal," *Physiological research,* vol. 65, 2016.

[39] X. Ming, R. Patel, V. Kang, S. Chokroverty, and P. O. Julu, "Respiratory and autonomic dysfunction in children with autism spectrum disorders," *Brain Dev,* vol. 38, no. 2, pp. 225-32, Feb 2016, doi: 10.1016/j.braindev.2015.07.003.

[40] C. Lory, G. Kadlaskar, R. McNally Keehn, A. L. Francis, and B. Keehn, "Brief report: Reduced heart rate variability in children with autism spectrum disorder," *Journal of autism and developmental disorders,* vol. 50, pp. 4183-4190, 2020, doi: 10.1007/s10803-020-04458-8.

[41] Y. C. Cheng, Y. C. Huang, and W. L. Huang, "Heart rate variability in individuals with autism spectrum disorders: A meta-analysis," *NEUROSCIENCE AND BIOBEHAVIORAL REVIEWS,* vol. 118, pp. 463-471, NOV 2020, doi: 10.1016/j.neubiorev.2020.08.007.

[42] S. Shahrestani, E. M. Stewart, D. S. Quintana, I. B. Hickie, and A. J. Guastella, "Heart rate variability during adolescent and adult social interactions: A meta-analysis," *Biological Psychology,* vol. 105, pp. 43-50, FEB 2015, doi: 10.1016/j.biopsycho.2014.12.012.

[43] I. Smeekens, R. Didden, and E. W. M. Verhoeven, "Exploring the Relationship of Autonomic and Endocrine Activity with Social Functioning in Adults with Autism Spectrum Disorders," *JOURNAL OF AUTISM AND DEVELOPMENTAL DISORDERS,* vol. 45, no. 2, pp. 495-505, FEB 2015, doi: 10.1007/s10803-013-1947-z.

[44] M. J. Maenner *et al.*, "Prevalence and characteristics of autism spectrum disorder among children aged 8 years—Autism and Developmental Disabilities Monitoring Network, 11 sites, United States, 2020," *MMWR Surveillance Summaries,* vol. 72, no. 2, p. 1, 2023.

[45] K. Y. C. Lai, E. C. W. Yuen, S. F. Hung, and P. W. L. Leung, "Autism Diagnostic Interview-Revised Within DSM-5 Framework: Test of Reliability and Validity in Chinese Children," (in English), *Journal of Autism and Developmental Disorders,* vol. 52, no. 4, pp. 1807-1820, 2022, doi: 10.1007/s10803-021-05079-5.

[46] N. E. Rosen, C. Lord, and F. R. Volkmar, "The diagnosis of autism: From Kanner to DSM-III to DSM-5 and beyond," *Journal of autism and developmental disorders,* vol. 51, pp. 4253-4270, 2021, doi: 10.1007/s10803-021-04904-1.

[47] P. Washington and D. P. Wall, "A Review of and Roadmap for Data Science and Machine Learning for the Neuropsychiatric Phenotype of Autism," *Annu Rev Biomed Data Sci,* May 3 2023, doi: 10.1146/annurev-biodatasci-020722-125454.

[48] S. J. Moon, J. Hwang, R. Kana, J. Torous, and J. W. Kim, "Accuracy of Machine Learning Algorithms for the Diagnosis of Autism Spectrum Disorder: Systematic Review and Meta-Analysis of Brain Magnetic Resonance Imaging Studies," *JMIR Ment Health,* vol. 6, no. 12, p. e14108, Dec 20 2019, doi: 10.2196/14108.

[49] A. Esteva *et al.*, "A guide to deep learning in healthcare," *Nature medicine,* vol. 25, no. 1, pp. 24-29, 2019, doi: 10.1038/s41591-018-0316-z.

[50] M. Terra, M. Baklola, S. Ali, and K. El-Bastawisy, "Opportunities, applications, challenges and ethical implications of artificial intelligence in psychiatry: a narrative review," (in English),

Egyptian Journal of Neurology, Psychiatry and Neurosurgery, vol. 59, no. 1, p. 80, 2023, doi: 10.1186/s41983-023-00681-z.

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

[52] X. E. Ma, H. X. Gu, and J. J. Zhao, "Atypical gaze patterns to facial feature areas in autism spectrum disorders reveal age and culture effects: A meta-analysis of eye-tracking studies," (in English), *AUTISM RESEARCH,* vol. 14, no. 12, pp. 2625-2639, DEC 2021, doi: 10.1002/aur.2607.

[53] N. R. Zurcher, "Novel Insights Into Facial Emotion Encoding in Autism Spectrum Disorder Through Deep Learning," (in English), *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging,* vol. 7, no. 7, pp. 631-632, 2022, doi: 10.1016/j.bpsc.2022.05.003.

[54] I. Rafique, K. Fatima, A. Dastagir, S. Mahmood, and M. Hussain, "Autism Identification and Learning Through Motor Gesture Patterns," 2019: Institute of Electrical and Electronics Engineers Inc., doi: 10.1109/ICIC48496.2019.8966740.

[55] H. Abbas, F. Garberson, E. Glover, and D. P. Wall, "Machine learning approach for early detection of autism by combining questionnaire and home video screening," *Journal of the American Medical Informatics Association,* vol. 25, no. 8, pp. 1000-1007, 2018, doi: 10.1093/jamia/ocy039.

[56] J. F. Cohen *et al.*, "Preferred reporting items for journal and conference abstracts of systematic reviews and meta-analyses of diagnostic test accuracy studies (PRISMA-DTA for Abstracts): checklist, explanation, and elaboration," *BMJ,* vol. 372, p. n265, Mar 15 2021, doi: 10.1136/bmj.n265.

[57] J. P. Salameh *et al.*, "Preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): explanation, elaboration, and checklist," *BMJ,* vol. 370, p. m2632, Aug 14 2020, doi: 10.1136/bmj.m2632.

[58] P. Manning-Courtney *et al.*, "Autism spectrum disorders," *Current problems in pediatric and adolescent health care,* vol. 43, no. 1, pp. 2-11, 2013, doi: 10.1016/j.cppeds.2012.08.001.

[59] P. F. Whiting *et al.*, "QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies," *Annals of internal medicine,* vol. 155, no. 8, pp. 529-536, 2011, doi: 10.7326/0003-4819-155-8-201110180-00009.

[60] J. J. Deeks, P. M. Bossuyt, M. M. Leeflang, and Y. Takwoingi, *Cochrane handbook for systematic reviews of diagnostic test accuracy*. John Wiley & Sons, 2022.

[61] C. Tang *et al.*, "Automatic identification of high-risk autism spectrum disorder: A feasibility study using video and audio data under the still-face paradigm," *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* vol. 28, no. 11, pp. 2401-2410, 2020, doi: 10.1109/TNSRE.2020.3027756.

[62] H. Drimalla *et al. Detecting autism by analyzing a simulated social interaction*, vol. 11051 LNAI, pp. 193-208, 2019.

[63] H. Drimalla *et al.*, "Towards the automatic detection of social biomarkers in autism spectrum disorder: introducing the simulated interaction task (SIT)," *npj Digital Medicine,* vol. 3, no. 1, p. 25, 2020, doi: 10.1038/s41746-020-0227-5.

[64] K. S. Lohan, E. Sheppard, G. Little, and G. Rajendran, "Toward Improved Child-Robot Interaction by Understanding Eye Movements," *Ieee Transactions on Cognitive and Developmental Systems,* vol. 10, no. 4, pp. 983-992, 2018, doi: 10.1109/tcds.2018.2838342.

[65] D. Alie, M. H. Mahoor, W. I. Mattson, D. R. Anderson, and D. S. Messinger, "Analysis of eye gaze pattern of infants at risk of autism spectrum disorder using Markov models," *2011 IEEE Workshop on Applications of Computer Vision (WACV),* pp. 282-287, 2011, doi: 10.1109/WACV.2011.5711515.

[66] N. Qiu *et al.*, "Application of the Still-Face Paradigm in Early Screening for High-Risk Autism Spectrum Disorder in Infants and Toddlers," *Frontiers in Pediatrics,* vol. 8, no. (Qiu, Zhai, Weng,

Li, Xiao, Fu, Xiao, Fang, Ke) Nanjing, Brain Hospital Affiliated to Nanjing Medical University, Nanjing, China(Tang) School of Biological Science & Medical Engineering, Southeast University, Nanjing, China(Huang) College of Telecommunica, p. 290, 2020, doi: 10.3389/fped.2020.00290.

[67] J. Zhang *et al.*, "RJAfinder: An automated tool for quantification of responding to joint attention behaviors in autism spectrum disorder using eye tracking data," *Frontiers in Neuroscience,* vol. 16, 2022, doi: 10.3389/fnins.2022.915464.

[68] M. Varma *et al.*, "Identification of Social Engagement Indicators Associated With Autism Spectrum Disorder Using a Game-Based Mobile App: Comparative Study of Gaze Fixation and Visual Scanning Methods," *Journal of Medical Internet Research,* vol. 24, no. 2, p. e31830, 2022, doi: 10.2196/31830.

[69] M. Alcaniz *et al.*, "Eye gaze as a biomarker in the recognition of autism spectrum disorder using virtual reality and machine learning: A proof of concept for diagnosis," *Autism research : official journal of the International Society for Autism Research,* vol. 15, no. 1, pp. 131-145, 2022, doi: 10.1002/aur.2636.

[70] I. Akin-Bulbul, I. Kok, and S. Ozdemir, "Development of a visual attention based decision support system for autism spectrum disorder screening," *INTERNATIONAL JOURNAL OF PSYCHOPHYSIOLOGY,* vol. 173, pp. 69-81, 2022, doi: 10.1016/j.ijpsycho.2022.01.004.

[71] A. Adham *et al.*, "Stratification of Children with Autism Spectrum Disorder through fusion of temporal information in eye-gaze scan-paths," *ACM Trans. Knowl. Discov. Data,* 2022, doi: 10.1145/3539226.

[72] Z. Zhao, H. Tang, X. Zhang, X. Qu, X. Hu, and J. Lu, "Classification of Children With Autism and Typical Development Using Eye-Tracking Data From Face-to-Face Conversations: Machine Learning Model Development and Performance Evaluation," *Journal of Medical Internet Research,* vol. 23, no. 8, p. e29328, 2021, doi: 10.2196/29328.

[73] P. R. Krishnappababu *et al.*, "Exploring Complexity of Facial Dynamics in Autism Spectrum Disorder," *IEEE Transactions on Affective Computing,* 2021, doi: 10.1109/TAFFC.2021.3113876.

[74] N. Kojovic, S. Natraj, S. P. Mohanty, T. Maillart, and M. Schaer, "Using 2D video-based pose estimation for automated prediction of autism spectrum disorders in young children," *Scientific reports,* vol. 11, no. 1, p. 15069, 2021, doi: 10.1038/s41598-021-94378-z.

[75] N. Zhang, M. Ruan, S. Wang, L. Paul, and X. Li, "Discriminative Few Shot Learning of Facial Dynamics in Interview Videos for Autism Trait Classification," *IEEE Transactions on Affective Computing,* pp. 1-1, 2022, doi: 10.1109/TAFFC.2022.3178946.

[76] J. T. Megerian *et al.*, "Evaluation of an artificial intelligence-based medical device for diagnosis of autism spectrum disorder," *npj Digital Medicine,* vol. 5, no. 1, p. 57, 2022, doi: 10.1038/s41746-022-00598-6.

[77] A. Di Nuovo, D. Conti, G. Trubia, S. Buono, and S. Di Nuovo, "Deep learning systems for estimating visual attention in robot-assisted therapy of children with autism and intellectual disability," *Robotics,* vol. 7, no. 2, 2018, doi: 10.3390/robotics7020025.

[78] C. Eunji *et al.*, "Detecting Gaze Towards Eyes in Natural Social Interactions and Its Use in Child Assessment," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.,* vol. 1, no. 3, p. Article 43, 2017, doi: 10.1145/3131902.

[79] E. Chong *et al.*, "Visual 3D Tracking of Child-Adult Social Interactions," *2017 the Seventh Joint Ieee International Conference on Development and Learning and Epigenetic Robotics,* pp. 399-406, 2017.

[80] G. Nie *et al.*, "Predicting Response to Joint Attention Performance in Human-Human Interaction Based on Human-Robot Interaction for Young Children with Autism Spectrum Disorder," in *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 27-31 Aug. 2018 2018, pp. 1-4, doi: 10.1109/ROMAN.2018.8525634.

[81] H. Javed, W. Lee, and C. H. Park, "Toward an Automated Measure of Social Engagement for Children With Autism Spectrum Disorder-A Personalized Computational Modeling Approach," *Frontiers in Robotics and Ai,* vol. 7, 2020, doi: 10.3389/frobt.2020.00043.

[82] E. Chong *et al.*, "Detection of eye contact with deep neural networks is as accurate as human experts," *Nature Communications,* vol. 11, no. 1, p. 6386, 2020, doi: 10.1038/s41467-020-19712-x.

[83] W. B. Liu, T. Y. Zhou, C. H. Zhang, X. B. Zou, M. Li, and Ieee, "Response to Name: A Dataset and A Multimodal Machine Learning Framework towards Autism Study," *2017 Seventh International Conference on Affective Computing and Intelligent Interaction,* pp. 178-183, 2017.

[84] G. Zhang, K. Kangsoo, B. Anjana, and B. Roghayeh, "An Automated Mutual Gaze Detection Framework for Social Behavior Assessment in Therapy for Children with Autism," Montréal, QC, Canada, 2021.

[85] C. Wu *et al.*, "Machine learning based autism spectrum disorder detection from videos," 2021, doi: 10.1109/HEALTHCOM49281.2021.9398924.

[86] J. Li *et al.*, "A two-stage multi-modal affect analysis framework for children with autism spectrum disorder," vol. 2897, pp. 7-14, 2021.

[87] D. Anagnostopoulou, N. Efthymiou, C. Papailiou, P. Maragos, and Ieee, "Engagement Estimation During Child Robot Interaction Using Deep Convolutional Networks Focusing on ASD Children," 2021.

[88] C. A. Molloy, D. S. Murray, R. Akers, T. Mitchell, and P. Manning-Courtney, "Use of the Autism Diagnostic Observation Schedule (ADOS) in a clinical setting," *Autism,* vol. 15, no. 2, pp. 143-162, 2011, doi: 10.1177/1362361310379241.

[89] Q. Guillon, N. Hadjikhani, S. Baduel, and B. Rogé, "Visual social attention in autism spectrum disorder: Insights from eye tracking studies," *Neuroscience & Biobehavioral Reviews,* vol. 42, pp. 279-297, 2014, doi: 10.1016/j.neubiorev.2014.03.013.

[90] S. Ramdoss, W. Machalicek, M. Rispoli, A. Mulloy, R. Lang, and M. O'Reilly, "Computer-based interventions to improve social and emotional skills in individuals with autism spectrum disorders: A systematic review," *Developmental neurorehabilitation,* vol. 15, no. 2, pp. 119-135, 2012, doi: 10.3109/17518423.2011.651655.

[91] R. Adolphs, L. Sears, and J. Piven, "Abnormal processing of social information from faces in autism," in *Autism*: Routledge, 2013, pp. 126-134.

[92] Z. A. P. Del Prette and A. Del Prette, *Social Competence and Social Skills: A Theoretical and Practical Guide*. Cham: Springer, 2021.

[93] M. Tomasello, M. Carpenter, J. Call, T. Behne, and H. Moll, "Understanding and sharing intentions: The origins of cultural cognition," *Behavioral and Brain Sciences,* vol. 28, no. 5, pp. 675-691, 2005, doi: 10.1017/S0140525X05000129.

[94] S. H. Spence, "Social skills training with children and young people: Theory, evidence and practice," *Child and Adolescent Mental Health,* vol. 8, no. 2, pp. 84-96, 2003, doi: 10.1111/1475-3588.00051.

[95] E. A. J. Croes, M. L. Antheunis, A. P. Schouten, and E. J. Krahmer, "The role of eye-contact in the development of romantic attraction: Studying interactive uncertainty reduction strategies during speed-dating," *Computers in Human Behavior,* vol. 105, p. 106218, 2020/04/01/ 2020, doi: 10.1016/j.chb.2019.106218.

[96] C. L. Kleinke, "Gaze and eye contact: a research review," *Psychological bulletin,* vol. 100, no. 1, p. 78, 1986, doi: 10.1037/0033-2909.100.1.78.

[97] A. P. Bayliss and S. P. Tipper, "Predictive gaze cues and personality judgments: Should eye trust you?," *Psychological Science,* vol. 17, no. 6, pp. 514-520, 2006, doi: 10.1111/j.1467-9280.2006.01737.x.

[98] A. Senju *et al.*, "Early social experience affects the development of eye gaze processing," *Current Biology,* vol. 25, no. 23, pp. 3086-3091, 2015, doi: 10.1016/j.cub.2015.10.019.

[99] A. Senju and M. H. Johnson, "The eye contact effect: mechanisms and development," *Trends in Cognitive Sciences,* vol. 13, no. 3, pp. 127-134, 2009, doi: 10.1016/j.tics.2008.11.009.
[100] R. Kuboshita, T. X. Fujisawa, K. Makita, R. Kasaba, H. Okazawa, and A. Tomoda, "Intrinsic brain activity associated with eye gaze during mother-child interaction," *scientific reports,* vol. 10, no. 1, p. 18903, Nov 3 2020, doi: 10.1038/s41598-020-76044-y.
[101] J. W. Tanaka and A. Sung, "The "Eye Avoidance" Hypothesis of Autism Face Processing," *Journal of Autism and Developmental Disorders,* vol. 46, no. 5, pp. 1538-1552, 2016, doi: 10.1007/s10803-013-1976-7.
[102] L. Milne, P. Greenway, A. Guedeney, and B. Larroque, "Long term developmental impact of social withdrawal in infants," *Infant Behavior and Development,* vol. 32, no. 2, pp. 159-166, 2009, doi: 10.1016/j.infbeh.2008.12.006.
[103] J. K. Buitelaar, "Attachment and Social Withdrawal in Autism: Hypotheses and Findings," (in English), *Behaviour,* vol. 132, no. 5-6, pp. 319-350, 01 Jan. 1995 1995, doi: 10.1163/156853995X00595.
[104] A. Guedeney, L. Marchand-Martin, S. J. Cote, and B. Larroque, "Perinatal risk factors and social withdrawal behaviour," *European child & adolescent psychiatry,* vol. 21, no. 4, pp. 185-191, 2012, doi: 10.1007/s00787-012-0250-4.
[105] F. R. Schneier, T. L. Rodebaugh, C. Blanco, H. Lewin, and M. R. Liebowitz, "Fear and avoidance of eye contact in social anxiety disorder," *Comprehensive Psychiatry,* vol. 52, no. 1, pp. 81-87, 2011/01/01/ 2011, doi: 10.1016/j.comppsych.2010.04.006.
[106] P. Waxer, "Nonverbal cues for depression," *Journal of Abnormal Psychology,* vol. 83, pp. 319-322, 1974, doi: 10.1037/h0036706.
[107] A. Senju and M. H. Johnson, "Atypical eye contact in autism: models, mechanisms and development," *Neuroscience & Biobehavioral Reviews,* vol. 33, no. 8, pp. 1204-14, Sep 2009, doi: 10.1016/j.neubiorev.2009.06.001.
[108] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders (5th ed.)* (DSM Library). 2013.
[109] H. Hops, "Children's social competence and skill: Current research practices and future directions," *Behavior Therapy,* vol. 14, no. 1, pp. 3-18, 1983, doi: 10.1016/S0005-7894(83)80084-7.
[110] J. Santer and C. Griffiths, *Free play in early childhood: A literature review*. National Children's Bureau, 2007.
[111] J. Hewes, "Seeking balance in motion: The role of spontaneous free play in promoting social and emotional health in early childhood care and education," *Children,* vol. 1, no. 3, pp. 280-301, 2014, doi: 10.3390/children1030280.
[112] D. Whitebread, "Free play and children's mental health," *The Lancet Child & Adolescent Health,* vol. 1, no. 3, pp. 167-169, 2017/11/01/ 2017, doi: 10.1016/S2352-4642(17)30092-5.
[113] S. D. Holloway and M. Reichhart-Erickson, "The relationship of day care quality to children's free-play behavior and social problem-solving skills," *Early Childhood Research Quarterly,* vol. 3, no. 1, pp. 39-53, 1988/03/01/ 1988, doi: 10.1016/0885-2006(88)90028-2.
[114] W. Greve and T. Thomsen, "Evolutionary advantages of free play during childhood," *Evolutionary Psychology,* vol. 14, no. 4, p. 1474704916675349, 2016, doi: 10.1177/1474704916675349.
[115] A. Anderson, D. W. Moore, R. Godfrey, and C. M. Fletcher-Flinn, "Social skills assessment of children with autism in free-play situations," *Autism,* vol. 8, no. 4, pp. 369-385, 2004, doi: 10.1177/1362361304045216.
[116] C. Jongerius, R. S. Hessels, J. A. Romijn, E. M. A. Smets, and M. A. Hillen, "The Measurement of Eye Contact in Human Interactions: A Scoping Review," *Journal of Nonverbal Behavior,* vol. 44, no. 3, pp. 363-389, 2020, doi: 10.1007/s10919-020-00333-3.

[117] J. White, J. Hegarty, and N. Beasley, "EYE CONTACT AND OBSERVER BIAS: A RESEARCH NOTE," *British Journal of Psychology,* vol. 61, no. 2, p. 271, 1970, doi: 10.1111/j.2044-8295.1970.tb01244.x.

[118] D. J. Knight, D. Langmeyer, and D. C. Lundgren, "Eye-Contact, Distance, and Affiliation: The Role of Observer Bias," *Sociometry,* vol. 36, no. 3, pp. 390-401, 1973, doi: 10.2307/2786340.

[119] N. Caruana, G. McArthur, A. Woolgar, and J. Brock, "Simulating social interactions for the experimental investigation of joint attention," *Neuroscience & Biobehavioral Reviews,* vol. 74, pp. 115-125, 2017/03/01/ 2017, doi: 10.1016/j.neubiorev.2016.12.022.

[120] U. J. Pfeiffer, K. Vogeley, and L. Schilbach, "From gaze cueing to dual eye-tracking: Novel approaches to investigate the neural correlates of gaze in social interaction," *Neuroscience & Biobehavioral Reviews,* vol. 37, no. 10, pp. 2516-2528, 2013, doi: 10.1016/j.neubiorev.2013.07.017.

[121] Q. Wang, L. Lu, Q. Zhang, F. Fang, X. B. Zou, and L. Yi, "Eye avoidance in young children with autism spectrum disorder is modulated by emotional facial expressions," *Journal of Abnormal Psychology,* vol. 127, no. 7, pp. 722-732, Oct 2018, doi: 10.1037/abn0000372.

[122] M. J. Wieser, P. Pauli, G. W. Alpers, and A. Mühlberger, "Is eye to eye contact really threatening and avoided in social anxiety?—An eye-tracking and psychophysiology study," *Journal of Anxiety Disorders,* vol. 23, no. 1, pp. 93-103, 2009/01/01/ 2009, doi: 10.1016/j.janxdis.2008.04.004.

[123] D. J. Ricks and M. B. Colton, "Trends and considerations in robot-assisted autism therapy," presented at the *IEEE International Conference on Robotics and Automation*, May, 2010.

[124] T. Bastos *et al.*, "Development of a Socially Assistive Robot Controlled by Emotions Based on Heartbeats and Facial Temperature of Children with Autistic Spectrum Disorder," 31 October 2020 2021, vol. 1290, no. 1290: Adv. Intell. Sys. Comput., pp. 228-239, doi: 10.1007/978-3-030-63092-8_15.

[125] Z. Ye *et al.*, "Detecting eye contact using wearable eye-tracking glasses," in *Ubicomp '12: The 2012 ACM Conference on Ubiquitous Computing*, September 2012: Association for Computing Machinery, pp. 699-704, doi: 10.1145/2370216.2370368.

[126] B. Noris, J.-B. Keller, and A. Billard, "A wearable gaze tracking system for children in unconstrained environments," *Computer Vision and Image Understanding,* vol. 115, no. 4, pp. 476-486, 2011, doi: 10.1016/j.cviu.2010.11.013.

[127] P. Müller, M. X. Huang, X. Zhang, and A. Bulling, "Robust eye contact detection in natural multi-person interactions using gaze and speaking behaviour," in *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, 2018: Association for Computing Machinery, pp. 1-10, doi: 10.1145/3204493.3204549.

[128] C. Lord, M. Rutter, P. DiLavore, S. Risi, K. Gotham, and S. Bishop, "Autism diagnostic observation schedule–2nd edition (ADOS-2)," *Western Psychological Corporation,* vol. 284, 2012.

[129] S. Tomar, "Converting video formats with FFmpeg," *Linux Journal,* vol. 2006, no. 146, p. 10, 2006. [Online]. Available: https://dl.acm.org/doi/fullHtml/10.5555/1134782.1134792.

[130] Y. Zhu and S. Newsam, "DenseNet for dense flow," in *IEEE International Conference on Image Processing (ICIP)*, Sept. 2017 2017, IEEE: IEEE, 2017, pp. 790-794, doi: 10.1109/ICIP.2017.8296389.

[131] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, "Performance of optical flow techniques," *International Journal of Computer Vision,* vol. 12, no. 1, pp. 43-77, 1994, doi: 10.1007/BF01420984.

[132] S. Sun, Z. Kuang, L. Sheng, W. Ouyang, and W. Zhang, "Optical flow guided feature: A fast and robust motion representation for video action recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 1390-1399, doi: 10.1109/CVPR.2018.00151.

[133] K. Lyu, H. Chen, Z. Liu, B. Zhang, and R. Wang, "3D human motion prediction: A survey," *Neurocomputing,* vol. 489, pp. 345-365, 2022, doi: 10.1016/j.neucom.2022.02.045.
[134] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *Computer Vision – ECCV 2016*, October 2016: Springer, pp. 20-36, doi: 10.1007/978-3-319-46484-8_2.
[135] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 27-30 June 2016: IEEE, pp. 770-778, doi: 10.1109/CVPR.2016.90.
[136] C.-F. R. Chen *et al.*, "Deep analysis of cnn-based spatio-temporal representations for action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 6165-6175, doi: 10.1109/CVPR46437.2021.00610.
[137] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv.org,* 2017, doi: 10.48550/arXiv.1412.6980.
[138] I. Loshchilov and F. Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts," *ArXiv.org,* 2016, doi: 10.48550/arXiv.1608.03983.
[139] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 618-626, doi: 10.1109/ICCV.2017.74.
[140] J. M. Franchak, K. S. Kretch, and K. E. Adolph, "See and be seen: Infant-caregiver social looking during locomotor free play," *Developmental Science,* vol. 21, no. 4, p. e12626, Jul 2018, doi: 10.1111/desc.12626.
[141] S. L. Rogers, C. P. Speelman, O. Guidetti, and M. Longmuir, "Using dual eye tracking to uncover personal gaze patterns during social interaction," *Sci. Rep.,* vol. 8, no. 1, pp. 1-9, 2018, doi: 10.1038/s41598-018-22726-7.
[142] P. Mundy and M. Sigman, "Joint attention, social competence, and developmental psychopathology," *Developmental psychopathology: Volume one: Theory and method,* pp. 293-332, 2015, doi: 10.1002/9780470939383.ch9.
[143] P. Mundy and M. Crowson, "Joint attention and early social communication: Implications for research on intervention with autism," *Journal of Autism and Developmental disorders,* vol. 27, pp. 653-676, 1997.
[144] P. Mundy and L. Newell, "Attention, joint attention, and social cognition," *Current directions in psychological science,* vol. 16, no. 5, pp. 269-274, 2007.
[145] P. Mundy, "A review of joint attention and social-cognitive brain systems in typical development and autism spectrum disorder," *Eur. J. Neurosci.,* vol. 47, no. 6, pp. 497-514, 2018, doi: 10.1111/ejn.13720.
[146] B. Ingersoll and L. Schreibman, "Teaching reciprocal imitation skills to young children with autism using a naturalistic behavioral approach: effects on language, pretend play, and joint attention," *J Autism Dev Disord,* vol. 36, no. 4, pp. 487-505, May 2006, doi: 10.1007/s10803-006-0089-y.
[147] T. Charman, S. Baron-Cohen, J. Swettenham, G. Baird, A. Cox, and A. Drew, "Testing joint attention, imitation, and play as infancy precursors to language and theory of mind," *Cognitive development,* vol. 15, no. 4, pp. 481-498, 2000.
[148] K. Bottema-Beutel, "Associations between joint attention and language in autism spectrum disorder and typical development: A systematic review and meta-regression analysis," *Autism Res,* vol. 9, no. 10, pp. 1021-1035, Oct 2016, doi: 10.1002/aur.1624.
[149] P. Mundy, "Joint attention and social-emotional approach behavior in children with autism," *Development and Psychopathology,* vol. 7, no. 1, pp. 63-82, 2009, doi: 10.1017/s0954579400006349.
[150] A. Estes *et al.*, "Behavioral, cognitive, and adaptive development in infants with autism spectrum disorder in the first 2 years of life," *J Neurodev Disord,* vol. 7, no. 1, p. 24, 2015, doi: 10.1186/s11689-015-9117-6.

[151] Y. Bruinsma, R. L. Koegel, and L. K. Koegel, "Joint attention and children with autism: a review of the literature," *Ment Retard Dev Disabil Res Rev,* vol. 10, no. 3, pp. 169-75, 2004, doi: 10.1002/mrdd.20036.

[152] C. Kasari, S. Freeman, and T. Paparella, "Joint attention and symbolic play in young children with autism: a randomized controlled intervention study," *J Child Psychol Psychiatry,* vol. 47, no. 6, pp. 611-20, Jun 2006, doi: 10.1111/j.1469-7610.2005.01567.x.

[153] T. Charman, "Why is joint attention a pivotal skill in autism?," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences,* vol. 358, no. 1430, pp. 315-324, 2003, doi: 10.1098/rstb.2002.1199.

[154] R. J. Landa, K. C. Holman, and E. Garrett-Mayer, "Social and communication development in toddlers with early and later diagnosis of autism spectrum disorders," *Archives of general psychiatry,* vol. 64, no. 7, pp. 853-864, 2007, doi: 10.1001/archpsyc.64.7.853.

[155] J. Green *et al.*, "Parent-mediated communication-focused treatment in children with autism (PACT): a randomised controlled trial," *Lancet,* vol. 375, no. 9732, pp. 2152-60, Jun 19 2010, doi: 10.1016/S0140-6736(10)60587-9.

[156] N. Stuart, A. Whitehouse, R. Palermo, E. Bothe, and N. Badcock, "Eye Gaze in Autism Spectrum Disorder: A Review of Neural Evidence for the Eye Avoidance Hypothesis," *J Autism Dev Disord,* vol. 53, no. 5, pp. 1884-1905, May 2023, doi: 10.1007/s10803-022-05443-z.

[157] M. L. Spezio, R. Adolphs, R. S. Hurley, and J. Piven, "Abnormal use of facial information in high-functioning autism," *J Autism Dev Disord,* vol. 37, no. 5, pp. 929-39, May 2007, doi: 10.1007/s10803-006-0232-9.

[158] M. B. Harms, A. Martin, and G. L. Wallace, "Facial emotion recognition in autism spectrum disorders: a review of behavioral and neuroimaging studies," *Neuropsychol Rev,* vol. 20, no. 3, pp. 290-322, Sep 2010, doi: 10.1007/s11065-010-9138-6.

[159] L. Nummenmaa and A. J. Calder, "Neural mechanisms of social attention," *Trends Cogn Sci,* vol. 13, no. 3, pp. 135-43, Mar 2009, doi: 10.1016/j.tics.2008.12.006.

[160] S. M. Anzalone *et al.*, "Quantifying patterns of joint attention during human-robot interactions: An application for autism spectrum disorder assessment," *Pattern Recognition Letters,* vol. 118, pp. 42-50, 2019, doi: 10.1016/j.patrec.2018.03.007.

[161] P. Nyström, E. Thorup, S. Bölte, and T. Falck-Ytter, "Joint Attention in Infancy and the Emergence of Autism," *Biological Psychiatry,* vol. 86, no. 8, pp. 631-638, 2019/10/15/ 2019, doi: 10.1016/j.biopsych.2019.05.006.

[162] S. M. Clifford and C. Dissanayake, "The early development of joint attention in infants with autistic disorder using home video observations and parental interview," *J Autism Dev Disord,* vol. 38, no. 5, pp. 791-805, May 2008, doi: 10.1007/s10803-007-0444-7.

[163] I. C. Stallworthy *et al.*, "Variability in Responding to Joint Attention Cues in the First Year is Associated With Autism Outcome," *J Am Acad Child Adolesc Psychiatry,* vol. 61, no. 3, pp. 413-422, Mar 2022, doi: 10.1016/j.jaac.2021.03.023.

[164] C. Ko, J. H. Lim, J. Hong, S. B. Hong, and Y. R. Park, "Development and Validation of a Joint Attention-Based Deep Learning System for Detection and Symptom Severity Assessment of Autism Spectrum Disorder," *JAMA Netw Open,* vol. 6, no. 5, p. e2315174, May 1 2023, doi: 10.1001/jamanetworkopen.2023.15174.

[165] T. W. Frazier *et al.*, "Validation of proposed DSM-5 criteria for autism spectrum disorder," *J. Am. Acad. Child Adolesc. Psychiatry,* vol. 51, no. 1, pp. 28-40.e3, 2012, doi: 10.1016/j.jaac.2011.09.021.

[166] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," ed: Taipei, Taiwan, 2003.

[167] E. Poljac, E. Poljac, and J. Wagemans, "Reduced accuracy and sensitivity in the perception of emotional facial expressions in individuals with high autism spectrum traits," *Autism,* vol. 17, no. 6, pp. 668-680, 2013.

[168] M. Malik *et al.*, "Heart rate variability: Standards of measurement, physiological interpretation, and clinical use," *European heart journal,* vol. 17, no. 3, pp. 354-381, 1996, doi: 10.1093/oxfordjournals.eurheartj.a014868.

[169] M. Malik, "Heart rate variability: Standards of measurement, physiological interpretation, and clinical use: Task force of the European Society of Cardiology and the North American Society for Pacing and Electrophysiology," *Annals of Noninvasive Electrocardiology,* vol. 1, no. 2, pp. 151-181, 1996.

[170] G. Perna, A. Riva, A. Defillo, E. Sangiorgio, M. Nobile, and D. Caldirola, "Heart rate variability: Can it serve as a marker of mental health resilience?: Special Section on "Translational and Neuroscience Studies in Affective Disorders" Section Editor, Maria Nobile MD, PhD," *Journal of Affective Disorders,* vol. 263, pp. 754-761, 2020, doi: 10.1016/j.jad.2019.10.017.

[171] M. A. Serhani, H. T. El Kassabi, H. Ismail, and A. Nujum Navaz, "ECG monitoring systems: Review, architecture, processes, and key challenges," *Sensors,* vol. 20, no. 6, p. 1796, 2020, doi: 10.3390/s20061796.

[172] A. Schäfer and J. Vagedes, "How accurate is pulse rate variability as an estimate of heart rate variability?: A review on studies comparing photoplethysmographic technology with an electrocardiogram," *International journal of cardiology,* vol. 166, no. 1, pp. 15-29, 2013, doi: 10.1016/j.ijcard.2012.03.119.

[173] O. Kwon *et al.*, "Electrocardiogram sampling frequency range acceptable for heart rate variability analysis," *Healthcare informatics research,* vol. 24, no. 3, pp. 198-206, 2018, doi: 10.4258/hir.2018.24.3.198.

[174] F. A. Oliveira Junior *et al.*, "Different acquisition systems for heart rate variability analysis may lead to diverse outcomes," *Braz J Med Biol Res,* vol. 55, p. e11720, 2022, doi: 10.1590/1414-431X2021e11720.

[175] G. R. Sandercock, P. D. Bromley, and D. A. Brodie, "The reliability of short-term measurements of heart rate variability," *International journal of cardiology,* vol. 103, no. 3, pp. 238-247, 2005, doi: 10.1016/j.ijcard.2004.09.013.

[176] G. R. H. Sandercock, P. D. Bromley, and D. A. Brodie, "The reliability of short-term measurements of heart rate variability," *International Journal of Cardiology,* vol. 103, no. 3, pp. 238-247, 2005/09/01/ 2005, doi: 10.1016/j.ijcard.2004.09.013.

[177] S. Shahrestani, E. M. Stewart, D. S. Quintana, I. B. Hickie, and A. J. Guastella, "Heart rate variability during social interactions in children with and without psychopathology: a meta-analysis," *Journal of Child Psychology and Psychiatry,* vol. 55, no. 9, pp. 981-989, SEPT 2014, doi: 10.1111/jcpp.12226.

[178] M. S. Milivojević, A. Gavrovska, and I. S. Reljin, "Python Based Physiological Signal Processing for Vital Signs Monitoring," 2017.

[179] D. Makowski *et al.*, "NeuroKit2: A Python toolbox for neurophysiological signal processing," *Behavior Research Methods,* pp. 1-8, 2021, doi: 10.3758/s13428-020-01516-y.

[180] P. Laguna, R. G. Mark, A. Goldberg, and G. B. Moody, "A database for evaluation of algorithms for measurement of QT and other waveform intervals in the ECG," in *Computers in cardiology*, 1997: IEEE, pp. 673-676, doi: 10.1109/CIC.1997.648140.

[181] D. Lakens, "Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses," *Soc Psychol Personal Sci,* vol. 8, no. 4, pp. 355-362, May 2017, doi: 10.1177/1948550617697177.

[182] D. Lakens, A. M. Scheel, and P. M. Isager, "Equivalence Testing for Psychological Research: A Tutorial," *Advances in Methods and Practices in Psychological Science,* vol. 1, no. 2, pp. 259-269, 2018/06/01 2018, doi: 10.1177/2515245918770963.

[183] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," *circulation,* vol. 101, no. 23, pp. e215-e220, 2000.

[184] G. B. Moody and R. G. Mark, "The MIT-BIH arrhythmia database on CD-ROM and software for use with it," in *[1990] Proceedings Computers in Cardiology*, 1990: IEEE, pp. 185-188.

[185] R. Thapa *et al.*, "Heart Rate Variability in Children With Autism Spectrum Disorder and Associations With Medication and Symptom Severity," *Autism Res,* vol. 14, no. 1, pp. 75-85, Jan 2021, doi: 10.1002/aur.2437.

[186] B. Singh, P. Singh, and S. Budhiraja, "Various Approaches to Minimise Noises in ECG Signal: A Survey," in *2015 Fifth International Conference on Advanced Computing & Communication Technologies*, 21-22 Feb. 2015 2015, pp. 131-137, doi: 10.1109/ACCT.2015.87.

[187] R. Thapa *et al.*, "Reduced heart rate variability in adults with autism spectrum disorder," *Autism Res,* vol. 12, no. 6, pp. 922-930, Jun 2019, doi: 10.1002/aur.2104.

[188] G. G. Berntson, K. S. Quigley, J. F. Jang, and S. T. Boysen, "An approach to artifact identification: Application to heart period data," *Psychophysiology,* vol. 27, no. 5, pp. 586-598, 1990, doi: 10.1111/j.1469-8986.1990.tb01982.x.

[189] D. S. Quintana, G. A. Alvares, and J. A. Heathers, "Guidelines for Reporting Articles on Psychiatry and Heart rate variability (GRAPH): recommendations to advance research communication," *Transl Psychiatry,* vol. 6, p. e803, May 10 2016, doi: 10.1038/tp.2016.73.

[190] C. Lory, G. Kadlaskar, R. M. Keehn, A. L. Francis, and B. Keehn, "Brief Report: Reduced Heart Rate Variability in Children with Autism Spectrum Disorder," *JOURNAL OF AUTISM AND DEVELOPMENTAL DISORDERS,* vol. 50, no. 11, pp. 4183-4190, NOV 2020, doi: 10.1007/s10803-020-04458-8.

[191] D. Giavarina, "Understanding Bland Altman analysis," (in eng), *Biochem Med (Zagreb),* vol. 25, no. 2, pp. 141-151, 2015, doi: 10.11613/BM.2015.015.

[192] J. K. Aggarwal and Q. Cai, "Human motion analysis: A review," *Comput Vis Image Underst,* vol. 73, no. 3, pp. 428-440, 1999, doi: 10.1006/cviu.1998.0744.

Appendices

Supplementary Materials For Chapter 2

Supplementary Table 1. Literature search term.

Database	Final Search Terms
Embase	1. exp autism/ 2. exp autism diagnostic observation schedule/ 3. exp Autism Diagnostic Interview Revised/ 4. exp autism assessment/ 5. exp Asperger syndrome/ 6. autis*.tw. 7. asperger*.tw. 8. ASD*.tw. 9. ADOS.tw. 10. 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 11. exp machine learning/ 12. exp artificial intelligence/ 13. exp algorithm/ 14. exp artificial neural network/ 15. exp automation/ 16. exp deep learning/ 17. exp computer vision/ 18. exp image processing/ 19. artificial intelligence.tw. 20. algorithm*.tw. 21. artificial neural network*.tw. 22. machine learning.tw. 23. deep neural network*.tw. 24. deep learning.tw. 25. exp computer vision/ 26. automatic* detection*.tw. 27. automation.tw. 28. image process*.tw. 29. pattern recognition.tw. 30. 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28 or 29 31. exp gaze/ 32. exp eye fixation/ 33. exp eye movement/ 34. exp attention/ 35. exp eye tracking/ 36. eye gaz*.tw. 37. eye trac*.tw. 38. eye move*.tw. 39. eye fixation*.tw. 40. attention.tw. 41. eye contact.tw. 42. 31 or 32 or 33 or 34 or 35 or 36 or 37 or 38 or 39 or 40 or 41 43. exp emotion/ 44. exp facial expression/ 45. exp face/

	46. exp facies/ 47. fac*.tw. 48. emotion*.tw. 49. 43 or 44 or 45 or 46 or 47 or 48 50. exp social disability/ 51. exp developmental disorder/ 52. exp communication disorder/ 53. exp social behavior/ 54. socia*.tw. 55. communication.tw. 56. development* disorder.tw. 57. exp play/ 58. exp play therapy/ 59. play*.tw. 60. interact*.tw. 61. 50 or 51 or 52 or 53 or 54 or 55 or 56 or 57 or 58 or 59 or 60 62. 42 or 49 63. 10 and 30 and 61 and 62
Medline	1. exp Autism Spectrum Disorder/ 2. exp Asperger syndrome/ 3. exp Autistic Disorder/ 4. Asperger*.tw. 5. autis*.tw. 6. ASD*.tw. 7. ADOS.tw. 8. 1 or 2 or 3 or 4 or 5 or 6 or 7 9. exp Imaging, Three-Dimensional/ 10. exp Algorithms/ 11. exp Pattern Recognition, Automated/ 12. exp Artificial Intelligence/ 13. exp Image Enhancement/ 14. exp Image Processing, Computer-Assisted/ 15. exp Image Interpretation, Computer-Assisted/ 16. exp Deep Learning/ 17. exp Diagnosis, Computer-Assisted/ 18. exp Artificial Intelligence/ 19. exp Machine Learning/ 20. exp Computational Biology/ 21. exp Neural Networks, Computer/ 22. automatic* detection*.tw. 23. deep learning.tw. 24. machine learning*.tw. 25. artificial neural network*.tw. 26. deep neural network*.tw. 27. artificial intelligence.tw. 28. algorithm*.tw. 29. pattern recognition.tw. 30. image process*.tw. 31. automation.tw. 32. 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 33. exp eye movements/

	34. exp fixation, ocular/ 35. exp Attention/ 36. eye fixation*.tw. 37. eye gaz*.tw. 38. eye trac*.tw. 39. eye move*.tw. 40. attention.tw. 41. eye contact.tw. 42. 33 or 34 or 35 or 36 or 37 or 38 or 39 or 40 or 41 43. exp Facial Expression/ 44. exp facies/ 45. exp Emotions/ 46. exp Facial Recognition/ 47. fac*.tw. 48. emotion*.tw. 49. 43 or 44 or 45 or 46 or 47 or 48 50. exp Developmental Disabilities/ 51. exp Social Behavior/ 52. communication.tw. 53. socia*.tw. 54. 50 or 51 or 52 or 53 55. development* disorder.tw. 56. exp Play Therapy/ 57. play*.tw. 58. exp Social Interaction/ 59. interact*.tw. 60. 50 or 51 or 52 or 53 or 54 or 55 or 56 or 57 or 58 or 59 61. 42 or 49 62. 8 and 32 and 60 and 61
Scopus	((TITLE-ABS-KEY (autis*)) OR (TITLE-ABS-KEY (asd*)) OR (TITLE-ABS-KEY (asperge*)) OR (TITLE-ABS-KEY (ados))) AND ((TITLE-ABS-KEY (fac*)) OR (TITLE-ABS-KEY (emotion*)) OR (TITLE-ABS-KEY ("eye move*")) OR (TITLE-ABS-KEY (attention)) OR (TITLE-ABS-KEY ("eye trac*")) OR (TITLE-ABS-KEY ("eye gaz*")) OR (TITLE-ABS-KEY (eye AND fixation*)) OR (TITLE-ABS-KEY (eye AND contact))) AND ((TITLE-ABS-KEY ("socia*")) OR (TITLE-ABS-KEY ("social behavio*")) OR (TITLE-ABS-KEY ("social deplay*")) OR (TITLE-ABS-KEY ("social respons*")) OR (TITLE-ABS-KEY (communication)) OR (TITLE-ABS-KEY ("developmental disabil*")) OR (TITLE-ABS-KEY (play*)) OR (TITLE-ABS-KEY (interact*))) AND ((TITLE-ABS-KEY (algorithm*)) OR (TITLE-ABS-KEY ("Pattern Recognition")) OR (TITLE-ABS-KEY ("machine learning")) OR (TITLE-ABS-KEY ("artificial intelligence")) OR (TITLE-ABS-KEY ("artificial neural network")) OR (TITLE-ABS-KEY ("deep learning")) OR (TITLE-ABS-KEY ("deep neural network")) OR (TITLE-ABS-KEY ("automat* detect*")) OR (TITLE-ABS-KEY (auto*)) OR (TITLE-ABS-KEY ("image process*"))) AND (LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "cp"))
Web of Science Core Collection	(ALL=(autis* OR asperger* OR ASD OR ADOS))AND (ALL=(fac* OR "eye gaze*" OR "eye contact" OR "eye trac*" OR "eye move*" OR "eye fixation*" OR emotion*)) AND

	(ALL=("deep learning" OR "machine learning" OR algorithm* OR auto* OR "artific ial intelligence" OR "artificial neural network" OR "deep neural network" OR image process* OR pattern recognition)) AND (ALL=("socia*" OR communication OR "development* disorder" OR play* OR inte ract*)) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years
ACM	AllField:(autis* OR asd* OR asperge* OR ADOS) AND AllField:(fac* OR emotion* OR "eye move*" OR "eye trac*" OR "eye gaz*" OR "eye fixation*" OR attention) AND AllField:(algorithm* OR "pattern recognition" OR "machine learning" OR "artificial intelligence" OR "artificial neural network" OR "deep learning" OR "deep neural network" OR "automat* detect*" OR "image process*" OR automation) AND AllField:("developmental disa*" OR communication OR "social delay*" OR "social respons*" OR "social behavio*" OR "socia*" OR play* OR interact*)
IEEE	(("All Metadata":autis* OR "All Metadata": asperge* OR "All Metadata": ASD OR "All Metadata": ADOS) AND ("All Metadata": fac* OR "All Metadata": emotion OR "All Metadata": "eye tracking" OR "All Metadata": "eye gaz*" OR "All Metadata": "eye movement" OR "All Metadata": attention OR "All Metadata": "eye fixation") AND ("All Metadata": "developmental disorder" OR "All Metadata": "socia*" OR "All Metadata": communication OR "All Metadata": play* OR "All Metadata": interact*) AND ("All Metadata": algorithm OR "All Metadata": "image processing" OR "All Metadata": "machine learning" OR "All Metadata": "artificial intelligence" OR "All Metadata": "pattern recognition" OR "All Metadata": "automatic detection" OR "All Metadata": automation OR "All Metadata": "deep learning" OR "All Metadata": "deep neural network" OR "All Metadata": "artificial neural network"))

Supplementary Table 2. Summary of the demographic characteristics of the included studies.

Authors	Autism Diagnostic Tool	Type of Assessment	Autism Severity Level	Age [range] (Mean (SD))	Male Participants	IQ Selected Tool, IQ Composite Score (Mean (SD))
A. Di Nuovo et al., 2018 [77]	ADI-R (CARS-2, PEP-3, VABS)	Diagnostic	ASD (grade 3), ID levels (4 mild to severe, 2 profound)	-	-	Leiter-R, WISC,
C. Eunji et al., 2018 [78]	ADOS, ADI-R	Diagnostic	-	[20 - 90 months]	-	-
C. Tang et al., 2020 [61]	ADI-R, ADOS, DSM-5	Diagnostic	high-risk	[8-24 months] (ASD: 19.65 (3.81) months, TD: 16.4 (4.7) months)	ASD: 36, TD: 32	full-scale intelligence quotient
E. Chong et al., 2017 [79]	-	-	-	(ASD: 43.8 months , TD: 30.8 months)	ASD: 4, TD: 5	-
G. Nie et al., 2018 [80]	DSM-IV-TR (ESCS test)	Diagnostic	-	(ASD: 2.66 (0.54) years)	-	-
H. Drimalla et al., 2019 [62]	ADOS and ADI-R (AQ Questionnaire)	Diagnostic	-	-	-	-
H. Drimalla et al., 2020 [63]	ADOS and ADI-R	Diagnostic	-	-	-	Verbal Intelligence (Wortschatz test)
H. Javed et al., 2020 [81]	-	-	-	(ASD: 8.2 (1.10) years, TD: 7.07 (2.56) years)	ASD: 5, TD: 8	-

(Continued)

Supplementary Table 2. *(Continued)*

Authors	Autism Diagnostic Tool	Type of Assessment	Autism Severity Level	Age [range] (Mean (SD))	Male Participants	IQ Selected Tool, IQ Composite Score (Mean (SD))
K. S. Lohan et al., 2018 [64]	ADOS	Diagnostic	-	(ASD: 128.63 (21.23) months, TD: 126.38 (13.61) months)	ASD: 27, TD: 29	FSIQ-4 (the Weschler abbreviated scale of intelligence), (90.54 (12.21))
E. Chong et al., 2020 [82]	ADOS-2 (CSBS-DP, M-CHAT)	Diagnostic	-	[18-60 months]	ASD: 55, TD: 36	-
D. Alie et al., 2011 [65]	ADOS, ADI-R, DSM-IV-TR criteria	Diagnostic	-	[6 months]	-	-
N. Qiu et al., 2020 [66]	M-CHAT, ADI-R and ADOS (2 years of age) (CSBS-DP, CARS and ABC)	Diagnostic	high-risk	[8-23 months] (ASD: 19.71 (3.43) months, TD: 16.4 (4.7) months)	ASD: 40, TD: 32	-
W. B. Liu et al., 2017 [83]	ADOS	Diagnostic	-	(ASD: 2.33 (2.5) years)	ASD: 16, TD: 19	-
N. Zhang et al., 2022 [75]	ADOS-2 Module 4 (CSS)	Diagnostic	-	[16-37 years] (ASD: 24.05 (5.14) years)	ASD: 26	Full-scale IQ (96.83 (13.48))
J. Zhang et al., 2022 [67]	ADI-R (ADOS)	Diagnostic	-	[4-7 years Child, and 10-32 years adult] (ASD: 5.4 (0.74) years, TD: 5.5(0.74) years)	ASD: 126, TD: 91 children	-
M. Varma et al., 2022 [68]	-	-	-	[2-15 years children]	ASD: 53, TD: 18	-
J. T. Megerian et al., 2022 [76]	clinicians	Diagnostic	-	[3.33 years (SD=1.15)]	-	-

(Continued)

Supplementary Table 2. *(Continued)*

Authors	Autism Diagnostic Tool	Type of Assessment	Autism Severity Level	Age [range] (Mean (SD))	Male Participants	IQ Selected Tool, IQ Composite Score (Mean (SD))
I. Akin-Bulbul et al., 2022 [70]	DSM-5 (BAY-LEY-III, M-CHAT-R/F)	Diagnostic	-	[26-36 months] (ASD: 34.85 months, TD: 32.90 months)	-	-
M. Alcaniz et al., 2022 [69]	ADOS-2 module 1 (ADI-R intervein)	Diagnostic	-	[4-7 years] (ASD: 5.26 (0.51) years, TD: 4.75 (0.77) years)	-	-
A. Adham et al., 2022 [71]	-	-	-	-	ASD: 44, TD: 15	-
Z. Zhao et al., 2021 [72]	DSM-4	Diagnostic	-	6-13 years	-	-
G. Zhang et al., 2021 [84]	SCQ	Screening	-	-	ASD: 7	-
C. Wu et al., 2021 [85]	-	-	-	6-36 months	-	-
J. Li et al., 2021 [86]	ADOS-2	Diagnostic	-	5-13 years	-	-
P. R. Krishnappababu et al., 2021 [73]	ADOS-T (M-CHAT-R/F)	Diagnostic	-	17-36 months	-	-
N. Kojovic et al., 2021 [74]	DSM-5, ADOS-G, ADOS-2	Diagnostic	-	-	-	-
D. Anagnostopoulou et al., 2021 [87]	-	-	-	(ASD: 10.6 years, TD: 8.6 years)	ASD: 5	-

Supplementary Table 3. Summary of the experimental tasks of the included studies.

Study	Type of Experiment	Experiment Duration Description
A. Di Nuovo et al., 2018 [77]	Therapy	14 encounters over one month, 3 sessions per week (6-8 min per child/session)
C. Eunji et al., 2017 [78]	Assessment +Therapy	Dataset 1: v-BOSCC (12-minute examiner-child interaction (2x5-minutes play segments with standardized sets of toys, and a 2-minute conversation segment)); Dataset2: ESCS (15-25 minute structured assessment uses standardized toys) and v-BOSCC (2x4-minutes play and 2-minute snake); Dataset 3: Rapid-ABC assessment (R-ABC), 2-to-4 minute play (5 activities); Dataset 4: Marcus therapy [1) pairing, 12 sessions; 2)Demands, 5 sessions; 3) Mands, 6 sessions], 10-minute each session, except two sessions (3 and 9 minutes)
C. Tang et al., 2020 [61]	Assessment	the mother amused baby without touch of body (2 minutes) + 1-minute still-face episode.
E. Chong et al., 2017 [79]	Assessment	ESCS (15-25 minutes)
H. Drimalla et al., 2019 [62]	Screening	Simulated Interaction Task (7minutes)
H. Drimalla et al., 2020 [63]	Screening	Simulated Interaction Task (7minutes)
H. Javed et al., 2020 [81]	Intervention	8-10 minutes per robot x2
K. S. Lohan et al., 2018 [64]	Intervention	-
E. Chong et al., 2020 [82]	Assessment	ESCS (15-25 minutes), BOSCC (10-12 minutes)
D. Alie et al., 2011 [65]	Intervention	5 minutes
N. Qiu et al., 2020 [66]	Screening	-
J. Zhang et al., 2022 [67]	Assessment	12 clips (5-7 second/clip)

(Continued)

Supplementary Table 3. *(Continued)*

Study	Type of Experiment	Experiment Duration Description
M. Varma et al., 2022 [68]	Assessment	90 seconds
M. Alcaniz et al., 2022 [69]	Assessment	24 minutes and 45 seconds
I. Akin-Bulbul et al., 2022 [70]	Screening	48 seconds for the stimulus (16x3), 2 second fixation before each stimulus
Z. Zhao et al., 2021 [72]	Screening	-
P. R. Krishppababu et al., 2021 [73]	well-child visit (could be both screening/assessment)	about 5 minutes
N. Kojovic et al., 2021 [74]	Assessment	1 hour
A. Adham et al., 2022 [71]	Assessment	-
W. B. Liu et al., 2017 [83]	Assessment	-
G. Nie et al., 2018 [80]	Assessment + Intervention	Least-to-Most-Based Robot-mediated Intervention (LTM-RI) (unknow session duration) and with/without ESCS (15-25 minutes)
D. Anagnostopoulou et al., 2021 [87]	Therapy	20 minutes
J. Li et al., 2021 [86]	Therapy	-
C. Wu et al., 2021 [85]	Screening	3 minutes
G. Zhang. et al., 2021 [84]	Therapy	15 minutes
N. Zhang et al., 2022 [75]	Assessment	1 hour
J. T. Megerian et al., 2022 [76]	Screening	5 minutes

Supplementary Table 4. Summary of the model training performance of the included studies.

Study	Input Feature Type	Algorithms	Performance
A. Di Nuovo et al., 2018 [77]	Face Detection	Naïve Classification	Acc:0.734, Precision: 0.706, Sensitivity: 0.678, Specificity: 0.776, NPV: 0.696, AUC: 0.738, F1: 0.692
		K-NN	Acc:0.754, Precision:0.729, Sensitivity:0.651, Specificity:0.812, NPV: 0.706, AUC: 0.793, F1:0.688
		Naïve Classification	Acc: 0.696, Precision: 0.663, Sensitivity: 0.759, Specificity: 0.642, NPV: 0.583, AUC: 0.868, F1: 0.708
		K-NN	Acc: 0.846, Precision: 0.745, Sensitivity: 0.858, Specificity: 0.767, NPV: 0.782, AUC: 0.931, F1: 0.798
			Acc: 0.709, Precision: 0.619, Sensitivity: 0.953, Specificity: 0.483, NPV: 0.902, AUC: 0.619, F1: 0.751
			Acc: 0.882, Precision: 0.833, Sensitivity: 0.830, Specificity: 0.873, NPV: 0.818, AUC: 0.913, F1: 0.831
C. Eunji et al., 2017 [78]	Eye Contact Detection	PiCNN	Precision: 0.750, Sensitivity: 0.800, AUC: 0.790, F1: 0.780, Matthews correlation coefficient: 0.770
		AlexNet	Precision: 0.710, Sensitivity: 0.770, AUC: 0.750, F1: 0.730, Matthews correlation coefficient: 0.720
		PEEC	Acc: 0.670, Precision: 0.590, AUC: 0.570, F1: 0.630, Matthews correlation coefficient: 0.620

(Continued)

Supplementary Table 4. *(Continued)*

Study	Input Feature Type	Algorithms	Performance
C. Tang et al., 2020 [61]	Head Movement Features	SVM	Acc: 0.590, Precision: 0.546, Sensitivity: 0.900, Specificity: 0.302, AUC: 0.577, F1:0.679
			Acc: 0.892, Precision: 0.919, Sensitivity: 0.850, Specificity: 0.930, AUC: 0.892, F1: 0.883
			Acc: 0.880, Precision: 0.857, Sensitivity: 0.900, Specificity: 0.861, AUC: 0.848, F1: 0.878
			Acc: 0.868, Precision: 0.837, Sensitivity: 0.900, Specificity: 0.837, AUC: 0.831, F1: 0.867
			Acc: 0.904, Precision: 0.944, Sensitivity: 0.850, Specificity: 0.954, AUC: 0.898, F1: 0.895
			Acc: 0.964, Precision: 0.974, Sensitivity: 0.950, Specificity: 0.977, AUC: 0.946, F1: 0.962
			Acc: 0.964, Precision: 0.974, Sensitivity: 0.950, Specificity: 0.977, AUC: 0.946, F1: 0.962
		K-NN	Acc: 0.542, Precision: 0.531, Sensitivity: 0.425, Specificity: 0.651, AUC: 0.541, F1: 0.472
			Acc: 0.807, Precision: 0.929, Sensitivity: 0.650, Specificity: 0.953, AUC: 0.779, F1: 0.765
			Acc: 0.868, Precision: 0.837, Sensitivity: 0.900, Specificity: 0.837, AUC: 0.852, F1: 0.867
			Acc: 0.880, Precision: 0.857, Sensitivity: 0.900, Specificity: 0.861, AUC: 0.858, F1: 0.878
			Acc: 0.807, Precision: 0.929, Sensitivity: 0.650, Specificity: 0.954, AUC: 0.779, F1: 0.765
			Acc: 0.940, Precision: 0.949, Sensitivity: 0.925, Specificity: 0.954, AUC: 0.925, F1:0.9367
			Acc: 0.928, Precision: 0.947, Sensitivity: 0.900, Specificity: 0.954, AUC: 0.914, F1: 0.923
E. Chong et al., 2017 [79]	Gaze Shift for Initiating Joint Attention	SVM	AUC: 0.780
			AUC: 0.800

(Continued)

Supplementary Table 4. *(Continued)*

Study	Input Feature Type	Algorithms	Performance
H. Drimalla et al., 2019 [62]	Facial expression, gaze, audio	Random Forest	Acc: 0.753, Precision: 0.774, Sensitivity: 0.667, Specificity: 0.829, AUC: 0.840, F1: 0.716
		SVM	Acc: 0.714, Precision: 0.719, Sensitivity: 0.639, Specificity: 0.780, AUC: 0.810, F1: 0.676
		Stacked CNN	Acc: 0.461, Precision: 0.391, Sensitivity: 0.250, Specificity: 0.650, AUC: 0.530, F1: 0.305
		Pooled CNN	Acc: 0.605, Precision: 0.594, Sensitivity: 0.528, Specificity: 0.675, AUC: 0.640, F1: 0.559
H. Drimalla et al., 2020 [63]	Facial AU features	Random Forest	Acc: 0.671, Precision: 0.647, Sensitivity: 0.611, Specificity: 0.721, AUC: 0.740, F1: 0.629
			Acc: 0.570, Precision: 0.528, Sensitivity: 0.528, Specificity: 0.605, AUC: 0.630, F1: 0.528
			Acc: 0.747, Precision: 0.722, Sensitivity: 0.722, Specificity: 0.767, AUC: 0.770, F1: 0.722
			Acc: 0.730, Precision: 0.727, Sensitivity: 0.667, Specificity: 0.791, AUC: 0.780, F1: 0.696
K. S. Lohan et al., 2018 [64]	Pupil Diameter Data	LSTM	Acc: 0.798
	AOI		Acc: 0.759
E. Chong et al., 2020 [82]	Cropped Face Regions	Deep model (smoothed) (ResNet-50 backbone)	Precision: 0.936, Sensitivity: 0.943, F1: 0.940
			Precision: 0.924, Sensitivity: 0.937, F1: 0.930
		Deep model (without transfer learning smoothed)	Precision: 0.917, Sensitivity: 0.915, F1: 0.916
D. Alie et al., 2011 [65]	Gaze Pattern	VMM	Acc: 0.938, Precision: 0.750, Sensitivity: 1, Specificity: 0.923, F1: 0.857
		HMM	Acc: 0.813, Precision: 0.5, Sensitivity: 0.333, Specificity: 0.923, F1: 0.400

(Continued)

Supplementary Table 4. *(Continued)*

Study	Input Feature Type	Algorithms	Performance
J. Zhang et al., 2022 [67]	valid sampling rate, number of RJA events	Logistic regression	Acc: 0.760, AUC: 0.818
M. Varma et al., 2022 [68]	Gaze fixation pattern	LSTM (window= 100, shift = 10)	Precision: 0.595, Sensitivity: 0.598
M. Alcaniz et al., 2022 [69]	(9 features) Number of AOIs seen, average number of AOIs seen per scene (and SD), number of frames with no AOIs, AOIs, persons, items	SVM	Acc: 0.860, Sensitivity: 0.910, Specificity: 0.820, AUC: 0.910, F1: 0.870
			Acc: 0.770, Sensitivity: 0.780, Specificity: 0.850, AUC: 0.830, F1: 0.790
			Acc: 0.840, Sensitivity: 0.940, Specificity: 0.710, AUC: 0.880, F1: 0.870
		Random Forest	Acc: 0.820, Sensitivity: 0.920, Specificity: 0.640, AUC: 0.840, F1: 0.860
		Naïve Bayes	Acc: 0.830, Sensitivity: 0.820, Specificity: 0.850, AUC: 0.840, F1: 0.840
		SVM	Acc: 0.780, Sensitivity: 0.900, Specificity: 0.610, AUC: 0.790, F1: 0.820
		XGBoost	Acc: 0.730, Sensitivity: 0.800, Specificity: 0.680, AUC: 0.770, F1: 0.770
		K-NN	Acc: 0.740, Sensitivity: 0.880, Specificity: 0.540, AUC: 0.670, F1: 0.790
			Acc: 0.740, Sensitivity: 0.880, Specificity: 0.540, AUC: 0.670, F1: 0.790
			Acc: 0.660, Sensitivity: 0.730, Specificity: 0.510, AUC: 0.640, F1: 0.700
		Naïve Bayes	Acc: 0.690, Sensitivity: 0.670, Specificity: 0.710, AUC: 0.720, F1: 0.710

(Continued)

Supplementary Table 4. *(Continued)*

Study	Input Feature Type	Algorithms	Performance
Z. Zhao et al., 2021 [72]	Gaze fixation time (three features)	SVM	Acc: 0.923, Precision: 1, Sensitivity: 0.842, Specificity: 1, AUC: 0.920, F1: 0.914
		LDA	Acc: 0.897, Sensitivity: 0.895, Specificity: 0.900, AUC: 0.920
		Decision Tree	Acc: 0.846, Sensitivity: 0.895, Specificity: 0.800, AUC: 0.860
		Random Forest	Acc: 0.846, Sensitivity: 0.842, Specificity: 0.850, AUC: 0.860
P. R. Krishppababu et al., 2021 [73]	Integrated entropy	Decision Tree	Acc: 0.783, Precision: 0.724, Sensitivity: 0.935, F1: 0.816
			Acc: 0.783, Precision: 0.705, Sensitivity: 0.888, F1: 0.786
			Acc: 0.766, Precision: 0.700, Sensitivity: 0.893, F1: 0.785
			Acc: 0.767, Precision: 0.703, Sensitivity: 0.896, F1: 0.788
			Acc: 0.750, Precision: 0.750, Sensitivity: 0.770, F1: 0.760
			Acc: 0.650, Precision: 0.751, Sensitivity: 0.825, F1: 0.786
			Acc: 0.700, Precision: 0.691, Sensitivity: 0.821, F1: 0.750
			Acc: 0.616, Precision: 0.625, Sensitivity: 0.517, F1: 0.566
			Acc: 0.791, Precision: 0.757, Sensitivity: 0.774, F1: 0.765
			Acc: 0.795, Precision: 0.751, Sensitivity: 0.851, F1: 0.798
			Acc: 0.750, Precision: 0.724, Sensitivity: 0.750, F1: 0.737
			Acc: 0.633, Precision: 0.613, Sensitivity: 0.655, F1: 0.633
N. Kojovic et al., 2021 [74]	Movement and 5seconds of video	VGG16 CNN + LSTM	Acc: 0.809, Precision: 0.784, Sensitivity: 0.854, Specificity: 0.765, F1: 0.818

(Continued)

Supplementary Table 4. *(Continued)*

Study	Input Feature Type	Algorithms	Performance
A. Adham et al., 2022 [71]	Eye Gaze scan path (Spatial)	DNN	Acc: 0.744, AUC: 0.6828
		CNN	Acc: 0.725, AUC: 0.721
			Acc: 0.763, AUC: 0.818
			Acc: 0.803, AUC: 0.839
W. B. Liu et al., 2017 [83]	Latency	Decision Tree	Acc: 0.907
			Acc: 0.907
			Acc: 0.930
J. Li et al., 2021 [86]	log-Mel spectrogram + Image	ResNet-18	Acc: 0.724, F1: 0.750
C. Wu et al., 2021 [85]	Image, Facial Key points Features	ResNet-18	Acc: 0.645, F1: 0.645
G. Zhang et al., 2021 [84]	social visual behaviour scores	Deep CNN (Three branches)	Mean Squared error loss:0.177
N. Zhang et al., 2022 [75]	LPQ-TOP scenes	SVM	Acc:0.917, F1:0.901
J. T. Megerian et al., 2022 [76]	video and questionnaires	gradient boosted decision tree	Sensitivity:0.984, Precision: 0.789

(Continued)

Supplementary Table 4. *(Continued)*

Study	Input Feature Type	Algorithms	Performance
D. Anagnostopoulou et al., 2021 [87]	Pose features (TD-Joint Attention Data set)	1D CNN	Acc: 0.730, Precision: 0.563, F1: 0.602
		ResNet-50	Acc: 0.745, Precision: 0.633, F1: 0.645
		LSTM (one layer)	Acc: 0.762, Precision: 0.744, F1: 0.742
		1D CNN	Acc: 0.774, Precision: 0.763, F1: 0.752
		2D CNN	Acc: 0.789, Precision: 0.774, F1: 0.765
		LSTM	Acc: 0.795, Precision: 0.780, F1: 0.769
		AlexNet	Acc: 0.804, Precision: 0.807, F1: 0.805
		1D CNN	Acc: 0.679, Precision: 0.681, F1: 0.680
		2D CNN	Acc: 0.709, Precision: 0.736, F1: 0.713
		LSTM	Acc: 0.718, Precision: 0.715, F1: 0.716
		AlexNet	Acc: 0.787, Precision: 0.824, F1: 0.734
		2D CNN	Acc: 0.758, Precision: 0.770, F1: 0.754
		2D CNN	Acc: 0.709, Precision: 0.736, F1: 0.713
		LSTM	Acc: 0.615, Precision: 0.621, F1: 0.605
		LSTM	Acc: 0.718, Precision: 0.715, F1: 0.716
		AlexNet	Acc: 0.672, Precision: 0.678, F1: 0.643
		AlexNet	Acc: 0.787, Precision: 0.824, F1: 0.734

(Continued)

Supplementary Table 4. *(Continued)*

Study	Input Feature Type	Algorithms	Performance
I. Akin-Bulbul et al., 2022 [70]	Fixation count, Fixation duration, Dwell time, Net Dwell Time, Diversion Duration (ReliefF feature selection)	Decision Tree	Acc: 0.725, AUC: 0.736, F1: 0.721
		Naïve Bayes	Acc: 0.825, AUC: 0.843, F1: 0.825
		Random Forest	Acc: 0.720, AUC: 0.759, F1: 0.716
		SVM	Acc: 0.775, AUC: 0.775, F1: 0.768
	Fixation count, Fixation duration, Dwell time, Net Dwell Time, Diversion Duration (Information Gain feature selection)	Decision Tree	Acc: 0.725, AUC: 0.736, F1: 0.721
		Naïve Bayes	Acc: 0.850, AUC: 0.845, F1: 0.848
		Random Forest	Acc: 0.725, AUC: 0.825, F1: 0.71
		SVM	Acc: 0.825, AUC: 0.825, F1: 0.822
	Fixation count, Fixation duration, Dwell time, Net Dwell Time, Diversion Duration (Wrapper feature selection)	Decision Tree	Acc: 0.725, AUC: 0.659, F1: 0.710
		Naïve Bayes	Acc: 0.800, AUC: 0.835, F1: 0.836
		Random Forest	Acc: 0.800, AUC: 0.794, F1: 0.795
		SVM	Acc: 0.750, AUC: 0.750, F1: 0.744

(Continued)

Supplementary Table 4. *(Continued)*

Study	Input Feature Type	Algorithms	Performance
I. Akin-Bulbul et al., 2022 [70]	Fixation count, Fixation duration, Dwell time, Net Dwell Time, Diversion Duration (ReliefF feature selection)	Decision Tree	Acc: 0.800, AUC: 0.800, F1: 0.792
		Naïve Bayes	Acc: 0.825, AUC: 0.820, F1: 0.825
		Random Forest	Acc: 0.875, AUC: 0.888, F1: 0.874
		SVM	Acc: 0.875, AUC: 0.875, F1: 0.874
	Fixation count, Fixation duration, Dwell time, Net Dwell Time, Diversion Duration (Information Gain feature selection)	Decision Tree	Acc: 0.675, AUC: 0.678, F1: 0.675
		Naïve Bayes	Acc: 0.800, AUC: 0.815, F1: 0.815
		Random Forest	Acc: 0.85, AUC: 0.874, F1: 0.848
		SVM	Acc: 0.800, AUC: 0.800, F1: 0.798
	Fixation count, Fixation duration, Dwell time, Net Dwell Time, Diversion Duration (Wrapper feature selection)	Decision Tree	Acc: 0.825, AUC: 0.825, F1: 0.819
		Naïve Bayes	Acc: 0.825, AUC: 0.858, F1: 0.824
		Random Forest	Acc: 0.800, AUC: 0.823, F1: 0.795
		SVM	Acc: 0.800, AUC: 0.800, F1: 0.795

(Continued)

Supplementary Table 4. *(Continued)*

Study	Input Feature Type	Algorithms	Performance
H. Javed et al., 2020 [81]	Movement + Expression	CNN (TD)	Acc: 0.680, F1: 0.650
		SVC (TD)	Acc: 0.640, F1: 0.580
		Random Forest (TD)	Acc: 0.740, F1: 0.740
		Decision Tree (TD)	Acc: 0.640, F1: 0.610
		K-NN (TD)	Acc: 0.740, F1: 0.730
		CNN (ASD)	Acc: 0.720, F1: 0.710
		SVC (ASD)	Acc: 0.600, F1: 0.580
		Random Forest (ASD)	Acc: 0.770, F1: 0.760
		Decision Tree (ASD)	Acc: 0.610, F1: 0.600
		K-NN (ASD)	Acc: 0.760, F1: 0.760
		CNN (ASD+ TD)	Acc: 0.650, F1: 0.620
		SVC (ASD + TD)	Acc: 0.590, F1: 0.540
		Random Forest (ASD + TD)	Acc: 0.740, F1: 0.710
		Decision Tree (ASD + TD)	Acc: 0.600, F1: 0.560
		K-NN (ASD + TD)	Acc: 0.710, F1: 0.710

Supplementary Table 4. *(Continued)*

Study	Input Feature Type	Algorithms	Performance
N. Qiu et al., 2020 [66]	behavioural characteristics	Naïve Bayes (FF episode)	Acc: 0.805, Precision: 0.811, Sensitivity: 0.750, Specificity: 0.837, F1: 0.779
		Naïve Bayes (SF episode)	Acc: 0.824, Precision: 0.838, Sensitivity: 0.775, Specificity: 0.860, F1: 0.805
		Random Forest (FF episode)	Acc: 0.807, Precision: 0.816, Sensitivity: 0.775, Specificity: 0.837, F1: 0.795
		Random Forest (SF episode)	Acc: 0.831, Precision: 0.861, Sensitivity: 0.775, Specificity: 0.884, F1: 0.816
		SVM (FF episode)	Acc: 0.812, Precision: 0.778, Sensitivity: 0.875, Specificity: 0.767, F1: 0.824
		SVM (SF episode)	Acc: 0.834, Precision: 0.829, Sensitivity: 0.850, Specificity: 0.837, F1: 0.840

Supplementary Table 5. Summary of feature extraction algorithms and tools used in the included studies.

Authors	Facial and Head Pose Estimation	Movement Feature	Gaze Estimation/ Eye Contact Algorithms	Vocal Features Extraction
A. Di Nuovo	Viola-Jones, MTCNN, Faster R-CNN,	-	-	-
C. Eunji	Faster-RCNN	-	PEEC Detector (Pose-Dependent Egocentric Eye Contact Detector), PiCNN Detector (Pose-implicit Convolutional Neural Networks), Modified AlexNet, Gaze Locking	-
C. Tang	MTCNN, HOG	-	-	openSMILE (low-level descriptors)
E. Chong	Omron OKAO Library, IntraFace (facial landmarks) + Perspective-n-Point Algorithm (head pose)	-	-	-
G. Nie	Gaussian Mixture Model, hard histogram of head pose, soft histogram of head pose	hard histogram of head pose motion, soft histogram of head pose motion	-	-
H. Drimalla	OpenFace, HOG	-	-	librosa library, proscodic (root-mean-square energy), spectral features (forty mel-frequency cepstral coefficients, MMCC)
H. Drimalla;	OpenFace	-	-	-
H. Javed	OpenPose	OpenPose, Laban Movement Analysis	-	-
K. S. Lohan	-	-	Pupil diameter and AOI	-
E. Chong	-	-	-	-
D. Alie	-	-	-	-

(Continued)

Supplementary Table 5. *(Continued)*

Authors	Facial and Head Pose Estimation	Movement Feature	Gaze Estimation/ Eye Contact Algorithms	Vocal Features Extraction
N. Qiu et al., 2020 [66]	-	-	Duration and frequency of eye contact, active social engagement, and social smiling during the Face-to-face episode and still-face episode	-
W. B. Liu et al., 2017 [83]	Dlib, landmarks (head pose) + PCA	-	-	Kaldi
N. Zhang et al., 2022 [75]	MTCNN, local phase quantization in three orthogonal planes (LPQ-TOP), histogram	-	-	-
J. Zhang et al., 2022 [67]	-	-	Eye Tracker	-
M. Varma et al., 2022 [68]	Amazon Rekognition, OpenFace	-	OpenFace, Coarse gaze annotation (16 predefined AOIs)	-
J. T. Megerian et al., 2022 [76]	-	-	-	-
M. Alcaniz et al., 2022 [69]	-	-	-	-
I. Akin-Bulbul et al., 2022 [70]	-	-	-	-
A. Adham et al., 2022 [71]	-	-	-	-
Z. Zhao et al., 2021 [72]	-	-	Tobii	-
G. Zhang et al., 2021 [84]	-	-	-	-
C. Wu et al., 2021 [85]	OpenFace 2.0	-	-	-
J. Li et al., 2021 [86]	MTCNN	-	-	log-Mel spectrograms
P. R. Krishnappababu et al., 2021 [73]	low-dimensional facial embedding	-	-	-
N. Kojovic et al., 2021 [74]	-	OpenPose	-	-
D. Anagnostopoulou et al., 2021 [87]	-	OpenPose	-	-

Supplementary Table 6. Quality assessment details of the eligible studies.

Study	Patient Selection						Index Test				Reference Standard				Flow and Timing				
	PS-Q1	PS-Q2	PS-Q3	ROB	PS-Q4	AC	IT-Q1	IT-Q2	ROB	AC	RS-Q1	RS-Q2	ROB	AC	FAT-Q1	FAT-Q2	FAT-Q3	FAT-Q4	ROB
A. Di Nuovo et al., 2018 [77]	Y	Y	Y	L	Y	L	N	N	L	L	Y	Y	L	L	Y	Y	Y	Y	L
C. Eunji et al., 2017 [78]	Y	N	Y	L	Y	L	N	Y	L	L	Y	Y	L	L	Y	Y	Y	Y	L
C. Tang et al., 2020 [61]	Y	N	Y	L	Y	L	N	Y	L	L	Y	Y	L	L	Y	Y	Y	Y	L
E. Chong et al., 2017 [79]	Y	N	Y	L	Y	L	N	N	L	L	Y	Y	L	L	Y	NA	NA	Y	L
G. Nie et al., 2018 [80]	Y	N	Y	NA	Y	L	N	Y	L	L	Y	Y	L	L	Y	NA	NA	NA	L
H. Drimalla et al., 2019 [62]	Y	N	Y	L	Y	L	N	Y	L	L	Y	Y	L	L	Y	Y	Y	Y	L
H. Drimalla et al., 2020 [63]	Y	N	Y	L	Y	L	N	Y	L	L	Y	Y	L	L	Y	Y	Y	Y	L
H. Javed et al., 2020 [81]	Y	N	NA	H	Y	L	N	N	L	L	NA	Y	NA	NA	NA	NA	NA	Y	L
K. S. Lohan et al., 2018 [64]	Y	N	Y	L	Y	L	N	N	L	L	Y	Y	L	L	Y	Y	Y	N	L
E. Chong et al., 2020 [82]	Y	N	Y	L	Y	L	N	Y	L	L	Y	Y	L	L	Y	Y	Y	N	L
D. Alie et al., 2011 [65]	Y	N	Y	H	Y	L	N	N	L	L	Y	Y	L	L	Y	Y	Y	Y	L
N. Zhang et al., 2022 [75]	Y	Y	Y	L	Y	L	N	Y	L	L	Y	Y	L	L	Y	Y	Y	Y	L
N. Qiu et al., 2020 [66]	Y	N	Y	L	Y	L	N	N	L	L	Y	Y	L	L	Y	Y	Y	Y	L
W. B. Liu et al., 2017 [83]	Y	N	NA	L	Y	L	N	N	L	L	Y	Y	L	L	Y	Y	Y	Y	L

(Continued)

Supplementary Table 6. *(Continued)*

Study	Patient Selection						Index Test				Reference Standard				Flow and Timing				
	PS-Q1	PS-Q2	PS-Q3	ROB	PS-Q4	AC	IT-Q1	IT-Q2	ROB	AC	RS-Q1	RS-Q2	ROB	AC	FAT-Q1	FAT-Q2	FAT-Q3	FAT-Q4	ROB
M. Varma et al., 2022 [68]	Y	N	Y	H	Y	L	N	Y	L	L	NA	Y	NA	L	Y	NA	NA	Y	L
J. Zhang et al., 2022 [67]	Y	N	Y	L	Y	L	N	Y	L	L	Y	Y	L	L	Y	Y	Y	Y	L
J. T. Megerian et al., 2022 [76]	Y	N	Y	H	Y	L	N	Y	L	L	Y	Y	NA	L	NA	N	NA	N	L
M. Alcaniz et al., 2022 [69]	Y	Y	Y	L	Y	L	N	N	L	L	Y	Y	L	L	Y	Y	Y	Y	L
I. Akin-Bulbul et al., 2022 [70]	Y	N	Y	L	Y	L	N	N	L	L	Y	Y	L	L	Y	Y	Y	Y	L
A. Adham et al., 2022 [71]	NA	N	NA	H	Y	L	N	Y	L	L	NA	Y	NA	NA	NA	NA	NA	N	L
Z. Zhao et al., 2021 [72]	Y	N	Y	L	Y	L	N	Y	L	L	Y	Y	L	L	Y	Y	Y	Y	L
G. Zhang et al., 2021 [84]	Y	Y	Y	L	Y	L	N	Y	L	L	Y	Y	L	L	Y	Y	Y	Y	L
C. Wu et al., 2021 [85]	Y	Y	Y	H	Y	NA	N	N	L	L	Y	Y	NA	L	NA	NA	NA	Y	L

(Continued)

Supplementary Table 6. *(Continued)*

Study	Patient Selection						Index Test				Reference Standard				Flow and Timing				
	PS-Q1	PS-Q2	PS-Q3	ROB	PS-Q4	AC	IT-Q1	IT-Q2	ROB	AC	RS-Q1	RS-Q2	ROB	AC	FAT-Q1	FAT-Q2	FAT-Q3	FAT-Q4	ROB
J. Li et al., 2021 [86]	Y	Y	Y	L	Y	L	N	Y	L	L	Y	Y	L	L	Y	Y	Y	N	L
N. Kojovic et al., 2021 [74]	Y	N	Y	L	Y	L	N	N	L	L	Y	Y	L	L	Y	Y	Y	Y	L
D. Anagnostopoulou et al., 2021 [87]	Y	N	NA	H	Y	L	N	N	L	L	Y	Y	L	L	Y	Y	Y	Y	L
P. R. Krishnappababu, et al., 2021 [73]	Y	N	Y	H	Y	L	N	N	L	L	Y	Y	L	L	Y	Y	Y	Y	L

Patient Selection, PS; Index Test, IT; Reference Standards: RS; Flow and Timing: FAT; Risk of Bias, ROB; Applicability Concerns, AC; YES, Y; NO, N; LOW, L; HIGH, H; UNCLEAR, NA.

Supplementary Table 7. Summary of the moderator analysis on the classification algorithms reported in the diagnostic accuracy tests studies.

Moderator	Estimate	SE	z-value	p-value	95% CI
CNN	1.371	0.634	2.162	0.0306	[0.128, 2.614]
HMM	1.695	1.569	1.080	0.280	[-1.381, 4.770]
K-NN	2.290	0.660	3.470	0.0005	[0.997, 3.584]
NB	2.774	0.703	3.946	<0.0001	[1.396, 4.152]
RF	2.941	0.594	4.954	<0.0001	[1.778, 4.105]
SVM	3.104	0.580	5.356	<0.0001	[1.968, 4.240]
VMM	4.847	1.998	2.426	0.0153	[0.931, 8.764]

Supplementary Table 8. Summary of the moderator analysis on the input modalities reported in the diagnostic accuracy tests studies.

Moderator	Estimate	SE	z-value	p-value	95% CI
Facial	2.789	0.915	3.049	0.0023	[0.996, 4.581]
Facial + Gaze + Vocal	3.014	0.919	3.280	0.001	[1.213, 4.814]
Gaze	1.911	0.887	2.155	0.0312	[0.173, 3.649]
Head Pose	-0.0968	0.980	-0.099	0.9214	[-2.018, 1.825]
Head Pose + Facial	3.153	1.057	2.984	0.0028	[1.082, 5.224]
Head Pose + Facial + Vocal	4.589	1.120	4.098	<0.0001	[2.394, 6.783]
Head Pose + Vocal	3.002	1.022	2.939	0.0033	[1.0, 5.004]
Posture	2.953	1.722	1.715	0.0864	[-0.422, 6.328]
Posture + Image	2.817	1.812	1.555	0.120	[-0.735, 6.368]
Vocal	3.139	0.918	3.420	0.0006	[1.340, 4.937]
Vocal + Facial	4.781	1.130	4.231	<0.0001	[2.566, 6.996]

Supplementary Table 9. Summary of the moderator analysis on the combined effect of classification algorithms and input modalities.

Moderator	Estimate	SE	z-value	p-value	95% CI
Facial	1.152	1.078	1.069	0.285	[-0.960, 3.264]
Facial + Gaze + Vocal	1.441	1.068	1.349	0.177	[-0.652, 3.535]
Gaze	0.230	1.089	0.211	0.833	[-1.904, 2.365]
Head Pose	-1.637	1.135	-1.442	0.149	[-3.861, 0.588]
Head Pose + Facial	1.547	1.198	1.291	0.197	[-0.802, 3.895]
Head Pose + Facial + Vocal	3.038	1.257	2.418	0.0156	[0.575, 5.500]
Head Pose + Vocal	1.368	1.166	1.173	0.241	[-0.918, 3.654]
Posture	1.229	1.800	0.683	0.495	[-2.299, 4.757]
Posture + Image	2.817	1.812	1.555	0.120	[-0.734, 6.368]
Vocal	1.512	1.080	1.400	0.162	[-0.605, 3.629]
Vocal + Facial	3.217	1.265	2.543	0.011	[0.738, 5.697]
HMM	1.464	2.272	0.645	0.519	[-2.988, 5.917]
K-NN	1.245	0.652	1.910	0.0561	[-0.0325, 2.523]
NB	1.564	0.693	2.258	0.024	[0.206, 2.922]
RF	1.766	0.549	3.220	0.0013	[0.691, 2.842]
SVM	1.860	0.538	3.458	0.0005	[0.806, 2.913]
VMM	4.617	2.587	1.786	0.0743	[-0.453, 9.687]

Supplementary Table 10. Summary of the moderator analysis on the combined effect of classification algorithms, input modalities and experimental tasks.

Moderator	Estimate	SE	z-value	p-value	95% CI
Facial	4.544	1.746	2.602	0.0093	[1.121, 7.967]
Facial + Gaze + Vocal	4.994	1.701	2.936	0.0033	[1.660, 8.327]
Gaze	3.532	1.636	2.159	0.0308	[0.326, 6.737]
Head Pose	1.722	1.802	0.956	0.339	[-1.810, 5.253]
Head Pose + Facial	4.903	1.844	2.660	0.0078	[1.290, 8.516]
Head Pose + Facial + Vocal	6.396	1.881	3.400	0.0007	[2.709, 10.082]
Head Pose + Vocal	4.723	1.824	2.590	0.0096	[1.149, 8.297]
Posture	3.678	1.778	2.069	0.0385	[0.194, 7.163]
Posture + Image	2.817	0.609	4.624	<0.0001	[1.623, 4.011]
Vocal	4.898	1.751	2.797	0.0052	[1.466, 8.330]
Vocal + Facial	6.575	1.887	3.484	0.0005	[2.876, 10.273]
HMM	-1.837	1.933	-0.950	0.342	[-5.626, 1.952]
K-NN	1.098	0.648	1.695	0.0902	[-0.172, 2.369]
NB	1.502	0.664	2.262	0.0237	[0.200, 2.803]
RF	1.765	0.475	3.718	0.0002	[0.834, 2.695]
SVM	1.733	0.529	3.275	0.0011	[0.696, 2.769]
VMM	1.316	2.296	0.573	0.567	[-3.184, 5.815]
Simulated Interaction Task	-4.773	1.666	-2.865	0.0042	[-8.039, 1.508]
Still-Face	-2.387	1.694	-1.409	0.159	[-5.708, 0.933]

Supplementary Materials For Chapter 3

Supplementary Table 11. A summary of the eye contact (and other) datasets for a single-view.

Dataset Type	Split No.	No. of Eye Contact Segments	No. of Other Segments	No. of ASD Participants	No. of TD Participants
Train	1	866	1332	67	19
	2	697	1243	67	21
	3	1068	1483	63	22
	4	1303	1612	66	22
	5	1244	1635	66	21
Validation	1	586	683	21	8
	2	443	509	24	6
	3	215	443	24	8
	4	208	380	25	8
	5	274	420	25	9
Test	1	274	420	25	9
	2	586	683	21	8
	3	443	509	24	6
	4	215	443	24	8
	5	208	380	25	8

Supplementary Table 12. Summary of the single-view predictive performance of eye contact (other) behaviours in the spatial domain.

Spatial Domain Single View	Top-1 Accuracy	Precision	Recall	F1
Split1	0.862	0.841 (0.880)	0.853 (0.870)	0.847 (0.875)
Split2	0.833	0.779 (0.883)	0.858 (0.815)	0.816 (0.847)
Split3	0.849	0.825 (0.863)	0.783 (0.892)	0.803 (0.877)
Split4	0.835	0.722 (0.922)	0.878 (0.810)	0.792 (0.863)
Split5	0.869	0.850 (0.883)	0.835 (0.893)	0.842 (0.888)

Supplementary Table 13. Summary of the single-view predictive performance of eye contact (other) behaviours in the temporal domain.

Temporal Domain Single View	Top-1 Accuracy	Precision	Recall	F1
Split1	0.801	0.784	0.764	0.774
		(0.813)	(0.830)	(0.822)
Split2	0.745	0.823	0.523	0.639
		(0.716)	(0.914)	(0.803)
Split3	0.821	0.859	0.654	0.742
		(0.805)	(0.930)	(0.863)
Split4	0.831	0.800	0.707	0.751
		(0.846)	(0.901)	(0.873)
Split5	0.791	0.727	0.801	0.763
		(0.846)	(0.784)	(0.813)

Supplementary Table 14. Summary of spatio-temporal fusion prediction performance of eye contact (other) behaviours for a single-view.

Spatial-Temporal Fusion Single View	Top-1 Accuracy	Precision	Recall	F1
Split1	0.890	0.879	0.873	0.876
		(0.898)	(0.903)	(0.900)
Split2	0.868	0.860	0.828	0.844
		(0.873)	(0.898)	(0.885)
Split3	0.882	0.900	0.788	0.840
		(0.872)	(0.943)	(0.906)
Split4	0.881	0.806	0.881	0.842
		(0.930)	(0.881)	(0.905)
Split5	0.890	0.869	0.869	0.869
		(0.906)	(0.906)	(0.906)

Supplementary Table 15. Summary of multi-view predictive performance of eye contact (other) behaviours in the spatial domain.

Spatial Domain Multi-view	Top-1 Accuracy	Precision	Recall	F1
Split1	0.934	0.918	0.937	0.940
		(0.948)	(0.932)	(0.927)
Split2	0.911	0.864	0.943	0.902
		(0.953)	(0.887)	(0.919)
Split3	0.937	0.922	0.918	0.920
		(0.947)	(0.949)	(0.948)
Split4	0.902	0.811	0.949	0.874
		(0.968)	(0.876)	(0.920)
Split5	0.937	0.917	0.934	0.925
		(0.951)	(0.939)	(0.945)

Supplementary Table 16. Summary of multi-view predictive performance of eye contact (other) behaviours in the temporal domain.

Temporal Domain Multi-view	Top-1 Accuracy	Precision	Recall	F1
Split1	0.873	0.875	0.838	0.856
		(0.872)	(0.902)	(0.887)
Split2	0.765	0.910	0.508	0.652
		(0.719)	(0.961)	(0.823)
Split3	0.851	0.925	0.677	0.782
		(0.820)	(0.964)	(0.886)
Split4	0.888	0.906	0.767	0.831
		(0.880)	(0.956)	(0.916)
Split5	0.876	0.833	0.882	0.857
		(0.911)	(0.872)	(0.891)

Supplementary Table 17. Summary of spatio-temporal fusion prediction performance of eye contact (other) behaviours for multi-view.

Spatial-Temporal Fusion Multi-view	Top-1 Accuracy	Precision	Recall	F1
Split1	0.950	0.938	0.952	0.945
		(0.960)	(0.949)	(0.954)
Split2	0.910	0.920	0.868	0.893
		(0.903)	(0.942)	(0.922)
Split3	0.930	0.946	0.873	0.908
		(0.921)	(0.967)	(0.943)
Split4	0.933	0.895	0.920	0.908
		(0.955)	(0.940)	(0.947)
Split5	0.952	0.926	0.963	0.944
		(0.973)	(0.944)	(0.958)

Supplementary Materials For Chapter 4

Supplementary Table 18. Definition of facial features extracted for ASD classification.

A) OpenFace2.0 Extracted Features

Feature Name	Feature Description
gaze_angle_x	Eye gaze direction in radians along the x-axis (looking left-right)
gaze_angle_y	Eye gaze direction in radians along the y-axis (looking up-down)
pose_Rx	Head pose rotation in radians along the x-axis (pitch)
pose_Ry	Head pose rotation in radians along the y-axis (yaw)
pose_Rz	Head pose rotation in radians along the z-axis (roll)

B) Hand-crafted Features (Ours)

Feature Name	Feature Description
gaze_angular_displacement_x	Eye-gaze angular displacement along the x-axis (looking left-right) by frame
gaze_angular_displacement_y	Eye-gaze angular displacement along the y-axis (looking up-down) by frame
gaze_velocity_x	Eye-gaze angular velocity along the x-axis (looking left-right) by frame
gaze_velocity_y	Eye-gaze angular velocity along the y-axis (looking up-down) by frame
gaze_acceleration_x	Eye-gaze angular acceleration along the x-axis (looking left-right) by frame
gaze_acceleration_y	Eye-gaze angular acceleration along the y-axis (looking up-down) by frame
pitch_Rx_angular_displacement	Angular displacement of the head pose rotation along the x-axis, calculated by frame (pitch)
yaw_Ry_angular_displacement	Angular displacement of the head pose rotation along the y-axis, calculated by frame (yaw)
roll_Rz_angular_displacement	Angular displacement of the head pose rotation along the z-axis, calculated by frame (roll)
pitch_Rx_velocity	Angular velocity of the head pose rotation along the x-axis, calculated by frame (pitch)
yaw_Ry_velocity	Angular velocity of the head pose rotation along the y-axis, calculated by frame (yaw)
roll_Rz_velocity	Angular velocity of the head pose rotation along the z-axis, calculated by frame (roll)
pitch_Rx_acceleration	Angular acceleration of the head pose rotation along the x-axis, calculated by frame (pitch)
yaw_Ry_acceleration	Angular acceleration of the head pose rotation along the y-axis, calculated by frame (yaw)
roll_Rz_acceleration	Angular acceleration of the head pose rotation along the z-axis, calculated by frame (roll)
mean_gaze_angular_displacement_x	Mean angular displacement of the eye gaze along the x-axis for all events of the target behaviours (looking left-right)
mean_gaze_angular_displacement_y	Mean angular displacement of the eye gaze along the y-axis for all events of the target behaviours (looking up-down)

(Continued)

Supplementary Table 18. *(Continued)*

Feature Name	Feature Description
mean_gaze_velocity_x	Mean angular velocity of the eye gaze along the x-axis for all events of the target behaviours (looking left-right)
mean_gaze_velocity_y	Mean angular velocity of the eye gaze along the y-axis for all events of the target behaviours (looking up-down)
mean_gaze_acceleration_x	Mean angular acceleration of the eye gaze along the x-axis for all events of the target behaviours (looking left-right)
mean_gaze_acceleration_y	Mean angular acceleration of the eye gaze along the y-axis for all events of the target behaviours (looking up-down)
mean_pose_angular_displacement_pitch_Rx	Mean angular displacement of the head pose rotation along the x-axis for all events of the target behaviours (pitch)
mean_pose_angular_displacement_yaw_Ry	Mean angular displacement of the head pose rotation along the y-axis for all events of the target behaviours (yaw)
mean_pose_angular_displacement_roll_Rz	Mean angular displacement of the head pose rotation along the z-axis for all events of the target behaviours (roll)
mean_pose_velocity_pitch_Rx	Mean angular velocity of the head pose rotation along the x-axis for all events of the target behaviours (pitch)
mean_pose_velocity_yaw_Ry	Mean angular velocity of the head pose rotation along the y-axis for all events of the target behaviours (yaw)
mean_pose_velocity_roll_Rz	Mean angular velocity of the head pose rotation along the z-axis for all events of the target behaviours (roll)
mean_pose_acceleration_pitch_Rx	Mean angular acceleration of the head pose rotation along the x-axis for all events of the target behaviours (pitch)
mean_pose_acceleration_yaw_Ry	Mean angular acceleration of the head pose rotation along the y-axis for all events of the target behaviours (yaw)
mean_pose_acceleration_roll_Rz	Mean angular acceleration of the head pose rotation along the z-axis for all events of the target behaviours (roll)

Supplementary Table 19. Comparison of mean cross-validation accuracy performance for model training using Lasso regression on different facial view data.

Joint Attention (All Features)	Frontal	Lateral	Occlusion
Lasso Regression + Linear SVM	0.807 (1.218e-2)	0.883 (8.912e-3)	0.889 (2.223e-2)
Lasso Regression + RBF SVM	0.860 (7.394e-3)	0.815 (1.006e-2)	0.884 (1.953e-2)
Lasso Regression + Sigmoid SVM	0.562 (4.180e-4)	0.688 (1.380e-3)	0.603 (1.970e-3)

Supplementary Table 20. Comparison table of features selected by Lasso regression.

Feature Name	Gaze	Pose	AU+Angle	Gaze+AU +Angle	Pose+AU +Angle	All Features	Frontal Features	Lateral Features	Occlusion Features
my_label	-	-	Y	Y	Y	Y	N	N	N
angle	-	-	Y	Y	Y	Y	Y	Y	Y
AU01_r (AU01_c)	-	-	YY	YY	YY	YY	YY	YY	YY
AU02_r (AU02_c)	-	-	YY	YN	YN	YY	YY	YY	YY
AU04_r (AU04_c)	-	-	YY	YY	YY	YY	YY	YY	YY
AU05_r (AU05_c)	-	-	YY	YY	YY	YY	YN	YY	YY
AU06_r (AU06_c)	-	-	YY	YY	YY	YY	YY	YY	YY
AU07_r (AU07_c)	-	-	YY	YY	YY	YY	YY	YY	YY
AU09_r (AU09_c)	-	-	YY	YY	YY	YY	YY	YY	YY
AU10_r (AU10_c)	-	-	YY	NY	YY	YY	YY	YY	YY
AU12_r (AU12_c)	-	-	YY	YY	YY	YY	YY	YY	YY
AU14_r (AU14_c)	-	-	YY	YN	YY	YY	NY	YY	YY
AU15_r (AU15_c)	-	-	YY	YY	YY	YY	YY	YY	YY
AU17_r (AU17_c)	-	-	YY	YY	YY	YY	YY	YY	YY
AU20_r (AU20_c)	-	-	YY	YY	YY	YY	YY	YY	YY
AU23_r (AU23_c)	-	-	YY	YY	YY	YY	YY	YY	YY
AU25_r (AU25_c)	-	-	YY	YY	YY	YY	YY	YY	YY
AU26_r (AU26_c)	-	-	YY	YY	YY	YY	YY	YY	YY
AU28_c	-	-	Y	N	Y	Y	Y	Y	Y
AU45_r (AU45_c)	-	-	YY	YY	YY	YY	YY	YY	YY

(Continued)

Supplementary Table 20. *(Continued)*

Feature Name	Gaze	Pose	AU+Angle	Gaze+AU +Angle	Pose+AU +Angle	All Features	Frontal Features	Lateral Features	Occlusion Features
gaze_acceleration_x	N	-	-	N	-	Y	N	N	N
gaze_acceleration_y	N	-	-	Y	-	Y	Y	N	Y
gaze_angle_x	Y	-	-	Y	-	Y	Y	Y	Y
gaze_angle_y	Y	-	-	Y	-	Y	Y	Y	Y
gaze_angular_displacement_x	Y	-	-	Y	-	Y	N	Y	Y
gaze_angular_displacement_y	Y	-	-	N	-	Y	Y	Y	Y
gaze_velocity_x	N	-	-	N	-	Y	N	Y	Y
gaze_velocity_y	N	-	-	N	-	N	N	N	N
pitch_Rx_acceleration	-	N	-	-	Y	Y	N	N	N
pitch_Rx_angular_displacement	-	Y	-	-	N	Y	Y	Y	Y
pitch_Rx_velocity	-	N	-	-	N	N	N	N	N
pose_Rx	-	Y	-	-	Y	Y	Y	Y	Y
pose_Ry	-	Y	-	-	Y	Y	Y	Y	N
pose_Rz	-	Y	-	-	Y	Y	Y	Y	Y
roll_Rz_acceleration	-	Y	-	-	Y	Y	Y	Y	N
roll_Rz_angular_displacement	-	Y	-	-	N	N	N	N	N
roll_Rz_velocity	-	N	-	-	N	N	N	N	N
yaw_Ry_acceleration	-	N	-	-	N	N	N	Y	N
yaw_Ry_angular_displacement	-	Y	-	-	Y	N	Y	Y	N
yaw_Ry_velocity	-	Y	-	-	Y	Y	N	Y	N

Y: included; N: not included.

Supplementary Table 21. Descriptive statistics (mean and standard deviation) of facial features extracted from frontal, lateral, occluded and all combined camera views per clinical group (ASD/TD) of joint attention episodes presented in free play.

	Frontal View		**Lateral View**		**Occluded View**		**All View**	
Feature Name	ASD (n=21)	TD (n=21)	ASD (n=24)	TD (n=19)	ASD (n=15)	TD (n=8)	ASD (n=24)	TD (n=22)
gaze_angle_x	-0.152 (0.214)	-0.0667 (0.217)	-0.193 (0.358)	-0.232 (0.381)	-0.160 (0.327)	-0.0627 (0.297)	-0.138 (0.220)	-0.130 (0.253)
gaze_angle_y	0.283 (0.126)	0.278 (0.144)	0.256 (0.198)	0.233 (0.188)	0.220 (0.159)	0.244 (0.126)	0.273 (0.150)	0.262 (0.137)
pose_Rx	0.264 (0.182)	0.272 (0.215)	0.324 (0.384)	0.201 (0.606)	0.186 (0.281)	0.220 (0.179)	0.281 (0.256)	0.268 (0.279)
pose_Ry	0.221 (0.271)	0.128 (0.267)	0.241 (0.490)	0.321 (0.523)	0.228 (0.404)	0.124 (0.412)	0.195 (0.287)	0.205 (0.328)
pose_Rz	-0.0221 (0.151)	0.0311 (0.111)	-0.200 (0.321)	-1.417 (0.491)	0.0843 (0.298)	0.0185 (0.194)	-0.0389 (0.227)	$7.27x10^{-3}$ (0.194)
gaze_angular_displacement_x	$9.12x10^{-3}$ (0.0248)	0.0118 (0.0334)	0.0131 (0.342)	-0.0717 (0.192)	-0.0353 (0.135)	0.0231 (0.0596)	$5.17x10^{-3}$ ($7.53x10^{-3}$)	$-1.98x10^{-3}$ (0.020)
gaze_angular_displacement_y	$2.75x10^{-3}$ ($7.82x10^{-3}$)	$-7.33x10^{-5}$ ($9.27x10^{-3}$)	$2.14x10^{-3}$ (0.0500)	-0.0192 (0.135)	$2.88x10^{-3}$ (0.0561)	$-9.0x10^{-3}$ (0.0211)	$2.34x10^{-3}$ ($4.06x10^{-3}$)	$3.04x10^{-3}$ (0.0108)
gaze_velocity_x	0.168 (0.555)	0.279 (0.838)	0.341 (8.452)	-1.723 (4.785)	-0.720 (3.425)	0.562 (1.460)	0.105 (0.149)	0.018 (0.168)
gaze_velocity_y	0.0531 (0.184)	$-1.15x10^{-3}$ (0.232)	0.0689 (1.217)	-0.507 (3.364)	0.200 (1.195)	-0.221 (0.545)	0.0525 (0.0937)	0.050 (0.152)
gaze_acceleration_x	3.633 (13.562)	6.783 (21.006)	8.996 (210.873)	-42.993 (119.639)	-16.819 (86.111)	13.845 (36.756)	2.495 (3.528)	0.431 (3.799)
gaze_acceleration_y	1.246 (4.552)	-0.0240 (5.797)	1.486 (30.405)	-12.801 (84.070)	5.550 (28.704)	-5.428 (13.756)	1.198 (2.271)	1.218 (3.627)

(Continued)

Supplementary Table 21. *(Continued)*

	Frontal View		Lateral View		Occluded View		All View	
Feature Name	ASD (n=21)	TD (n=21)	ASD (n=24)	TD (n=19)	ASD (n=15)	TD (n=8)	ASD (n=24)	TD (n=22)
pitch_Rx_angular_displacement	-1.471x10^{-3} (0.0158)	4.542x10^{-3} (0.0502)	0.025 (0.102)	-0.0742 (0.348)	1.83x10^{-3} (0.0770)	-0.0281 (0.580)	2.22x10^{-3} (0.0112)	3.42x10^{-3} (0.0139)
yaw_Ry_angular_displacement	-0.0110 (0.0326)	-0.0131 (0.0408)	-0.0348 (0.446)	0.0822 (0.275)	0.0468 (0.149)	-0.0257 (0.0829)	-7.03x10^{-3} (9.76x10^{-3})	3.07x10^{-3} (0.0282)
roll_Rz_angular_displacement	0.0102 (0.0193)	0.0128 (0.0504)	-0.0895 (0.236)	-0.0752 (0.252)	5.490x10^{-3} (0.0785)	0.0214 (0.0337)	1.58x10^{-3} (0.0103)	-8.78x10^{-4} (3.67x10^{-3})
pitch_Rx_velocity	-0.0439 (0.404)	0.108 (1.255)	0.534 (2.461)	-1.885 (8.698)	0.260 (1.630)	-0.702 (1.467)	0.0377 (0.270)	0.0503 (0.205)
yaw_Ry_velocity	-0.201 (0.715)	-0.309 (1.019)	-0.899 (11.003)	1.962 (6.860)	0.913 (3.769)	-0.631 (2.044)	-0.147 (0.201)	-0.0236 (0.219)
roll_Rz_velocity	0.243 (0.477)	0.316 (1.262)	-2.143 (5.939)	-1.924 (6.294)	-0.115 (1.612)	0.579 (0.854)	0.0355 (0.238)	-0.038 (0.109)
pitch_Rx_acceleration	-1.110 (10.175)	2.647 (31.391)	12.540 (61.236)	-47.291 (217.410)	7.614 (39.513)	-17.359 (36.871)	0.753 (6.654)	1.190 (4.892)
yaw_Ry_acceleration	-4.364 (17.296)	-7.467 (25.535)	-23.049 (274.517)	48.968 (171.498)	20.975 (94.945)	-15.521 (51.524)	-3.518 (4.803)	-0.564 (4.927)
roll_Rz_acceleration	6.029 (11.884)	7.812 (31.575)	-53.048 (148.600)	-48.289 (157.303)	-5.025 (38.265)	14.348 (21.615)	0.797 (5.892)	-1.071 (2.774)
AU01_r	0.872 (0.653)	1.354 (0.908)	1.332 (0.678)	1.218 (0.797)	0.911 (0.388)	1.126 (0.715)	1.0 (0.466)	1.288 (0.694)
AU02_r	0.134 (0.245)	0.282 (0.490)	0.334 (0.437)	0.421 (0.734)	0.280 (0.368)	0.105 (0.134)	0.166 (0.206)	0.268 (0.423)
AU04_r	0.275 (0.403)	0.473 (0.613)	0.537 (0.574)	0.646 (0.610)	0.391 (0.443)	0.363 (0.452)	0.414 (0.333)	0.506 (0.542)

(Continued)

Supplementary Table 21. *(Continued)*

	Frontal View		**Lateral View**		**Occluded View**		**All View**	
Feature Name	ASD (n=21)	TD (n=21)	ASD (n=24)	TD (n=19)	ASD (n=15)	TD (n=8)	ASD (n=24)	TD (n=22)
AU05_r	0.571 (0.370)	0.456 (0.521)	0.905 (0.599)	1.329 (0.923)	0.597 (0.613)	0.607 (0.612)	0.590 (0.296)	0.680 (0.604)
AU06_r	0.212 (0.263)	0.405 (0.457)	0.362 (0.324)	0.349 (0.318)	0.372 (0.353)	0.207 (0.174)	0.275 (0.259)	0.431 (0.423)
AU07_r	0.198 (0.294)	0.356 (0.343)	0.334 (0.408)	0.516 (0.840)	0.418 (0.624)	0.446 (0.528)	0.298 (0.399)	0.382 (0.308)
AU09_r	0.146 (0.190)	0.222 (0.204)	0.141 (0.181)	0.212 (0.287)	0.162 (0.184)	0.240 (0.256)	0.159 (0.176)	0.239 (0.205)
AU10_r	0.517 (0.456)	0.560 (0.477)	0.540 (0.442)	0.754 (0.649)	0.431 (0.444)	0.362 (0.156)	0.492 (0.349)	0.570 (0.428)
AU12_r	0.754 (0.601)	0.974 (0.720)	0.513 (0.364)	0.623 (0.525)	0.559 (0.365)	0.574 (0.364)	0.626 (0.462)	0.908 (0.703)
AU14_r	0.527 (0.489)	0.704 (0.615)	0.596 (0.493)	0.588 (0.743)	0.471 (0.447)	0.448 (0.243)	0.480 (0.342)	0.688 (0.543)
AU15_r	0.295 (0.244)	0.301 (0.269)	0.523 (0.298)	0.425 (0.339)	0.634 (0.585)	0.598 (0.434)	0.414 (0.324)	0.368 (0.250)
AU17_r	0.0545 (0.123)	0.150 (0.232)	0.511 (0.474)	0.327 (0.366)	0.429 (0.321)	0.383 (0.273)	0.220 (0.166)	0.218 (0.314)
AU20_r	0.506 (0.432)	0.772 (0.592)	0.348 (0.363)	0.282 (0.363)	0.419 (0.385)	0.776 (0.449)	0.414 (0.309)	0.579 (0.396)
AU23_r	0.132 (0.176)	0.128 (0.149)	0.158 (0.188)	0.303 (0.680)	0.145 (0.148)	0.212 (0.205)	0.117 (0.098)	0.146 (0.121)
AU25_r	0.642 (0.447)	0.776 (0.646)	0.490 (0.502)	0.581 (0.614)	0.613 (0.351)	0.570 (0.333)	0.652 (0.419)	0.772 (0.607)
AU26_r	0.406 (0.355)	0.418 (0.328)	0.707 (0.654)	0.548 (0.558)	0.575 (0.379)	0.373 (0.302)	0.526 (0.345)	0.478 (0.348)

(Continued)

Supplementary Table 21. *(Continued)*

	Frontal View		Lateral View		Occluded View		All View	
Feature Name	ASD (n=21)	TD (n=21)	ASD (n=24)	TD (n=19)	ASD (n=15)	TD (n=8)	ASD (n=24)	TD (n=22)
AU45_r	0.890 (0.488)	0.921 (0.466)	0.247 (0.271)	0.120 (0.139)	0.830 (0.703)	0.658 (0.616)	0.807 (0.445)	0.747 (0.422)
AU01_c	0.565 (0.330)	0.650 (0.345)	0.861 (0.221)	0.719 (0.326)	0.691 (0.319)	0.676 (0.288)	0.709 (0.264)	0.650 (0.294)
AU02_c	0.736 (0.300)	0.674 (0.324)	0.871 (0.230)	0.809 (0.274)	0.797 (0.229)	0.730 (0.274)	0.765 (0.218)	0.724 (0.268)
AU04_c	0.489 (0.363)	0.567 (0.359)	0.497 (0.332)	0.540 (0.384)	0.429 (0.363)	0.406 (0.320)	0.537 (0.296)	0.538 (0.310)
AU05_c	0.263 (0.206)	0.299 (0.300)	0.344 (0.341)	0.263 (0.342)	0.413 (0.308)	0.293 (0.215)	0.295 (0.212)	0.286 (0.260)
AU06_c	0.107 (0.153)	0.187 (0.278)	0.217 (0.277)	0.183 (0.292)	0.134 (0.235)	0.128 (0.178)	0.137 (0.172)	0.229 (0.267)
AU07_c	0.0849 (0.173)	0.119 (0.137)	0.222 (0.291)	0.266 (0.316)	0.152 (0.238)	0.129 (0.148)	0.177 (0.230)	0.136 (0.127)
AU09_c	0.214 (0.298)	0.266 (0.252)	0.0928 (0.210)	0.0712 (0.150)	0.134 (0.220)	0.279 (0.258)	0.140 (0.199)	0.232 (0.214)
AU10_c	0.331 (0.344)	0.446 (0.365)	0.289 (0.339)	0.218 (0.243)	0.259 (0.308)	0.392 (0.302)	0.268 (0.266)	0.364 (0.312)
AU12_c	0.533 (0.376)	0.564 (0.353)	0.245 (0.269)	0.274 (0.286)	0.262 (0.207)	0.379 (0.322)	0.410 (0.268)	0.534 (0.324)
AU14_c	0.449 (0.321)	0.539 (0.359)	0.360 (0.334)	0.288 (0.276)	0.295 (0.272)	0.340 (0.264)	0.371 (0.254)	0.460 (0.313)
AU15_c	0.557 (0.324)	0.728 (0.274)	0.574 (0.317)	0.419 (0.363)	0.508 (0.328)	0.764 (0.323)	0.465 (0.249)	0.652 (0.245)
AU17_c	0.0616 (0.145)	0.104 (0.193)	0.368 (0.313)	0.256 (0.343)	0.349 (0.314)	0.277 (0.252)	0.183 (0.177)	0.141 (0.220)

(Continued)

Supplementary Table 21. *(Continued)*

	Frontal View		**Lateral View**		**Occluded View**		**All View**	
Feature Name	ASD (n=21)	TD (n=21)	ASD (n=24)	TD (n=19)	ASD (n=15)	TD (n=8)	ASD (n=24)	TD (n=22)
AU20_c	0.820 (0.242)	0.772 (0.254)	0.409 (0.311)	0.410 (0.362)	0.584 (0.315)	0.730 (0.264)	0.670 (0.207)	0.717 (0.247)
AU23_c	0.293 (0.259)	0.347 (0.324)	0.156 (0.192)	0.187 (0.280)	0.223 (0.201)	0.333 (0.261)	0.227 (0.174)	0.293 (0.229)
AU25_c	0.376 (0.316)	0.577 (0.366)	0.354 (0.351)	0.335 (0.343)	0.443 (0.330)	0.562 (0.180)	0.411 (0.280)	0.539 (0.324)
AU26_c	0.338 (0.289)	0.465 (0.351)	0.258 (0.225)	0.246 (0.322)	0.339 (0.312)	0.219 (0.207)	0.345 (0.240)	0.406 (0.317)
AU28_c	0.0521 (0.145)	0.0985 (0.231)	0.540 (0.308)	0.604 (0.368)	0.400 (0.387)	0.408 (0.270)	0.240 (0.204)	0.227 (0.258)
AU45_c	0.510 (0.291)	0.586 (0.369)	0.174 (0.260)	0.173 (0.249)	0.557 (0.394)	0.530 (0.349)	0.490 (0.250)	0.519 (0.342)

Publications

https://doi.org/10.1111/jpc.14832

MEASUREMENT OF BODY FAT IN PREMATURE AND TERM INFANTS BY NEAR-INFRARED SPECTROSCOPY

Sun Jian[1], Alam Raquib-ul[2], Zhou Luping[2], McEwan Alistair[1]

[1]School of Biomedical Engineering, The University of Sydney, Sydney, Australia,

[2]School of Electrical and Information Engineering, The University of Sydney, Sydney, Australia

Email: jian.sun@sydney.edu.au, rala7323@uni.sydney.edu.au, luping.zhou@sydney.edu.au, alistair.mcewan@sydney.edu.au

Background

Fat mass (FM) is an important indicator to measure the nutritional status of infants. This study aims to investigate the most relevant features from near-infrared spectroscopy (NIRS) in FM prediction by Lasso regression support vector machine (SVM).

Methods

Premature and term infants from birth (<= 48 hours) to 2 years were recruited in Soweto, South Africa. NIRS spectra (wavelengths 850 nm-1090 nm, 100 features) were collected at subscapular, flank, thighs and triceps. The patient features of weight, sex, length, skin-tone and age are also included. FM was measured by DXA, Deuterium Dilution and Anthropometry. To predict FM, both linear and nonlinear SVM models were built, and the most relevant features were selected by utilising the sparsity of feature combination coefficients generated by Lasso regression.

Results

Infants (n = 370) with a uniform gender distribution were selected for analysis. Using all 105 features, the linear SVM model obtains an R-squared of 0.794. The Bland Altman limits of agreement of 0 (-0.989, 1.027) and 0 (-0.835, 0.795) are for linear and non-linear SVM respectively. Using the 16 features (5 patient and 11 NIRS features) selected by Lasso regression, the linear SVM model obtains a R-squared of 0.7997. The limits of agreement of 0 (-0.802, 0.828) and 0 (-0.803, 0.794) are for linear and non-linear SVM respectively.

Conclusions

Using the 16 features selected by Lasso regression, SVM models achieved comparable or even better prediction performance than using all 105 features, suggesting a more efficient way for NIRS spectra measurement for body fat prediction.

Statistical Validation of An Automated Method for Calculating Time Domain Heart Rate Variability on The QT Dataset

Carter Sun, *Student Member, IEEE*, Adam J. Guastella, Kelsie A. Boulton, Rinku Thapa, Alistair McEwan, *Senior Member, IEEE*

***Abstract*— Heart rate variability measures calculated from electrocardiography recordings reflect social competence. Clinical assessments of social skills have found that reduced heart rate variability is related to differences in the development of social skills in children and increase the risk of mental disorders. Limited by widespread manual signal processing and R-peak detection in current clinical assessments, most literature reports only short-term baseline studies with fewer studies reporting social interaction settings with prolonged recording. There is an urgent need for an automated physiological signal processing toolbox to detect R-peaks and perform heart rate variability measurements in social settings. This paper proposes a modified automated Neurokit2 toolbox with signal processing procedures similar to the MindWare software that requires manual inspection of R-peak locations. We calculate time domain heart rate variability metrics from the publicly available QT database by PhysioNet collected at resting states and under stress tests, mimicking social interaction stress scenarios. Statistical analysis conveys that heart rate variability metrics calculation applying both signal processing approaches using the Neurokit2 toolbox are statistically equivalent in comparison to the hand-labelled R-peaks from the QT database (n= 10 in the normal sinus rhythm group, and n= 6 in the ST Change group). Such validation results are crucial for the adoption of automated toolboxes for heart rate variability measures in social interaction assessments, where more movement and mood changes of participants are expected.**

***Clinical Relevance*— This contributes to the body of evidence of the reliability of the Neurokit2 toolbox for automatic cleaning of prolonged cardiac electrophysiological signals and calculation of heart rate variability in time-domain characterization in social interaction stress assessment.**

I. Introduction

Physiological monitoring has long been an important part of tracking the physical condition in healthcare services and personal training. Electrical activity of the heart can be recorded through electrocardiogram (ECG), reflecting cardiovascular condition. Heart rate variability (HRV) is a group of measures obtained from fluctuations in the duration of inter-beat-interval (IBI) and reflects the autonomic nerves' working condition, which can be calculated from ECG signals [1, 2]. Reduced HRV indicates a higher risk of mental disorders and cardiovascular diseases [3]. As a gold standard for recording cardiac variability, non-invasive ECG devices have been revolutionized from bedside 12-lead dry electrode ECGs to wearable 2-lead ECGs, and Photoplethysmography (PPG) using optical LED sensors. Technological innovations seek to make products that are portable, easy to use and low-cost to record ECG signals. The inconsistency of data processing algorithms in some parts due to differences in the design of the signal acquisition systems has been a long-standing challenge. Some differences are due to the nature of the signals. For example, ECG is a direct measure of the electrical activity of the heart, but PPG measures the mechanical response at some distance from the heart, with a lag from the ECG beats that vary with time (pulse transit time).

There is currently no consensus on the optimal sampling frequency for digital ECG data recording. With lower sampling frequencies such as 50-100Hz, signal interpolation [4] is required as the signal is more sensitive to motion artifacts; while higher frequencies (e.g., above 500Hz) usually require better computational resources to evaluate the electrical activity of the heart. The key to calculating HRV metrics, however, is to accurately locate the QRS complexes [5] from ECG signals because the QRS complex in ECG indicates ventricular depolarizations with high amplitude changes in electrical potential. Calculating the change in duration between adjacent R-peaks gives an indication of HRV in the time domain. The lower sampling frequency ECG recording affects frequency domain analysis in HRV. Different sample frequencies may have an unpredictable impact on the R-peak detection process, resulting in inconsistent HRV calculations [5].

Manual ECG signal quality and R-peak detection inspection is impractical for long-term datasets (for instance, several hours to days). Therefore, many studies have analyzed only short-term data (e.g., about 5 minutes) for HRV calculations. The most commonly used signals to validate the reliability of HRV calculations are resting state ECGs, while fewer studies have used ECG data collected under stress to validate algorithms [6]. In studies diagnosing social competence and quality of life in individuals with mental disorders, the response to individuals under social stress is a key component of such assessments [7, 8]. ECG cleaning and examination of the location of the R-peaks are often performed manually. An automated R-peak detection and HRV calculation software is urgently needed for ECGs collected in social interaction assessments.

Several automatic physiological signal processing toolboxes have been proposed in the literature, however, these have not been widely applied to psychological studies for the diagnosis of mental disorders. In this paper, we present a

C. Sun and A. McEwan are with the School of Biomedical Engineering, USYD, Sydney, Australia. A. J. Guastella, K. A. Boulton, R. Thapa and C. Sun are with the Brain and Mind Centre, USYD, Sydney, Australia.
E-mail for correspondence(s): jian.sun@sydney.edu.au; alistair.mcewan@sydney.edu.au; adam.guastella@sydney.edu.au

modified Neurokit2 toolbox [9] for ECG signal processing and HRV calculation. We are validating it against HRV statistics calculated on Neurokit2 using R-peaks that were identified and manually annotated by the QT database published on PhysioNet. The QT database [10] consists of expert-annotated ECG signals for the measure of QT interval. The annotation includes the onset and offset of QRS complexes and T-waves, which indicates cardiac repolarization of the heart. Using the QT database, we analyzed the differences of the modified Neurokit2 toolbox compared to its default setup in cleaning ECG data collected from resting states, and during stress tests.

The rest of the paper is structured as follows. Section II describes the data collection and signal processing algorithms, feature extraction and statistical comparisons in general. Section III provides a relevant discussion of the experimental results, and the conclusions of this paper are given in Section IV.

II. METHODOLOGY

In this section, we describe the dataset, data pre-processing and statistical analysis used to validate the Neurokit2 toolbox for automatic detection of R-peaks from ECG recordings and calculation of HRV metrics.

A. Dataset

Two-lead ECG Holter recordings published in the QT database were used in this study to examine the validity of automated physiological feature extraction algorithm proposed by Neurokit2.

Assessing the social interaction skills required participants to be free of any cardiac condition and use of cardiovascular medication. During the assessment, at the request of the assessor, participants are first asked to watch a pre-recorded video as a baseline in which they are asked to stay calm and make less movement. Healthy participants usually show normal sinus rhythm (NSR) when remaining seated, and participants may show ST-segment elevation or depression during social interaction tasks that require them to be active. Therefore, in this study we only considered ECG recordings from the NSR and the ST Change subsets from the QT database. All recordings were sampled or converted at 250Hz. The NSR dataset consists of 10 records in total 2 hours and 30 minutes long, 300 annotated beats. The ST change dataset represents the ECG recorded by a group of participants during an exercise stress test, and it contains 6 recordings in total 1 hour and 30 minutes long with 206 annotated beats. The WFDB library was used to read the annotated R-peak locations and ECG signals for each recording. Only the annotated sections of the ECG recordings were used in this study.

B. ECG Signal Pre-processing

Before extracting HRV metrics, ECG signal pre-processing is needed first to reduce the baseline wander and muscle noises, which are caused by body movement, and powerline interference that occurs in the power supply (50 or 60Hz) [11]. The 50Hz power supply is used in Europe, Asia, Australia and New Zealand, whereas 60Hz is used in the North and South America. The Neurokit2 toolbox with its default settings removes the baseline wandering in the raw signal below 0.5Hz using a 5th order high-pass Butterworth filter. A powerline smoothing algorithm is used with a moving average kernel. The window size equals to one period of the powerline frequency, with a default setting of 50Hz.

To obtain nearly consistent ECG data cleaning processes with the MindWare software used in our previous study [12], in this paper we created a custom Python script based on the original Neurokit2 package. A 9th order band-pass Butterworth filter between 0.25 and 40 Hz was applied to remove the baseline and muscle noise, as in

$$H(j\omega)=\frac{1}{1+\left(\frac{\omega}{\omega_0}\right)^{2n}}. \quad (1)$$

Where ω is the radian frequency, ω_0 represents the band-pass frequency and the order n.

A notch filter was then applied to remove powerline noise (in s-domain), as in

$$H(s)=\frac{s^2+\omega_0^2}{s^2+\frac{\omega_0}{Q}s+\omega_0^2}. \quad (2)$$

The band-stop frequency was set to 60Hz after visual inspection using a power spectrum density plot. Where ω_0 is the band-stop frequency, and the quality factor Q is set to 30.

C. R-peak Detection

The Neurokit2 toolbox extracts QRS complexes using the steepness of the absolute gradient of the ECG signal, and the position of the local maxima in the QRS complex was labelled as the R-peaks, which is consistent between our methods and the default settings in Neurokit2.

D. Feature Extraction

From the detected R-peaks, the IBI series representing the mean of the time interval between two heartbeats was calculated. A pool of time domain components of HRV was derived from the IBI series.

Time domain features were analyzed to reflect HRV over consecutive R-R intervals. In this study, we report MeanNN (mean of R-R intervals in milliseconds), SDNN (standard deviation of R-R intervals in milliseconds), RMSSD (root mean square of successive R-R interval differences), NN50 (the number of pairs of successive R-R intervals that differ by more than 50ms) and pNN50 (the proportion of successive R-R intervals that are larger than 50ms) metrics [13].

E. Statistical Analysis

Time domain feature analysis was conducted using SPSS (version 28). Descriptive statistics and One-Sample Kolmogorov-Smirnov Tests were first conducted to assess data normality. The data were found not to be normally distributed using a two-tailed t-test (p-value < 0.05). Two one-sided tests (TOST) [14, 15] were conducted to test equivalences between the two signal processing methods implemented on the Neurokit2 toolbox (Neurokit2 default settings, and modified settings conforming to the manual labelling tool, MindWare) and the ground truth labels provided by the QT database. The null hypothesis was that the mean differences of the two signal processing approaches were not equivalent, and that the observed true differences were outside of the equivalence range. A sensitivity analysis was conducted using GPower (version 3.1) to determine the smallest effect size of interest (SESOI) [15] for the equivalence bounds. The

alpha level was set to 0.05. The "TOSTtwo" function in R (version 4.2.2) was applied to conduct the TOST test.

III. Results and Discussion

A. ECG Signal Pre-processing and R-peak Detection

All ECG recordings included in this study were filtered through two signal processing methods separately (that is, the Neurokit2 default method and our modified Neurokit2 signal processing approach similar to the MindWare software). As the databases were collected from both the U.S. (the normal sinus rhythm group) and European research sites (the ST Change subsets), powerline noise may differ between sites. To apply a correct powerline noise filtering algorithm using our modified Neurokit2 toolbox, power spectrum density of the original ECG signal was first visualized. Fig. 1(a) shows an example where a 60Hz powerline reference was recorded in the original ECG signal. In our modified Neurokit2 toolbox, a powerline interference at 60Hz and motion artifacts were filtered, resulting in a signal with less noise (Fig. 1(b)). The Neurokit2's default signal processing approach applied powerline smoothing at 50Hz.

B. Between Group Differences in HRV metrics

A total of 506 annotated R-peaks from the 16 selected ECG recordings provided by the QT database were used to calculate time domain HRV metrics and consider it as the ground truth labels. Corresponding time domain HRV metrics calculated from the two signal processing methods applying Neurokit2 toolbox were also calculated. Normality of these calculated metrics were tested using the One-Sample Kolmogorov-Smirnov Test. The SDNN and RMSSD values from the NSR dataset, as well as the SDNN value from the ST change dataset were found significant (p-value < 0.05), indicating that the assumption of a normal distribution of the data was violated.

The TOST tests revealed that all time-domain HRV metrics were statistically significant (p-value < 0.05), suggesting equivalences between the time domain HRV metrics of the QT database (NSR and ST change datasets) and the two automated signal processing approaches using Neurokit2, as shown in Table I-IV. However, the mean differences for some time-domain HRV metrics were large (for instance, the RMSSD and NN50 features in both datasets), may be due to the limitations of the small sample size. Therefore, these tests should be repeated on a larger dataset to report the reliability of HRV metrics calculations by the described methods in ECG signals recorded in social interaction settings. These differences may also stem from the fact that the ECG acquisition devices integrate digital filters before recording the data and, therefore, no significant powerline noise is recorded in the ECG signal. The signal may have been over-filtered after the application of a 50Hz/60Hz notch filter (Fig. 1(c) and Fig. 1(d)). Further investigation would be needed to determine other potential causes for the observed differences in time-domain HRV measures between the two algorithms.

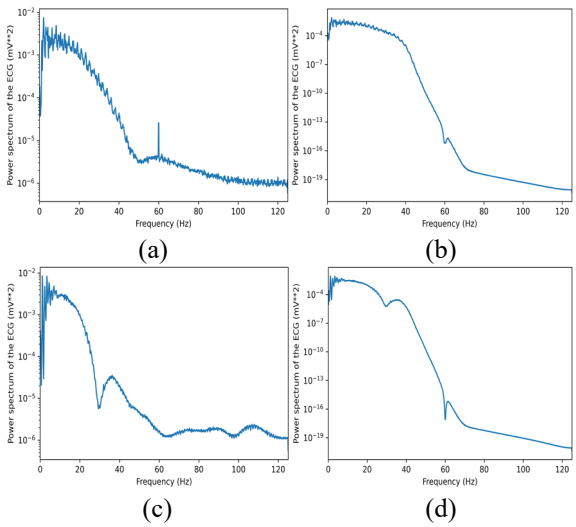


Figure 1. Power spectrum density plots. (a) Before signal processing (powerline noise at 60 (Hz)); (b) after signal processing using our modified Neurokit2 toolbox; (c) before signal processing (no powerline noise evident), and (d) showing an over-processed signal.

TABLE I. Two one-sided tests for Equivalence Between Time Domain Heart Rate Variability Metrics Calculated By The QT Databases And The Neurokit2 Toolbox With Default Settings In The Normal Sinus Rhythm Group

Normal Sinus Rhythm	**QT – Neurokit2 Default**			
	Mean Diff (SE)	***p-value (lower/upper bound)***	***TOST [90% CI]***	***Equivalence Bounds***
MeanNN	0.072 (0.096)	0.010/ 0.011	[-91.516, 91.660]	[-133.531, 133.531]
SDNN	-0.031 (0.330)	0.011/ 0.010	[-33.941, 33.879]	[-49.439, 49.439]
RMSSD	0.227 (0.606)	0.010/ 0.011	[-34.380, 34.834]	[-50.456, 50.456]
NN50	4.700 (1.545)	0.009/ 0.012	[-148.896, 158.296]	[-223.935, 223.935]
pNN50	0.453 (0.139)	0.010/ 0.011	[-18.092, 18.998]	[-27.038, 27.038]

TABLE II. Two one-sided tests for Equivalence Between Time Domain Heart Rate Variability Metrics Calculated By The QT Databases And The Neurokit2 Toolbox With Our Modified Settings In The Normal Sinus Rhythm Group

Normal Sinus Rhythm	**QT – Neurokit2 Ours**			
	Mean Diff (SE)	***p-value (lower/upper bound)***	***TOST [90% CI]***	***Equivalence Bounds***
MeanNN	-0.020 (0.018)	0.011/ 0.011	[-91.595, 91.554]	[-133.511, 133.511]
SDNN	0.336 (0.084)	0.010/ 0.011	[-33.597, 34.27]	[-49.474, 49.474]
RMSSD	0.944 (0.310)	0.010/ 0.012	[-33.755, 35.644]	[-50.590, 50.590]
NN50	5.500 (2.669)	0.009/ 0.012	[-148.755, 159.755]	[-224.896, 224.896]
pNN50	0.529 (0.257)	0.009/ 0.012	[-18.097, 19.156]	[-27.157, 27.157]

TABLE III. TWO ONE-SIDED TESTS FOR EQUIVALENCE BETWEEN TIME DOMAIN HEART RATE VARIABILITY METRICS CALCULATED BY THE QT DATABASES AND THE NEUROKIT2 TOOLBOX WITH DEFAULT SETTINGS IN THE ST CHANGE GROUP

ST Change	**QT – Neurokit2 Default**			
	Mean Diff (SE)	***p-value (lower/upper bound)***	***TOST [90% CI]***	***Equivalence Bounds***
MeanNN	-0.058 (0.132)	0.010/ 0.010	[-222.298, 222.182]	[-335.539, 335.539]
SDNN	-0.645 (0.489)	0.011/ 0.010	[-45.814, 44.523]	[-68.196, 68.196]
RMSSD	-6.310 (2.233)	0.019/ 0.006	[-39.159, 26.540]	[-49.597, 49.597]
NN50	1.833 (4.028)	0.010/ 0.011	[-90.329, 93.996]	[-139.148, 139.148]
pNN50	0.264 (0.366)	0.010/ 0.011	[-9.800, 10.327]	[-15.194, 15.194]

TABLE IV. TWO ONE-SIDED TESTS FOR EQUIVALENCE BETWEEN TIME DOMAIN HEART RATE VARIABILITY METRICS CALCULATED BY THE QT DATABASES AND THE NEUROKIT2 TOOLBOX WITH OUR MODIFIED SETTINGS IN THE ST CHANGE GROUP

ST Change	**QT – Neurokit2 Ours**			
	Mean Diff (SE)	***p-value (lower/upper bound)***	***TOST [90% CI]***	***Equivalence Bounds***
MeanNN	-0.148 (0.201)	0.011/ 0.010	[-222.381, 222.085]	[-335.529, 335.529]
SDNN	-0.659 (0.522)	0.011/ 0.010	[-45.857, 44.538]	[-68.239, 68.239]
RMSSD	-6.538 (2.387)	0.019/ 0.006	[-39.794, 26.718]	[-50.210, 50.210]
NN50	1.167 (3.643)	0.010/ 0.011	[-91.975, 94.308]	[-140.626, 140.626]
pNN50	0.167 (0.331)	0.010/ 0.011	[-10.019, 10.352]	[-15.377, 15.377]

IV. CONCLUSIONS

In this paper, we presented a modified pipeline for automated ECG signal processing and HRV metrics calculation using the Neurokit2 toolbox. The modified algorithm is similar to the MindWare HRV analysis tool used for manual cleaning and manual inspection of R-peak detection in social interaction studies. Current algorithms developed for automatic R-peak detection and heart rate variability calculation performed well for 2-lead electrode ECG signals, accounting for issues such as motion artifacts, and irregular ECG patterns. Our results show that time domain HRV metrics calculated by the Neurokit2 toolbox are statistically equivalent to the annotated QT database. It is therefore feasible to calculate HRV metrics on prolonged ECG recordings in the social interaction context. The practical significance and impact of the differences in HRV measures in clinical medicine should be explored in future studies.

ACKNOWLEDGMENT

The authors would like to thank Andrew Goodwin for the assistance during the study. The authors would like to thank Alex Shaw at the Sydney Informatics Hub for statistical consultations.

REFERENCES

[1] M. Malik et al., "Heart rate variability: Standards of measurement, physiological interpretation, and clinical use," European heart journal, vol. 17, no. 3, pp. 354-381, 1996, doi: 10.1093/oxfordjournals.eurheartj.a014868.

[2] M. Malik, "Heart rate variability: Standards of measurement, physiological interpretation, and clinical use: Task force of the European Society of Cardiology and the North American Society for Pacing and Electrophysiology," Annals of Noninvasive Electrocardiology, vol. 1, no. 2, pp. 151-181, 1996.

[3] G. Perna, A. Riva, A. Defillo, E. Sangiorgio, M. Nobile, and D. Caldirola, "Heart rate variability: Can it serve as a marker of mental health resilience?: Special Section on "Translational and Neuroscience Studies in Affective Disorders" Section Editor, Maria Nobile MD, PhD," Journal of Affective Disorders, vol. 263, pp. 754-761, 2020, doi: 10.1016/j.jad.2019.10.017.

[4] O. Kwon et al., "Electrocardiogram sampling frequency range acceptable for heart rate variability analysis," Healthcare informatics research, vol. 24, no. 3, pp. 198-206, 2018, doi: 10.4258/hir.2018.24.3.198.

[5] F. A. Oliveira Junior et al., "Different acquisition systems for heart rate variability analysis may lead to diverse outcomes," Braz J Med Biol Res, vol. 55, p. e11720, 2022, doi: 10.1590/1414-431X2021e11720.

[6] G. R. H. Sandercock, P. D. Bromley, and D. A. Brodie, "The reliability of short-term measurements of heart rate variability," International Journal of Cardiology, vol. 103, no. 3, pp. 238-247, 2005/09/01/ 2005, doi: 10.1016/j.ijcard.2004.09.013.

[7] S. Shahrestani, E. M. Stewart, D. S. Quintana, I. B. Hickie, and A. J. Guastella, "Heart rate variability during social interactions in children with and without psychopathology: a meta-analysis," J. Child Psychol. Psychiatry, vol. 55, no. 9, pp. 981-989, SEPT 2014, doi: 10.1111/jcpp.12226.

[8] S. Shahrestani, E. M. Stewart, D. S. Quintana, I. B. Hickie, and A. J. Guastella, "Heart rate variability during adolescent and adult social interactions: A meta-analysis," Biological Psychology, vol. 105, pp. 43-50, FEB 2015, doi: 10.1016/j.biopsycho.2014.12.012.

[9] D. Makowski et al., "NeuroKit2: A Python toolbox for neurophysiological signal processing," Behavior Research Methods, pp. 1-8, 2021, doi: 10.3758/s13428-020-01516-y.

[10] P. Laguna, R. G. Mark, A. Goldberg, and G. B. Moody, "A database for evaluation of algorithms for measurement of QT and other waveform intervals in the ECG," in Computers in cardiology, 1997: IEEE, pp. 673-676, doi: 10.1109/CIC.1997.648140.

[11] B. Singh, P. Singh, and S. Budhiraja, "Various Approaches to Minimise Noises in ECG Signal: A Survey," in 2015 Fifth International Conference on Advanced Computing & Communication Technologies, 21-22 Feb. 2015 2015, pp. 131-137, doi: 10.1109/ACCT.2015.87.

[12] R. Thapa et al., "Reduced heart rate variability in adults with autism spectrum disorder," Autism Res, vol. 12, no. 6, pp. 922-930, Jun 2019, doi: 10.1002/aur.2104.

[13] D. S. Quintana, G. A. Alvares, and J. A. Heathers, "Guidelines for Reporting Articles on Psychiatry and Heart rate variability (GRAPH): recommendations to advance research communication," Transl Psychiatry, vol. 6, p. e803, May 10 2016, doi: 10.1038/tp.2016.73.

[14] D. Lakens, "Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses," Soc Psychol Personal Sci, vol. 8, no. 4, pp. 355-362, May 2017, doi: 10.1177/1948550617697177.

[15] D. Lakens, A. M. Scheel, and P. M. Isager, "Equivalence Testing for Psychological Research: A Tutorial," Advances in Methods and Practices in Psychological Science, vol. 1, no. 2, pp. 259-269, 2018/06/01 2018, doi: 10.1177/2515245918770963.