Toddler mental development interventions: Can machine learning play a part?

A thesis submitted in fulfillment of the requirements

for the degree of Interdisciplinary Dual Degree Programme

by

Akshat Gautam

190110004

under the guidance of

Prof. Sharat Chandran

to the

Centre for Machine Intelligence and Data Science

Indian Institute of Technology Bombay

June 2023

Page intentionally left blank

ABSTRACT

Name of student: **Akshat Gautam** Roll no: **190110004**

Degree for which submitted: **Interdisciplinary Dual Degree Programme**

Department: **Centre for Machine Intelligence and Data Science**

Thesis title: **Toddler mental development interventions: Can machine learning play a part?**

Name of Thesis Supervisor: **Prof. Sharat Chandran**

Month and year of thesis submission: **June 2023**

Monitoring physical development and detection of ailments in infants, toddlers, and very young children is critical for parents and is reasonably well understood. On the other hand, monitoring mental development is challenging, and may require the administration of 'tests' by trained professionals, and thus expensive. This is particularly difficult for low-income citizens in countries like India and Malawi, resulting in a "detection gap" in these regions. A tablet-based assessment easily administered at home by non-specialist workers may then be considered.

The main contribution of this project is the development of a machine learning mechanism to generate *developmental scores* across three principal domains: motor, social, and cognitive. As a first step, we convert assessment data created by non-specialist workers into meaningful features that capture a child's performance. This involves the use of, and adjustments of data analysis algorithms and computer vision algorithms to 'child settings'. We then predict scores in a supervised setting based on expensive and time-consuming labeled data from a psychometric test known as the Griffith Mental Development Scale. As an alternative, we employ Item Response Theory to generate developmental scores in an unsupervised setting.

In conclusion, we hope that the tablet-based tasks could be used to produce developmental scores, leading the way to a standardized method for monitoring mental development.

Acknowldgements

First and foremost, I would like to express my deepest gratitude to my guide, Prof. Sharat Chandran, for his unwavering support and invaluable guidance throughout this research. This work would not have been possible without his frequent meetings, which helped me understand the complexities of a project of this magnitude, and the countless hours of discussion. His mentorship has been instrumental not only for this academic endeavor but also for my future career.

I would also like to extend my thanks to Prof. Bhismadev Chakrabarti and the entire STREAM team. This project was made possible by the dedicated data collection team, who spent countless hours in the field gathering data. The constant discussions and inputs during the fortnightly meetings with the team were crucial at every step of the project.

I am also very thankful to Shubham Chitnis, who provided significant assistance with both code and data collection from a distance. His constant inputs and advice greatly contributed to my work. Lastly, I would like to thank my family and friends, for their unwavering support and motivation.

Akshat Gautam

Honor Code

I, *Akshat Gautam,190110004* declare that this written submission represents my ideas in my own words, and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I declare that I have properly and accurately acknowledged all sources used in the production of this report. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented, fabricated, or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the Institute and can also evoke penal action from the sources that have thus not been properly cited or from whom proper permission has not been taken when needed.

Akshat Gautam

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Continuous monitoring of both physical and mental development in toddlers and young children is essential. Early interventions for children experiencing slow development can significantly impact their functioning later in life. Organizations like the WHO provide standardized procedures and methods for tracking physical health. For example, growth charts, such as height and weight charts, allow parents or caregivers to measure a child's height and weight and calculate z-scores using these standardized charts. These z-scores serve as a means to track physical development. Additionally, tools like the WHO's anthropometric survey offer a set of z-scores targeting various areas of physical development using measurements such as head circumference, skinfold thickness, and more.

In contrast, monitoring mental development is not as straightforward. It typically involves administering time-intensive and expensive psychometric tests conducted by trained mental health professionals. Unfortunately, these tests are often administered too late, as parents or caregivers only seek them out after noticing atypical symptoms. This reliance on mental health professionals is a significant issue in low-income countries like India and Malawi. The lack of resources and accessibility to these tests in these regions results in a "detection gap."

Nationwide findings have revealed significant proportions of children in India between the ages of 2 and 9 years affected by one or more neurodevelopmental disorders (NDD): 10% in hilly areas, 13% in urban areas, and 18% in rural areas [11]. With such a large number of children living in remote areas, it is imperative to develop a tool that can help bridge this detection gap. STREAM (Scalable Transdiagnostic Early Assessment of Mental Health) is an open-source tablet-based assessment that can be administered by non-specialist workers to determine mental development in areas such as social, motor, and cognitive skills, thus paving the way for early interventions.

1.2 Introduction to STREAM

STREAM is a digital tablet-based platform designed to measure social, motor, and cognitive development, which can be administered by non-specialist workers. STREAM integrates two previously proven assessment tools, START and DEEP. However, unlike the earlier versions of these tools, the aim of the STREAM project is not merely to serve as a screening tool but to generate standardized developmental scores across various developmental domains.

START is an open-source tablet-based assessment tool comprising a set of tasks designed to measure development primarily in the social and motor domains. This tool includes multiple tasks, whose details will be explained in the upcoming chapters. Originally, START was designed to screen for autism in children aged 2.5 to 6 years. The work in this thesis will be focused on START.

DEEP is a tablet-based platform that features various gamified tasks aimed at understanding and measuring cognitive development. It includes a set of 14 games that focus on different areas of cognitive development.

1.3 Problem Description

The output generated after a child undergoes assessment consists of raw data in the form of videos, images, and Excel files. This data cannot be directly used for any analysis. Hence, the first challenge this work addresses is converting raw data into appropriate features that capture a child's performance in each task. Each task requires different processing methodologies, involving various computer vision and data analysis techniques, and generates one or more features for each task.

The overall aim of the STREAM project is to generate developmental scores. Therefore, the next challenge is to generate these scores using the extracted features. We demonstrate that machine learning can transform these features into developmental scores by using psychometric test scores as supervised labels.

However, as mentioned in earlier sections, obtaining psychometric scores is expensive and time-consuming. Thus, we aim to avoid relying on psychometric tests to generate scores. We address this problem by using Item Response Theory to generate scores in an unsupervised paradigm. Overall, this study explores the process of transitioning from tablet-based assessments to developmental scores.

1.4 Contribution

The main contributions of this thesis are as follows:

- Development of pre-processing scripts in R and Python to process raw data using various techniques and generate meaningful features.
- Generation of developmental scores using a machine learning model with psychometric scores as supervision.
- Application of Item Response Theory to generate developmental scores without relying on psychometric scores.

Chapter 2

Background and Related Work

We will first examine the various tasks in the START component of the STREAM project. Each task will be briefly described, including the features extracted from it and the developmental domain it targets. Next, we will discuss the details of the various psychometric tests used in this study. Following that, we will cover the basics of Item Response Theory (IRT). Finally, we will explore the various methods used to extract distance from the camera using video data, as this will form the basis of feature extraction in the wheel task.

Task	**Short Description**	**Domain**	**Feature Extracted**
Preferential Looking Task	Social and non-social videos presented on a split screen	Social	Proportion of time looking at social video
Button Task	Two buttons appear, one shows social video when clicked, the other non-social	Social	% social, % non-social
Wheel Task	A black and white wheel appears on the screen	Social	Min, Max, Median, and SD of distance from camera; Proportion of total trial completed
Motor Following Task	A butterfly follows a trajectory along the screen which the child traces	Motor	RMSE, jerk, acceleration, speed, weighted frequency gain
Bubble Popping Task	A series of bubbles appear on the screen to be popped	Motor	Distance of touch from bubble center
Coloring Task	An outline of a car which the child colors	Motor	Number of crossovers, points inside/outside
Synchrony Task	Child matches the drum beat appearing on screen by tapping	Motor	Frequency of tapping
Delayed Gratification Task	A star appears; if the child waits, they get 3 stars	Cognition	Proportion of time waited; Proportion of frames face visible

Table 2.1: Description of tasks, their domains, and features extracted in the STREAM project

2.1 START

Screening **T**ools for **A**utism **R**isk using **T**echnology (START) [3] is an open-source autism screening tool that was extensively tested in the Delhi-NCR region. 2.1 gives the detail description of various tasks with their target domains and extracted features.

The original aim of the START paper was to classify children into ASD (autism spectrum disorder), TD (typically developing), and NDD (neurodevelopmental disorder) using the START tablet-based assessment. However, STREAM shifts its focus from being a screening tool to generating developmental scores for standardizing mental development.

2.2 Psychometric Tests

As mentioned, the current method for monitoring mental development involves administering psychometric tests by trained professionals. These tests are recognized as a standard and scientific method used to assess individuals' mental capabilities and behavioral styles. Typically, these tests adhere to specific protocols that must be followed by the administering professional. However, these tests can be costly due to licensing fees, the expense of professional services, and other associated costs. In the upcoming sections, we will focus on two specific tests: GMDS and MDAT.

2.2.1 Griffith's Mental Development Scale (GMDS)

(a) (b)

Figure 2.1: (a)Assessor taking GMDS test (b)Sample GMDS kits

GMDS is widely recognized as the gold standard for measuring mental development in young children aged 0-6 years. In medicine and medical statistics, the gold standard, criterion standard, or reference standard refers to the diagnostic test or benchmark that is considered the best available under reasonable conditions. The latest version of this test, Griffiths 3, was launched in 2016 and comprises 321 items. This test is given by certified professional in a set environment Fig. 2.1a Each item consists of tasks that the child is asked to perform, with outcomes recorded as pass or

fail. Some items may require certain objects, thus GMDS test requires a kit Fig. 2.1b. GMDS uses these items to assess mental development across five different domains. Table 2.2 provides a brief description of these domains and the number of items assessed in each.

Domain	**Description**	**No of Items**
Foundations of learning	Assesses critical aspects of learning during early childhood	63
Language and communication	Measures overall language development, including expressive and receptive language, and social communication	63
Eye and hand coordination	Considers fine motor skills, manual dexterity, and visual perception	67
Personal-social-emotional	Measures sense of self, independence, social interactions, and emotional development	65
Gross motor	Assesses postural control, balance, and gross body coordination	63

Table 2.2: GMDS Domains and Item Counts

Given the 321 items across the five domains, it would be challenging for a child to complete all tasks. Therefore, the procedure of the tests is designed to streamline the process. The test begins at a level specified by the manual, starting with items of appropriate difficulty. The difficulty increases until the child continuously fails five items, referred to as the 'ceiling.' All items more difficult than this level are automatically marked as failed. Conversely, the level is decreased until the child consecutively passes five items, known as the 'basal.' All levels below this are marked as passed. The total number of passes in a domain is known as the domain score. This procedure is applied to all five domains. Sample GMDS output scores for three children are provided in Table 2.3.

ChildID	**Domain A**	**Domain B**	**Domain C**	**Domain D**	**Domain E**
MW-0113	33	43	44	53	47
IN-1653	49	51	52	56	60
IN-1682	31	43	46	45	46

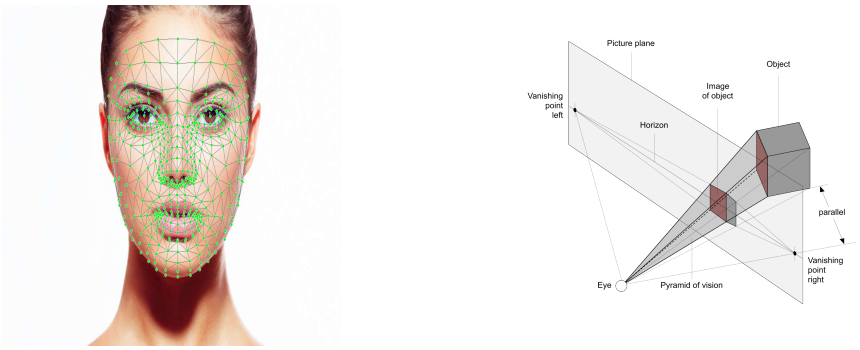
Table 2.3: Sample GMDS Output Scores for 3 Children

2.2.2 Malawi Development Assessment Tool(MDAT)

MDAT is a psychometric test similar to GMDS but designed for low-income countries like Malawi and India. The language and items in the test are tailored to better suit children residing in these regions. MDAT consists of 136 items, divided equally across four domains: Gross Motor, Fine Motor, Language, and Social, with 34 items in each domain

2.3 Face Mesh

An important term that the reader will frequently encounter in this work is facial landmarks. Landmarks are points of interest like eye margins, center of the nose, elbow and shoulder joints, etc. The skeleton of such landmarks can prove to be a suitable proxy for certain applications where privacy is essential. We deal with face-mesh, which is a dense representation of a typical human face fitted on the subject in consideration. There are multiple off-the-shelf face-mesh detection tools available for getting this representation. The mediapipe face-mesh detection tool is particularly suited for our application. Performance wise, the detection system is able to capture faces well in varying lighting conditions, for videos captured in both lab and in-the-wild settings. In total, 468 facial landmarks are captured as (x, y, z) triplets. Here x and y values are normalized to the pixel space with 0,0 being the bottom-left point and 1,1 the top-right one. The normalized z coordinate is a little tricky as it is not the actual depth. Instead, the depth values are transformed to match the x coordinate scale. The choice of scale for the z axis makes the conversion to metric distances non-trivial.



(a) (b)

Figure 2.2: (a)Medipipe Facemesh (b)Pyramid of vision

Given the normalized landmarks, we employ a view frustum transformation made available by the mediapipe team to calculate metric landmarks. At the end of this, we will have coordinates that are distanced from each other as they would in the real world. The question is what origin are these coordinates in reference to. We believe the x and y coordinate origin is at the center of the imaging window. The z coordinate however, is aligned with the middle of the subject's skull. The convention followed by mediapipe is such that the closer a landmark is to the image plane, its z values will be lesser (not the absolute value, but the actual value). Consecutive frames from the video are passed through this system to generate an array of metric coordinates with respect to some origin and corresponding points in the pixel space.

2.4 Perspective-n-Point

It is essential that we move from an arbitrary origin that changes across frames when using facemesh to something more consistent. As our final goal is to calculate distance from the camera, it makes sense to have the camera as the origin. We have multiple face landmark coordinates (3D) with respect to a coordinate system and their pixel coordinates (2D) on the captured image. Our task is to find a matrix that transforms a point in the original world coordinate system to the camera coordinate system. A classic algorithm that deals with this is the Perspective-n-Point (PnP) algorithm. The original work by [4] coined the term Perspective-n-Point, an algorithm to obtain camera pose with n known 3D-2D correspondence. More recent literature target iterative and non-iterative solutions for the general PnP problem or propose solutions for a subset (P3P, P4P). [8] provide details about the PnP landscape in their work. There are tradeoffs involved in choosing between iterative and non-iterative techniques. Iterative methods tend to be more accurate because of the refinement steps involved in their pipeline; however, their convergence is highly dependent on the choice of initial guess. Closed-form solutions offer speed and simplicity but often involve assumptions that might not hold in specific scenarios. We use the Efficient PnP (EPnP) algorithm, in specific opencv's solvepnp(fig 2.3) based on the work by [6]. In addition to the complexity being O(n), EPnP offers an accurate non-iterative solution for $n \geq 4$ correspondence pairs. However, a pinhole camera model is enforced, and the camera's intrinsic parameters are required. Additionally, the original implementation does not take distortion into account and expects points to be sufficiently far from the camera.

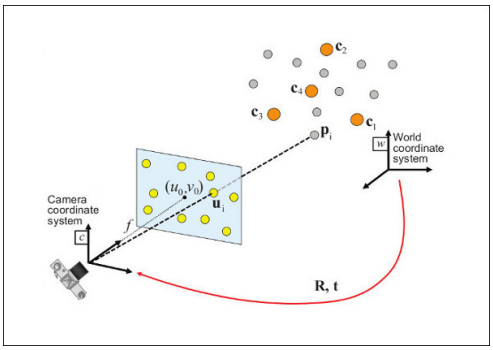


Figure 2.3: Rotation and Translation matrix by PnP algorithm

2.5 Related Work: Distance from camera

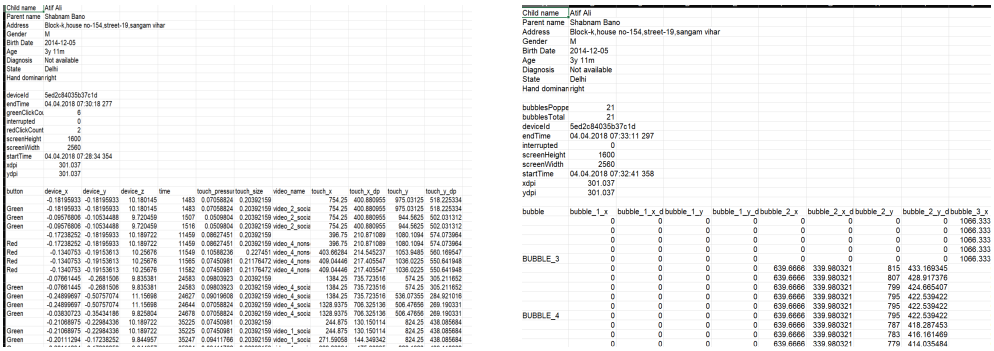
The problem of estimating distance of face from the camera using video and images has been extensively studied. Numerous works address this challenge by calculating anatomical features using facial landmarks[10, 9]. Image pyramids and template matching are implemented by [9] to

perform face and eye detection. An empirical formula is calculated relating distance between the eyes and distance from the camera. Similarly, [10] relate faces sizes extracted by the Viola Jones detector with the camera distance. [1] harness Mediapipe face landmarks to get the iris radius and find a logarithmic correlation between the radius and camera distance. All three methods are based on the assumption that the same camera is used for experimentation. This doesn't align with what we see in the real world, where different cameras will project these anatomical features differently on the image plane causing any and all real world distances to get altered. Additionally, their findings rely on the premise of similar face shape and size across the populace. However we know that this is not the case, especially when working with individuals across various age groups, including infants and adults. These empirical relations will also fail when dealing with non-frontal faces, which are commonly encountered. Specific methodologies also leverage deep learning techniques for distance calculation. [2] use a VGG-16 network pre-trained on ImageNet for transfer learning on images with known distances. Their method suffers with the inability to use the same model across different camera focal lengths. Furthermore, their dataset exhibits limitations, encompassing a relatively narrow age group representation, specific lighting conditions, and ethnicities. These limitations could introduce bias into the model's performance. PnP methods have been used previously in the literature for camera distance calculation. [5] perform distance estimation on unseen faces using EPnP. For a test image, they use its 2D facial (fiducial) landmarks and match those with 3D landmarks from exemplar faces to get an approximate distance measure. Similarly, [7] estimate head poses for owls, chameleons, and humans using an iterative variant of PnP.

Chapter 3

Pre-processing

This chapter is dedicated to the preprocessing of START data, aimed at generating metrics suitable for machine learning and statistical analysis. After a child completes a game on the tablet, raw data is stored as xlsx files. These xlsx files are processed by R scripts. The upcoming sections will detail the processing steps for each of these tasks.



Child name	Atif Ali
Parent name	Shabnam Bano
Address	Block-k,house no-154,street-19,sangam vihar
Gender	M
Birth Date	2014-12-05
Age	3y 11m
Diagnosis	Not available
State	Delhi
Hand dominan	right
deviceId	5ed2c84035b37c1d
endTime	04.04.2018 07:30:18.277
greenClickCo	6
interrupted	0
redClickCount	2
screenHeight	1600
screenWidth	2560
startTime	04.04.2018 07:28:34.354
xdpi	301.037
ydpi	301.037

button	device_x	device_y	device_z	time	touch_pressu	touch_size	video_name	touch_x	touch_x_dp	touch_y	touch_y_dp
	-0.18195933	-0.18195933	10.180145	1483	0.07058824	0.20392159		754.25	400.880955	975.03125	518.225334
Green	-0.18195933	-0.18195933	10.180145	1483	0.07058824	0.20392159	video_2_socia	754.25	400.880955	975.03125	518.225334
Green	-0.09576806	-0.10534488	9.720459	1507	0.0509804	0.20392159	video_2_socia	754.25	400.880955	944.5625	502.031312
Green	-0.09576806	-0.10534488	9.720459	1516	0.0509804	0.20392159	video_2_socia	754.25	400.880955	944.5625	502.031312
	-0.17238252	-0.18195933	10.189722	11459	0.08627451	0.20392159		396.75	210.871089	1080.1094	574.073864
Red	-0.17238252	-0.18195933	10.189722	11459	0.08627451	0.20392159	video_4_nons	396.75	210.871089	1080.1094	574.073864
Red	-0.1340753	-0.19153613	10.25676	11549	0.10588236	0.227451	video_4_nons	403.66284	214.545237	1053.9485	560.169547
Red	-0.1340753	-0.19153613	10.25676	11565	0.07450981	0.21176472	video_4_nons	409.04448	217.405547	1036.0225	550.641948
Red	-0.1340753	-0.19153613	10.25676	11582	0.07450981	0.21176472	video_4_nons	409.04448	217.405547	1036.0225	550.641948
	-0.07661445	-0.2681506	9.835381	24583	0.09803923	0.20392159		1384.25	735.723516	574.25	305.211652
Green	-0.07661445	-0.2681506	9.835381	24583	0.09803923	0.20392159	video_4_socia	1384.25	735.723516	574.25	305.211652
Green	-0.24899697	-0.50757074	11.15698	24627	0.09019608	0.20392159	video_4_socia	1384.25	735.723516	536.07355	284.921016
Green	-0.24899697	-0.50757074	11.15698	24644	0.07058824	0.20392159	video_4_socia	1328.9375	706.325136	506.47656	269.190331
Green	-0.03830723	-0.35434186	9.825804	24678	0.07058824	0.20392159	video_4_socia	1328.9375	706.325136	506.47656	269.190331
	-0.21068975	-0.22984336	10.189722	35225	0.07450981	0.20392159		244.875	130.150114	824.25	438.085684
Green	-0.21068975	-0.22984336	10.189722	35225	0.07450981	0.20392159	video_1_socia	244.875	130.150114	824.25	438.085684
Green	-0.20111294	-0.17238252	9.844957	35247	0.09411766	0.20392159	video_1_socia	271.59058	144.349342	824.25	438.085684

(a)

Child name	Atif Ali
Parent name	Shabnam Bano
Address	Block-k,house no-154,street-19,sangam vihar
Gender	M
Birth Date	2014-12-05
Age	3y 11m
Diagnosis	Not available
State	Delhi
Hand dominan	right
bubblesPoppe	21
bubblesTotal	21
deviceId	5ed2c84035b37c1d
endTime	04.04.2018 07:33:11.297
interrupted	0
screenHeight	1600
screenWidth	2560
startTime	04.04.2018 07:32:41.358
xdpi	301.037
ydpi	301.037

bubble	bubble_1_x	bubble_1_x_d	bubble_1_y	bubble_1_y_d	bubble_2_x	bubble_2_x_d	bubble_2_y	bubble_2_y_d	bubble_3_x
	0	0	0	0	0	0	0	0	1066.333
	0	0	0	0	0	0	0	0	1066.333
	0	0	0	0	0	0	0	0	1066.333
	0	0	0	0	0	0	0	0	1066.333
BUBBLE_3	0	0	0	0	0	0	0	0	1066.333
	0	0	0	0	639.6666	339.980321	815	433.169345	
	0	0	0	0	639.6666	339.980321	807	428.917376	
	0	0	0	0	639.6666	339.980321	799	424.665407	
	0	0	0	0	639.6666	339.980321	795	422.539422	
	0	0	0	0	639.6666	339.980321	795	422.539422	
BUBBLE_4	0	0	0	0	639.6666	339.980321	795	422.539422	
	0	0	0	0	639.6666	339.980321	787	418.287453	
	0	0	0	0	639.6666	339.980321	783	416.161469	
	0	0	0	0	639.6666	339.980321	779	414.035484	

(b)

Figure 3.1: (a)Sample input for button task (b)Sample input for bubble task

3.1 Button-Task

The Button task involves displaying two buttons on the tablet screen. One button, when pressed, plays a social video, while the other plays a non-social video. Figure 3.1a provides a sample input XLSX file used in the R script.

ChildID	Social	Non-social	Interrupt	Trials	Soc_prop
151	4	4	0	8	0.5
152	5	3	0	8	0.625
153	6	2	0	8	0.75

Table 3.1: Sample pre-processed output of button task

The script begins by determining which button (red or green) is linked to the social video. This information is then utilized to calculate the number of social and non-social clicks and subsequently determine the proportion of social clicks. The sample output for the button task after processing is shown in the table 3.1.

3.2 Bubble-Task

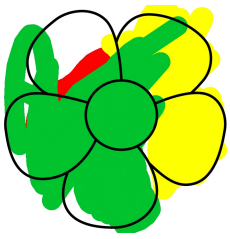
In this task, bubbles appear on the screen over time, and the child is supposed to pop the bubbles by touching them. The tablet device records the force applied on the screen. Refer to Figure 3.1b for sample input.

ChildID	Mean Force	Mean DisX	Mean DisY	Interrupt	Bubble Popped
151	0.2066	31.7619	55.2857	0	21
152	0.2085	30.5238	35.8571	0	21
153	0.1719	23.1429	49.3333	0	21

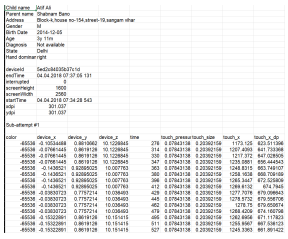
Table 3.2: Sample pre-processed output of bubble task

R script takes the mean of the force applied and calculates the mean distance between the child's touch point and the bubble's center. The output after pre-processing should look like table 3.2

3.3 Colouring-Task



(a) Subfigure A



(b) Subfigure B

Figure 3.2: (a)Coloured flower by a child (b)Sample input for colour task

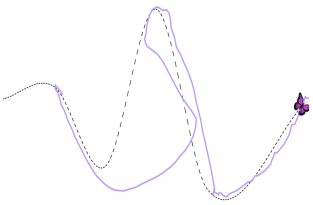
During the Coloring Task, children are presented with simple outline figures and are instructed to carefully fill them with colors. This task utilizes both an XLSX file and a colored JPG image as input, as depicted in Figure 2.

Child ID	Interrupted	Points Inside	Points Outside	Crossover Counts	Proportion
477	0	1047.5	26	11	0.536
478	0	1055.5	107	38	0.647

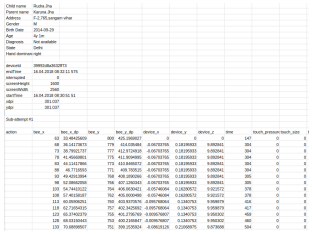
Table 3.3: Sample pre-processed output of colouring task

In the data processing phase, R employs the sp.polygon function to determine the number of times the child crossed over the outline in the figure. Furthermore, the colored image is used to calculate the proportion of the image that has been colored. This is achieved by converting the image into a binary format and subsequently computing the proportion of the colored area. Table 3.3 shows the sample output for this task.

3.4 Motor Following Task



(a) Subfigure A



(b) Subfigure B

Figure 3.3: (a)Child tracing butterfly path (b)Sample input for Motor task

This task involves instructing the child to trace the path of a target butterfly. The target butterfly follows random trajectories with varying velocities in both the x and y axes.

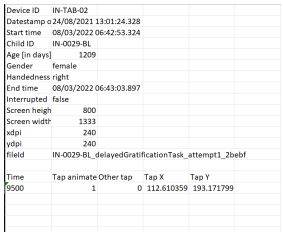
ChildID	rmse_mn	weighted_x_freq_gain_mn	weighted_y_freq_gain_mn	jerk
480	595.98	2.814	37.241	0.020537
484	368.06	1.7563	7.68	0.020204
485	879.18	1.7761	8.3876	0.032486

Table 3.4: Sample pre-processed output of motor following task

To analyze this task, we employ a R script to compute the root mean square error between the child's touch trajectories and those of the butterfly. Additionally, we calculate the jerk of the child's traced path. A Discrete Fourier Transform (DFT) is applied to determine the frequency gain in this context. For sample output, refer to Table 3.4.

3.5 Delayed Gratification Task

(a) Subfigure A



Device ID	IN-TAB-02
Datestamp o	24/08/2021 13:01:24.328
Start time	08/03/2022 06:42:53.324
Child ID	IN-0029-BL
Age [in days]	1209
Gender	female
Handedness	right
End time	08/03/2022 06:43:03.897
Interrupted	false
Screen heigh	800
Screen width	1333
xdpi	240
ydpi	240
fileId	IN-0029-BL_delayedGratificationTask_attempt1_2bebf

Time	Tap animate	Other tap	Tap X	Tap Y
9500	1	0	112.610359	193.171799

(b) Subfigure B

Figure 3.4: (a)DGT task (b)Sample input for DGT task

In this task, a star appears on the screen. Children are instructed to wait until the circle in the middle turns completely blue to earn 3 stars instead of 1. This task involves processing an Excel file and a video recorded through the front camera separately. 3.4 shows sample dgt task screen and sample input xlsx file.

ChildID	dgt proptime	dgt faceprop
480	0.8	1.0
484	1.0	1.0
485	1.0	0.2

Table 3.5: Sample pre-processed output of motor following task

To analyze the Excel file, R scripts read the start and end times of the task and calculate the difference, which gives the total time the child waited. By dividing this by the total task time of 180 seconds, we obtain the proportion of time the child waited.

Mediapipe Facemesh is used to determine the proportion of frames in which the child's face is present. If a frame contains more than one face, it is excluded from our calculation.Refer to table 3.5 for sample processed output.

Chapter 4

Obtaining metric distances

This chapter outlines our proposed methodology for calculating the distance from the camera to the subject's face in metric units. Leveraging the capabilities of the Mediapipe FaceMesh toolkit and the EPnP algorithm, we have devised a robust pipeline to accomplish this task. To substantiate our method's effectiveness, we conducted a series of experiments, which are elaborated upon in this chapter.

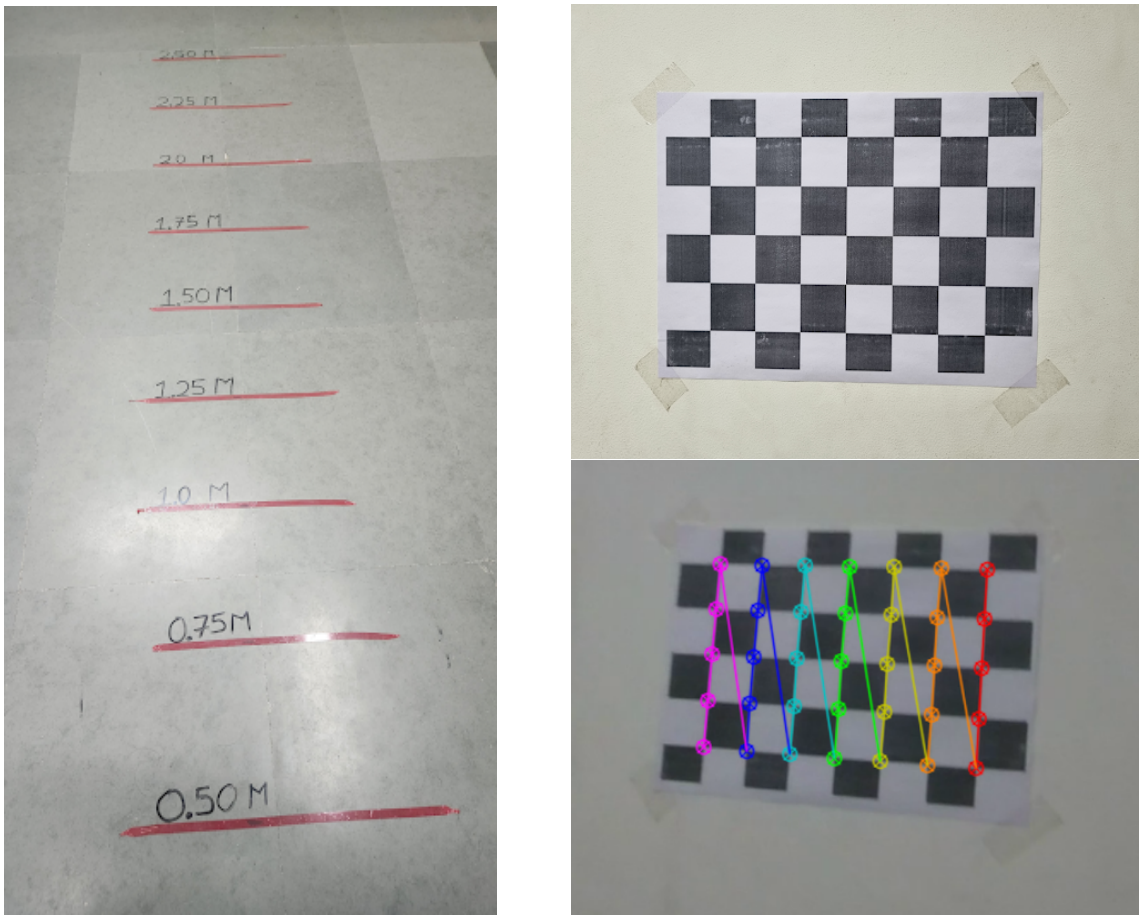
4.1 Method

Our methodology commences with a vital step: camera calibration, which allows us to derive the camera's internal matrix. We captured multiple views of a standard chessboard with the capturing device during this process, as illustrated in Figure 4.1b. To obtain the internal matrix and distortion coefficients, we employed OpenCV's findChessboardCorners and calibrateCamera functions.

We utilized mediapipe's faceMesh to extract 468 normalized landmarks (Xn, Yn, Zn). These coordinates were used in generating metric landmarks (Xm, Ym, Zm) by perspective camera frustum (PCF) transformation. Notably, this transformative step also relies on utilizing the internal camera matrix. The EPnP algorithm was employed for each video frame to minimize the projection error between the normalized 2D coordinates (Xn, Yn) and the metric 3D coordinates (Xm, Ym, Zm). Furthermore, it determines the rotation and translation between the camera's coordinate system and the world coordinate system. We specifically extracted the z-component (Tz) of the translation vector as our distance metric.

4.2 Experimental Setup

The experimental setup for all trials took place within our lab. We marked the floor using a standard 3-meter measuring tape to establish reference points for distance measurements. Markings were done at nine equally spaced points starting from 2.5 m to 0.5 m away from the camera. Each test subject held a cue card displaying their actual distance from the camera. We intentionally included subjects with varying heights to ensure a comprehensive assessment of our algorithm's performance. We fixed the recording device securely to a tripod for consistent and stable readings. The device that we used is Lenovo TAB4 10 PLUS (this being the tablet used for recording START videos). All videos for both calibration and experimentation were taken with the front camera.



(a) (b)

Figure 4.1: Left: Ground truth distance markings, Right: Chessboard frames for camera calibration

In the first variation of our experiment, the test subject was instructed to stand at each of the four designated reference points while holding the corresponding cue card for a duration of 3 seconds. The subject was first positioned at the furthest distance, 1.25 meters away from the camera, and asked to progress towards the closest reference point, 0.5 meters from the camera. The estimated distances at each reference marking for both the subjects can be seen in Fig 4.3. We also tabulate distance-wise median predictions and corresponding errors in Tables 4.1, 4.2.

Figure 4.2: Subjects 1 and 2 holding measurement cues

Real Distance (in meters)	Median Predicted Distance (in meters)	Median Error (in meters)
1.25	1.1957	0.0543
1.00	0.9535	0.0465
0.75	0.7214	0.0286
0.50	0.5146	0.0146

Table 4.1: Results for Subject 1

For the second variation, the subject stood at a constant distance of 0.75m, which is the closest reference mark. The subject then performed in-planar head rotations, that is, facing front and moving the head sideways, while staying at the same camera distance. The results of this experiment for both the subjects can also be seen in Fig 4.4.

For the last variation, the subject moved 0.25m on both sides horizontally, keeping the perpendicular distance from the camera at 0.75m (as showcased in Figs 4.5, 4.6).

Markings beyond 1.25m are reserved for a later section. The FaceMesh model that we use fails to detect faces confidently beyond this point. Although this distance upper bound is suitable for our use case, we mention a way in which similar computations can be carried out for subjects further away.

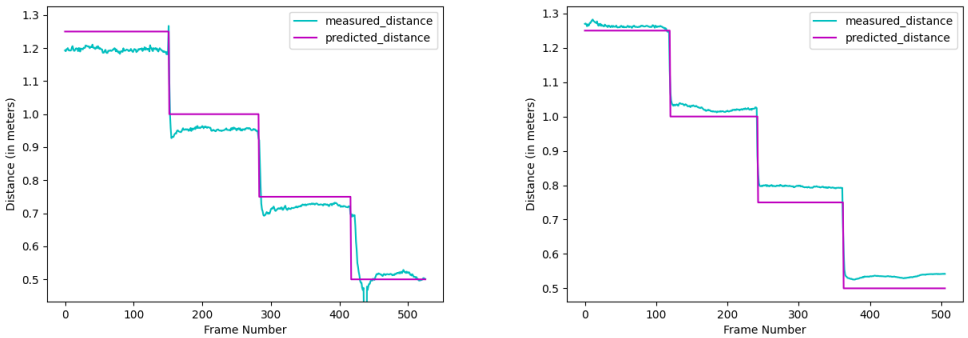


Figure 4.3: First variation: Measured and Predicted distances for Subjects 1 (left) and 2 (right)

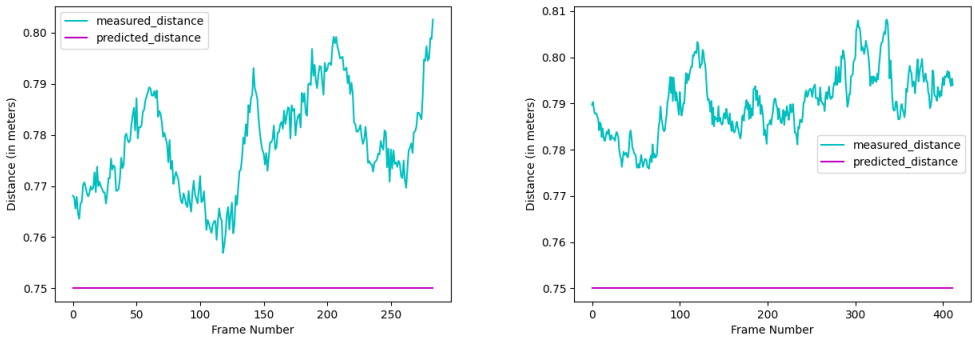


Figure 4.4: Second variation: Measured and Predicted distances for Subjects 1 (left) and 2 (right)

Figure 4.5: Subjects holding the cue and moving left to right at constant 0.75 meters distance

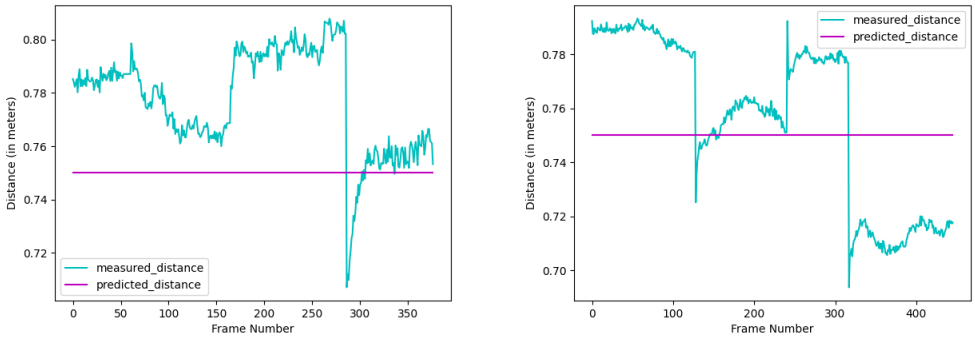


Figure 4.6: Last variation: Measured and Predicted distances for Subjects 1 (left) and 2 (right)

Real Distance (in meters)	Median Predicted Distance (in meters)	Median Error (in meters)
1.25	1.2623	0.0123
1.00	1.0215	0.0215
0.75	0.7966	0.0466
0.50	0.5347	0.0347

Table 4.2: Results for Subject 2

4.3 Discussion

4.3.1 First Variation

In the first case, the test subject just comes closer to camera in a straight line. As illustrated by the figure 4.3 and table4.1,4.2, it is evident that our method effectively estimates the correct distance with an acceptable margin of error, which is less than 6 cm. This slight error could have arisen from various factors encountered during our experiments. For instance, achieving precise alignment with the marking for every distance is challenging. Throughout our experiments, we made diligent efforts to ensure that the midpoint of the foot was aligned with the marking line while standing.

4.3.2 Second and Third Variation

In the second scenario, analyzing the results becomes challenging due to the continuous alteration of the distance caused by the ongoing head motion within the plane. Ideally, one would expect a constant distance since the movement is occurring within the same plane. However, there appears to be a slight decrease in distance (around 2-3 cm) when the subject turns their head to either side (Figure 4.4).

This effect becomes more pronounced in the final experiment (Figure 4.6) when the subject moves 0.25 m to the side. A substantial drop in the measured distance (approximately 4-5 cm) is observed. This phenomenon is consistent for both subjects, although the extent of change varies, and it occurs on both sides.

4.3.3 Method Limitations and Future Improvements

One of the limitations mentioned earlier arises from the face mesh's incapacity to detect faces at distances exceeding 1.5 meters. An alternative approach would involve using MediaPipe's own MediaPose model, which is designed to detect body landmarks rather than facial landmarks and works for longer distances. However, our experiments have demonstrated that, while we were able to predict distances using MediaPose, the results were noisier compared to the more accurate output produced by facemesh.

The second issue is addressed in the third variation of our experiments. We have observed that when a person within the frame moves to the sides or to the extremes of the image frame, the measured distances decrease. This phenomenon can be attributed to several factors. First, it may be due to camera distortion, where the individual, even when maintaining the same perpendicular distance from the camera, appears closer to the human eye when moving to the side. This issue could potentially be resolved by using a higher-quality camera that does not exhibit this distortion effect or by employing more sophisticated calibration methods to determine better distortion coefficients. For example, exploring 3D calibration techniques as an alternative to using a chessboard could be beneficial. The second reason for this ambiguity may stem from inherent assumptions associated with the pinhole camera model in the EPNP algorithm, assumptions that are not entirely accurate when dealing with modern cameras. To mitigate this, we could explore alternative variants of the EPNP algorithm that have been developed to address this particular issue.

Chapter 5

Using distance on START videos

Our work's utility is demonstrated through its application on a specific set of videos, namely, the "wheel task" videos from the START project. In the "wheel task," a black and white wheel is displayed on the tablet screen, and children are instructed to watch the video while their faces are recorded using its camera.

5.1 Method

Upon successfully verifying the functionality of our method through experiments, we applied it to the "Wheel Task" video dataset, as previously mentioned. This dataset comprises 111 videos featuring different children, each categorized into one of three groups: Autism Spectrum Disorder (ASD), Intellectual Disability (ID), or Typically Developing (TD). The ASD and ID groups were recruited from a tertiary clinic and diagnosed by a specialist clinician following the criteria outlined in the Diagnostic and Statistical Manual of Mental Disorders (5th ed.; DSM-V). In contrast, the TD group was recruited from the community. For our task, we merged the ASD and ID groups to differentiate them from the TD group. We obtained a metric distance measurement per frame for each video, following its passage through our pipeline. This yielded a distance vector of dimension n_frames x 1, where n_frames represents the variable number of frames within each video, which can vary among children. Upon conducting an in-depth analysis, we identified the median and standard deviation as effective discriminating features, as shown in Fig 5.1

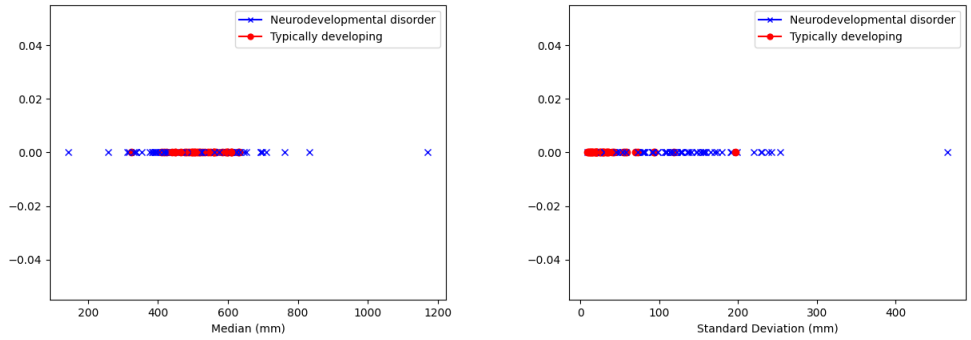
We subsequently utilized these two discriminative features as inputs for our machine learning model, which was specifically designed for a binary classification task. To assess the model's performance, we implemented a 5-fold cross-validation approach, where training was conducted on 4 of the folds, and testing was performed on the remaining fold. The final accuracy metric is calculated as the average of all five test folds. We leveraged various classification algorithms,

including Logistic Regression, Support Vector Machine (SVM), and Random Forest, to classify the data. The outcomes of our classification experiments are summarized in Table 5.1. Remarkably, the highest accuracy we achieved was 81%, which was obtained using Logistic Regression, with a corresponding F1 score of 0.74. Given the data's imbalanced nature, featuring 73 children with NDD and only 38 TD children, the F1 score assumes particular significance as an evaluation metric.

Algorithm	**Accuracy (%)**	**F1 Score**
Random Forest	78.46	0.67
Logistic Regression	81.23	0.74
SVM	73.08	0.55

Table 5.1: Summary of Classification Results

5.2 Discussion



(a) (b)

Figure 5.1: (a) Plot of median distances (b) Plot of Standard deviation

Based on the observations from the plots, several significant patterns emerge. Typically developing (TD) participants tend to display lower standard deviations, and their median distances rarely deviate to extreme values. This distinctive behavior could potentially serve as a basis for classification. On the other hand, children with neurodevelopmental disorders (NDD), particularly those with conditions like autism spectrum disorder (ASD) and intellectual disabilities (ID), exhibit different characteristics. It appears that some NDD children struggle to remain still while focusing on the tablet screen. The median plots also suggest that some NDD children tend to get very close to the screen when interacting with the wheel.

The decision to combine ID and ASD into a single category is based on the recognition that there is overlap between these conditions. Some children with ASD may also exhibit features of intellectual disabilities. Currently, our dataset suffers from data imbalance, with a larger number of non-TD samples. In the future, data sampling techniques can be explored to address this imbalance. Additionally, we can consider binary classification between just ASD and TD, which

would lead to a more balanced dataset. To enhance our analysis, we can move beyond simple features like mean and standard deviation and explore more complex features, such as using the distances between all frames as time series data. This could open the door to employing deep learning models like LSTM and RNN for improved classification.

Chapter 6

Features to Psychometric Scores

In the previous chapters, we discussed how raw data from each task can be converted into features. For each child, we would have multiple features across different tasks. By extracting these features for multiple children, we obtain a feature tensor as shown in figure 6.1. These features capture a child's abilities in different subdomains. Now that we have extracted the features, this chapter will explore how these features can be used to generate developmental scores, which is the main aim of this work.

Child 1	Coloring Task Feature	Wheel Task Feature	Button Task Feature	.....
Child 2	Coloring Task Feature	Wheel Task Feature	Button Task Feature	.....
Child 3	Coloring Task Feature	Wheel Task Feature	Button Task Feature	.....

Figure 6.1: Representation of feature tensor

6.1 Predicting GMDS scores

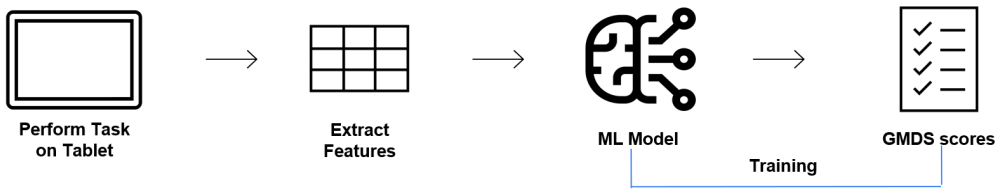


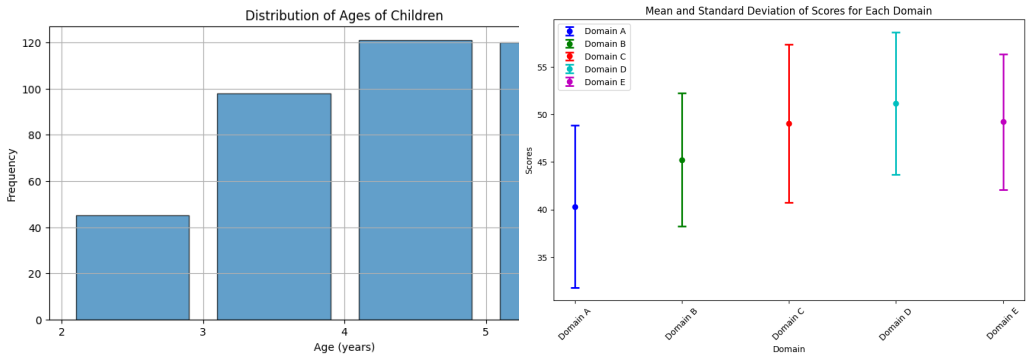
Figure 6.2: Training model to predict GMDS scores

As discussed earlier, the Griffiths Mental Development Scales (GMDS) is a gold standard tool for measuring mental development, but it is difficult to administer and expensive. Our objective is to generate scores similar to those of the GMDS using the extracted features. To achieve this,

we train machine learning models with the features as inputs and the 5-domain GMDS scores as training labels 6.2. Consequently, for a new data point, the trained model should be able to predict GMDS-like scores using only the features extracted from the tablet assessments.

6.2 Dataset

To train the machine learning model, we need both the features from the tablet assessments and the GMDS scores for each child. Therefore, a subset of the total children in STREAM undergoes both tablet assessments and GMDS scoring. We focus on data points that have no missing features, meaning every task is attempted and has a corresponding GMDS score. This results in a dataset of 384 data points. As mentioned, since we are focusing on START tasks, the age of children is distributed between 2-6 years 6.3a



(a) (b)

Figure 6.3: (a) Plot of Age distribution (b) Plot of mean and std dev of GMDS scores

For the model, the input will have a dimension of 56, which includes 54 features across 6 tasks, as well as the child's age and gender. The training label will have a dimension of 5, corresponding to the GMDS scores across 5 domains. To understand the distribution of GMDS scores refer figure 6.2b

6.3 Training Setup

We will use regression machine learning models for our experiments. The list of models used can be seen in Table 6.1. All models are used in their baseline configuration. Due to the scarcity of data points, we will perform 5-fold cross-validation. All results will be averaged over the 5 test folds.

No.	Model
1	Linear Regression
2	Ridge Regression
3	Random Forest
4	Gradient Boosting
5	AdaBoost
6	Decision Tree
7	Support Vector Regression
8	KNN Regressor
9	XGBoost

Table 6.1: List of regression models used in the experiments

6.4 Results

The metrics used to evaluate our models are the R^2 score, Mean Absolute Percentage Error (MAPE), and Mean Square Error (MSE). All values are averaged over the 5 test folds.

6.4.1 Comparing different models

Table 6.2 shows the comparison between different regression models.

Model	R^2 Score	MSE	MAPE
Linear Regression	0.57	12.74	4.86%
Ridge Regression	0.58	12.66	4.85%
Random Forest	**0.60**	**11.97**	**4.61%**
Gradient Boosting	0.60	11.95	4.65%
AdaBoost	0.50	14.87	5.54%
Decision Tree	0.23	22.88	5.89%
Support Vector Regression	0.53	14.14	5.02%
KNN Regressor	0.40	18.12	5.98%
XGBoost	0.55	13.31	4.75%

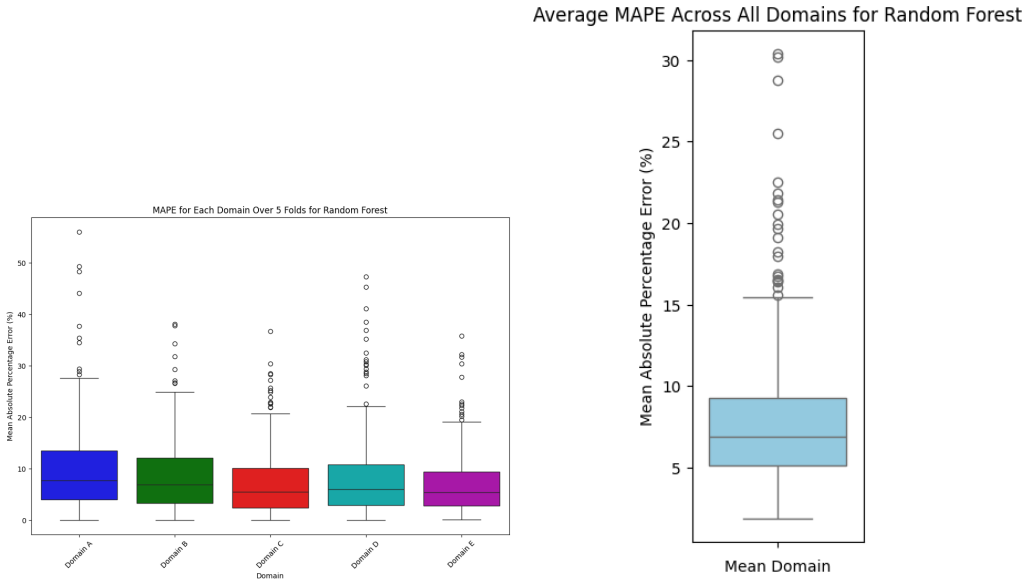
Table 6.2: Model performance metrics

Gradient Boosting and Random Forest show good performance, with Random forest achieving the best values in all 3 metrics.

6.4.2 Analysing best model

We will now analyze the actual predicted GMDS values and their percentage errors relative to ground truth values using the best model, which is the random forest. The MAPE values are presented in a box plot. Figure 6.4a illustrates the MAPE for test samples across the 5 domains, while Figure 6.4b depicts the averaged MAPE for test samples across these domains. Given that we are employing 5-fold cross-validation, each sample will be included in the test fold approximately

once during the process, thus number of data points in the plot would be 384.



(a) (b)

Figure 6.4: (a) Box plot of MAPE across 5-domains (b) Box plot of MAPE averaged over 5-domains

Figure 6.4 a illustrates the median error across all domains, which ranges from approximately 5% to 8%. Observing the average MAPE plot in Figure 6.4 b, we note that for the majority of data points, the MAPE values fall between 0% and 15%. These box plots provide insights into the variability of our model's predictions compared to the actual ground truth values.

6.4.3 More experiments

In addition to game features, Age and Sex are provided as inputs to the model. To determine the impact of these features on the model's predictive performance, we will evaluate the model's performance without including Age and Sex as inputs. Additionally, we will incorporate Country as an input feature and assess its effect on the model's performance.

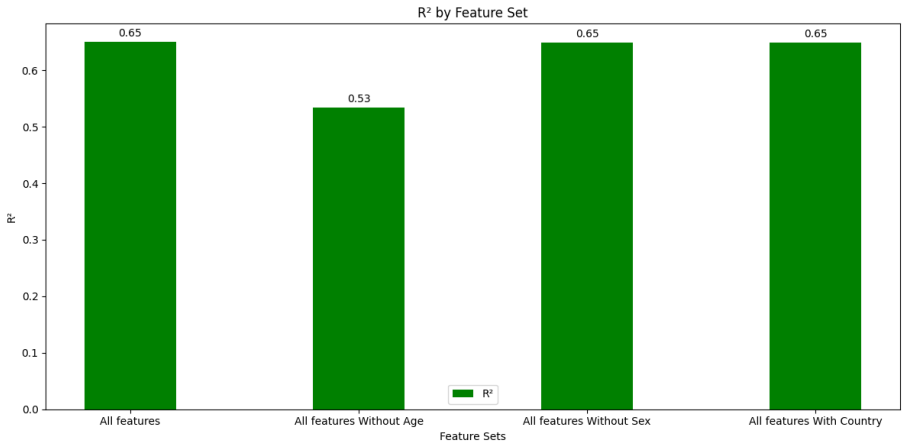
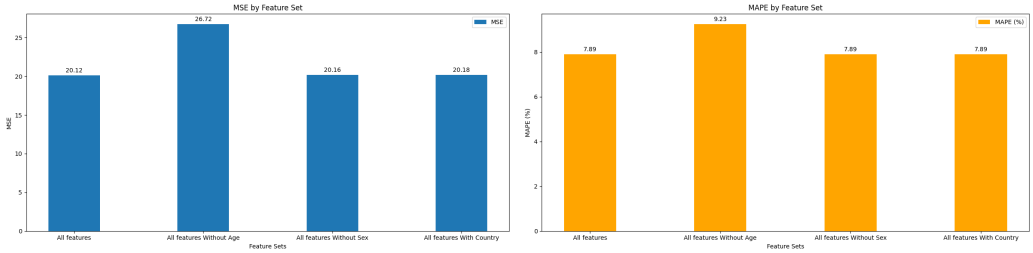


Figure 6.5: Model performance with different feature sets

Examining the R^2 values in the figure 6.5, we observe that removing age as a feature significantly decreases the model's performance. This outcome is expected, given the high correlation between GMDS scores and age. Conversely, neither removing sex nor adding country to the feature set affects the model's performance. Other metrics like MAPE and MSE show the same trend 6.6.



(a) (b)

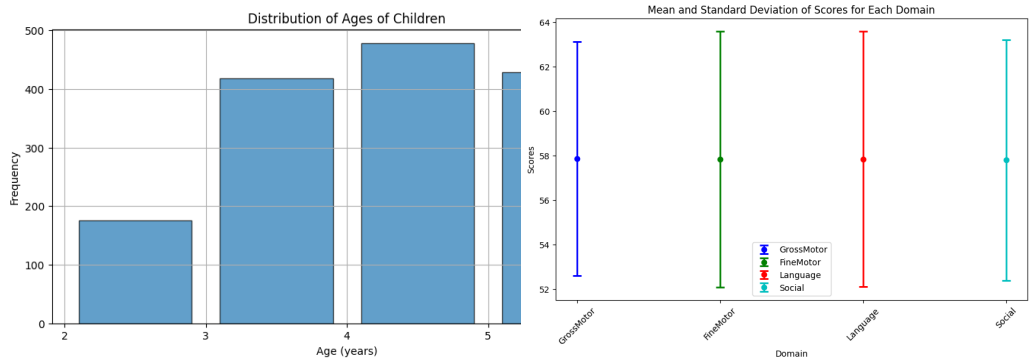
Figure 6.6: MSE and MAPE across different feature sets

6.5 Predicting MDAT scores

MDAT is another psychometric test, similar to GMDS, but specifically designed for low-income countries like Malawi. Both the language and tasks are tailored to suit the local context. The experimental setup will be the same, utilizing tablet assessment features to predict MDAT scores. In addition to being more suitable for low-income countries, MDAT has been administered to every child, providing us with a larger dataset for training.

6.6 Dataset

We will once again examine the number of data points that include both features and MDAT scores, which totals 1,459. Figure 6.7illustrates the distribution of age and MDAT scores.



(a) (b)

Figure 6.7: (a) Plot of Age distribution (b) Plot of mean and std dev of MDAT scores

For the model, each data point will once again have an input dimension of 56, consisting of 54

features across 6 games, along with Age and Sex. However, the training label in this case will have a size of 4 instead of 5, as there are 4 domains in MDAT, as shown in Figure 6.7b.

6.7 Training Setup

We will use the same set of regression models listed in Table 6.1 for GMDS, along with a neural network due to the availability of more data. The experimental setup for the machine learning models will continue to involve 5-fold cross-validation.

6.8 Results

The metrics used to evaluate our models are the $\mathbf{R^2}$ score, Mean Absolute Percentage Error (MAPE), and Mean Square Error (MSE). All values are averaged over the 5 test folds.

6.8.1 Comparing different models

Table 6.3 shows the comparison between different regression models.

Model	**R^2 Score**	**MSE**	**MAPE**
Linear Regression	0.57	12.74	4.86%
Ridge Regression	0.58	12.66	4.85%
Random Forest	**0.60**	11.97	**4.61%**
Gradient Boosting	**0.60**	**11.95**	4.65%
AdaBoost	0.50	14.87	5.54%
Decision Tree	0.23	22.88	5.89%
Support Vector Regression	0.53	14.14	5.02%
KNN Regressor	0.40	18.12	5.98%
XGBoost	0.55	13.31	4.75%

Table 6.3: Model performance metrics

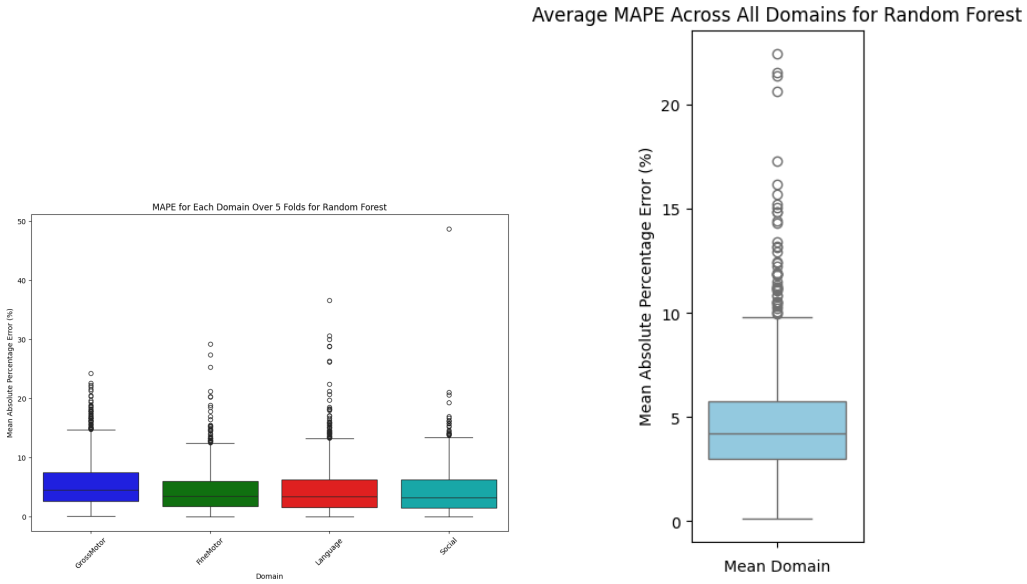
Random forest shows the best performance for R^2 Score and MAPE where as Random forest shows best values for R^2 Score and MSE. We will use random forest for further experiments.

6.8.2 Analyzing the best model

We will once again utilize the best model, i.e., the random forest, and plot the MAPE values for test samples. This will provide insight into our prediction accuracy.

Figure 6.8 a demonstrates that for all four domains, our model can predict MDAT scores with a median error of less than 5 percent. In Figure 6.8 b, which shows the MAPE values averaged

over the four domains, it is evident that our model can predict most samples with an error rate between 0-10



(a) (b)

Figure 6.8: (a) Box plot of MAPE across 5-domains (b) Box plot of MAPE averaged over 5-domains

6.9 Summary

The goal of this chapter is to demonstrate that the features extracted from tablet assessments effectively capture a child's abilities. This is accomplished by showing that these features can be used to predict the scores of psychometric tests. We illustrate this by predicting two types of psychometric scores, GMDS and MDAT, using the extracted features. The advantage of these experiments is that they indicate the potential for tablet-based home assessments to replicate, to some extent, the results of costly and time-consuming psychometric tests.

Chapter 7

Conclusion

In this report, we demonstrated the process of transitioning from tablet-based assessments to developmental scores. This approach is significant because obtaining developmental scores using traditional psychometric tests is challenging. In the initial part of the report, we outlined the preprocessing steps required to transform data generated from assessments into meaningful features that provide insights into a child's abilities and development.

We showed that these features could be converted into developmental scores using machine learning models, with psychometric test results serving as training labels. This indicates that a trained model can generate developmental scores akin to those from GMDS for a new child, without the child undergoing a psychometric test.

Finally, Item Response Theory (IRT) was utilized to generate scores in an unsupervised setting, eliminating the dependence on GMDS data as training labels. This is crucial because administering the GMDS test anew in a different region to obtain training labels would be impractical. Through our work, we have shown that an easy and accessible tablet-based assessment can effectively monitor mental development.

Bibliography

[1] Syed Ausaf Hussain, Waseemullah, and Najeed Ahmed Khan. Face-to-camera distance estimation using machine learning. In *2022 3rd International Conference on Innovations in Computer Science Software Engineering (ICONICS)*, pages 1–8, 2022.

[2] Enrique Bermejo, Enrique Fernandez-Blanco, Andrea Valsecchi, Pablo Mesejo, Oscar Ibáñez, and Kazuhiko Imaizumi. Facialscdnet: A deep learning approach for the estimation of subject-to-camera distance in facial photographs. *Expert Systems with Applications*, 210:118457, 2022.

[3] Indu Dubey, Rahul Bishain, Jayashree Dasgupta, Supriya Bhavnani, Matthew K Belmonte, Teodora Gliga, Debarati Mukherjee, Georgia Lockwood Estrin, Mark H Johnson, Sharat Chandran, Vikram Patel, Sheffali Gulati, Gauri Divan, and Bhismadev Chakrabarti. Using mobile health technology to assess childhood autism in low-resource community settings in india: An innovation to address the detection gap. *Autism*, 0(0):13623613231182801, 0. PMID: 37458273.

[4] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, jun 1981.

[5] Arturo Flores, Eric Christiansen, David Kriegman, and Serge Belongie. Camera distance from face images. In George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Baoxin Li, Fatih Porikli, Victor Zordan, James Klosowski, Sabine Coquillart, Xun Luo, Min Chen, and David Gotz, editors, *Advances in Visual Computing*, pages 513–522, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[6] Vincent Lepetit, Francesc Moreno-Noguer, and P. Fua. Epnp: An accurate o(n) solution to the pnp problem. *International Journal of Computer Vision*, 81:155–166, 2009.

[7] Shay Ohayon and Ehud Rivlin. Robust 3d head tracking using camera pose estimation. volume 1, pages 1063–1066, 01 2006.

[8] Shiye Pan and Xinmei Wang. A survey on perspective-n-point problem. In *2021 40th Chinese Control Conference (CCC)*, pages 2396–2401, 2021.

[9] Khandaker Abir Rahman, Md. Shafaeat Hossain, Md. Al-Amin Bhuiyan, Tao Zhang, Md. Hasanuzzaman, and H. Ueno. Person to camera distance measurement based on eye-distance. In *2009 Third International Conference on Multimedia and Ubiquitous Engineering*, pages 137–141, 2009.

[10] Mohamed Tahir Ahmed Shoani, Shamsudin H. M. Amin, and Ibrahim M. H. Sanhoury. Determining subject distance based on face size. In *2015 10th Asian Control Conference (ASCC)*, pages 1–6, 2015.

[11] Donald Silberberg, Narendra Arora, Vinod Bhutani, Maureen Durkin, Shefalli Gulati, Mkc Nair, and Jennifer Pinto-Martin. Neuro-developmental disorders in india - from epidemiology to public policy (p7.324). *Neurology*, 82(10 Supplement), 2014.