

Wrist movement classification for adaptive mobile phone based rehabilitation of children with motor skill impairments

Kayleigh Schoorl and Tamara Pinos Cisneros and Albert Ali Salah and Ben Schouten

Abstract Rehabilitation exercises performed by children with cerebral palsy are tedious and repetitive. To make them more engaging, we propose to use an exergame approach, where an adaptive application can help the child remain stimulated and interested during exercises. In this paper, we describe how the mobile phone sensors can be used to classify wrist movements of the user during the rehabilitation exercises to detect if the user is performing the correct exercise and illustrate the use of our approach in an actual mobile phone application. We also show how an adaptive difficulty system was added to the application to allow the system to adjust to the user. We present experimental results from a pilot with healthy subjects that were constrained to simulate restricted wrist movements, as well as from tests with a target group of children with cerebral palsy. Our results show that wrist movement classification is successfully achieved and results in improved interactions.

Key words: cerebral palsy, rehabilitation, hand therapy, activity recognition, machine learning, applied games

Kayleigh Schoorl
Utrecht University, Utrecht, the Netherlands, e-mail: kayleighschoorl@gmail.com

Albert Ali Salah
Utrecht University, Utrecht, the Netherlands e-mail: a.a.salah@uu.nl

Tamara Pinos Cisneros
University of Twente and Amsterdam University of Applied Sciences, the Netherlands e-mail: t.v.pinos.cisneros@hva.nl

Ben Schouten
Eindhoven University of Technology, the Netherlands e-mail: bschouten@tue.nl

1 Introduction

Children with cerebral palsy often need to perform repetitive movements during these therapy sessions. It can be a difficult task to keep them motivated to continue doing these exercises regularly, especially when these need to be performed at home in between sessions with their therapists. When they are intrinsically motivated, they participate in therapy for their own satisfaction and are more likely to improve and keep up with their therapy sessions and exercises [26]. Moreover, children with cerebral palsy might often encounter frustrating moments and failure during their therapy, which underlines the need for keeping therapy as motivational and engaging as possible to minimize frustrations and uphold motivation [10]. It is therefore important for a therapist to accommodate an environment that both enhances self-directed engagement in therapeutic exercises and also provides intrinsic motivation to perform the therapy exercises at home.

Since the exercises recommended for cerebral palsy are often not inherently experienced as being enjoyable, applied games can be a powerful tool to motivate children to gain more intrinsic motivation, and can contribute to a higher degree of personalization and adaptivity to accommodate their individual needs [27, 7]. Digital developments make different types of learning incorporated into video games possible, allowing for games that are both engaging to play and can help children to learn [31]. They can help decrease the negative aspects of therapy, such as frustration and boredom, and increase the positive aspects, such as happiness and enjoyment, and can therefore be a suitable tool for providing intrinsic motivation for children to perform their therapeutic exercises.

The hand and wrist rehabilitation of cerebral palsy patients focuses on a type of movement called “dorsiflexion”. The dorsiflexion of the wrist and hand involves bending the hand upward towards the top of the forearm; the downward bending movement towards the bottom of the forearm is called “palmar flexion”. The main exercise for the patients is shaking the hand up and down, moving between dorsiflexion and palmar flexion repeatedly. The mobile rehabilitation application requires such movements for successful completion of its tasks.

In this paper, we describe how a mobile application that has been developed for facilitating hand therapy [1] can be extended to become adaptive by classifying hand and wrist dorsiflexion movements of its users on the fly. For this purpose, we test the suitability of a small number of classifiers, based on the sensors on the phone. We test and refine our approach first on a pilot study with healthy subjects, whose wrists we constrain in several ways to simulate patient conditions. Our original target group for this application is children from ages 7 to 12, who have hemiparesis as a result of cerebral palsy and are within level I or II on the Manual Ability Classification Scale or MACS [9]. After the pilot, we also conduct an experiment with eight children with cerebral palsy, and present our results.

We briefly introduce the mobile phone based exergame application here, which serves as our experimental platform. The phone is inserted in a soft, animal-shaped sheath, and the phone display becomes the face of the animal, which is capable of expressing emotions, as can be seen in Figure 1. In the game, the animal is an

assistant for performing magic tricks and comes with a separate booklet in which the magic tricks are explained. The performing of the trick requires some dexterity, and wrist movements are performed during the procedure. The mobile application core used for this research has been previously developed by a gaming company using the Unity3D game engine, and runs on Android phones [1].



Fig. 1 The monster case used for the Magic Monster application and a screenshot from the mobile app.

In this study, we implement a movement recognition system to accurately recognize dorsiflexion movements during the use of this mobile application, as well as an intuitive adaptive difficulty system. Since the patients need to perform specific movements as part of their hand therapy, the accuracy of these movements needs to be assessed to evaluate therapy progression. Also, different patients have different needs and might not have the same range of motion in their wrists as other patients. Consequently, it is necessary that the application should adjust to the user, and keep adjusting if the capabilities of the user change over time.

The remainder of the paper is structured as follows. First, we discuss related research in Section 2. Next, we present our data collection approach for the movement classification system in Section 3. In Section 4 we describe the features we use, and the approaches tested for classification of dorsiflexion. The experimental results are summarized in Section 5. In Section 6, we present the adaptive difficulty system introduced to the application. Section 7 reports results obtained with the target patient group. Finally, we provide a discussion and conclusions in Section 8.

2 Related work

The potential of personal informatics (PI) systems for people with motor disabilities has been discussed in [18]. Here we focus on games that target children with motor skill impairments specifically, as well as provide some references for mobile phone based activity recognition, which is a very broad area due to the potential richness of the activities that can be detected. This paper, however, is the first for detecting dorsiflexion movements in the literature.

2.1 Games targeted at children with motor skill impairments

While many applied games in the field of health have been developed in the past, the selection of applied games that specifically target children with motor skill impairments is not as substantial [2, 22]. In [11], a number of smart toys and game applications are listed to monitor and enhance fine motor skills of children, but these are not geared towards rehabilitation. Researchers from the University of Amsterdam and the Amsterdam University of Applied Sciences developed a game for detecting delays in motor skill development on a commercially available toy called the *Futurocube* [28]. Using this game, they were able to predict the degree of motor skill impairment in the participants with increasing accuracy as the difficulty of the game rose. More recent work with the *Futurocube* has focused on using different machine learning techniques for classifying children with motor skill impairments [4].

[16] developed a smart toy for motor skill assessment in form of a board game. The goal of this game is to move tokens (depicting mice) across the board and turn them around as carefully as possible so as to not wake up ‘the cat’. The tokens contain accelerometers, which are used to assess the smoothness of movements by computing the mean squared jerk feature. The results show that children with better fine motor skills can move the tokens more smoothly using one hand, making the game suitable for detecting motor skill impairments. Moreover, the game was perceived as fun and exciting to play by the children participating in the experiment. More recently, [17] presented a mobile application for analyzing the fine motor performance in children based on a similar idea, by making them move an object along several paths, varying in difficulty from straight to zigzag-shaped.

A smart toy that was developed for practicing fine motor skills in [30] encompassed several different exercises to practice different types of movements: tracing a path while holding a pen, placing clothespins on colored fabric, and carrying a metal ring from left to right while avoiding touching the filament, respectively. In most of these applications, the difficulty of the game is adjusted by having multiple challenges with different difficulty levels. In this paper, however, we measure and adapt the difficulty to the child. Measuring performance has the double purpose of observing rehabilitation progress longitudinally.

2.2 Human activity recognition using smartphone sensors

A modern smartphone contains a variety of sensors, including an accelerometer, a magnetometer, and a gyroscope, which can be highly useful for recognizing the activities of a user [23, 25, 32]. Previous research has largely focused on human movement recognition using the sensors available in smartphones [24, 15, 20, 21]. The most common application is the detection of everyday activities including sitting, walking, and walking up the stairs. For these types of applications, the phone is usually passively carried around. Some of these applications addressed disabilities or pathologies impairing movements of the limbs, such as gait [35].

Deep learning methods have become widely used for activity recognition in recent years [3]. Some examples of different types of networks previously used for activity recognition include basic multilayer perceptrons (MLPs) [33], convolutional neural networks (CNNs) [13], and recurrent neural networks (RNNs) [19].

CNNs are mostly used for visual data, such as images and video, but have also been used on time series data such as sensor data collected using smartphone sensors [13, 33]. One study proposed the use of CNNs for activity recognition using the accelerometer data in a mobile phone [13]. As input for the network, they used both the raw accelerometer data and a selection of extracted basic statistical features that describe the global properties of the time-series data. The combination of both seemed to achieve the best performance. For mobile phone applications, shallow architectures and lower computational costs are desirable, as battery life is an important consideration.

For supervised activity recognition tasks, the temporal dimension is important, as a large part of the challenge is to properly segment the activity boundaries and to provide continuous class labels. For this purpose, RNNs [19] and long short-term memory (LSTM) networks were used [6]. LSTM layers can also be combined with convolutional layers to create more complex models for human activity recognition using raw mobile phone sensor data [34].

In our application, we examine a range of simple to complex models to classify dorsiflexion movements. What is more important than the selection of the machine learning approach is the insights into how a cerebral palsy patient would handle a mobile phone during the exergame, and how that will affect the performance of the classifiers. We first deal with the problem of data collection, as most ML approaches need as much data as possible to provide good classification results.

3 Data collection

To train a model for distinguishing dorsiflexion from other types of hand movements, we collected data for a variety of movements, as well as dorsiflexion movements from different starting positions. We have defined 28 distinct movement classes: the first 10 classes describe dorsiflexion movements from different starting positions, and the other 18 classes describe other kinds of movements and non-movements which

function as the true negatives in the dataset, including rotations, shaking, and the phone sitting still. We use the accelerometer and the gyroscope to record the data along three perpendicular axes (x, y, z). The phone that was used for data collection, and all subsequent training and evaluation, was a Samsung Galaxy A52.

We have created a custom mobile application for data collection, where the user is shown an animation showing how to perform a movement, along with auditory instructions and a text description of the movement. This interface selects ten movement classes per session randomly for each user. Additionally, we recorded the hand movements of the user using an external video camera, for improving annotation quality. A total of 37 data collection sessions with 20 unique subjects were completed. Before participating in the experiment, the subjects were informed of the purpose and methods of the study and signed a written consent form. The data collection and storage procedures were reviewed and approved by a medical ethics review committee¹.

Two annotators manually annotated the beginning and end times of the movements using the annotation tool ELAN [5]. Only a binary label was introduced (i.e. dorsiflexion vs. no dorsiflexion) during the annotations.

The resulting dataset consists of a total number of 337 segments, out of which 113 are classified as dorsiflexion movements. The dataset is divided into a training (15 subjects) and test (5 subjects) set for model training and evaluation purposes. The test set contains 53 segments, representing 15.7% of the total dataset.

Figures 2 and 3 show two examples of what the data looks like, for one dorsiflexion movement and one non-dorsiflexion movement (i.e. rotating the phone), respectively. The collected accelerometer and gyroscope data are plotted for each of the three axes. From these plots, it can be seen that the recorded sensor data for these movements look similar at first sight, especially when looking at the gyroscope data.

4 Classification of Dorsiflexion

4.1 Feature extraction and selection

We investigated both feature extraction and feature selection to train the machine learning models. For the former, we extracted seven low level descriptors (i.e. mean, min, max, standard deviation, variance, skew, and kurtosis, respectively) from each of the axes of the sensors, resulting in $6 \times 7 = 42$ features. Each feature was normalized using min-max normalization.

For feature selection, we used the Minimum Redundancy Maximum Relevance (mRMR) algorithm [8]. This method, attempts to find an optimal number of features by eliminating redundant features. For each feature, a score is computed based on

¹ Details left out for blind review. The authors declare that there was no influence or involvement from the funding organization in the design, data collection, analysis, interpretation, writing, or submission of this study.

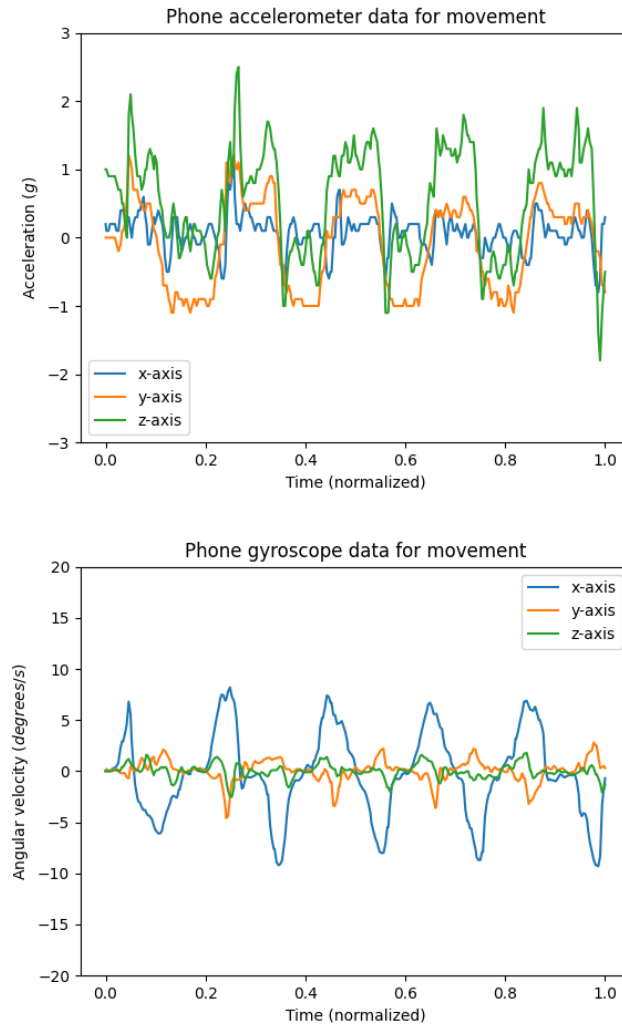


Fig. 2 Recorded sensor data for movement class 1: dorsiflexion with the phone upright and the screen turned away from the hand palm.

the maximum relevance with regard to the output variable, and minimal redundancy with regard to the already selected features. The algorithm iterates until k features have been selected. One advantage of mRMR is the explainability of the selected features, which is important because we want to be able to explain the logic behind the selected thresholds for the dorsiflexion movements.

There are various options for score functions to use with mRMR, and we used a function that combines the F-test statistic and the Pearson correlation coefficient to

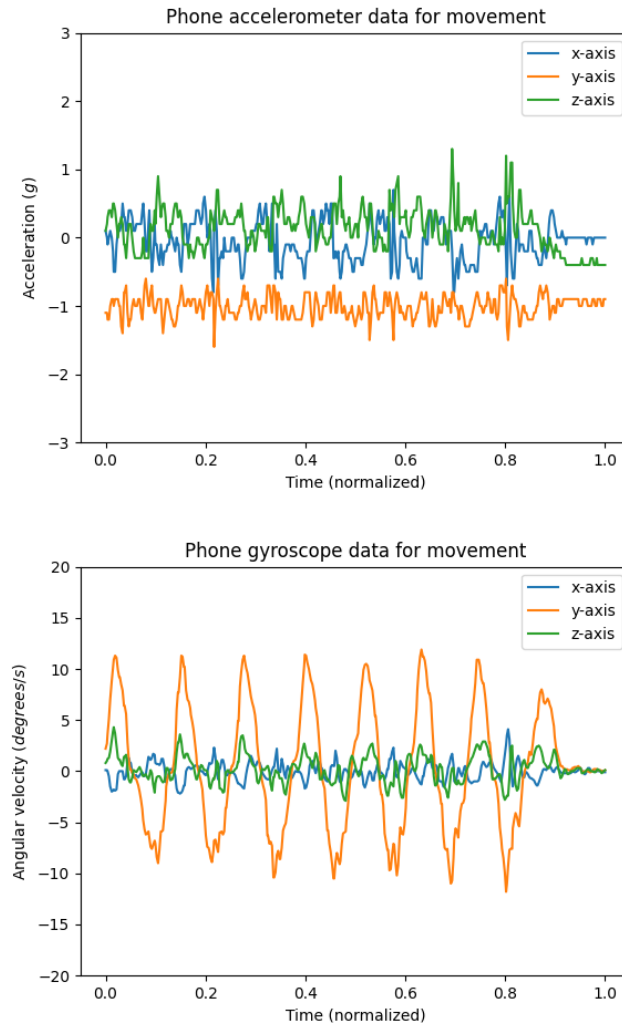


Fig. 3 Recorded sensor data for movement class 21: rotating the phone to the left and right side alternately, with the phone held horizontally and the screen turned upward.

compute a score for each feature [36]. This method is simple and fast, yet provides comparable results to other implementations of mRMR. The relevance of a feature is computed as the F-statistic between the feature and the target output. The redundancy of the feature is computed as the average Pearson correlation between the feature and the features selected during previous iterations.

4.2 Classification

We tested and compared several machine learning approaches of different complexity for solving the classification problem. Since the mobile phone presents a limited energy budget, simple approaches are preferred to computation-heavy approaches. We tested k-nearest neighbors (k-NN), support vector machines (SVM), multi-layer perceptrons (MLP), convolutional neural networks (CNN), long-short-term-memory (LSTM), and Bidirectional LSTM, respectively. To evaluate the results, we used classification accuracy, precision, recall, and the F-score.

A basic k-NN classifier is used as a baseline algorithm with the 42 features obtained via feature extraction, as described earlier. We used the Euclidean distance as the distance function, and leave-one-out cross-validation on the training partition was used to select the value for k that provided the highest accuracy, which was set as 1.

Next, we performed mRMR for determining the features to be used for the final model. The highest training set accuracy was observed when using 21 features out of 42, ultimately resulting in an accuracy of 0.997 on the training set for $k = 1$ and 21 features. Therefore, these are the values that are used for evaluation on the test set. We have tested dynamic time warping (DTW) in conjunction with kNN, but DTW is too time-intensive to use in a real-time application, and was not taken into further consideration.

The second model we trained is a support vector machine (SVM) with a linear kernel, using the same features that were used with 1-NN. Using mRMR on the training set resulted in the selection of only 2 features out of the 42, with a cross-validation accuracy of 0.993. We also trained four neural network (NN) models.

4.2.1 Multilayer perceptron (MLP)

The first NN we have trained is a multilayer perceptron, with two hidden layers (with 128 and 256 neurons, respectively), a dropout layer, and a classification layer. This network takes as input the same features that are also used for the k-NN and the SVM. We first used mRMR to select the appropriate number of features, computing the accuracy using leave-one-out cross-validation on the training set for each possible number of features. The network was trained with the ADAM optimizer for 100 epochs max [14]. The highest scoring number of features was 37, with an accuracy of 0.958.

4.2.2 Convolutional neural network (CNN)

While CNNs are most commonly used for image recognition and classification tasks, they can also be applied to activity recognition tasks using sensor data. The network structure we used was based on a CNN used previously for activity recognition [13]. The first layer is a 1-dimensional convolutional layer with 196 filters. The kernel

size used is 1×16 . This convolutional layer is followed by a max pooling layer with pool size 4 and a flattening layer. Next is a dense layer with 1024 neurons, followed by the classification layer.

4.2.3 Long Short-Term Memory

We tested an RNN with LSTM layers [12]. LSTMs have been shown to benefit from at least two or more LSTM layers, which is why our model had two LSTM layers, with 128 and 256 nodes respectively. They are followed by a dropout layer and the classification layer.

4.2.4 Bidirectional Long Short-Term Memory

The BLSTM was constructed similarly to the LSTM, but instead of two LSTM layers, it contains two BLSTM layers, also with 128 and 256 nodes respectively [29]. Similar to the LSTM, the BLSTM layers are followed by a dropout layer and the classification layer.

5 Experimental Results

The experimental results of each trained model can be seen in Table 1. The accuracy, precision, recall, and F-score are shown for each of the models, reported on the independent test set. The precision, recall, and F-score are also shown for each class separately. The highest scores are shown in bold text for clarity.

The CNN model performed well overall, but the precision and recall results indicate that most of the time dorsiflexion movements are indeed classified as such, but there are also false positives. This is made evident by the fact that the F-score for these models is lower. The more complex NN models show more balanced scores and higher F-scores, with CNN scoring the highest overall, suggesting that it is a suitable machine learning model for classifying dorsiflexion in a binary classification task.

6 Adapting the Game to the User

When a movement has been positively identified as a dorsiflexion movement, we want to check at what level the user is shaking the phone, using an adaptive difficulty system. The mobile application is supplied with two indicators for this purpose, namely, the range of motion and speed. To achieve adaptive difficulty, the application needs to be able to adjust the sensitivity to the user with regards to these two measures,

		KNN*	SVM*	MLP*	CNN	LSTM	BLSTM
Overall	Accuracy	0.943	0.925	0.962	0.970	0.961	0.964
	Precision	0.932	0.913	0.952	0.963	0.949	0.953
	Recall	0.956	0.941	0.971	0.971	0.968	0.971
	F-score	0.940	0.921	0.960	0.967	0.957	0.961
Dorsiflexion	Precision	0.864	0.826	0.905	0.938	0.903	0.911
	Recall	1.000	1.000	1.000	0.974	0.991	0.991
	F-score	0.927	0.905	0.950	0.956	0.945	0.949
Non-dorsiflexion	Precision	1.000	1.000	1.000	0.987	0.995	0.995
	Recall	0.912	0.882	0.941	0.968	0.945	0.950
	F-score	0.954	0.938	0.970	0.977	0.970	0.972

* uses extracted features as input, as opposed to raw sensor data

Table 1 The accuracy, precision, recall, and F-score for dorsiflexion recognition per model on the test set.

and slowly decrease the sensitivity to make it harder to get the correct movement. This feature helps to get the user to progress in their therapy and improve their motor skills.

6.1 Rule-based adaptive difficulty

We implemented two decision rules to determine at which points the thresholds for both the range of motion and the speed should change. To determine when to change the difficulty, the past performance of the player and the percentage of correct movements over the last 10 registered shakes is analysed. If at least 90% of the movements reached the threshold, the threshold increases; if no more than 60% of the movements remained under the threshold, the threshold decreases.

6.2 Calibration

Since we wanted to adjust the dorsiflexion recognition sensitivity to the user, the application first goes through a calibration phase when it is used for the first time. During this step, the player performs dorsiflexion to activate the different stages of the ‘magic trick’, and the adaptive system uses the first five movements to determine the thresholds for both the range of motion and the speed. Besides the automatic calibration, the user can also reset the thresholds and restart the calibration from the settings screen, or adjust the thresholds manually if they wish. This is a precaution against the application not functioning according to the preferences of the user, who may prefer easier or more challenging settings.

6.3 Range of motion

In order to train the system, a pilot was conducted without accessing the target patient group.

Children with hemiparesis as a result of cerebral palsy often wear an arm brace to support their wrist, such as the one shown in Figure 4. However, besides supporting the wrist, this arm brace also restricts the movement of the wrist. Subsequently, we used arm braces to simulate restricted wrist movements for different ranges of motion.

Data were collected for the following conditions of dorsiflexion, all from the same starting position to make comparing the movements easier:

- Free movement without the arm brace
- Half-restricted movement with a plastic ruler inserted in the arm brace, resulting in the range of motion seen in Figure 4
- Fully-restricted movement with a metal ruler inserted in the arm brace, resulting in the range of motion seen in Figure 5

These three degrees of movement are considered the three different movement classes. For each of these movements, we recorded 20 samples. We performed feature extraction and selection as with the previously described experiments. The idea is to obtain features that are suitable to recognize varying strengths of dorsiflexion, which can be used to set a threshold to determine the range of motion of the dorsiflexion movement.

A visualization of some of the collected sensor data, for each level of restriction, can be seen in Figures 6, 7, and 8.

We have considered the mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE) to measure the performance of regression models. RMSE is expressed in the same units as the original y values, and penalizes higher error values more compared to MAE. Therefore, we have used RMSE for evaluating the models.

For each set of features, leave-one-out cross-validation was used to compute the RMSE for that number of features. Using a single feature (namely, the standard deviation of the x -axis of the gyroscope) ultimately results in the lowest RMSE. On the test set, the model results in an RMSE of 1.419. Since a single feature is powerful enough to determine the range of motion for the most part, the adaptive difficulty system was linked to this indicator, together with the simple threshold adjustment rules described earlier. This provided a simple and reliable method for evaluating the range of motion of the user.

6.4 Speed

We collected data for dorsiflexion movements with the phone upright for three degrees of speed, classified as slow, medium, and fast, respectively. These classes



Fig. 4 Dorsiflexion while wearing an arm brace with a plastic ruler inserted.

were used to provide some qualitative insights, and examples of the collected sensor data can be seen in Figures 9, 10, and 11. A clear difference can be observed between different movement speeds. Most notably, the maximum amplitude of both the accelerometer and gyroscope signals goes up as speed increases along the axis on which the movement occurred. As can be observed, the signals are similar to the different ranges of motion as seen in Figures 6, 7, and 8. However, what sets the speed apart from the range of motion, is the frequency at which the minimum and maximum amplitudes alternate.

To compute the frequency, there are some different options, including:

- Computing a fast Fourier transform (FFT) and taking the lowest frequency.
- Calculating the number of zero crossings or the number of times the sensor value oscillates about 0. The faster the sensor value oscillates around 0, the higher the frequency.



Fig. 5 Dorsiflexion while wearing an arm brace with a metal bar ruler inserted.

In our implementation, we used the number of zero crossings for the speed. While this is a rigid scale and does not allow for a gradual increase in difficulty, it does provide clear levels for the user to achieve, and it is intuitively interpreted.

7 Tests with the Target Group

Using the implemented dorsiflexion recognition and adaptive difficulty systems as described above, we performed an evaluation test with the target group to check the accuracy of the dorsiflexion recognition system. We collected data from eight children with cerebral palsy during therapy sessions at two clinics that specialize in treating posture and movement-related disorders. We traveled to the clinics and

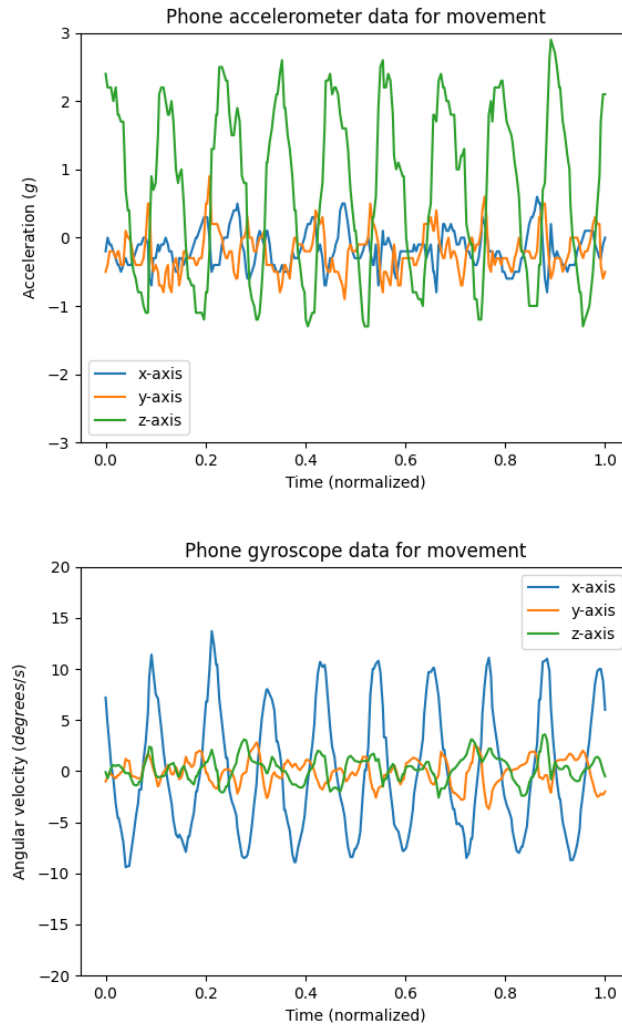


Fig. 6 Sensor data for dorsiflexion, with the wrist unrestricted.

performed one-on-one sessions with each participant. Each participant was given a demonstration of how the magic monster is supposed to be used and was then asked to play with the magic monster application for three minutes. We obtained ethical approval from XXX prior to the study. The sessions were recorded using a video camera. The parents of the children gave written consent before the session with the magic monster, after informing them of the purpose and data collection procedure of the experiment.

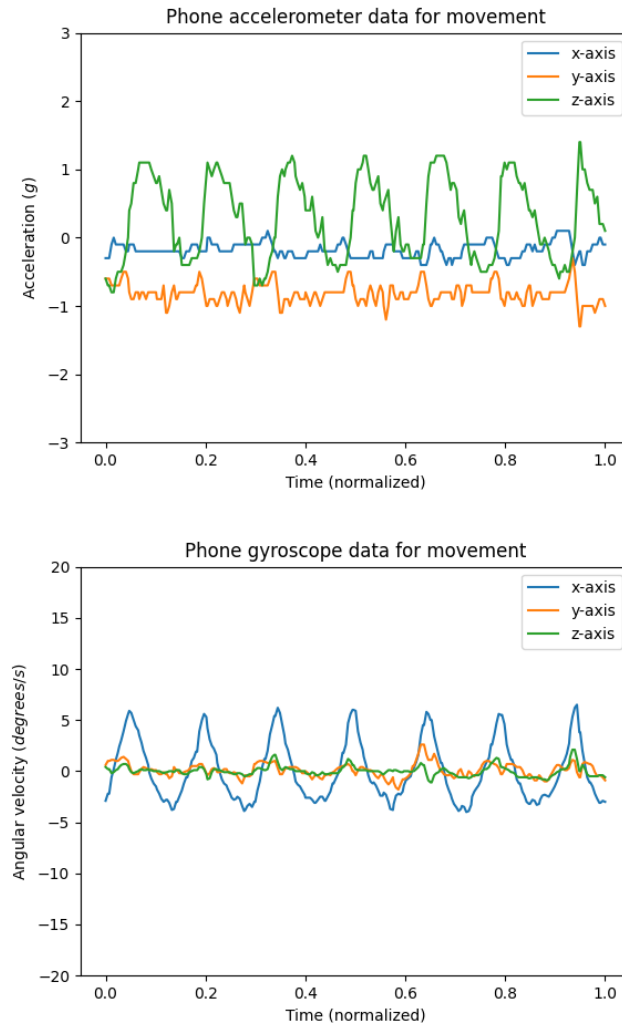


Fig. 7 Sensor data for dorsiflexion, while wearing the arm brace with a ruler inserted.

One of the findings of this study, besides the data collection, was that the magic monster can sometimes be difficult and too big to hold for children with cerebral palsy. This is why a strap was added to the back of the monster, which wraps around the hand to hold the monster in place. This strap was added to the monster halfway through the testing phase, and could potentially have influenced the results.

Similarly to the previous data annotation that was done for training the machine learning algorithm, two people manually annotated the beginning and end times of the recorded movements in all of the recorded videos using ELAN [5]. The annotated

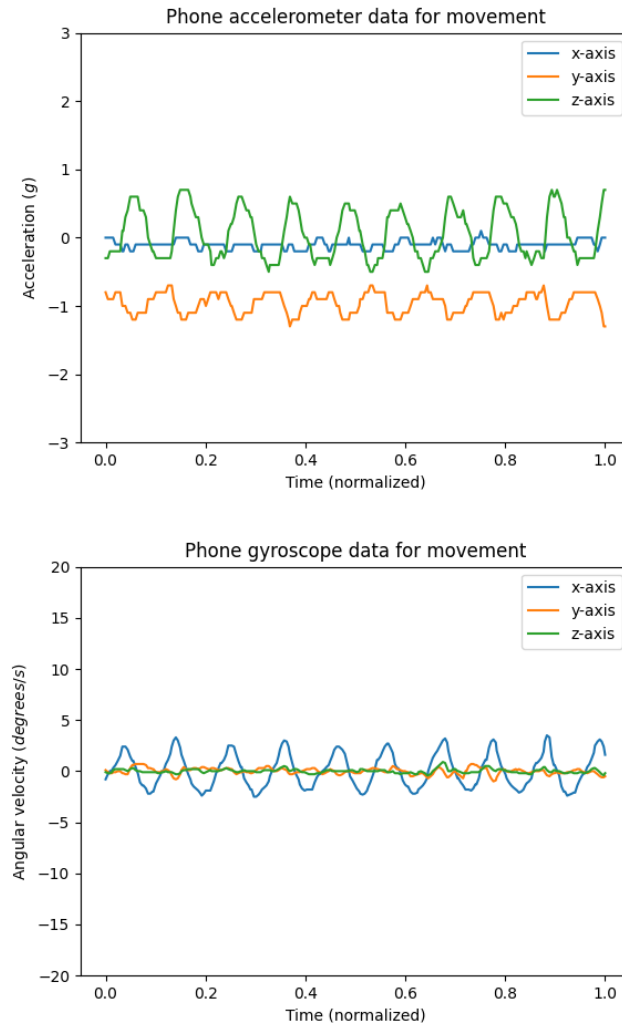


Fig. 8 Sensor data for dorsiflexion, while wearing the arm brace with a metal bar inserted.

dorsiflexion movements were then manually compared to the sensor data recorded using the phone. The mobile application checks if it detects dorsiflexion once per second to prevent performance issues by checking more often. The ground truth is obtained by inspection of the videos. The confusion matrix can be seen in Table 2.

To further evaluate the results, we computed the accuracy (0.948), precision (0.929), recall (0.846), and F-scores (0.885). Overall, the scores are high, with recall being lower than precision overall, meaning that not all dorsiflexion movements annotated as such are indeed being recognized as dorsiflexion. One limitation is that

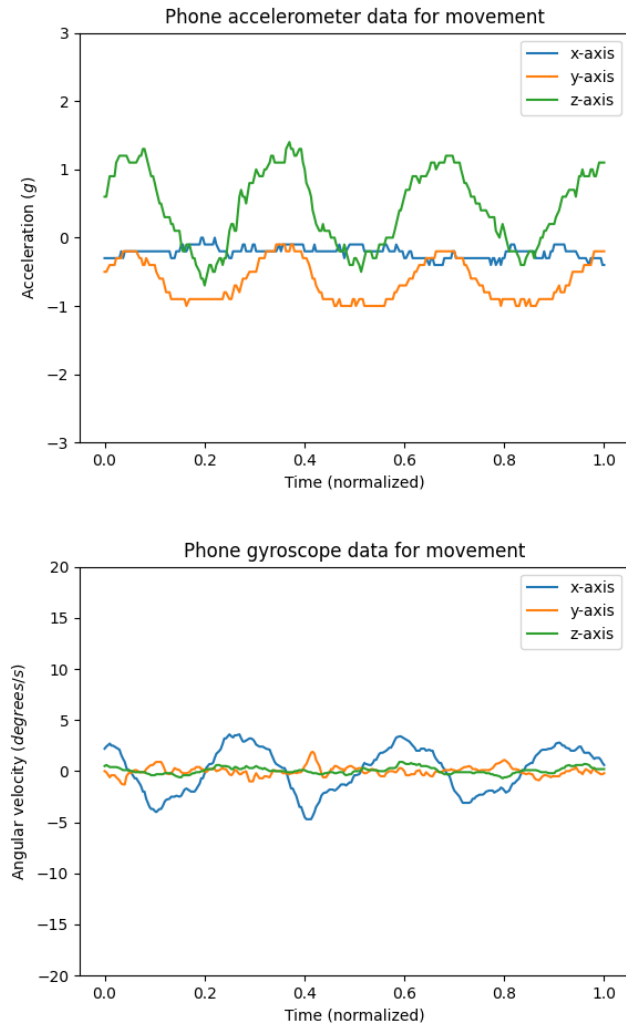


Fig. 9 Sensor data for dorsiflexion, performed with low speed.

the target group is very small and it is difficult to obtain data from children with cerebral palsy. Our training set did not completely reflect the variation in the target group, and was collected from adult subjects.

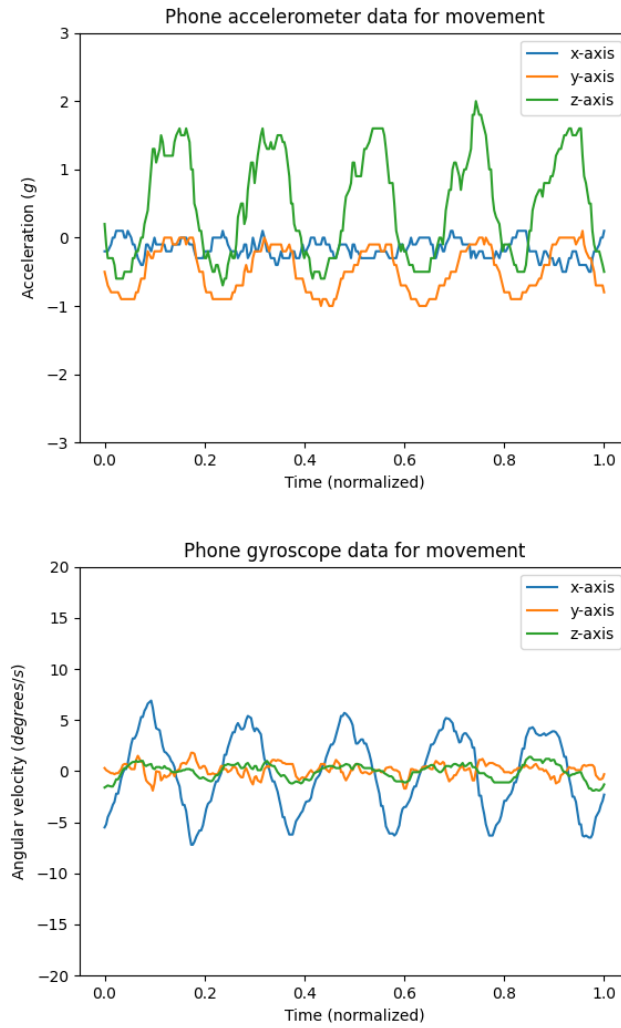


Fig. 10 Sensor data for dorsiflexion, performed at medium speed.

8 Conclusions

In this study, we proposed a simple approach to detect dorsiflexion movements, and incorporated it into an application for hand and wrist rehabilitation of children with cerebral palsy. Our approach uses accelerometer and gyroscope data and uses a convolutional deep neural network on simple features extracted from these sensors to classify dorsiflexion movements.

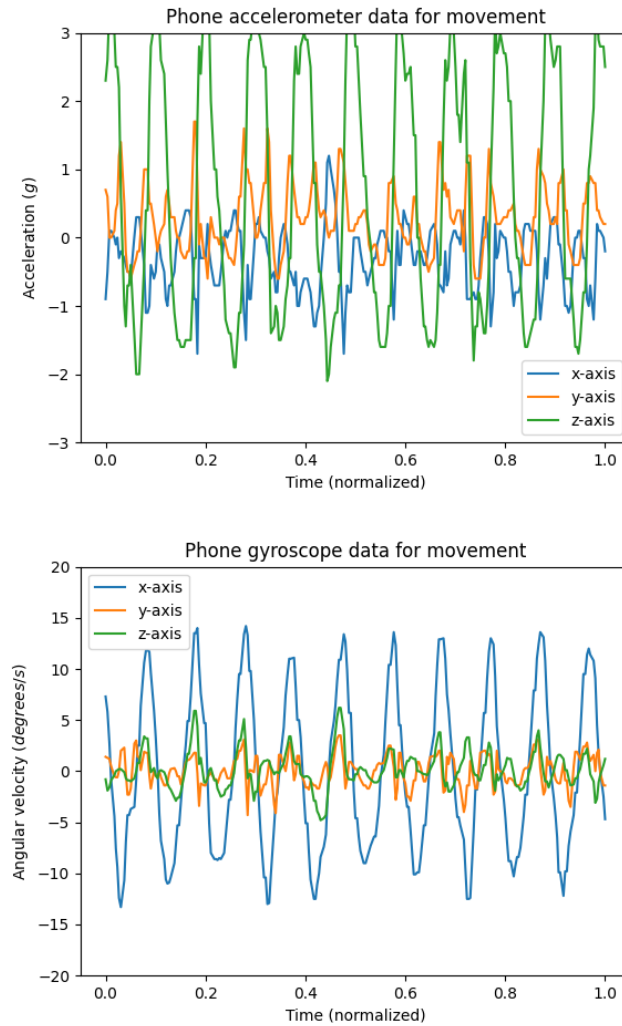


Fig. 11 Sensor data for dorsiflexion, performed at fast speed.

Based on the classifier's output, the speed and range of motion were computed into simple indicators, which were added to the exergame to let it adapt to the user. We performed a series of experiments to measure the efficacy of the approach, and also tested the adaptive exergame on children with cerebral palsy.

The study has several shortcomings. More extensive data collection is needed for understanding the user experience aspects with the target group. This could be done with either the data collection application developed for this study, or with the magic monster application itself. While application data can be collected relatively

		Annotated (real) values	
		Positive	Negative
Predicted by AI	Positive	252	10
	Negative	35	931

Table 2 Confusion matrix for the binary dorsiflexion recognition.

easily, the video data that need to be collected for annotation and quality assessment purposes is much more difficult to obtain, and more sensitive to handle. However, once the annotations are completed, the video data can be destroyed, and only the anonymous gyroscope and accelerometer data can be shared with the annotations.

The data collection was mostly focused on training an activity recognition model, and the adaptive difficulty adjustment of the system was developed with much less data. However, the logic of adaptation is fairly simple, and in general, less data will be sufficient for its development. Nonetheless, there is a need for a longitudinal user study, since the effects of having an adaptive difficulty can only be measured if the application is used for a longer period of time, and against a control group that uses a non-adaptive version of the application. Longitudinal studies can also assess the appeal of the application, and find out whether the novelty wears off, or whether the application loses its likeability after the magic trick is fully mastered.

The current data collection and evaluation were done using only one phone model. For future research, it would be fruitful to collect data using different models and compare the results, since the gyroscopes included in different phones might show slightly different behaviors. An additional shortcoming of the evaluation tests with the target group is that a strap was added to the back of the monster halfway through the data collection, potentially influencing the results.

We have not explored data augmentation strategies, but since data collection is both costly and time-consuming, data augmentation might help increase the number of available data points and increase the versatility of trained models. Data augmentation could potentially help increase the performance and reliability of the trained models when applying them to real-time applications.

Children with cerebral palsy stand to benefit from innovative solutions for aiding them in the arduous rehabilitation process. We believe the developed application can serve as a valuable tool, and the sub-modules can be integrated into other applications.

Acknowledgments

This study was funded by Utrecht University, the Amsterdam University of Applied Sciences, Eindhoven University of Technology as part of the Smart Technologies Empowering Citizens project, an in collaboration with Phillips and Ijsfontein and the HUMAN-AI fund. The authors declare that there was no influence or involvement

from the funding organization in the design, data collection, analysis, interpretation, writing, or submission of this study.

References

- [1] Max Alberts, Ellen AM de Ridder, Joris AJ Lodewijks, Tamara V Pinos Cisneros, Kayleigh Schoorl, Albert Ali Salah, and Ben AM Schouten. 2022. Designing a Smartphone Exergame for Children with Cerebral Palsy in the Home Environment. In *Extended Abstracts of the 2022 Annual Symposium on Computer-Human Interaction in Play*. 183–188.
- [2] Ines Ayed, Adel Ghazel, Antoni Jaume-i Capo, Gabriel Moyà-Alcover, Javier Varona, and Pau Martínez-Bueso. 2019. Vision-based serious games and virtual reality systems for motor rehabilitation: A review geared toward a research methodology. *International journal of medical informatics* 131 (2019), 103909.
- [3] Ferhat Bozkurt. 2021. A Comparative Study on Classifying Human Activities Using Classical Machine and Deep Learning Methods. *Arabian Journal for Science and Engineering* (2021), 1–15.
- [4] Annette Brons, Antoine de Schipper, Svetlana Mironcika, Huub Toussaint, Ben Schouten, Sander Bakkes, and Ben Kröse. 2021. Assessing Children's Fine Motor Skills With Sensor-Augmented Toys: Machine Learning Approach. *Journal of Medical Internet Research* 23, 4 (2021), e24237.
- [5] Hennie Brugman, Albert Russel, and Xd Nijmegen. 2004. Annotating Multimedia/Multi-modal Resources with ELAN. (2004), 2065–2068. Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands. <https://archive.mpi.nl/tla/elan> (visited: 2022-03-22).
- [6] Yuwen Chen, Kunhua Zhong, Ju Zhang, Qilong Sun, Xueliang Zhao, et al. 2016. LSTM networks for mobile human activity recognition. In *Proceedings of the 2016 International Conference on Artificial Intelligence: Technologies and Applications, Bangkok, Thailand*. 24–25.
- [7] Tamara Pinos Cisneros, Ben Kröse, Ben Schouten, and Geke Ludden. 2020. Hand rehabilitation for children with cerebral palsy: From clinical settings to home environment. *Editors: Kirsty Christer, Claire Craig & Paul Chamberlain* (2020), 65.
- [8] Chris Ding and Hanchuan Peng. 2005. Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology* 3, 02 (2005), 185–205.
- [9] Ann-Christin Eliasson, Lena Krumlinde-Sundholm, Birgit Rösblad, Eva Beckung, Marianne Arner, Ann-Marie Öhrvall, and Peter Rosenbaum. 2006. The Manual Ability Classification System (MACS) for children with cerebral palsy: scale development and evidence of validity and reliability. *Developmental medicine and child neurology* 48, 7 (2006), 549–554.

- [10] Peter Fikar, Florian Güldenpfennig, and Roman Ganhör. 2018. The use (fulness) of therapeutic toys: Practice-derived design lenses for toy design. In *Proceedings of the 2018 Designing Interactive Systems Conference*. 289–300.
- [11] İpek Gürbüzsöl, Tilbe Göksun, and Aykut Coşkun. 2022. Eliciting parents' insights into products for supporting and tracking children's fine motor development. In *Interaction Design and Children*. 544–550.
- [12] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [13] Andrey Ignatov. 2018. Real-time human activity recognition from accelerometer data using Convolutional Neural Networks. *Applied Soft Computing* 62 (2018), 915–922.
- [14] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [15] Daniela Micucci, Marco Mobilio, and Paolo Napoletano. 2017. Unimib shar: A dataset for human activity recognition using acceleration data from smartphones. *Applied Sciences* 7, 10 (2017), 1101.
- [16] Svetlana Mironcika, Antoine de Schipper, Annette Brons, Huub Toussaint, Ben Kröse, and Ben Schouten. 2018. Smart toys design opportunities for measuring children's fine motor skills development. In *Proceedings of the Twelfth International Conference on Tangible, Embedded, and Embodied Interaction*. 349–356.
- [17] David Moreno Naya, Francisco J Vazquez-Araujo, Paula M Castro, Jamile Vivas Costa, Adriana Dapena, and Luz González Doniz. 2021. Utilization of a Mobile Application for Motor Skill Evaluation in Children. *Applied Sciences* 11, 2 (2021), 663.
- [18] Tamanna Motahar and Jason Wiese. 2022. A Review of Personal Informatics Research for People with Motor Disabilities. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–31.
- [19] Abdulmajid Murad and Jae-Young Pyun. 2017. Deep recurrent neural networks for human activity recognition. *Sensors* 17, 11 (2017), 2556.
- [20] Henry Friday Nweke, Ying Wah Teh, Mohammed Ali Al-Garadi, and Uzoma Rita Alo. 2018. Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges. *Expert Systems with Applications* 105 (2018), 233–261.
- [21] Henry Friday Nweke, Ying Wah Teh, Ghulam Mujtaba, and Mohammed Ali Al-Garadi. 2019. Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions. *Information Fusion* 46 (2019), 147–170.
- [22] Zoey E Page, Stephanie Barrington, Jacqueline Edwards, and Lisa M Barnett. 2017. Do active video games benefit the motor skill development of non-typically developing children and adolescents: A systematic review. *Journal of science and medicine in sport* 20, 12 (2017), 1087–1100.
- [23] Liangying Peng, Ling Chen, Zhenan Ye, and Yi Zhang. 2018. Aroma: A deep multi-task learning based simple and complex human activity recognition

- method using wearable sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–16.
- [24] Ivan Miguel Pires, Nuno M Garcia, Nuno Pombo, and Francisco Flórez-Revuelta. 2016. From data acquisition to data fusion: a comprehensive review and a roadmap for the identification of activities of daily living using mobile devices. *Sensors* 16, 2 (2016), 184.
- [25] Nishkam Ravi, Nikhil Dandekar, Preetham Mysore, and Michael L Littman. 2005. Activity recognition from accelerometer data. In *AAAI*, Vol. 5. 1541–1546.
- [26] Richard M Ryan and Edward L Deci. 2000. Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology* 25, 1 (2000), 54–67.
- [27] Albert Ali Salah, Ben AM Schouten, Stefan Göbel, and Bert Arnrich. 2014. Playful interactions and serious games. *Journal of Ambient Intelligence and Smart Environments* 6, 3 (2014), 259–262.
- [28] Jörg Sander, Antoine de Schipper, Annette Brons, Svetlana Mironcika, Huub Toussaint, Ben Schouten, and Ben Kröse. 2017. Detecting delays in motor skill development of children through data analysis of a smart play device. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*. Association for Computing Machinery, 88–91.
- [29] Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing* 45, 11 (1997), 2673–2681.
- [30] Luis Serpa-Andrade, Isaac Ojeda-Zamalloa, Angel Perez-Muñoz, Vladimir Robles-Bykbaev, and Adriana Leon-Pesantez. 2021. Intelligent Environment to Support Fine Motor Learning in Children with and Without Motor Disorder. In *International Conference on Applied Human Factors and Ergonomics*. Springer, 332–337.
- [31] David Williamson Shaffer and James Paul Gee. 2006. *How computer games help children learn*. Springer.
- [32] Muhammad Shoaib, Stephan Bosch, Ozlem Durmaz Incel, Hans Scholten, and Paul JM Havinga. 2014. Fusion of smartphone motion sensors for physical activity recognition. *Sensors* 14, 6 (2014), 10146–10176.
- [33] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. 2019. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters* 119 (2019), 3–11.
- [34] Kun Xia, Jianguang Huang, and Hanyu Wang. 2020. LSTM-CNN architecture for human activity recognition. *IEEE Access* 8 (2020), 56855–56866.
- [35] Hanbin Zhang, Chenhan Xu, Huining Li, Aditya Singh Rathore, Chen Song, Zhisheng Yan, Dongmei Li, Feng Lin, Kun Wang, and Wenyao Xu. 2019. Pdmov: Towards passive medication adherence monitoring of parkinson’s disease using smartphone-based gait assessment. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 3, 3 (2019), 1–23.
- [36] Zhenyu Zhao, Radhika Anand, and Mallory Wang. 2019. Maximum relevance and minimum redundancy feature selection methods for a marketing machine

learning platform. In *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 442–452.