Trustworthy AI Assessment of Quality of movements in Trunk Control Rehabilitation Exercises for children

Joseph Sherman

Delft University of Technology

TU Delft

Trustworthy AI Assessment of Quality of movements in Trunk Control Rehabilitation Exercises for children

by

Joseph Sherman

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Monday October 31st, 2022 at 09:00 AM.

Student number:	528558	
Project duration:	December 1, 2021 – October 31, 2022	
Thesis committee:	Dr. Laura M. Crespo,	TU Delft, supervisor
	Dr. Arkady Zgonnikov ,	TU Delft, supervisor
	Dr. Robert Pangalila,	Erasmus MC

This thesis is confidential and cannot be made public until October 31, 2024.

Preface

Ad maiorem Dei gloriam et bonnum hominum (For the greater glory of God and the good of humankind)

This thesis marks the completion of my master's education in Robotics at the 3mE faculty of TU Delft. The report documents the findings of my master's thesis work. I am happy to have had the opportunity to work on such a thesis, aligned with my personal motto (above), where I could explore the application of scientific and technical principles for the medical domain

Countless people supported me during my journey on this thesis. I am very grateful to my supervisors at TU Delft, Laura M. Crespo and Arkady Zgonnikov and Dr. Robert Pangalila from Erasmus MC, who provided invaluable feedback on my framing and analysis, and have been a constant source of support and motivation throughout the ebbs and flows of my thesis journey.

Several other external experts gave helpful advice and valuable insights that helped me develop my thesis. I am very grateful to all the physiotherapists who provided me with their valuable time, and expertise for the interviews, without which this thesis would have lacked its practical and clinical nature. I would like to thank in particular, Nerea Garcia, who spent countless hours providing me with valuable knowledge on physiotherapy, feedback on my method and without whose contribution, the thesis would not have had the progress it currently has achieved. I would also like to thank all the parents and their children who participated in the experiments and further helped in the success of this thesis. My grattitude also extends to the members of the Motor Learning and Neurorehabilitation Laboratory, who created a friendly atmosphere for all the masters students and always made time for discussing and guiding us in our research. I am also grateful to the faculty and staff of 3mE, TU Delft for their support throughout the masters journey.

Pursuing this master's education is a privilege I enjoyed because of the love and sacrifices of my family (Michael Suresh, Anuja Suresh Sharona Fernando). Thank you for being the firm ground on which I could bloom. A big thanks to Tanvi, Rik, Georgios, Amber and all my friends in India and Delft, who supported me throughout these last two years. Lastly I remember with gratitude Late Fr. Dick Verbakel, who helped me profoundly over the last two years in Delft.

Joseph Sherman
Delft, October 2022

Abstract

Neurological disorders in the nervous and neuromuscular systems affect approximately 260 million people annually and among these 255 million would benefit from rehabilitation [1]. Patients with neurological disorders usually require multi-dimensional rehabilitation, involving physical, cognitive, psychological, and medical help. Children with trunk control problems arising due to some of these neurological disorders also require such multi-dimensional rehabilitation. A major part of this is administered to the patient through the activities of a physiotherapist in the clinical context.

But the limited number of physiotherapists result in exercises often being prescribed for patients as in-home rehabilitation. During in-home rehabilitation, the patient and the primary care-giver may not be able to comply with the prescription without feedback from a physiotherapist.

To address this challenge, this paper proposes an automated method for assessing movement quality of children during trunk control rehabilitation exercises. We adopted a Human-centered AI approach to the development of our system. We identified the needs of physiotherapists for assessing patient's functional abilities through semi-structured interviews with six physiotherapists. As a result, we co-designed and developed an Artificially Intelligent decision support system that automatically assesses the quality of motion. We created a trunk-control rehabilitation exercise movement dataset based on a protocol co-designed by the authors and the physiotherapists. The data was collected from 15 typically developing children (mean age 7 years, range 4–10 years) using a ZED-mini stereo-camera and the quality scores as ground-truth were obtained from a physiotherapist. The exercises involved reaching targets kept on a table in front while being seated away from the table on a stool.

We investigated the performance of Random Convolutional Kernel transform and XCM, two state-of-the-art multivariate time-series classification algorithms on this dataset and achieved a quality prediction f1-score of 65% on the test dataset and similar promising results on the detection of compensatory movements in the exercise motion. In addition, to increase the trust-worthiness of our AI solution, we have provided explanations on the predictions of the black-box algorithms, which can aid the users of the system to understand the causal relationships between the input and output to the AI algorithm.

Table of Contents

Nomenclature

Abbreviations

Abbreviation	Definition
ADL	Activities of daily living
AI	Artificial Intelligence
AS	Antepulsion with shoulder
BL	Bottom lifting
BSEF	Increase of balance surface by extending feet
CNN	Convolutional Neural Network
CP	Cerebral Palsy
DS	Desk Support
DNN	Deep Neural Network
FS	Feet Swing
FRJ	Feet going up return journey
FMF	Feet moving forward
GDPR	General Data Protection Regulations
GRAD-CAM	Gradient-weighted Class Activation Mapping
HAR	Human Activity Recognition
HM	Head movement
LIME	Local Intrerpretable Model-Agnostic Explanations
LSTM	Long Short Term Memory
MD	Movement Disorders
NASB	Non-Active arm swing backward
ND	Neurological Disorder
NR	Neurorehabilitation
OLS	Ordinary least square
PT	Physiotherapist
QOM	quality of movement
ROCKET	Random Convolutional Kernel transform
RNN	Recurrent Neural Network
ROM	Range of Motion
SP	Speed
SDK	Software Development Kit
SSH	Seat Support with hand
SVM	Support Vector Machines
TCMS	Trunk Control Measurement Scale
ASL	Too much support on lap
TI	Trunk inclination
XAI	Explanable Artificial Intelligence

1

Scientific Paper

Trustworthy AI Assessment of Quality of movements in Trunk Control Rehabilitation Exercises for children

Joseph Sherman
Dept. of Cognitive Robotics
Delft University of Technology
Delft, The Netherlands
mjoesherman@gmail.com

***Abstract*—Neurological disorders in the nervous and neuromuscular systems affect approximately 260 million people annually and among these 255 million would benefit from rehabilitation [4]. Patients with neurological disorders usually require multi-dimensional rehabilitation, involving physical, cognitive, psychological, and medical help. Children with trunk control problems arising due to some of these neurological disorders also require such multi-dimensional rehabilitation. A major part of this is administered to the patient through the activities of a physiotherapist in the clinical context.**

But the limited number of physiotherapists result in exercises often being prescribed for patients as in-home rehabilitation. During in-home rehabilitation, the patient and the primary caregiver may not be able to comply with the prescription without feedback from a physiotherapist.

To address this challenge, this paper proposes an automated method for assessing movement quality of children during trunk control rehabilitation exercises. We adopted a Human-centered AI approach to the development of our system. We identified the needs of physiotherapists for assessing patient's functional abilities through semi-structured interviews with six physiotherapists. As a result, we co-designed and developed an Artificially Intelligent decision support system that automatically assesses the quality of motion. We created a trunk-control rehabilitation exercise movement dataset based on a protocol co-designed by the authors and the physiotherapists. The data was collected from 15 typically developing children (mean age 7 years, range 4–10 years) using a ZED-mini stereo-camera and the quality scores as ground-truth were obtained from a physiotherapist. The exercises involved reaching targets kept on a table in front while being seated away from the table on a stool.

We investigated the performance of Random Convolutional Kernel transform and XCM, two state-of-the-art multivariate time-series classification algorithms on this dataset and achieved a quality prediction f1-score of 65% on the test dataset and similar promising results on the detection of compensatory movements in the exercise motion. In addition, to increase the trust-worthiness of our AI solution, we have provided explanations on the predictions of the black-box algorithms, which can aid the users of the system to understand the causal relationships between the input and output to the AI algorithm.

***Index Terms*—trunk-control rehabilitation, automated assessment, trustworthy AI**

I. Introduction

Neurological disorders (ND) in the nervous and neuromuscular systems affect approximately 260 million people annually and among these 255 million would benefit from neurorehabilitation (NR) [4]. Patients with ND usually require multidimensional rehabilitation, including physical, cognitive, psychological, and medicinal help. A major part of this rehabilitation is administered to the patient through the activities of a physiotherapist (PT) in the clinical context.

PTs assess the disorder, assign and implement appropriate interventions and finally evaluate the effects of the intervention during the physical therapy of patients [19]. The PTs experience high workload during therapy due to the physical use of their body to support the patients, the efforts applied to keep the patient motivated and the efforts applied to understand the complexity of the process [25] for proper goal setting. Adding to these, the limited number of PTs often necessitate some of the therapy to be prescribed as in-home rehabilitation so that the PTs can offer adequate care to all their patients. But during in-home rehabilitation, the patient and the primary caregiver may not be able to comply with the prescription without feedback from a physiotherapist. To overcome the difficulties understanding the performance levels of users and facilitating in-home monitoring of therapy, we have explored the use of Artificial intelligence (AI) to aid the PTs in their treatment of patients. A focus of the present study, is also to increase the trustworthiness of AI solutions in physiotherapy.

The main contributions of the paper are: (1) A human-centered approach for AI development in physiotherapy; (2) The creation of trunk-control rehabilitation exercise dataset; and (3) evaluation of state-of-the-art in multi-variate time series classification and explainable AI on the above dataset.

The article is organized as follows: The next section provides an overview of related work. Section III introduces the methodology adopted in this study involving the interaction with the PTs, the data collection protocol and the development and evaluation of the AI models. The results of the proposed solution on a dataset of trunk-control rehabilitation exercises is presented in Section IV. The last two sections discuss the results and conclude with the main findings of the study.

II. Related Work

Children with MD, such as Cerebral Palsy (CP) frequently show impaired trunk-control affecting performance of activities of daily living (ADL) such as sitting, walking and reaching [22]. However, a majority of research on assessment and treatment is focused on the upper and lower extremities, while the literature on trunk control in children is scarce [14].

A. Assessments

A few tools exist to measure trunk control in children, such as Seated Postural Control Measure (SPCM) [9], Spinal Alignment and Range of Motion Measure (SAROMM) [1] and Segmental Assessment of Trunk Control (SATCo) [15]. These measures usually only evaluate the postural characteristics of the children in a static position and lack the ability to evaluate dynamic motions, which constitute most of the actions of a child while performing ADL. Trunk Control Measurement Scale (TCMS)[11] is an assessment scale that is specifically designed to evaluate the trunk control in children and assess the movements of children under three scenarios: static sitting, dynamic sitting and dynamic reaching. Scores on various motions within each scenario is defined to incorporate the presence or absence of several compensation movements to calculate a final score indicating the level of trunk control in the patient. Such an assessment is more suitable for guiding the development of an automated system to assess the quality of movements in trunk rehabilitation exercises in children and will be used in our study.

B. Automation in neurorehabilitation

AI has provided many promising results in problems that were previously considered very difficult for a computer and even for humans to solve. The application of AI in NR has the potential to decrease the provider and patient burden as well as personalising the services offered by the PT [13]. NR tasks that could be automated using AI are risk analysis of disorders, video analysis, interaction with patient and prediction of therapy outcomes [24]. Deep Neural Networks (DNN) are good at feature learning and are achieving state-of-the-art results in the field of Human Activity recognition (HAR). They are shown to frequently outperform models trained on hand-crafted domain-specific features [31]. Two main DNN architectures that have been reported to perform well for the physiotherapy domain tasks are, Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). In [20], RNNs were used for classifying time-series exercise motions into motion primitives such as reach, reposition, transport or stabilize. Using physiotherapist-labelled Inertial Measurement Unit (IMU) data of spine, pelvis and hands corresponding to each motion primitive from exercise motions of 41 post-stroke patients, an RNN was trained to decode 6-second motion windows into motion primitives. This was used to count the number of primitives and total exercise cycles performed by the patients accurately. In [18], the authors use a RNN on motion sensor raw data to predict the quality of movement. However the ground truth for quality scores are not given by a PT and thus the output from this study is not compared to a PT's evaluation and lacks clinical assessment. In [17], the authors proposed an architecture in which kinematic data is used by binary classifiers to predict the correctness on previously-determined performance components. The confidence of these classifiers for their predictions is then used to estimate a performance score. The performance components are defined for exercises involving the upper limbs of post-stroke adult patients. As seen, although some attempts have been made at automating the assessment of quality of movements, most of the solutions are proposed for studies involving adults and no literature was found on solutions targeted at assessing the quality of movements in trunk-control rehabilitation exercises for children.

C. Trustworthy AI

Current European General Data Protection Regulations (GDPR) allow users to enquire about AI decisions and the majority of AI techniques discussed in the physiotherapy domain show black-box characteristics and lack of transparency, which restrict its applicability and acceptance in the medical context [12, 23]. Additionally, the application of AI in NR, would likely result in a shift in emphasis for therapists from prediction to judgement, where value and context-based human decision-making exists alongside potentially more accurate algorithmic predictions, which are capable of learning salient features of exercise motion in high-dimensional movement data, currently beyond the capabilities of human cognition. Thus, knowing when to trust the outputs of such clinical decision support systems will become all the more important. Intepretability would serve the purpose of highlighting whether the models have learnt hidden biases in the dataset instead of the actual phenomenon to be studied [27]. This also has the potential to enhance the PTs knowledge regarding the unique aspects of a patients' state highlighted from such interpretable models, when the predictions differ from the PTs judgement, but they have established that the AI is trustworthy.

Intrepretability can be established by using non-black box solutions which are easy to interpret by the user. An example of a simple interpretable model is one where if-else statements are used to check whether the values of selected features in data fit a criteria. By using human interpretable conditions for the various features, it can be easily seen which features led to the final classification of the input data to a specific class (in the case of a classification problem). It is used in the work by Dhiman et al [6] for adapting the parameters of an exercise, where the game difficulty is changed based on the measured anxiety values. However, models using if-else rules and fuzzy logic, as seen in [21] and [2], require the decision boundaries (thresholds) to be known apriori due to the complex nature of the task. Moreover, these models deliver on interpretability but the assumptions of hard boundaries on known features simplifies the problem too much. Explainable AI (XAI) explores simple interpretable AI models that can be used to explain the decisions of more complex black-box methods.

Prediction explanation is a local explainability method where the predictions of an AI model for a given input instance is explained using other AI methods. In the case of classification, these methods can explain which part of the input influenced the model's prediction most. These methods can further be divided into: i)*self-explaining* and ii) *post-hoc* models. Self-explaining models have components in them that learn the relationships between input data and predictions during the training and give a explanation along with the prediction output. In the case of Hendricks et al. [10], a CNN and RNN based architecture classifies a generic image containing objects into the type of object present in the image and outputs the prediction with a textual explanation of the prediction. These methods provide easily-interpretable explanations for the predictions. However such methods were scarcely used in the literature as it cannot be applied to previously trained models for comparing competitive methods of classification.

Post-hoc methods can be applied to already trained models and thus have the additional advantage of facilitating comparative studies of models performances with interpretability as the criteria. The post-hoc methods can be further classified into i)*Propagation-based*, ii) *Perturbation-based*. Local Intrerpretable Model-Agnostic Explanations (LIME) is a Perturbation-based method that performs an approximation of a black box model with a simpler interpretable model to explain how a black box model makes a single prediction. In Dindorf et al. [7], a SVM classifier was used to classify between healthy and various other pathologies based on the spine posture data readings and it was found that LIME's relevance ratings on the various regions of the spine led to understanding relations previously unknown. The analysis from LIME also contributed towards correcting the models performance by comparing the relevance ratings on features for a rightly classified subject with a mis-classified subject, thus bringing out understandings into what might be going wrong in the relations of the model. While perturbation-based methods have shown the potential to provide explanations, the explanations are based on the outputs from a simpler model which approximates the behaviour of the more complex model and thus faithful explanations are not guaranteed. In Shahtalebi et al. [27] 'Gradient-weighted Class Activation Mapping' (Grad-CAM) [26] a propagation-based methodology, is used to discover parts of the input motion data of the hand that contribute to prediction of Parkinson's disease or Essential tremors in the subject. The method has the advantage of showing the relevant features in the input space and no approximation is involved and thus explanations are guaranteed to be faithful to the behaviour of the model.

Recently, XCM, an explainable CNN for multivariate time-series classification has reported promising results both in classification and faithful explanations for its predictions on both small and large public datasets [8]. XCM utilises the GRAD-CAM approach to provide prediction explanations of a black-box neural network model. This method is able to identify the features and timestamps of the input data that are important for the prediction. It has outperformed the current most accurate state-of-the art algorithm in both synthetic and real-world dataset while providing faithful and more informative explanations, but has not yet been applied to physiotherapy applications.

The gap in literature for explainable (and thus trustworthy) and accurate prediction algorithms for assessing the quality of movements in trunk-control rehabilitation exercises specific to children will be addressed in this study.

III. Methodology and Data

We have adopted a multistage methodology to produce the automated decision support system. Each stage will be explained chronologically in the following subsections.

A. Interviews

We follow a human-centered AI approach in this study to guide the development of the automated system for decision making to enable prop. Human centred AI approach refers to the involvement of people in making AI designs / algorithms for those people [30].

Two rounds of interactions with the future users, the PTs, were conducted to understand the current process in practice at clinics, facilitate requirements elicitation and guide the experiment design specification for our study. In the first round of interactions, semi-structured interviews were conducted to get knowledge on rehabilitation of children with trunk-control disorders in semi-structured interviews with the physiotherapists. In total, five physiotherapists (PT 1-5) of varying work experiences were interviewed, see Table I. These interviews were conducted remotely via MS Teams and were later transcribed using Atlas.ti software. Explicit permission to record the interviews was sought from each of the five physiotherapists and the required ethical approval was obtained for this interaction. The interviews lasted for about 1 – 1,5 hours and were semi-structured in nature. We asked them questions about the children they encounter, how they set up the therapy, what factors affect the therapeutic exercises and finally how they validate their therapy.

In the second round of interactions, focused discussion with two PTs (PT5 and PT6) were conducted with the aid of a digital questionnaire. The questionnaire contained proposals for assessment factors, based on the factors found in [16] for assessing the quality of upper-limb body movements in post-stroke patients. It also contained mechanisms to record the relative importance of different kinds of compensations in the trunk control exercises derived from TCMS [11] discussed before. Based on the results of the interviews, the exercise protocol and factors important for assessing the performance of the participants were established.

B. Experiment Design

The experiment design was guided using the results of the interviews with the PTs. The participants for these experiments were TD children - 4 to 10 years old (mean 7 years). However, to replicate the clinical condition, the exercises were modified

TABLE I
DETAILS OF PARTICIPATING PHYSIOTHERAPISTS

	Number of years of experience
PT1	3
PT2	10
PT3	12
PT4	8
PT5	15
PT6	2

according to the suggestions of the PTs, to simulate children with impairment. The exercises involved three sets of sitting

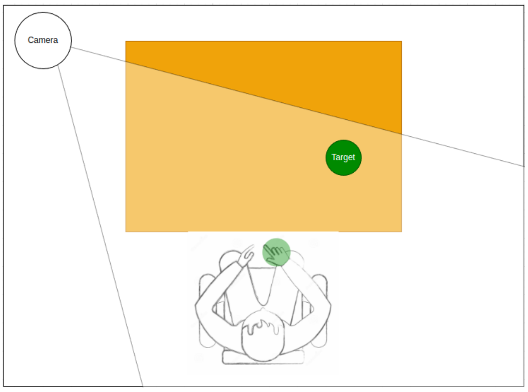


Fig. 1. Exercise Set-up graphic

and reaching tasks where the children were first made to sit on a normal stool and reach forward to pick up an object. Next, a raised platform was placed under their left-foot and they were asked to sit on pillows placed on the right side of the seat. The position of the children was destabilised this way and required them to actively control their trunk to be stable throughout the motion. In the final iteration, the children were instructed to prevent a paper kept on of their lap from falling to the ground using their arms. The paper had a weight attached to it, which necessitated the use of force to prevent it from falling. This was done to replicate the motion of children with trunk control issues as they would often take the support of their lap using their hand to compensate for the lack of trunk control. The research study related to the data collection of the exercises and interviews was approved by the Human Research Ethics Committee at the Technische Universiteit Delft under the identification code 2113. A written informed consent for participation in the research study was approved by the board, and was obtained from all participants (and legal guardians) in the study.

Each of these sets, were done twice with either hand being used. A 3D stereo-camera recorded the motion of the children in RGBD format. The setup of the experiments can be seen in Fig 2 A PT was asked to evaluate the quality of these exercise videos on a scale of 0-2 with 0 being very poor performance and 2 being good performance. Additionally, the score on speed as well as the indication of 13 compensatory motions were also collected. The speed was scored on a scale of 0-2, with 0 being very slow and 2 for very fast and the presence or absence of the 13 compensation movement was indicated either by 1 or 0 respectively. The choices for factors of assessment and their scales were established from the interviews (see Sec. IV-A2. The evaluations on these factors were used as labels for the AI model, explained in Stage 4.

C. Data Pre-processing

The data from the ZED camera were in RGBD format - Red, Green, Blue (color components constituting the visual image) and Depth (distance of object from camera lens). The ZED SDK was used for human pose estimation which provided 3D coordinates of 34 body parts per frame. On average, the exercise motion lasted approximately 3 seconds (45 frames). The SDK executes a human detector on the 2D RGB image to recognise and localise 34 different body parts in the image. The depth information from the corresponding localised pixel location of each body part is used to create a 3D depth-map of the image and the 3D coordinates of the body parts are extracted from this.

The coordinates of the body parts are available in the camera's frame of reference. For better human interpretability of these coordinates in 3d space, the frame of reference is changed from the camera's frame to another frame where an imaginary person is viewing the 3d points from behind each participant's pelvis. This position corresponds to the most common position of the physiotherapist during trunk-control exercise sessions where they are located behind the patient to monitor and aid them. to do this, a pre-processing stage is used to shift the origin of all coordinates to the location of the body part - "pelvis" in the first frame of the recording and the axis system is rotated so that the x and y axes are parallel to the edges of the table, while the z is anti-parallel to the direction of gravity. The 3d coordinates of three points on the table (sufficient to define a 3d surface) are extracted semi-automatically using the SDK interface. These points are used to compute the unit vectors corresponding to the direction of the edges of the table. The z axis reference is directly obtained from the camera's internal IMU sensor.

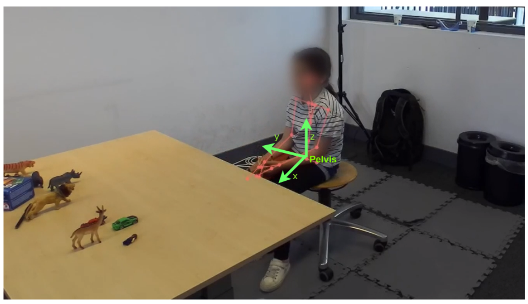


Fig. 2. Participant and Pelvis frame in camera frame

A pilot data collection study was conducted with two participants where it was visually observed that the position of the camera with respect to the participant, had an influence on the tracking abilities of the SDK, where body movements of the participant along the camera axis showed poor body-tracking results. An optimal performance was observed while placing the camera diagonal to the participant as this reduces the occurrences of alignment between the direction of motion during the exercises and the camera axis. However, due to the presence of the table in front of the participant, the table occluded the lower limbs of the participant and the tracking of the knee and lower body parts were often lost or erroneous due to this. Therefore, 18 of the 34 data points of the body belonging to the lower extremities were excluded from further stages to avoid noisy data in the pipeline.

Time-series data are a sequence of data points occurring in a successive order over a period of time. In our case, 3d coordinates of the body parts are available for each frame and occur in a successive order over a period of 1-8 seconds, with an average of three seconds and a maximum of eight seconds. In the literature review, it was also reported that most state-of-the-art time series classification algorithms require time series of equal length.Two methods are reported to address this issue: padding and interpolation of values of all time series to match the longest series. Interpolation of values for shorter series to match the length of longest series have the adverse effect of losing the information contained in the temporal dimension of the data. Thus, padding with zero's was used to make all time series of equal length.

The participants were of different age groups and thus their physical features were different from each other. Most relevant to the physical aspects of our experiment protocol, the body lengths of the participants varied and with them the ability to reach farther or nearer varied across subjects. An analysis was done to determine whether normalisation of the input features is required to account for this variance in the dataset, so that the classification algorithms do not learn unintended biases in the dataset pertaining to the relationship between target-distances and quality scores.

D. Development of AI model

For each time-series in our dataset, we have a corresponding quality score as the label and also labels for nine compensation movements observed by the PT. This poses a time-series classification problem. Further as there are more than one feature in the data (we have a total of 30 features - 10 body parts with 3 features each), the task is designed as a multi-variate time-series classification.

Random Convolutional Kernel transform (ROCKET) is a machine learning algorithm that uses convolutional kernels over the data to create feature maps. This is similar to other methods such as CNNs and RNNs found in literature, however, in these other methods, the kernel parameters (weights) are learned in a training stage. In ROCKET, a large number of randomly initialised kernels are used to transform the data, through convolutions, into an effective set of features. It has produced state-of-the-art accuracies with a fraction of the computational expense of existing state-of-the-art methods [5]. Similar to XCM, discussed at the end of Sec. II-C, there is no literature indicating ROCKET's performance on physiotherapy applications, while ROCKET and XCM are reported to have shown the best prediction performance on the benchmark time-series classification datasets and thus we select these algorithms for our investigation. We use the original implementation of ROCKET in the sktime library. XCM additionally provides explanations on its predictions using GRAD-CAM. We have used these algorithms in our analysis. The sktime implementation available publicly for ROCKET ([28]) and the code provided by the authors of XCM [32] was used. In both cases, the original parameters of the models were used in our analysis. The computation was performed on a machine with Nvidia GTX-1660 6Gb GPUs and Intel i7-9750H processor.

E. Evaluation of AI model

1) Evaluation of classification performance: The underlying data has unequal class distribution across all prediction labels (explained in Sec. IV-A2 as some quality scores and compensation were more numerous than the others in the dataset as seen in Table II & III. To evaluate the AI models meaningfully, f1 score is selected as it accounts for the imbalance, by looking at both precision and recall of the model performance. Cross validation tests are carried out on the dataset to account for the lack of large data and to evaluate the stability of model performance across different training and validation sets.

TABLE II
CLASS IMBALANCE IN DATASET FOR 3-CLASS LABELS

Assessment Factor	***train*** 0	1	2	***test*** 0	1	2
QOM	25%	42%	33%	20%	24%	**56**%
SP	32%	45%	23%	**51**%	26%	23%

TABLE III
CLASS IMBALANCE IN DATASET FOR BINARY LABELS

Assessment Factor	***train*** negative	positive	***test*** negative	positive
NASB	92%	**7**%	95%	**5**%
HM	52%	48%	43%	57%
SSH	95%	**5**%	94%	**6**%
DS	58%	42%	53%	47%
BL	58%	42%	80%	**20**%
TI	55%	45%	60%	40%
ASL	47%	52%	36%	63%
AS	24%	**75**%	19%	**80**%

The dataset is split across subjects to ensure that the model is not trained on the test subjects data to provide realistic results. This supports the real-world scenario where a model trained on some subjects will be used to evaluate the movements of an unseen subject. Data from the first 12 subjects are used for training the AI model and the last three are used to test the performance of the trained model. Within the training data, a further 10-2 split is performed to create a validation set to perform 6-fold cross-validation tests to evaluate its robustness to being trained and tested on a different subset of subjects each time.

2) Evaluation of explanations: Evaluation of the interpretation given by the XAI models is a challenge as unlike labels for data, there is no ground truth for the explanations given by the model. To evaluate the explanations, statistical methods and expert judgment are commonly used [29, 7, 3]. In this study, we use the judgment of the PT6 to evaluate our explanations. First, we evaluate the fidelity of explanations from the GRAD_CAM outputs on a model trained to predict the hand that was used during the experiments. It is expected that this is a relatively easier way to evaluate the explanations given by this method by checking if the GRAD-CAM highlights the expected hand-related (right or left) features for a given input series.

In the case of detecting for presence of individual compensation movements, we use a mapping between the compensation movements and the body parts related to these movements. This mapping, provided by PT6, is used as ground-truth to check if the GRAD-CAM highlights the expected body part(s)-related features for the input series.

For predictions of quality, only the prediction performance are evaluated and not the explanations, as the ground truth was not available for the explanations at the time of the study. However, during the interviews, PT6 was also asked about the relation between the quality score and the scores on the other factors. The presence of some of these factors were reported (see Table IV by the PT as being more important than the others for their judgement on poorer quality scores.

TABLE IV
IMPORTANCE OF OTHER ASSESSMENT FACTORS FOR PREDICTING QOM

Factor	Important
SP	No
NASB	No
FS	No
HM	No
SSH	Yes
DS	Yes
BL	Yes
FRJ	Yes
TI	No
FMF	No
ASL	Yes
AS	No
BSEF	No

To validate the relative importance data from the PT regarding the relationship between quality score and compensatory movements and score on the other variables, Ordinary least square (OLS) multivariate regression is used. The associations between 13 independent variables (speed and nine compensatory movements) and one independent variable, quality of movement is obtained from this analysis and compared to the PT's report on important factors. OLS is used to get an overview of how the quality of movement (QOM), which is our dependent variable is affected by 13 independent variables defined later in Sec. IV-A2. The details of the model are given below:

Step 1: Multivariate model of variables To check how our independent variables affect the dependent variable, we first make an empirical with 737 data points, each representing a motion performed by the subjects as shown in the previous experiment. Quality of Movement (QOM) is shown as a factor of independent variables as

$$\sum_{j=1}^{n} \alpha_n * d_n$$

where α is weight of determinant d_n and n ranges from 1 to 13. Therefore, QOM = $\alpha_1 * SP + \alpha_2 * NASB + \alpha_3 * FS + \alpha_4 * HM + \alpha_5 * SSH + \alpha_6 * DS + \alpha_7 * BL + \alpha_8 * FRJ + \alpha_9 * TI + \alpha_{10} * FMF + \alpha_{11} * ASL + \alpha_{12} * AS + \alpha_{13} * BSEF + constant$

Step 2: Multivariate regression analysis
Using the model in step 1, Ordinary least square (OLS) multivariate linear regression analysis will be performed and the statistical significance of the results at confidence levels of 90, 95, 99 is investigated. It is assumed that the data fits model assumptions of regression analysis. These assumptions are validated in the diagnostic checks.

The results from the OLS regression, discussed in the next section, will be evaluated against the PT's indications of the relative importance of the compensation movements to their judgement for a poorer quality score. If the results overlap, we can hypothesise that the presence of certain compensation movements (which can be automatically classified) can be used as explanations for the quality score from the AI models.

IV. RESULTS

A. Interview Analysis Results

1) First Interviews: The results of the first round of interviews with five PTs will be reported in this subsection. These involve the patient demographics, evaluation protocol, design of interventions, assessment of movements and the expectations of the therapists from automation and technology.

Age: The therapists work with children as young as new born children until the age of 18 years. PT1 works with children from 2-4 years old while PT3 works with children from 2 – 18 years. The other PTs work with children from 0-4 years. But they occasionally work with older children as well.

Types of patient case : PT1 said that the patients (children) differ a lot in their disorders. Some have low cognitive abilities, while some have a bit higher cognitive ability. PT4 and PT5 also have the same opinion. All the therapists agreed on the fact that every child has different needs and they need customized care because the complexity of disorder among children can vary a lot. In this thesis, my focus will mainly be on children having trunk instability disorders.

Evaluation: PT1 and PT2 mentioned that in the evaluation stage, they observe the patient for 1-2 months. In this observation phase, they see what child does in its free movement, how the child freely plays, what is their natural movement, for how much time can the child sit/stand on its own without any support, what is the extent of reach of the child's grabbing or reaching motions and so on. This is to understand or get a sense of the extent of the child's physical and mental impairment and understand why they are not able to move properly. PT1 further mentioned that such observation also helps the therapists to understand if the impairment is existing or getting enhanced due to behavioral traits or any physical pain.

In agreement with the answers of PT1 and PT2, PT3, PT4 and PT5 further added that during such observations, the therapists also pay special attention to how the child is changing between different postures and what their highest level of current motor functions are. During this phase, the therapists also talk with the parents about what they want their child to learn, about the interests of the child and their behaviour at home. This helps the PTs estimate the current capabilities of the child and potential goals for the rehabilitation scheme.

Next, in the evaluation phase, the therapists perform standardized tests to check the motor functions of the child, form an understanding regarding their muscle tone and to check the child's trunk control. PT4 mentioned that tests like Gross Motor Function Measure (GMFM) and standardized lichamelijk onderzoek (SLO) can help the therapists understand the current capabilities of the child and make an estimate about their expected development. For younger babies and toddlers, these test are often replaced by Bailey Scale of Infant and Toddler development (version 3). Sometimes the therapists also use an instrument called the "goniometer" to measure the joint angles of children in their knees, shoulders, hips and other places in the body. However, most of the therapists said that the goniometer measurements are not always accurate as the measurement depends on the skill of PT, rendering the measurements with low repeatability both with the same PT or between PTs. Although the above assessments tools are used for assessment, the therapists often rely on their experience and visual observation in making the initial evaluation of the children instead of solely using standardized tests or instruments, due to suspicions of the usefulness of the tests and instruments given shortcomings such as low repeatability discussed before.

Finally, these tests forming the evaluation stage are usually repeated once or twice a year to check if the child's initial status has improved or if there is an emergence of a new impairment that needs to be treated.

Intervention : In this part that follows the evaluation phase, the therapists decide the actual exercises that they want the child to perform to improve their impairment.

These tasks and their initial level of difficulty is based on the therapists initial evaluation of the child. PT3, PT4, PT5 said that they often begin with the intervention phase by setting certain goals. These goals can be for example to make the children perform certain daily activities with minimum support. PT1 said that to improve the trunk balance, the therapists develop exercises based on their knowledge of physiotherapy. These exercises often involve sitting and reaching tasks where sitting can be with and without a support structure under the child's feet. The child can be asked to reach an object while sitting or while resting its knees on the floor. PT2, PT3, P4 and PT5 also mentioned involvement of sitting exercises that use the measure of time, during which the child was able to balance themselves, to assess the performance in the exercise. Most of these exercises are functional in nature, simulating activities of daily living. For example, the children are made to sit on a chair that simulates a toilet seat with and without an armrest.

As per PT2, sometimes they modify the standard trunk control exercises (designed for healthy children) in terms of repetitions and intensity to suit their patients because there is very low data availability for children with disorders affecting the trunk region, with the exception of Cerebral Palsy. The difficulty of these exercises is changed (increased, decreased or kept the same) based on many factors. The therapists mentioned that they don't use any concrete measures to switch the difficulty levels. Often a lot of trial and error is involved. They see how the children react to a particular difficulty level. If they see signs of stress or pain or anxiety, then they lower the difficulty or sometimes modify the exercise. If they think that the children are getting too comfortable, indicating gain in trunk control, with a particular level of difficulty then they often increase the level to motivate the children to come out of their comfort zone.

All the PTs agreed that these exercises are often performed in a fun, cooperative but quiet environment without many distractions. The PTs said that they need to find out the interests of children in order to motivate them to perform the exercises. For some children it can be toys, for others it can be visual or auditory stimulus or something totally different. Based on the interest of the children, the therapist use such objects of interest to design the exercises. The use of toys is noted for the experiment design in this study, to increase motivation.

Assessment of Movements : During this phase, the PTs make an evaluation if the intervention or exercise that the child is doing is correct or wrong. PT1 said that they focus on total body motion of the child, instead of evaluating every bio-mechanical aspect of the motion. They use their knowledge of existing disabilities in the children while judging whether the particular movement in a exercise trial is correct or wrong. PT3 and PT4 said that they also take help from their colleagues to confirm and evaluate the actions, when in doubt about some movements exhibited by the child. No assessment tests or instruments were reported as aids in the assessment of the performance of the exercises. All the PTs agreed that this section of the therapy is highly subjective and inter-PT repeatability on the evaluation of the same exercise movement will be low. The current study addresses this gap in the rehabilitation scheme, were assessments of performance are made while the children are engaged in trunk-balance exercises.

Role of automation Finally, the therapists were also asked where they think would technology / automation will help in their work. PT1 said that they would like some help with relaxing and calming the kids during exercise that they find too exciting. Further they added that a system to detect anxiety in children can be very helpful as this can help the therapist to modify the exercise accordingly. PT1 and PT4 said that a system that tells them what is wrong with the exercise, how well an exercise is done or what is the current training progress or what level of support is required can be helpful.

PT2 and PT3 added that sensor technology can also help. They said that pressure plates could help to know while sitting how much pressure the child is putting on its left and right or to understand what is normal and what is not normal. They also talked about the use of sensor technology to compare their patients with the healthy population. Finally, PT1 also added that while sensor technology is good and can be very helpful, in their experience, most of the children will reject placement of sensors on their body or any contact with sensors. Thus non-intrusive sensor modalities are preferred over more intrusive sensor modalities to increase acceptance among children.

Furthermore, PT5 added that Interactive games, virtual reality, use of touch screens on wall or floor with games will be important to motivate the children in performing the exercises. PT5 also stated that they would like a system or software that can measure small changes / errors / improvements in the exercises that cannot be measured with the naked eye. All the therapists said that the evaluation of exercise movements is generally very customized, very expertise-dependent and the elicitation of the evaluation procedure into explicit rules is very difficult.

2) Second Interviews: The second round of interviews with two PTs (PT5 and 6) resulted in the finalisation of the exercise design and the assessment factors to evaluate the movements. The final set of factors consist of: Quality of movement (QOM) corresponding to the subjective evaluation of the PT for the exercise movement, Speed (SP) and the presence of 12 compensation movements which can impair the quality of movement. Non-Active arm swing backward (NASB) is defined as the extension of the non-active (non-reaching) arm backward in order to maintain balance when the trunk muscles are not sufficient for the same. Feet Swing (FS) is similarly defined to NASB as the swinging motion of the feet to gain balance. Head movement (HM) is movement of neck muscles that is not expected in un-affected movement for the task at hand. Seat Support with hand (SSH) is when the subject maintains balance using the support of the seat by touching it using their non-active hand. Desk Support (DS) is when the balance is achieved by resting on the table in front of them while reaching either with the active or non-active hand. Bottom lifting (BL) is when the pelvic region is lifted off the seat. Feet going up return journey (FRJ) is the action of balancing the body during the backward motion of the trunk while returning from the target. Trunk inclination (TI) is the inclination in the trunk while executing the exercise movement. Feet moving forward (FMF) is a forward extension of the knees during the motion. Too much support on lap (ASL) is when the weight distribution of the body is skewed towards one side, Antepulsion with shoulder (AS) is the antepulsion motion of the shoulder where the shoulder is extended to increase the range of motion. Increase of balance surface by extending feet (BSEF) is the motion of the feet whereby the balance area on the ground is increased by widening the stance. Based on the interview results and findings from the available literature, the assessment factors and the scale on which they are evaluated are given in Table V.

TABLE V
ASSESSMENT FACTORS FOR EXERCISES

Assessment Parameter	***classes***		
QOM	Good-2	Normal-1	Poor-0
SP	Fast-2	Normal-1	Slow-0
NASB	Yes-1	No-0	
FS	Yes-1	No-0	
HM	Yes-1	No-0	
SSH	Yes-1	No-0	
DS	Yes-1	No-0	
BL	Yes-1	No-0	
FRJ	Yes-1	No-0	
TI	Yes-1	No-0	
FMF	Yes-1	No-0	
ASL	Yes-1	No-0	
AS	Yes-1	No-0	
BSEF	Yes-1	No-0	

B. Pre-processing

To account for the variance in target distance arising from the different body lengths of the participant, the experiment was designed such that for each subject three target distances were identified according to the ease with which they could obtain the target: easy- for target distances where the participant didn't have to move their trunk at all, hard- for target distances where the participant had to be in full trunk and hand extension to reach the target on the table and medium-where the participants were optimally challenged between the two extremes. This introduced the hypothesis that the "easy" and "hard" target distances would have an imbalanced distribution of quality scores. It was expected that "easy" targets would be skewed with better quality scores (2 out of 2) whereas the "hard" targets would be skewed with poorer quality scores (0 out of 2). To test this hypothesis, we counted the frequency of 0, 1 & 2 scores in the easy, medium and hard target distances categories. The easy, medium and hard target distances bins were established by obtaining the height-normalised target distance in each exercise motion from the position of the pelvis of each participant and binning each of these with an equal-width binning strategy between the shortest and the longest target distances in the dataset.

Table VI shows the frequencies of the quality scores normalised by their total occurrences in the entire dataset respectively in each target distance bin. It was observed that ”hard” target distance category showed the expected unbalance in quality scores, where harder targets got more ”poorer” scores (0 out of 2) while surprisingly the easy and medium bins were both well-balanced. As some imbalance still exists in the dataset (found in the ”hard” distance category), there is a possibility that the classification algorithms use the target distance embedded in the input features to predict the quality scores. This is undesirable as it is of more interest to enable the classification algorithms to find correlations in the data other than that of target distance on the quality scores, as even within harder targets, there still exists some variance in the quality scores, albeit in an imbalanced manner. With these results, it was determined to normalise all the input features of each exercise trial using the height normalised target distance.

TABLE VI
EQUAL-WIDTH BINNING FOR TARGET-DISTANCE QUALITY SCORES

		target_distance		
		0-easy	*1-medium*	*2-hard*
QOM	0	0.32	0.34	**0.64**
	1	0.33	0.34	**0.23**
	2	0.35	0.32	**0.14**

C. AI classification

The performance of two algorithms, ROCKET and XCM for all the assessment factors are reported in this section. Additionally, the performance on predicting the hand used and the region of the target is also reported. The cross validation f1-scores for the two algorithms are shown in Table. VII. For both algorithms, across all assessment factors, the cross-validation test on the validation set resulted in low variance in the f1-scores across validation folds as indicated by the values of standard deviation in the table. The f1-scores on the training set in the folds are close to 1.0 in all the runs, indicating overfitting of the model to the training set. The classification performance in the cross-validation tests is highest in the case of Hand and Target predictions and identical for both the algorithms. The performance on predicting quality of movements is higher in XCM with an f1-score of 0.56 while ROCKET has an f1-score of 0.49.

The final classification performance is obtained by training the model on the entire dataset and testing on the test set. The results for both algorithms are reported in Table. Notably, XCM has a better prediction performance on quality than ROCKET with a f1-score of 0.66 and 0.56 respectively. The highest classification performance is seen once again with the Hand and target classification tasks, but in this, XCM has a higher performance for Hand with a f1-score of 0.95 over 0.92 of ROCKET, while ROCKET classifies target region better with a f1-score of 0.86 over 0.81 of XCM. For the other assessment factors, the reported performance is slightly above the chance levels corresponding to the respective factors. In case of Speed, the chance performance is 0.33, while the chance level for all the compensation factors is 0.5.

TABLE VII
CROSS VALIDATION F1-SCCORES

Assessment Factor	***ROCKET*** *mean*	*std*	***XCM*** *mean*	*std*
QOM	0.49	0.05	0.56	0.07
SP	0.59	0.07	0.57	0.08
NASB	0.63	0.1	0.65	0.11
HM	0.57	0.06	0.64	0.05
SSH	0.59	0.22	0.69	0.12
DS	0.66	0.07	0.69	0.07
BL	0.63	0.04	0.71	0.07
TI	0.54	0.08	0.54	0.09
ASL	0.67	0.05	0.68	0.04
AS	0.51	0.06	0.59	0.07

TABLE VIII
F1-SCORES ON TEST DATA

Assessment Factor	***f1_ROCKET***	***f1_XCM***
QOM	0.56	**0.66**
SP	**0.47**	0.38
NASB	**0.62**	0.48
HM	0.61	**0.67**
SSH	0.61	**0.66**
DS	**0.72**	0.69
BL	**0.69**	0.64
TI	0.54	**0.73**
ASL	**0.67**	0.54
AS	**0.65**	0.63

D. Explanations of AI model

XCM is a explainable AI algorithm which provides prediction explanations using GRAD-CAM on individual predictions by highlighting the features and time stamps relevant for the classification of the input data into one of the classes. The input data, in our case consists of 3d coordinates of 10 body part; pelvis, chest, neck, left shoulder, left elbow, left hand, right shoulder, right elbow, right hand and head.

For the first analysis, two sets of GRAD-CAM images on the task of predicting the hand is shown in Fig. 5 and Fig. 8. The y-axis consists of the 30 features (10 body parts with 3 coordinates each) in the order listed before and the x-axis is the temporal dimension. The red rectangles on the images indicate the region on the image corresponding to features of the correct body parts associated with the prediction label. In this case, the left and right side features are shown in the region encapsulated by the red rectangle. An example of a correct explanation for predicting ”left” as the correct active hand is seen in Fig. 3, while an incorrect explanation can be seen in Fig. 4. Similarly, the explanations for predicting right as the correct active hand can be seen in Fig. 8. It can be seen that the highlighted features for these two classification instances for correct explanations correspond to the features belonging to the respective sides of the body, but in the other GRAD-CAM images they do not correspond and thus are not correct explanations for the predictions.

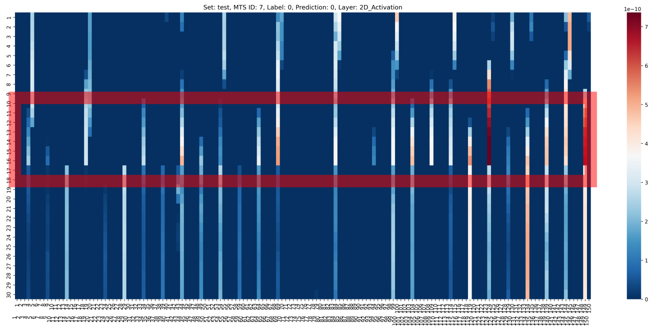


Fig. 3. Example of good explanation

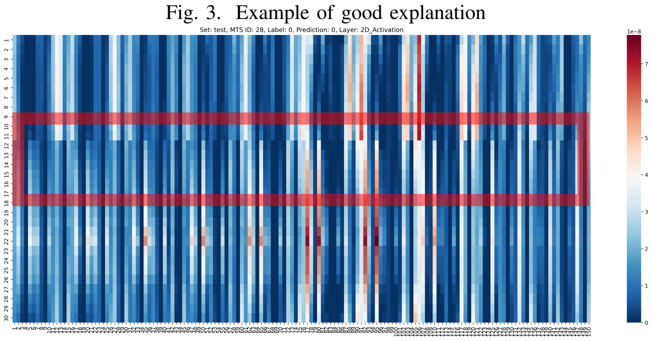


Fig. 4. Example of bad explanation

Fig. 5. GRAD-CAM examples for predicting ”left” as correct active hand

Secondly, Fig. 11, shows two GRAD-CAM images for a prediction on the absence and presence of BL compensation movement in the exercise movement respectively. The green rectangle indicates the region of the image corresponding to the body part ”pelvis”, which PT6 reported as being significant for their judgement on the presence of this compensation movement. It can be seen that both predictions highlight the same features as being significant for the decision to classify the presence or absence of the BL compensation movement.

In addition to the GRAD-CAM prediction explanations, it was hypothesised that the presence of compensation movements in the exercise motion can be used as explanation for poorer quality scores. For this, PT6 first provided their judgement on relative importance data regarding the relationship between quality score and the other assessment factors. As stated in the Interviews section of this chapter, PT6 judged how some compensatory variables are more relevant / significant than the others. They mentioned that SSH, DS, BL, FRJ and ASL were more significant than SP, HM, TI, FS, FMF, NASB, AS, BSEF. To cross-validate the judgement of PT6, first, an OLS analysis was conducted to investigate the associations between quality and the other factors. It was performed also to check the statistical significance of the independent variables in comparison to each other. The results of this analysis are

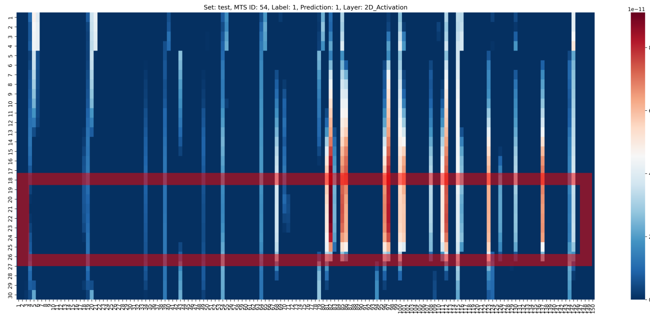


Fig. 6. Example of good explanation

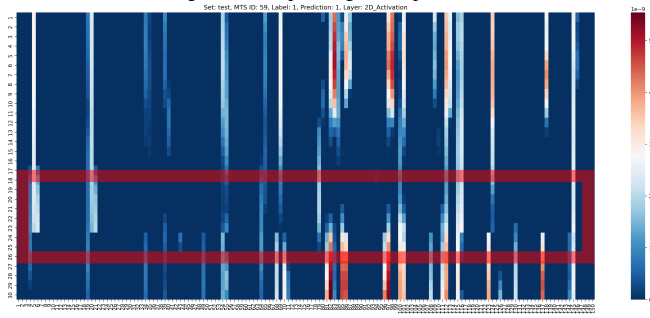


Fig. 7. Example of bad explanation

Fig. 8. GRAD-CAM examples for predicting ”right” as correct active hand

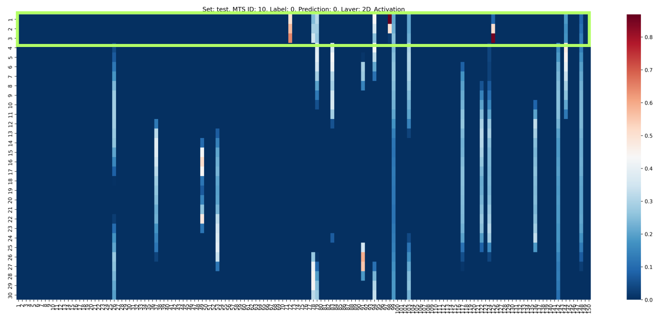


Fig. 9. Predicting the absence of BL compensation

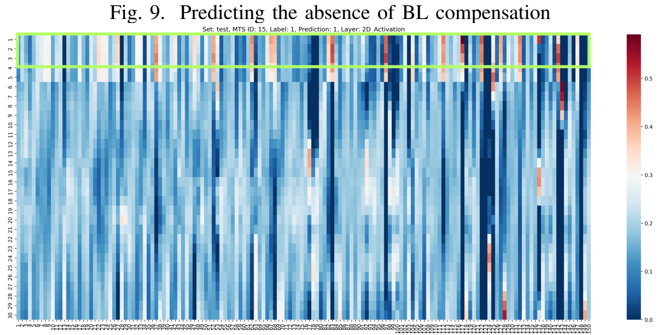


Fig. 10. Predicting the presence of BL compensation

Fig. 11. GRAD-CAM examples for predicting the presence and absence of BL compensation

reported below.

OLS analysis The results on the OLS analysis indicated that all the independent variables are statistically significant and the model has R2 and adjusted R2 values of about 62% each. Thus, it can also be said that about 62% of the Quality of Movement is influenced or can be captured by these independent variables. The positive / negative signs of the coefficients of independent variables can be used to estimate the nature of relationship between QOM and the independent variables. For example, higher speed values can be approximated to higher value of QOM or higher TI (presence of TI) can be negatively associated with the QOM (poor QOM). Notably, the overall empirical model remains statistically significant.

Next, the multicollinearity assumption of linear regression has been validated for the empirical model via diagnostic tests to check if the data has any multicollinearity. A check for multicollinearity was done using the Variance Inflation Factor (VIF) test. Other than the variable AS, no other independent variable had a VIF value of greater than 3. Thus, only AS showed the presence of multicollinearity. After removing AS, a second OLS multivariate regression analysis was performed with 12 independent variables. The results of this analysis are shown in column (2) of table X. It has been shown that other than variables SP, FS, and FMF all other variables remain statistically significant at p values less than 0,05. These remaining nine statistically significant independent variables are all negatively associated with QOM meaning that presence of these movement compensations can negatively affect the quality of movement. The nine statistically significant variables are NASB, HM, SSH, DS, BL, FRJ, TI, ASL, BSEF. This means that for instance, the presence of NASB or HM can negatively affect the QOM. The overall model remains significant as well. In this instance, the R2 and adjusted R2 values are about 65 percent. Thus, 65% of the changes to QOM can be explained by the independent variables. The results of this OLS regression were then compared with the judgement of a physiotherapist, who was also involved with the data labeling process of the exercise trials. The result is shown in Table IX where the matching assessment factors influencing the quality score are highlighted.

TABLE IX
OLS RESULTS ON FACTOR IMPORTANCE FOR PREDICTING QOM

Factor	PT	OLS
SP	No	No
NASB	No	Yes
FS	No	No
HM	No	Yes
SSH	Yes	Yes
DS	Yes	Yes
BL	Yes	Yes
FRJ	Yes	Yes
TI	No	Yes
FMF	No	No
ASL	Yes	Yes
AS	No	No
BSEF	No	Yes

V. Discussion

A. Interviews

The interviews with the physiotherapists yielded some critical information about the evaluation and treatment of children with trunk-control issues. The most common exercises that are developed to improve trunk instability and impairment in the patients are the sitting and reaching exercises. Since many of our daily activities, like sitting on a toilet seat, or sitting on a chair involve sitting and reaching for other objects, trunk-control rehabilitation exercises are designed to simulate these activities. Such exercises are desired to enable the children to become independent in these activities or reduce the extent of support that is required.

The exercises are personalized to the patients in terms of movements, complexity levels and a fun and safe learning environment is ensured. Thus, it is important to note that different objects like toys, puzzles can be used to motivate the children to perform these exercises in a fun way. Based on the answers of the physiotherapists, it is important to note that there is no one-size fits all solution in terms of rehabilitation treatments for the children. The treatments are customized and very often the experience and tacit knowledge of the physiotherapist plays a crucial role in the treatment. For example, an experienced physiotherapist might have better judgement of the quality of motion performed by a patient or an experienced PT can adapt the exercise complexity more accurately than a less-experienced PT. Therefore, in such cases, the AI model can serve as a baseline for the PTs to judge the quality of movement. The PTs also indicated that such a system, if also explainable, can help them in identifying problematic motion sequences during the exercises. The analysis of these interviews also provided critical information for development of the exercise trials which involved mainly sitting and reaching tasks of varying difficulties as these exercises target the trunk area for children experiencing trunk instability.

B. Pre-processing

The classification algorithms used in this study require time series of equal-length. A limitation of the study is the use of padding with zero's for making all the time series data of equal length. Padding with other values, such as the mean of last few data points or the value at the extreme should be investigated in a future study. The effect of different types of interpolation techniques can also be included in a future study to study the impact on the prediction of speed of movement.

An ablation study was performed on the normalisation preprocessing step in the pipeline, reported in Table. X. It is seen that the performance of the algorithms for normalised data was generally higher than that of non-normalised data. Specifically for QOM predictions, the performance of model using normalized data is much higher than the non-normalised counterpart. This can be seen as a result in support of normalisation of the input data based on participant's height-adjusted target distance from the pelvis. Non-normalised data could have led to the learning of associations such as that of the maximum value of the features (corresponding to the Range of Motion) on the quality score. It was reported earlier that such a relationship could be seen only in one (hard) out of the three target-distance bins (easy, medium and hard) and even in this, it is important to note the number of target-distance instances in that particular bin were lesser than the other the instances in the other two bins. Thus a small portion of the dataset showed this correlation between the target-distance and a poorer quality score and if the model would have picked up on this correlation and applied it to all the instances, that could explain the lower prediction performances.

TABLE X
EFFECT OF NORMALISATION ON CLASSIFICATION

Assessment Factor	*non-norm*	*norm*
QOM	0.41	**0.66**
SP	0.48	0.38
NASB	0.56	0.48
HM	0.61	**0.67**
SSH	0.72	0.66
DS	0.67	**0.69**
BL	0.64	**0.64**
TI	0.63	**0.73**
ASL	0.67	0.54
AS	0.57	**0.63**

C. Classification

The classification performance for quality by both the algorithms are promising. We have achieved a f1-score of 0.66 in 3-class classification problem by using the XCM algorithm for predicting the quality of exercise movements in trunk control exercises performed by children. However, in the cross validation tests, it was seen that the f1-score on the training set is generally close to 1.0, indicating an over-fit of the model to the training data. This could be an indicator of the model learning the salient features and noise in the training data to such an extent that the performance on unseen data suffers. Adding more data with different variations could be one of the approaches to alleviating this issue to help the model generalise better. Generally, increasing the number of patients should result in reducing the over-fitting. The over-fit on training data could also be an indicator of too high model complexity and regularisation on the model parameters can be investigated to increase generalisation over unseen data.

Another limitation of the study, is that the labels for the classification task are collected only from one PT and thus investigations on noisy labels were not possible. Collection of labels from multiple PTs to investigate the consistency of predictions between PTs, as reported in [16], could help in highlighting the variations possible in judging the same input data by two different decision makers and also serve to compare the AI's prediction to another source of ground truth. Multiple rounds of label collection from the same PT for the same input data after a time interval between the two label collection events could also provide interesting insights about the repeatability of predictions by a PT and comparing the PT's first assessment with the AI's performance (trained on the first labels) as well as the second labels by the PT, can produce interesting insights into trustability of the PTs labels as input for this classification task. In a future work, it is also desirable to be able to use the kinematic data from all the body parts, including the lower limb for analysis, as according to the PT and OLS analysis, some critical information for the predicting the quality is also embedded within the compensations related to the features involving the lower body.

D. Explainability

The explainability results reported using GRAD-CAM show the potential of this method to provide faithful explanations for prediction on assessment factors in physiotherapy. Using the highlighted relevant features and timestamps in the input data for a prediction, PTs can form an understanding of which part of the data were of importance to the AI model's decisions. The use of clinically interpretable features in the input data to the model and the availability of the explanation in the input-space greatly increases the interpretability of this approach and has the potential to increase the clinical relevance and acceptability of such solutions. GRAD-CAM only indicates which features were relevant for the classification by the AI model, and not how those features are used for the classification. This is noted as a limitation as even in prediction tasks with a very high classification performance, such as predicting the hand which was used during the exercise, some of the highlighted features did not correspond to the expected set of features. This highlights that although the indicated features were important for classification, we do not gain any new insights as to why the features were important. If a method can also provide the reason why a set of features are important, then it can lead to greater interaction between the AI solution and the users and increase trust between them. This is were other XAI approaches may be of benefit and should be further investigated.

The presence of other assessment factors for explaining (poor) quality score is also investigated in this study, as an alternative to XAI methods, such as GRAD-CAM reported before. For supporting this hypothesis, we obtained the relevance of each of the other assessment factors for predicting a poor quality score and compared it to the results of OLS regression analysis on the associations between the other factors and quality. According to the physiotherapist, the independent variables of SP, NASB, FS, HM, TI, FMF, AS and BSEF are not very relevant to the quality of motion. This judgement partially matches the results of OLS regression which marks SP, FMF, AS and FS are statistically non-significant at p¡0,05. However, the variables NASB, HM, TI and BSEF which were marked as "not very relevant" by the physiotherapist are found to be statistically significant by the model. The divergence in the relevance results could be attributed to the limitation of the OLS regression method applicability for categorical variables, which is also to be noted as a limitation in this study, and the subjective perception of PT6 on their reasons for their decision making. To overcome these limitations, it is worth analysing in a future work, the associations between the variables using a OLS regression model suited to categorical variables and also include more PTs' input for this analysis to increase the data points and check for consistency between PTs regarding the relevant factors for predicting quality. A limitation of the current study is the experience of the PT who provided the ground truth for this experiment (PT6 - 2 years). In future, more experienced PTs can be involved for obtaining the ground truth and explanation evaluations. Once the limitations are overcome, the final set of assessment factors can be used as potential explanations on quality scores as indicated by the input of the PT and the statistical analysis.

VI. Conclusion

To aid the PTs in their therapy for children with trunk-control problems, in this thesis we have developed a trustworthy AI model that will assess the quality of rehabilitation exercises. By conducting interviews and focused discussions with six PTs, we have gathered important knowledge on trunk-control rehabilitation in children, specifically regarding the assessment of quality during exercise movements, previously not available in literature.

From the interviews we understood that the task of physiotherapist during exercises is complex and the assessments are made based on knowledge of the patients' impairments and the physiotherapist's intuitions build through training and experience over the years. The interviews yielded that sitting and reaching are the most common exercises in trunk rehabilitation for children. Quality, speed and 13 compensatory movements during trunk rehabilitation exercises involving sitting and reaching were identified and the scales of scoring on them were established through these interviews. The PTs indicated that for capturing motion data of the children, it will be critical not to use intrusive sensors, for practical acceptance by the children.

Using the information gained from the interview, a data collection experiment was designed where 3 sets of sitting and reaching exercise motions were performed by TD children. Two state-of-the-art methods for multi-variate time series classification were trained on this data. XCM, a CNN-based method performed best on the task of quality prediction with an f1-score of 65% on the test data. This is, to the knowledge of the author, the first time state-of-the-art multi-variate time series classification algorithms have been applied for assessment of movement quality in physiotherapy exercises involving children. Additionally, this work demonstrates the potential of XAI as a tool to increase the trustworthiness of AI solutions for physiotherapy. The explanations provided by XCM show potential to enable the users of the system to gain insights on regions of the input that are important for a specific prediction. This will enable trust building between the system and the users by providing open and interpretable decisions.

We believe that the methodology for eliciting PT knowledge and requirements for an AI model, use of non-intrusive sensor technology to collect kinematic data, the performance obtained on the multi-variate time series algorithms and explanations on predictions are have addressed a significant gap in literature and clinical practice for automated assessments of physiotherapy exercises, specifically for trunk-control rehabilitation for children.

References

[1] Doreen Bartlett and Barbara Purdie. "Testing of the spinal alignment and range of motion measure: a discriminative measure of posture and flexibility for children with cerebral palsy". eng. In: *Developmental Medicine and Child Neurology* 47.11 (Nov. 2005), pp. 739–743. ISSN: 0012-1622. DOI: 10 . 1017 / S0012162205001556.

[2] Giuseppe Caggianese et al. "Serious Games and In-Cloud Data Analytics for the Virtualization and Personalization of Rehabilitation Treatments". en. In: *IEEE Transactions on Industrial Informatics* 15.1 (Jan. 2019). 5 citations (Semantic Scholar/DOI) [2022-01-07], pp. 517–526. ISSN: 1551-3203, 1941-0050. DOI: 10.1109/TII.2018.2856097. URL: https://ieeexplore.ieee.org/document/8410810/ (visited on 08/09/2021).

[3] Mariana Campos, Rafael Caldas, and Fernando Buarque. "eXplainable and Individualizable for Physiotherapeutic Decision Support for the Elderly". In: *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. ISSN: 2577-1655. Oct. 2021, pp. 1876–1881. DOI: 10.1109/SMC52423.2021.9659272.

[4] Alarcos Cieza et al. "Global estimates of the need for rehabilitation based on the Global Burden of Disease study 2019: a systematic analysis for the Global Burden of Disease Study 2019". en. In: *The Lancet* 396.10267 (Dec. 2020), pp. 2006–2017. ISSN: 01406736. DOI: 10.1016/S0140-6736(20)32340-0. URL: https://linkinghub.elsevier.com/retrieve/pii/S0140673620323400 (visited on 04/06/2022).

[5] Angus Dempster, François Petitjean, and Geoffrey I. Webb. "ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels". en. In: *Data Mining and Knowledge Discovery* 34.5 (Sept. 2020), pp. 1454–1495. ISSN: 1573-756X. DOI: 10.1007/s10618-020-00701-z. URL: https://doi.org/10.1007/s10618-020-00701-z (visited on 08/31/2022).

[6] Ashish Dhiman et al. "Design of adaptive haptic-enabled virtual reality based system for upper limb movement disorders: A Usability Study". en. In: *2016 6th IEEE International Conference on Biomedical Robotics and Biomechatronics (BioRob)*. 6 citations (Semantic Scholar/DOI) [2022-01-07]. Singapore, Singapore: IEEE, June 2016, pp. 1254–1259. ISBN: 978-1-5090-3287-7. DOI: 10.1109/BIOROB.2016.7523803. URL: http://ieeexplore.ieee.org/document/7523803/ (visited on 08/02/2021).

[7] Carlo Dindorf et al. “Classification and Automated Interpretation of Spinal Posture Data Using a Pathology-Independent Classifier and Explainable Artificial Intelligence (XAI)”. en. In: *Sensors* 21.18 (Jan. 2021). Number: 18 Publisher: Multidisciplinary Digital Publishing Institute, p. 6323. ISSN: 1424-8220. DOI: 10.3390/s21186323. URL: https://www.mdpi.com/1424-8220/21/18/6323 (visited on 02/03/2022).

[8] Kevin Fauvel et al. “XCM: An Explainable Convolutional Neural Network for Multivariate Time Series Classification”. In: *Mathematics* 9.23 (Dec. 2021). arXiv:2009.04796 [cs, stat], p. 3137. ISSN: 2227-7390. DOI: 10.3390/math9233137. URL: http://arxiv.org/abs/2009.04796 (visited on 09/29/2022).

[9] S. E. Fife et al. “Development of a clinical measure of postural control for assessment of adaptive seating in children with neuromotor disabilities”. eng. In: *Physical Therapy* 71.12 (Dec. 1991), pp. 981–993. ISSN: 0031-9023. DOI: 10.1093/ptj/71.12.981.

[10] Lisa Anne Hendricks et al. “Grounding Visual Explanations”. In: 2018, pp. 264–279. URL: https://openaccess.thecvf.com/content_ECCV_2018/html/Lisa_Anne_Hendricks_Grounding_Visual_Explanations_ECCV_2018_paper.html (visited on 08/29/2022).

[11] Lieve Heyrman et al. “A clinical tool to measure trunk control in children with cerebral palsy: The Trunk Control Measurement Scale”. eng. In: *Research in Developmental Disabilities* 32.6 (Dec. 2011), pp. 2624–2635. ISSN: 1873-3379. DOI: 10.1016/j.ridd.2011.06.012.

[12] Andreas Holzinger et al. “What do we need to build explainable AI systems for the medical domain?” In: *arXiv:1712.09923 [cs, stat]* (Dec. 2017). arXiv: 1712.09923 version: 1. URL: http://arxiv.org/abs/1712.09923 (visited on 02/02/2022).

[13] Vera C. Kaelin et al. “Artificial Intelligence in Rehabilitation Targeting the Participation of Children and Youth With Disabilities: Scoping Review”. EN. In: *Journal of Medical Internet Research* 23.11 (Nov. 2021). Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada, e25745. DOI: 10.2196/25745. URL: https://www.jmir.org/2021/11/e25745 (visited on 01/19/2022).

[14] Katrijn Klingels et al. “Upper limb motor and sensory impairments in children with hemiplegic cerebral palsy. Can they be measured reliably?” eng. In: *Disability and Rehabilitation* 32.5 (2010), pp. 409–416. ISSN: 1464-5165. DOI: 10.3109/09638280903171469.

[15] Karen Kott and Sharon Held. “Refinement, Reliability, and Validity of the Segmental Assessment of Trunk Control”. en-US. In: *Pediatric Physical Therapy* 22.3 (2010), p. 258. ISSN: 0898-5669. DOI: 10.1097/PEP.0b013e3181eb6561. URL: http://journals.lww.com/pedpt/fulltext/2010/22030/Refinement,_Reliability,_and_Validity_of_the.3.aspx (visited on 10/24/2022).

[16] Min Hun Lee et al. “Co-Design and Evaluation of an Intelligent Decision Support System for Stroke Rehabilitation Assessment”. en. In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW2 (Oct. 2020). 7 citations (Semantic Scholar/DOI) [2022-01-07], pp. 1–27. ISSN: 2573-0142. DOI: 10.1145/3415227. URL: https://dl.acm.org/doi/10.1145/3415227 (visited on 07/12/2021).

[17] Min Hun Lee et al. “Learning to assess the quality of stroke rehabilitation exercises”. en. In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 20 citations (Semantic Scholar/DOI) [2022-01-07]. Marina del Ray California: ACM, Mar. 2019, pp. 218–228. ISBN: 978-1-4503-6272-6. DOI: 10.1145/3301275.3302273. URL: https://dl.acm.org/doi/10.1145/3301275.3302273 (visited on 07/15/2021).

[18] Yalin Liao, Aleksandar Vakanski, and Min Xian. “A Deep Learning Framework for Assessing Physical Rehabilitation Exercises”. en. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 28.2 (Feb. 2020), pp. 468–477. ISSN: 1534-4320, 1558-0210. DOI: 10.1109/TNSRE.2020.2966249. URL: https://ieeexplore.ieee.org/document/8957502/ (visited on 01/27/2022).

[19] Edwin Daniel Ona Simbana et al. “Review of Automated Systems for Upper Limbs Functional Assessment in Neurorehabilitation”. en. In: *IEEE Access* 7 (2019), pp. 32352–32367. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2019.2901814. URL: https://ieeexplore.ieee.org/document/8653273/ (visited on 01/18/2022).

[20] Avinash Parnandi et al. “PrimSeq: a deep learning-based pipeline to quantitate rehabilitation training”. In: *arXiv:2112.11330 [cs]* (Dec. 2021). arXiv: 2112.11330. URL: http://arxiv.org/abs/2112.11330 (visited on 01/19/2022).

[21] Michele Pirovano et al. “Self-adaptive games for rehabilitation at home”. en. In: *2012 IEEE Conference on Computational Intelligence and Games (CIG)*. 97 citations (Semantic Scholar/DOI) [2022-01-07]. Granada, Spain: IEEE, Sept. 2012, pp. 179–186. ISBN: 978-1-4673-1194-6 978-1-4673-1193-9 978-1-4673-1192-2. DOI: 10.1109/CIG.2012.6374154. URL: http://ieeexplore.ieee.org/document/6374154/ (visited on 08/06/2021).

[22] *Postural Control : A Key Issue in Developmental Disorders — Wiley*. en-ie. URL: https://www-wiley-com.tudelft.idm.oclc.org/en-ie/Postural+Control+%3A+A+Key+Issue+in+Developmental+Disorders-p-9781898683575 (visited on 10/24/2022).

[23] Oliver Radley-Gardner, Hugh Beale, and Reinhard Zimmermann, eds. *Fundamental Texts On European Private Law*. en. Hart Publishing, 2016. ISBN: 978-1-78225-864-3 978-1-78225-865-0 978-1-78225-866-7 978-1-78225-867-4. DOI: 10.5040/9781782258674.

URL: http://www.bloomsburycollections.com/book/fundamental-texts-on-european-private-law-1 (visited on 02/02/2022).

[24] Michael Rowe, David A. Nicholls, and James Shaw. "How to replace a physiotherapist: artificial intelligence and the redistribution of expertise". en. In: *Physiotherapy Theory and Practice* (June 2021), pp. 1–9. ISSN: 0959-3985, 1532-5040. DOI: 10.1080/09593985.2021.1934924. URL: https://www.tandfonline.com/doi/full/10.1080/09593985.2021.1934924 (visited on 06/16/2022).

[25] Lesley Scobbie, Diane Dixon, and Sally Wyke. "Goal setting and action planning in the rehabilitation setting: development of a theoretically informed practice framework". In: *Clinical Rehabilitation* 25.5 (May 2011). Publisher: SAGE Publications Ltd STM, pp. 468–482. ISSN: 0269-2155. DOI: 10.1177/0269215510389198. URL: https://doi.org/10.1177/0269215510389198 (visited on 06/16/2022).

[26] Ramprasaath R. Selvaraju et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. ISSN: 2380-7504. Oct. 2017, pp. 618–626. DOI: 10.1109/ICCV.2017.74.

[27] Soroosh Shahtalebi et al. "A deep explainable artificial intelligent framework for neurological disorders discrimination". en. In: *Scientific Reports* 11.1 (May 2021). Number: 1 Publisher: Nature Publishing Group, p. 9630. ISSN: 2045-2322. DOI: 10.1038/s41598-021-88919-9. URL: http://www.nature.com/articles/s41598-021-88919-9 (visited on 02/03/2022).

[28] *sktime-rocket*. URL: https://github.com/sktime/sktime (visited on 10/24/2022).

[29] Djordje Slijepcevic et al. "On the Explanation of Machine Learning Predictions in Clinical Gait Analysis". In: *arXiv:1912.07737 [cs, stat]* (Aug. 2020). arXiv: 1912.07737. URL: http://arxiv.org/abs/1912.07737 (visited on 02/01/2022).

[30] Stanford University, United States of America and Jan Auernhammer. "Human-centered AI: The role of Human-centered Design Research in the development of AI". en. In: Sept. 2020. DOI: 10.21606/drs.2020.282. URL: https://dl.designresearchsociety.org/drs-conference-papers/drs2020/researchpapers/89 (visited on 10/23/2022).

[31] Jindong Wang et al. "Deep learning for sensor-based activity recognition: A survey". en. In: *Pattern Recognition Letters* 119 (Mar. 2019), pp. 3–11. ISSN: 01678655. DOI: 10.1016/j.patrec.2018.02.010. URL: https://linkinghub.elsevier.com/retrieve/pii/S016786551830045X (visited on 02/14/2022).

[32] XAIseries. *XCM: An Explainable Convolutional Neural Network for Multivariate Time Series Classification*. original-date: 2021-12-08T16:17:50Z. Oct. 2022. URL: https://github.com/XAIseries/XCM (visited on 10/24/2022).

2

Appendix A: Body pose tracking

The experimental protocol designed through interviews involved the capture of exercise motion data using a 3d stereo camera. The ZED mini camera was used with its accompanying SDK, which provided the 3d coordinates of 34 body part locations in the human body. The output can be seen in the figure below.

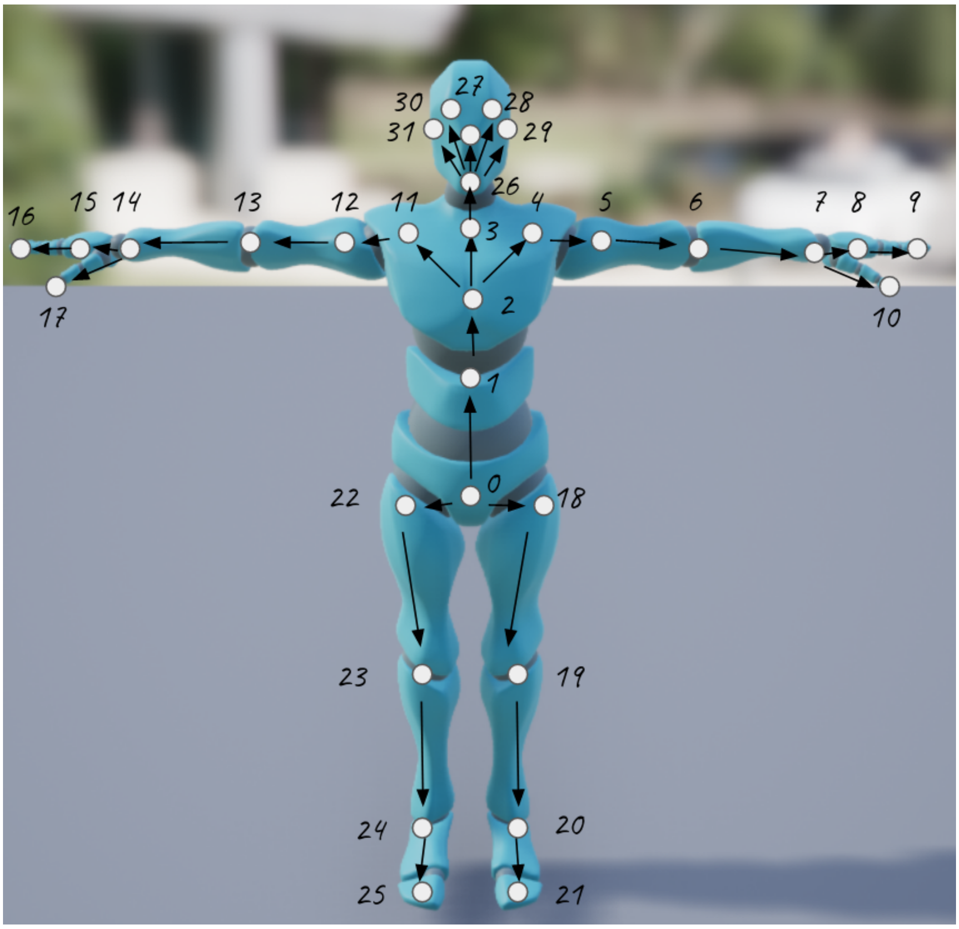


Figure 2.1: ZED body tracking output

Table 2.1: Body Part Index Mapping

Index	**Body Part**
0	PELVIS
1	NAVAL_SPINE
2	CHEST_SPINE
3	NECK
4	LEFT_CLAVICLE5
5	LEFT_SHOULDER
6	LEFT_ELBOW
7	LEFT_WRIST
8	LEFT_HAND
9	LEFT_HANDTIP
10	LEFT_THUMB
11	RIGHT_CLAVICLE
12	RIGHT_SHOULDER
13	RIGHT_ELBOW
14	RIGHT_WRIST
15	RIGHT_HAND
16	RIGHT_HANDTIP
17	RIGHT_THUMB
18	LEFT_HIP
19	LEFT_KNEE
20	LEFT_ANKLE
21	LEFT_FOOT
22	RIGHT_HIP
23	RIGHT_KNEE
24	RIGHT_ANKLE
25	RIGHT_FOOT
26	HEAD
27	NOSE
28	LEFT_EYE
29	LEFT_EAR
30	RIGHT_EYE
31	RIGHT_EAR
32	RIGHT_EAR
33	RIGHT_EAR

3
Appendix B: Frame of Reference

The camera was placed in a room with the participant diagonally in front. The data from the camera is in the frame of the reference of the camera. The system of axes can be seen in Fig. 3.1

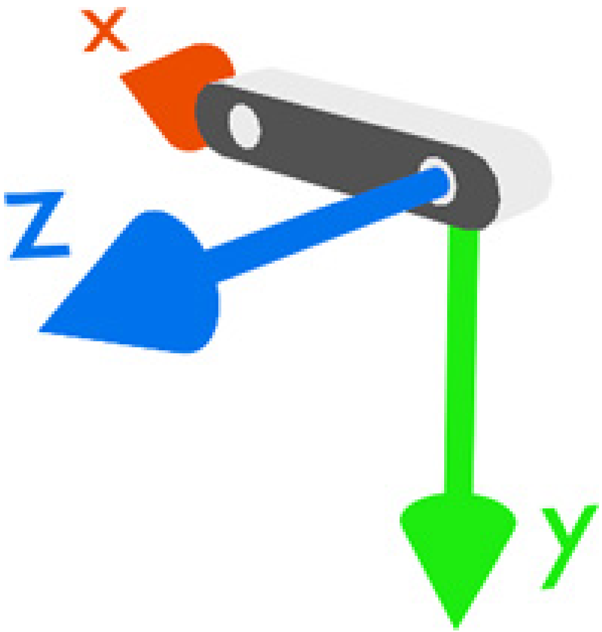


Figure 3.1: ZED Mini Axis system

Since, the positions readings of the tracked body parts are not easily interpretable in this axis system, we transformed the points to coincide to the "pelvis" frame defined at the location of the pelvis in the first captured frame in a trial. The axes of this new reference system were designed such that the x and y axes were oriented along the edges of the table and the z was anti-parallel to the table.

The final axis can be seen in Fig. 3.2

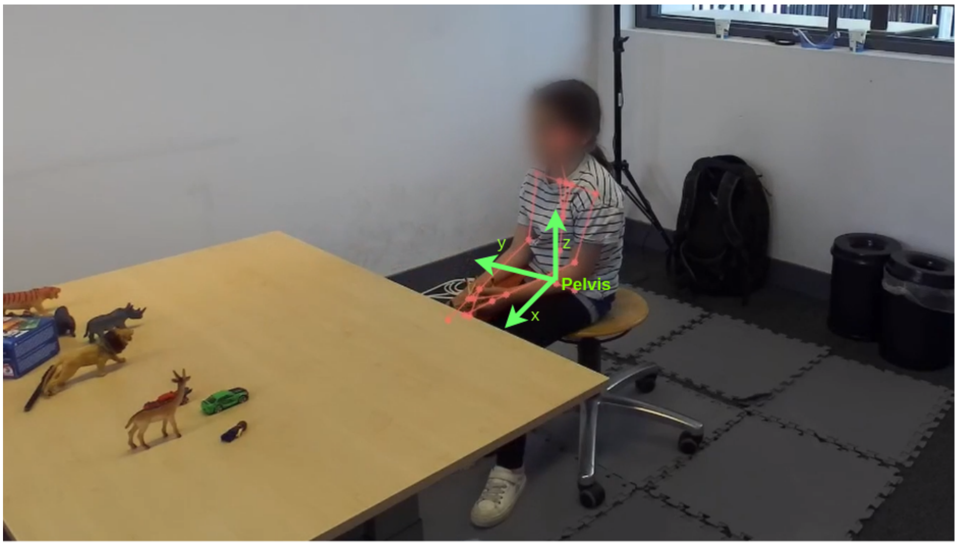


Figure 3.2: Participant and Pelvis frame in camera frame

4

Appendix C: Questionnaire for interviews

The following questionnaire was used to guide the interviews and focused discussions with the PTs which resulted in the elicitation of their expert knowledge.

Kinematic exercise assessment and automated difficulty adaptation in trunk stability training

This is a questionnaire for the project "Interpretable AI for Physiotherapy Decision Making". The goal is the development of an Artificially Intelligent (AI) model, which assists physiotherapists in task-level goal setting during trunk stability training in children. Motion Sensors will track the skeletal joints of the patient, while performing activities, disguised as a game, while sitting in front of a table. The position and velocity of the virtual or physical object on the table will be decided by the AI taking into account the inputs from the physiotherapist during model development. The AI system is intended to enable the rehabilitation of the trunk muscles in children - with a focus on static and dynamic reaching tasks while being seated. This survey is designed to gather clinical needs and inform the development of the AI model to facilitate the decision making of experts. The following questions relate to a 6-10 years old child with trunk control issues.

Consent Form

1. Your participation is voluntary and you can cancel the questionnaire at any time without giving a reason. The survey is anonymous and you will not provide any personal information other than your occupation. The data collected will be used to develop an AI system for trunk stability rehabilitation and can also be used for a scientific publication. Your data will be stored anonymously in a data storage system. By clicking on the button below, you consent to your participation.

   Check all that apply.

   - [ ] Agree
   - [ ] Disagree

Personal Information

2. In which clinic do you work?

   ____

3. In which area do you work professionally?

   Check all that apply.

   - [ ] Physiotherapy
   - [ ] Occupation Therapy
   - [ ] Other: ____________________

4. Have you worked with children with trunk stability issues?

   Mark only one oval.

   - ( ) Yes
   - ( ) No

5. How many years of experience do you have?

   ____

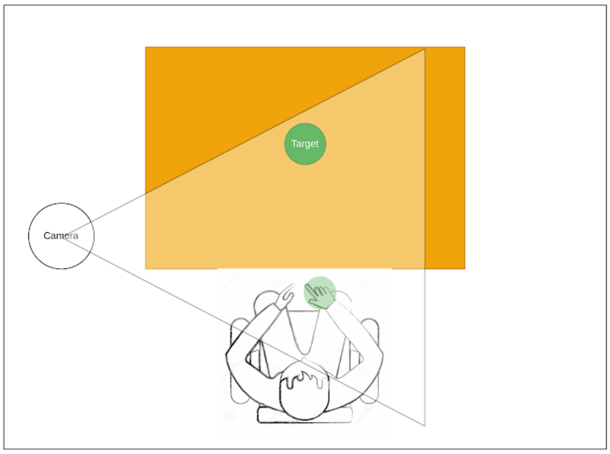
Trunk Control Exercises

Consider the following setting for the rest of the survey:
The child sits on a bench without back or arm support and with their feet firmly resting on the ground. The child is tasked with multiple reaching tasks having the following components:
- Starting position with arms on the lap
- Reach the virtual/physical object on the table
- Track/pick the object and stop at another location on the table
- Return to starting position with arms on the lap

Note: Each child or groups of children will have varying pathologies affecting their movements and different tasks may be relevant for each child/group. In this survey, we are attempting to gather all types of movements and assessment factors. Thus even if some of the options might not be relevant to a group of patients in your experience, and yet relevant for another group of patients, all of whom have trunk stability issues, please rate the factor as relevant. The AI system is envisioned to be under the PT control in its current version and hence during deployment, the PT can choose to omit certain assessment factors from consideration according to their judgement for the particular patient. However, it is important for the development of the AI to capture all the relevant factors for the entire population.

Please rate the following tasks (given the current setting) according to their clinical relevance for increasing trunk control.

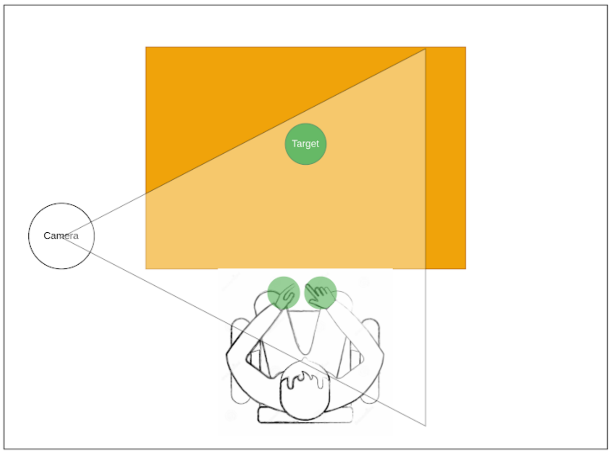
6. Reaching with one hand to target(s) in the anterior region along the sagittal plane (reaching forward)



Mark only one oval.

	1	2	3	4	5	
Not relevant	◯	◯	◯	◯	◯	Highly relevant

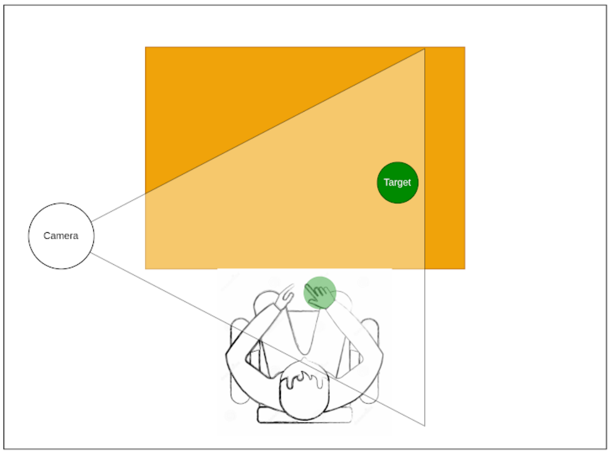
7. Reaching with both hands to target(s) in the anterior region along the sagittal plane (reaching forward)



Mark only one oval.

	1	2	3	4	5	
Not relevant	◯	◯	◯	◯	◯	Highly relevant

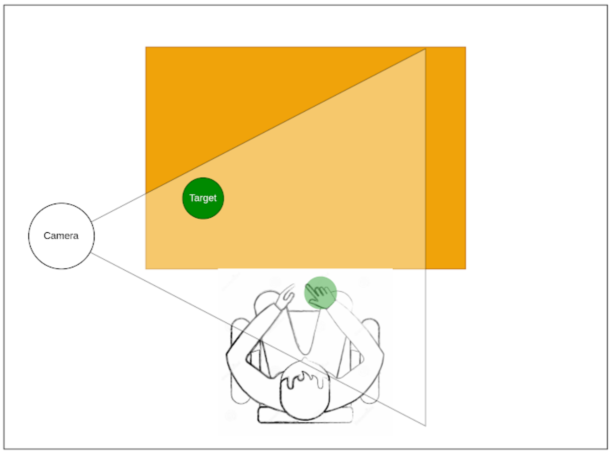
8. Reaching sideways with one hand (for e.g. right) to target(s) in the anterior region on the same side (for e.g. front right region)



Mark only one oval.

	1	2	3	4	5	
Not relevant	◯	◯	◯	◯	◯	Highly relevant

9. Reaching sideways with one hand (for e.g. right) to target(s) in the anterior region on the opposite side (left) (reaching across the sagittal plane)



Mark only one oval.

	1	2	3	4	5	
Not relevant	⬭	⬭	⬭	⬭	⬭	Highly relevant

10. Do you have a suggestion about any other activity in the current setting that might be relevant besides the ones listed above?

____

____

____

____

____

Assessment Factors

Assessment Factors are defined here as those factors that are essential for the physiotherapist to form an opinion on the quality of movements by the child during the training stage in the treatment. Thus when the patient is periodically visiting the clinic to perform exercises with the PT, the PT while constantly monitoring the performance is expected to be forming opinions on these factors.

These factors are expected to influence the modifications done by the PT on the choice of parameters of game during the session. These parameters could be the location of the reaching target, the speed of a moving target, removing ground support and the like.

The AI system is envisioned to assess the performance of the patient on these factors to imitate the PT decision making and find new insights from the data.

Which of these factors do you think are important to assess during the performance of an action (such as the ones listed in the previous section - Reaching Forward, Reaching Sideways, Reaching across midline)?

11. Range of Motion : the amount of movement in joint(s) to achieve a particular motion (for e.g. does the final wrist position and target position coincide/ has the trunk flexed an appropriate amount for the task)

Mark only one oval.

	1	2	3	4	5	
Not important	⬭	⬭	⬭	⬭	⬭	Very important

12. Smoothness : the presence of jerky movement patterns (for e.g. was there a smooth coordination or was the movement influenced by tremors)

Mark only one oval.

	1	2	3	4	5	
Not important	⬭	⬭	⬭	⬭	⬭	Very Important

13. Sequence of Motion : the sequence in which different parts of the body start to move to achieve a particular motion (for e.g. was the motion initiated throughout the body at the same time or was the trunk first engaged and then the shoulder)

Mark only one oval.

	1	2	3	4	5	
Not important	◯	◯	◯	◯	◯	Very Important

14. Compensation : the presence of compensatory movements to achieve the target (for e.g. shoulders elevated, sideward lean, etc.)

Mark only one oval.

	1	2	3	4	5	
Not Important	◯	◯	◯	◯	◯	Very Important

15. Do you have a suggestion about any other factors that might be relevant besides the ones listed above? If yes, please also provide a short description of it.

____
____
____
____
____

Task assessment design

In this section, 3 task scenarios will be described where a participant will be instructed to perform an activity in a certain way. The participant age, as stated before, will be 6-10 years. The parameters that can be changed to influence the rehabilitation of the participant will also be listed. All 3 tasks follow the same structure, where you have to first comment on the Assessment Factors and then define the scores for each factor. The adaptation in the game parameters which will result in increasing the score on those factors, thereby resulting in positive rehabilitation for the patient.

If you had selected some factor(s) as irrelevant to assess trunk rehabilitation performance, skip the questions on those factor(s). If you had mentioned any other factor as important, you can add whatever information could be relevant for this factor at the end of each task.

In this exercise, the child is sitting in the setting described in the previous section.

Task 1:
The child is instructed to reach from the resting position (both arm on the lap) to a target in front using one hand and then return to the resting position.

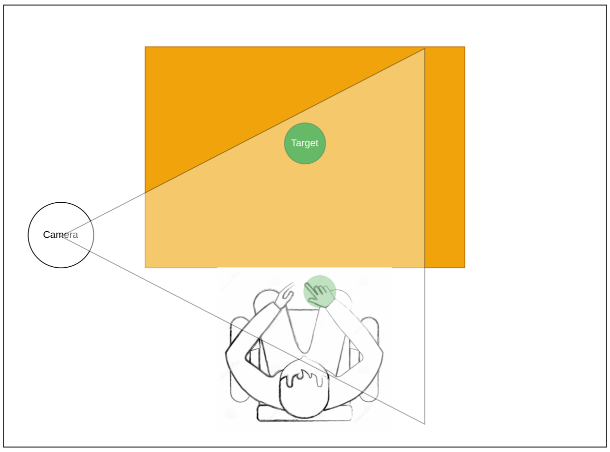
Task 2:
The child is instructed to reach from the resting position (both arm on the lap) to a target on the right using the right hand and then return to the resting position.

Task 3:
The child is instructed to reach from the resting position (both arm on the lap) to a target on the left side using the right hand and then return to the resting position.

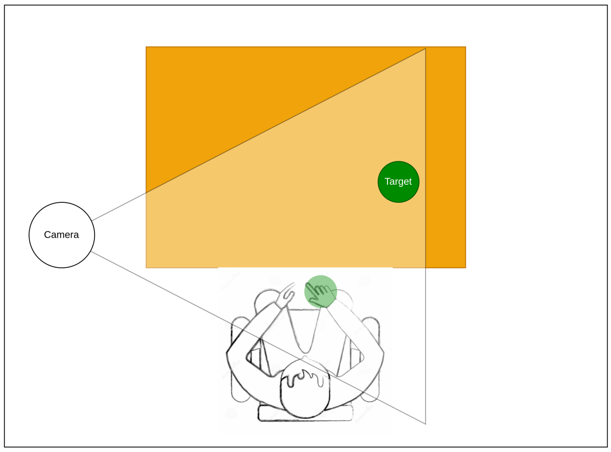
Mutable Game Parameters (others can be added):
- The position of the target in the sagittal plane can be increased or decreased as a factor of the fore-arm length of the child (away or towards in front of the person)
- The position of the target in the frontal axis can be increased or decreased as a factor of the fore-arm length of the child (away or towards the body on the left or right)
- The bench height can be adjusted so as to lift the feet off the ground, removing ground support

You can choose to not answer any question, if you feel it is irrelevant or not applicable. After having spend time answering questions in task 1, it is likely that you would feel many questions in task 2 are similar and can be answered similarly, in which case, you do not need to repeat your answers, but just indicate that it is similar to the answer you gave in the previous tasks.

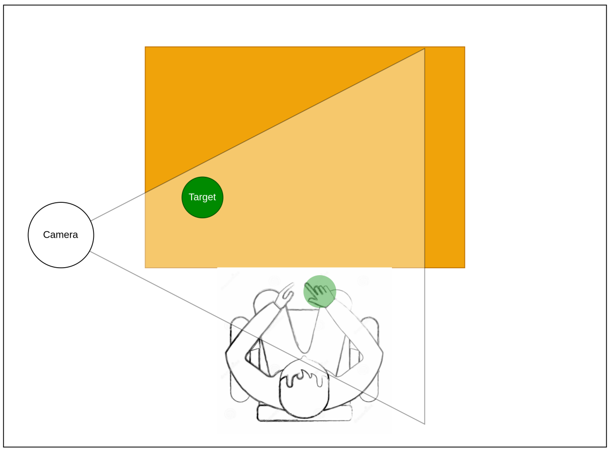
Task 1



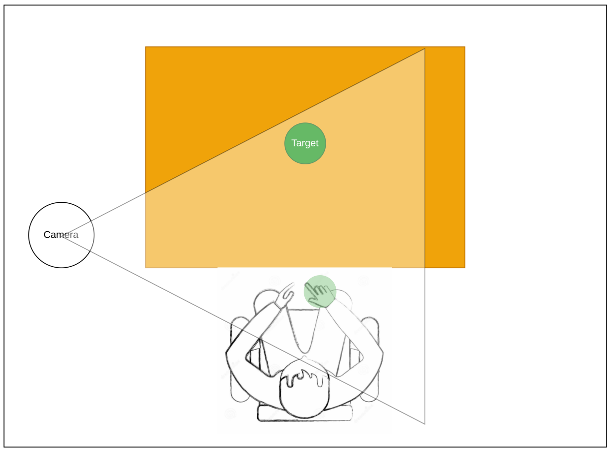
Task 2



Task 3

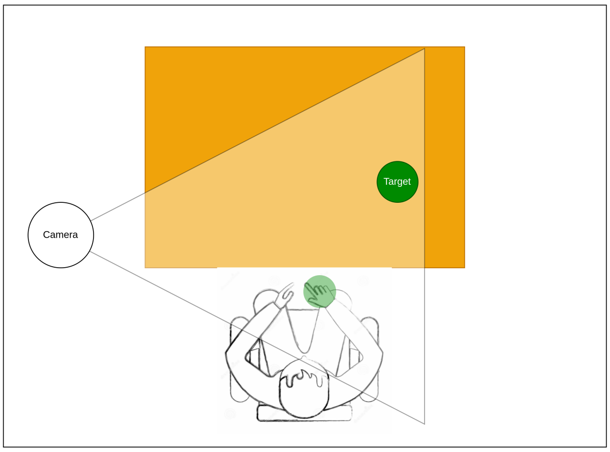


Task 1

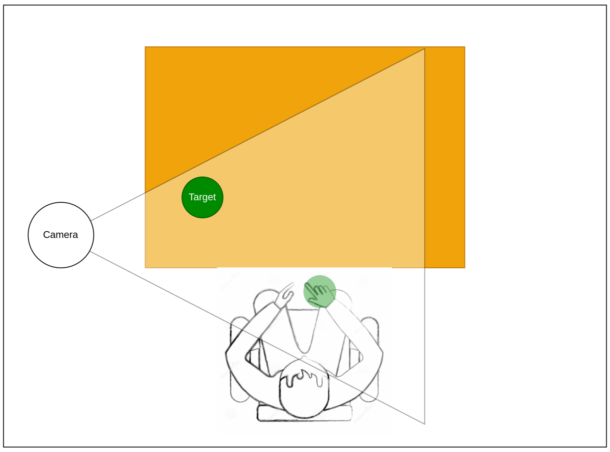


Range of Motion	Range of Motion: For each task, rate the following skeletal joint parameters for their relevance to your judgement on the range of motion exhibited by the child (in the context of trunk stability rehabilitation).

Task 2



Task 3



16. Wrist Position - Task 1 Forward Reaching

Mark only one oval.

	1	2	3	4	5	
Not important	⬭	⬭	⬭	⬭	⬭	Very important

17. Wrist Position - Task 2 same side reaching

Mark only one oval.

	1	2	3	4	5	
Not important	⬭	⬭	⬭	⬭	⬭	Very important

18. Wrist Position - Task 3 opposite side reaching

Mark only one oval.

	1	2	3	4	5	
Not important						Very important

19. Elbow flexion/extension - Task 1 Forward Reaching

Mark only one oval.

	1	2	3	4	5	
Not important						Very important

20. Elbow flexion/extension - Task 2 same side reaching

Mark only one oval.

	1	2	3	4	5	
Not important						Very important

21. Elbow flexion/extension - Task 3 opposite side reaching

Mark only one oval.

	1	2	3	4	5	
Not important						Very important

22. Shoulder flexion/extension - Task 1 Forward Reaching

Mark only one oval.

	1	2	3	4	5	
Not important						Very important

23. Shoulder flexion/extension - Task 2 same side reaching

Mark only one oval.

	1	2	3	4	5	
Not important						Very important

24. Shoulder flexion/extension - Task 3 opposite side reaching

Mark only one oval.

	1	2	3	4	5	
Not important						Very important

25. Trunk Flexion/Extension - Task 1 Forward Reaching

Mark only one oval.

	1	2	3	4	5	
Not important						Very important

26. Trunk Flexion/Extension - Task 2 same side reaching

Mark only one oval.

	1	2	3	4	5	
Not important	◯	◯	◯	◯	◯	Very important

27. Trunk Flexion/Extension - Task 3 opposite side reaching

Mark only one oval.

	1	2	3	4	5	
Not important	◯	◯	◯	◯	◯	Very important

28. For range of motion, if applicable, what other body parts movements are significant for assessment in these tasks?

29. On a scale of 0-2, where 0 is very poor ROM and 2 being a perfect ROM, what characteristics in the above joint parameters would qualify for a score of 0 for Task 1? (for e.g., if the wrist position is still near the lap and/or the elbow is retracted)

30. On a scale of 0-2, what characteristics in the above joint parameters would qualify for a score of 1 for Task 1?

31. On a scale of 0-2, what characteristics in the above joint parameters would qualify for a score of 2 for Task 1?

32. On a scale of 0-2, what characteristics in the above joint parameters would qualify for a score of 0 for Task 2?

33. On a scale of 0-2, what characteristics in the above joint parameters would qualify for a score of 1 for Task 2?

34. On a scale of 0-2, what characteristics in the above joint parameters would qualify for a score of 2 for Task 2?

35. On a scale of 0-2, what characteristics in the above joint parameters would qualify for a score of 0 for Task 3?

36. On a scale of 0-2, what characteristics in the above joint parameters would qualify for a score of 1 for Task 3?

37. On a scale of 0-2, what characteristics in the above joint parameters would qualify for a score of 2 for Task 3?

38. For task 1, If the action doesn't reach its highest ROM rank (2), what would you do so that the child can increase their ROM? If you can increase ROM performance by changing the game parameters, how would you change them (for e.g. bring the goal closer)? If applicable, please distinguish the changes necessary for a score of 0 and 1 respectively.

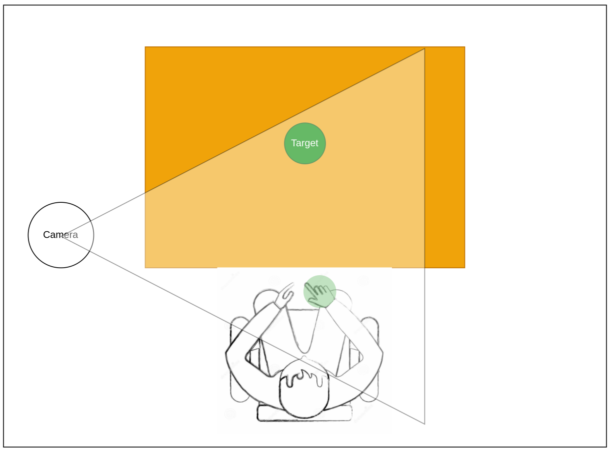
39. For task 2, If the action doesn't reach its highest ROM rank (2), what would you do so that the child can increase their ROM?

40. For task 3, If the action doesn't reach its highest ROM rank (2), what would you do so that the child can increase their ROM?

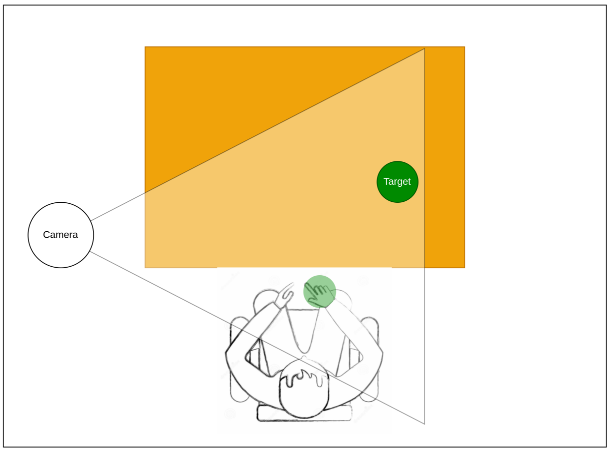
Smoothness Factor

If you had answered smoothness to be a significant factor to assess, please answer the following questions. If not please skip this section.

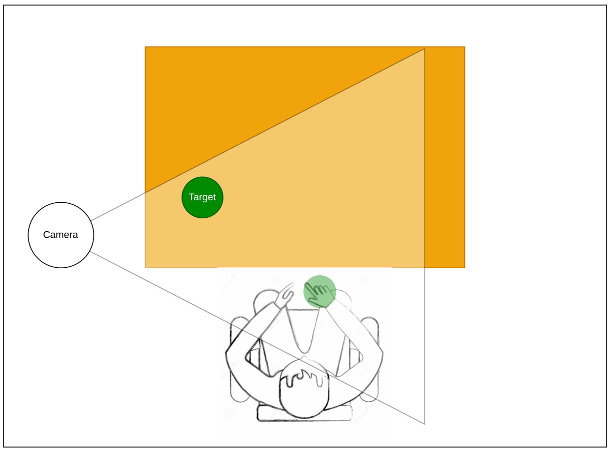
Task 1



Task 2



Task 3



41. For Smoothness, if applicable, which body parts movements are significant for assessment (in the context of trunk stability rehabilitation)?

42. On a scale of 0-2, where 0 is very poor performance and 2 being a perfect performance for Smoothness, what characteristics in the above parameters would qualify for a score of 0 in Task 1 (forward reaching)?

43. On a scale of 0-2, what characteristics in the above parameters would qualify for a score of 1 in Task 1 ?

44. On a scale of 0-2, what characteristics in the above parameters would qualify for a score of 2 in Task 1?

45. For Task 1, if the action doesn't reach its highest Smoothness rank (2), what would you do so that the child can increase their Smoothness? Additionally, If you can increase Smoothness performance by changing the game parameters, how would you change them (for e.g. make the target dynamic and move fast or slow)? If applicable, please distinguish the changes necessary for a score of 0 and 1 respectively.

46. On a scale of 0-2, where 0 is very poor performance and 2 being a perfect performance for Smoothness, what characteristics in the above parameters would qualify for a score of 0 in Task 2 (same side reaching)?

47. On a scale of 0-2, what characteristics in the above parameters would qualify for a score of 1 in Task 2 ?

48. On a scale of 0-2, what characteristics in the above parameters would qualify for a score of 2 in Task 2?

49. For Task 2, if the action doesn't reach its highest Smoothness rank (2), what would you do so that the child can increase their Smoothness?

50. On a scale of 0-2, where 0 is very poor performance and 2 being a perfect performance for Smoothness, what characteristics in the above parameters would qualify for a score of 0 in Task 3 (opposite side reaching)?

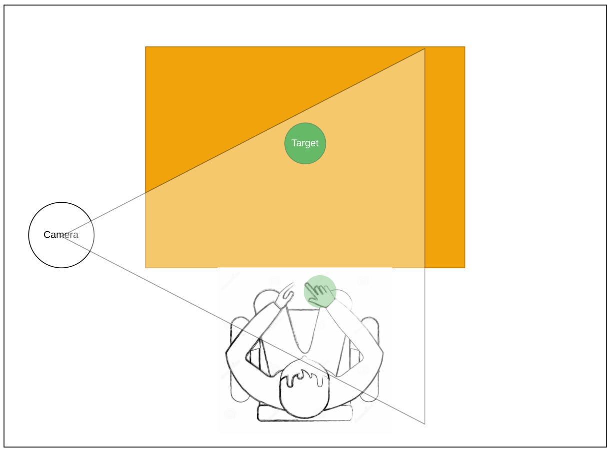
51. On a scale of 0-2, what characteristics in the above parameters would qualify for a score of 1 in Task 3?

52. On a scale of 0-2, what characteristics in the above parameters would qualify for a score of 2 in Task 3?

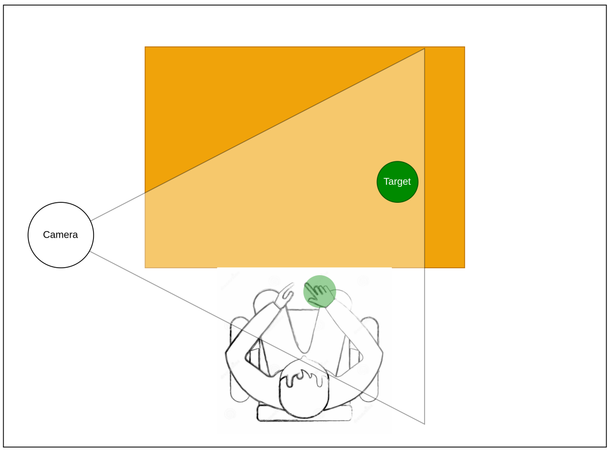
53. For Task 3, if the action doesn't reach its highest Smoothness rank (2), what would you do so that the child can increase their Smoothness?

Sequence of Movements	If you had answered sequence to be a significant factor to assess, please answer the following questions. If not please skip this section.

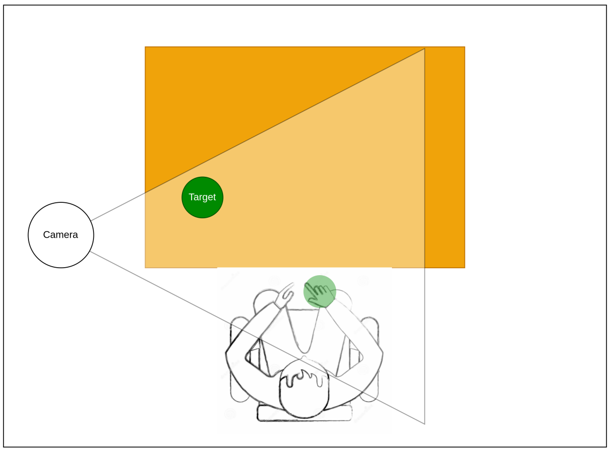
Task 1



Task 2



Task 3



54. For Sequence of movements, if applicable, which body parts movements are significant for assessment (in the context of trunk stability rehabilitation)?

55. On a scale of 0-2, where 0 is very poor performance and 2 being a perfect performance for Sequence, what characteristics in the above parameters would qualify for a score of 0 in Task 1 (forward reaching)?

56. On a scale of 0-2, what characteristics in the above parameters would qualify for a score of 1 in Task 1 ?

57. On a scale of 0-2, what characteristics in the above parameters would qualify for a score of 2 in Task 1?

58. For Task 1, if the action doesn't reach its highest Sequence rank (2), what would you do so that the child can increase their Sequence score? Additionally, If you can increase Sequence performance by changing the game parameters, how would you change them (for e.g. move target closer to engage one joint at a time)? If applicable, please distinguish the changes necessary for a score of 0 and 1 respectively.

59. On a scale of 0-2, what characteristics in the above parameters would qualify for a score of 0 in Task 2 (same side reaching)?

60. On a scale of 0-2, what characteristics in the above parameters would qualify for a score of 1 in Task 2 ?

61. On a scale of 0-2, what characteristics in the above parameters would qualify for a score of 2 in Task 2?

62. For Task 2, if the action doesn't reach its highest Sequence rank (2), what would you do so that the child can increase their Sequence score?

63. On a scale of 0-2, what characteristics in the above parameters would qualify for a score of 0 in Task 3 (opposite side reaching)?

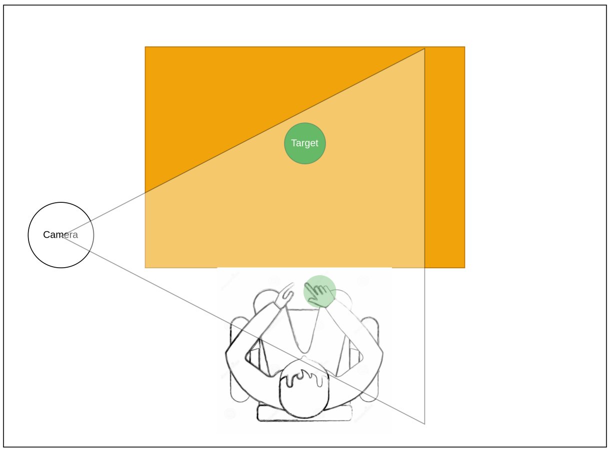
64. On a scale of 0-2, what characteristics in the above parameters would qualify for a score of 1 in Task 3?

65. On a scale of 0-2, what characteristics in the above parameters would qualify for a score of 2 in Task 3?

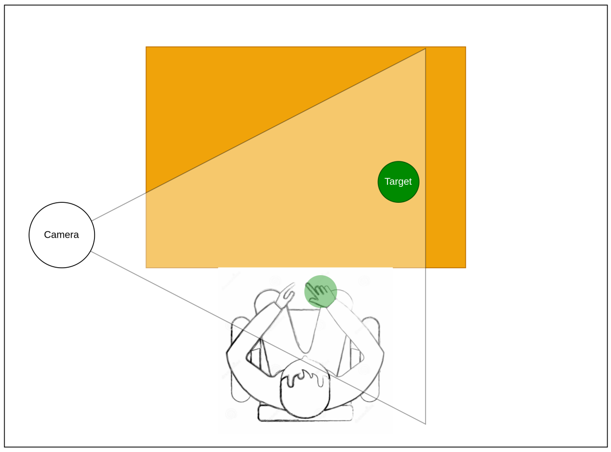
66. For Task 3, if the action doesn't reach its highest Sequence rank (2), what would you do so that the child can increase their Sequence score?

Compensation	For these activities, the child could perform the task with some compensatory movements in the body. In the following set of questions, please rate the relevance of the identified compensatory movements as applicable to the current task. These movements could be because of pathological impediments in which case you could judge them to be acceptable, but there are some compensatory movements which can be corrected by rehabilitation. In the case that the movement you identified as relevant could be treated by rehabilitation, please mention how you would change the game parameters to achieve that objective.

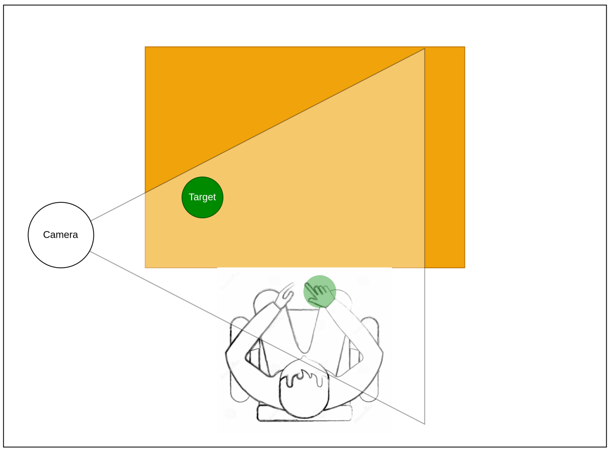
Task 1



Task 2



Task 3



67. Trunk extension (Backward Lean)

Mark only one oval.

	1	2	3	4	5	
Not important	⬭	⬭	⬭	⬭	⬭	Very important

68. How can the game parameters be changed to correct the above?

69. Trunk Lateral Flexion (Sideward Lean)

Mark only one oval.

	1	2	3	4	5	
Not important	⬭	⬭	⬭	⬭	⬭	Very important

70. How can the game parameters be changed to correct the above?

71. Knee Flexion

Mark only one oval.

	1	2	3	4	5	
Not important	⬭	⬭	⬭	⬭	⬭	Very important

72. How can the game parameters be changed to correct the above?

73. Shoulder elevation

Mark only one oval.

	1	2	3	4	5	
Not important	⬭	⬭	⬭	⬭	⬭	Very important

74. How can the game parameters be changed to correct the above?

75. Trunk Rotation

Mark only one oval.

	1	2	3	4	5	
Not important	⬭	⬭	⬭	⬭	⬭	Very important

76. How can the game parameters be changed to correct the above?

77. Other hand position (on lap or bench)

Mark only one oval.

	1	2	3	4	5	
Not important	⬭	⬭	⬭	⬭	⬭	Very important

78. How can the game parameters be changed to correct the above?

79. Movement of legs

Mark only one oval.

	1	2	3	4	5	
Not important	⬭	⬭	⬭	⬭	⬭	Very important

80. How can the game parameters be changed to correct the above?

81. Are there other movements that qualify as Compensatory movements for these activity and if yes, how can the game parameters be changed for correcting it?

Other factors

82. If in the section on identifying relevant factors for assessment you had mentioned any new factors, you can mention some more information on guidelines to assess that factor for this task and how to change the task based on the scores for that factor.

Nature of Participants

83. If the initial trials for the development were to be conducted with healthy participants of similar age (6-10 yrs old), do you have any suggestions to modify the game/body posture/restrict body part movements in the healthy participant to make their movements representative of movements made by a child with trunk control issues?

Any other feedback

84. Do you have any other information you would like to share with us?

Further Contact

85. Would like to be contacted in the future for giving your assessment on the tasks done by a participant in the settings described above? The assessment by the PT will be used to build the AI system.

Mark only one oval.

- Yes
- No

86. If yes, please share your contact details here (the contact details will be stored in our group's data storage and would not be used in any publication or made publicly available in any way)

Thanks for having participated in this survey.

Please click on submit.

References

[1] Alarcos Cieza et al. “Global estimates of the need for rehabilitation based on the Global Burden of Disease study 2019: a systematic analysis for the Global Burden of Disease Study 2019”. In: *The Lancet* 396.10267 (Dec. 2020), pp. 2006–2017. ISSN: 01406736. DOI: 10.1016/S0140-6736(20)32340-0. URL: https://linkinghub.elsevier.com/retrieve/pii/S0140673620323400 (visited on 04/06/2022).