Machine Learning (2026) 115:25
https://doi.org/10.1007/s10994-025-06950-7

Machine Learning to Identify a New Digital Biomarker to Monitor Everyday Upper Limb Use in Children with Unilateral Cerebral Palsy

Silvia Filogna[1] · Giuseppe Prencipe[2] · Alina Sîrbu[4] · Elena Beani[1,3] · Davide Marchi[2] · Giordano Scerra[2] · Giuseppina Sgandurra[1,3]

Received: 12 April 2025 / Revised: 22 September 2025 / Accepted: 9 December 2025

Abstract
Children with Unilateral Cerebral Palsy (UCP) often experience reduced spontaneous use of the non-dominant upper limb in daily life. Traditional clinical assessments, such as the Assisting Hand Assessment (AHA), provide valuable but clinical-environment-related and thus episodic measures of functional performance in structured settings. There remains a critical need for tools that enable continuous, ecologically valid, and objective monitoring of upper limb activity in real-world environments. In this study, we introduce the *Daily AHA Biomarker (DAB)*, a novel digital biomarker derived from wearable sensor data, designed to estimate AHA scores based on spontaneous motor behavior. The DAB is intended to be used in clinical practice to monitor the movement of the upper extremities of subjects with UCP in their naturalistic environments. Using bilateral wrist-worn accelerometers, we collected multiday time-series data from 80 children (54 with UCP and 26 with Typical Development). Our Machine Learning pipeline combines time-series classification and regression to predict AHA scores from unstructured, daily living-recorded data. The final DAB indicator showed high predictive accuracy ($R^2 = 0.709$) and a strong correlation with the clinical AHA score and the Manual Ability Classification System (MACS) level.

Keywords Ensemble methods · Digital biomarker · Upper limb · Monitoring daily life activities · Unilateral cerebral palsy · Children

1 Introduction

Unilateral Cerebral Palsy (UCP) is the most common form of Cerebral Palsy (CP), accounting for approximately one-third of all cases (Himmelmann et al., 2018). It is characterized by motor impairments predominantly affecting one side of the body, often resulting

Editors: Riccardo Guidotti, Anna Monreale, Dino Pedreschi.

Extended author information available on the last page of the article

Published online: 23 January 2026

in limited spontaneous use of the non-dominant upper limb. These limitations can significantly impact children's autonomy in daily life activities and overall quality of life. As such, evaluating and monitoring upper limb function is a central goal for providing personalized pediatric neurorehabilitation.

Traditional assessment tools, such as the Assisting Hand Assessment (AHA, Krumlinde-Sundholm and Eliasson 2003) and Melbourne Assessment 2 (MA-2, Randall 1999), provide validated and standardized measures of the non-dominant hand and arm use during bimanual tasks as well as their individual use, respectively. However, these tools are typically administered in controlled clinical settings and offer only a snapshot of the child's functional abilities. This can reduce the measurement of spontaneous and context-dependent behaviors, which, if reliable, can be better indicators of real-world functioning. On the other hand, daily life assessment is typically described with functioning classification metrics such as the Manual Ability Classification System (MACS) (Eliasson et al., 2006), which outlines children's independence in daily life activities, and by parent-reported questionnaires, making these measures inherently subjective.

Recent advances in wearable sensor technology, particularly the use of wrist-worn accelerometers, have enabled quantitative continuous, non-invasive movement monitoring in naturalistic environments such as at home or school. These technologies have the potential to extend the assessment beyond the clinic and into everyday life. Several studies have shown correlations between sensor-derived metrics (e.g., asymmetry indexes, movement, intensity) and clinical scores such as the AHA or MA-2; however, most of these studies rely on linear statistical analyses, focus on structured tasks or limited monitoring periods, and stop short of developing predictive models.

At the same time, Machine Learning (ML) techniques have been increasingly applied to time series data from wearable devices, most commonly for activity recognition or energy expenditure estimation (Sabry et al., 2022). Although valuable, these approaches rarely target clinically meaningful outcomes, especially in children with motor impairments, and are often limited to the classification of general activity types rather than functional assessments.

In this study, we address these gaps by introducing a novel digital biomarker, the *Daily AHA Biomarker* (DAB), designed to estimate AHA scores from accelerometer data collected during everyday activities. Our ML pipeline transforms time-series data from bilateral wrist sensors into an interpretable indicator that is aligned with standard AHA scoring. The pipeline combines time series classification, aggregation of model outputs, and regression to predict AHA values in an ecologically valid and personalized manner.

By bridging the gap between wearable sensing, machine learning, and clinical neurorehabilitation, this work aims to support clinicians with tools for remote, continuous monitoring and outcome tracking, paving the way for more adaptive and individualized care for patients with UCP.

Overall, the main contributions of this work are as follows: we introduce the DAB, a novel digital indicator derived from wearable actigraph data that provides an ecologically valid estimate of AHA scores and enables continuous monitoring of upper limb use in children with UCP. To support this, we design a machine learning pipeline specifically tailored for time-series accelerometer data, combining classification, clustering, and regression models to handle multiday, noisy signals and accurately predict clinically meaningful outcomes. Furthermore, we present a unique dataset that integrates clinical AHA evaluations, MACS classifications, and one of the largest collections of bilateral actigraph recordings

in children with UCP and typically developing peers, gathered in both clinical and real-life settings. Finally, we propose a clinically interpretable framework for remote and scalable monitoring, demonstrating that wearable-sensor-based ML models can provide robust and actionable insights to support personalized rehabilitation strategies and the integration of digital biomarkers into pediatric neurorehabilitation practice.

The rest of the paper is organised as follows: Sect. 2 reviews the existing literature on wearable sensors and machine learning applications for upper limb monitoring in children. Section 3 details the methodology employed to develop the DAB, including participant recruitment, data preprocessing, and machine learning models. Section 4 presents the experimental results, while Sect. 5 provides a critical discussion of the findings and their clinical implications, highlighting the main differences with other studies available in the literature on the topic. Finally, Sect. 6 concludes the paper and outlines directions for future work. For the sake of clarity, in Table 1 we report a glossary of all acronyms used in the paper.

2 Background

This section outlines the existing literature and research work relevant to the topic addressed in this paper. Our goal is to contextualize the current state of the field, identify the key developments, and summarize the main methodological approaches and perspectives that have shaped the field thus far. This background is intended to establish the conceptual and technical groundwork for understanding the motivation, the relevance and positioning of our study. A direct comparison between the aspects of the existing literature most pertinent to our contributions and the novel elements introduced in this work is deferred to Sect. 5 (please refer to Tables 2, 3 and 4 for a synthetic overview of the main works in the literature related to our study.)

Table 1 Glossary of acronyms

Acronym	Full term
CP	Cerebral Palsy
UCP	Unilateral Cerebral Palsy
AHA	Assisting Hand Assessment
MA-2	Melbourne Assessment 2
MACS	Manual Ability Classification System
ML	Machine Learning
DAB	Daily AHA Biomarker
IMU	Inertial Measurement Unit
AI	Asymmetry Index
TD	Typical Development
AC	Activity Counts
D	Dominant
ND	Non Dominant
TS	Time Series
MVS	Mean Validation Score

Table 2 ML approaches in literature (part I)

Reference	Objective	Methodology	Results
Ahmadi et al. (2020a)	Assessment of physical activities (PA) in children with CP	Random Forest to classify activities (out of three) using accelerometer at wrist, hip, and ankle and their combination	In lab settings, the models had accuracy above 78%; in simulated free-living scenario, accuracy dropped ranging from 51% to 61.5%
Ruch et al. (2011)	Hip and wrist sensors to recognize the mode of children's PA in free-living conditions	Tested different models: k-Nearest Neighbor (best single classifier overall), Normal Density Discriminant Function, Custom Decision Tree, Majority Vote (meta-classifier)	The procedure was able to classify 90% of stationary activities, 83% of walking, 81% of running and 61% of jumping activities (overall recognition rate of 67% when combining classifiers)
de Vries et al. (2011)	Hip and wrist sensors to classify children's PA type (8 types)	Four Artificial Neural Network classification models, based on feed-forward network	Overall accuracy between 57% and 76%
Trost et al. (2012)	Classification of five PA types, in a lab setting	Feed-forward neural networks with a single hidden layer	Overall accuracy: 81.3% (10 s window) to 88.4% (60 s window), depending on the kind of activity
Ahmadi et al. (2020b)	Energy expenditure estimation in free-living scenarios, using hip and wrist accelerometers	Random Forest, Feed-forward neural network, Support Vector Machine, Logistic Regression	RMSE (Root Mean Square Error) range: 0.58–0.73 kcals/min (0.92–1.15 METs, Metabolic Equivalent of Task), and MAPE (Mean Absolute Percent Error) range: 25%–36%
Ahmadi et al. (2020d)	PA classification in lab using raw data from a single accelerometer placed at the hip	Random Forest and Logistic Regression	Random Forest overall accuracy: 95.7% and overall F-score: 95.8%; Logistic Regression overall accuracy: 94.3% and overall F-score: 94.6%
Chowdhury et al. (2018)	Energy expenditure prediction with wearable on wrist and hip (quite restricted population of 8 children), in a lab scenario	Convolutional Neural Network, various conventional supervised ML (such as Multiple Linear Regression, SVM), Simplified Regression	Accuracy in terms of RMSE (kcal/min): CNN 0.54–0.66, Conventional ML 0.55–0.62, Simplified Regression 0.82–0.89

2.1 Introduction to the Problem

The assessment of motor function in children with CP has traditionally relied on certain standardized measures, such as the semi-structured AHA (Krumlinde-Sundholm & Eliasson, 2003) or MA-2 (Randall, 1999), which are more widely adopted for evaluating movement in clinical practice (da Silva et al., 2022). Although widely used and clinically validated, these scales are time-consuming, require trained raters, and are subject to inter- and intra-rater variability, which limits their feasibility for frequent or continuous monitoring. As a result, there has been a growing interest in the development of digital technologies capable of providing objective, reproducible, and scalable measures of movement. In particular, wearable sensors, when combined with ML techniques, have significantly enhanced quantitative movement analysis. These improvements have increased the amount of meaningful information that can be extracted from wearable devices; consequently, ML have emerged as a promising tools for deriving new digital biomarkers that could complement traditional

Table 3 ML approaches in literature (part II)

Reference	Objective	Methodology	Results
Zheng et al. (2013)	Activity Recognition (out of 7) with accelerometer (hip or wrist), in a controlled environment	SVM ensemble, Single SVM, Feed-forward with 1 hidden layer, 1-Nearest Neighbor with Euclidean (EUC) or Dynamic Time Warping (DTW)	ANNs performed reasonably well, but below SVM. Accuracy: 94% for hip, 89% for wrist
Coyle-Asbil et al. (2024)	Energy expenditure prediction in semi-structured familiar environments using triaxial accelerometers from different brands	CNNs, Linear Regression, Random Forest (RF), Fully Connected Neural Network (FcNN)	CNNs consistently had lowest RMSE values, improving over the other models, on the internal dataset, demonstrating the benefit of automated feature learning; however, it failed to generalize well across populations and accelerometer brands
Trost et al. (2016)	Energy expenditure in children with CP in lab setting. with accelerometer (hip)	Decision Tree (DT) models	Overall accuracy: about 80% across impairment levels
Mathew et al. (2024)	Measuring functional hand use of children with UCP in a semi-structured play and clinical tasks, wearing dual wrist-worn accelerometers (using activity counts)	Random Forest, k-Nearest Neighbor, Support Vector Machine. Video-labelled functional vs. non-functional hand use was used as ground truth	Random Forest was the best performer, with an F1 score of about 0.8
Chowdhury et al. (2017)	Test whether ensemble machine learning methods improve recognition of PA from wrist accelerometer data compared to single classifiers	Single classifiers: Binary Decision Tree, kNN, SVM, ANN; Conventional ensembles: Bagging, Boosting (Adaboost), Random Forest; Custom ensemble: Combined BDT, kNN, SVM, ANN using three fusion rules (Weighted Majority Voting - WMV, Naïve Bayes, Behaviour Knowledge Space)	Ensemble methods outperformed single classifiers across all tested datasets. Random Forest is a reliable conventional method. Custom ensemble (WMV fusion) achieved the highest recognition accuracy overall

Table 4 ML approaches in literature (part III)

Reference	Objective	Methodology	Results
Ahmadi and Trost (2022)	Compares the accuracy of PA intensity predictions for preschool children in a free living scenario, provided by already published ML classification models. Accelerometers worn on hip and non-dominant wrist	Random Forest classifiers trained separately for hip and wrist placements	Hip RF classifier overall accuracy: 81%, Wrist RF classifier overall accuracy: 75%
Goodlich et al. (2020)	Classification of six standardized tasks for children with CP in a controlled laboratory environment. Accelerometers placed on wrist, hip, thigh	Decision Tree, Random Forest, and Support Vector Machine	Single-sensor models: 59%–79% (Best = Wrist RF), Hip/Thigh alone: 62–66%. Two-sensor models: 75%–92% (Best = Wrist + Hip RF), Wrist + Thigh 88–90%, Hip + Thigh 75%. Three-sensor models: 88%–90% (Best = SVM, RF 88%)

clinical evaluations for CP and enabling long-term monitoring in daily life contexts (Idri et al., 2018).

2.2 Wearable Sensors in Children: From Traditional Metrics to Advanced Analysis

Accelerometers and, more generally, inertial measurement units (IMUs) are currently the most widely used sensors for assessment of movement in children . A large body of literature focuses on the use of wearable technologies to monitor physical activity in both children with Typical Development (TD) (Au et al., 2024; Sousa et al., 2023), and those with neurodevelopmental conditions (González Barral & Servais, 2025). Concerning the studies related to the CP, most of them were conducted in a controlled environment (Ahmadi et al., 2018, 2020a; Beani et al., 2019) with respect to home-based monitoring. In fact, a part of the literature explores the integration of sensor-derived metrics with clinical assessments in children, attempting to establish a direct link between wearables' parameters, obtained from upper limb movements, and standardized clinical assessment scores (Braito et al., 2018). For instance, Beani *et al.* first investigated the use of two ActiGraph wGT3X (one for each wrist) during the AHA in subjects with TD and with UCP. By using a novel index, the *Asymmetry Index* (AI), the authors demonstrated that accelerometry is a valid tool for measuring the asymmetry in the use of dominant and non-dominant hands in clinics (Beani et al., 2019) and home settings (Beani et al., 2025), as expected from the clinical score. Likewise, Burgess et al. (2024), evaluated the validity of ActiGraph wGT3X during the assessment with the Both Hands Assessment (BoHA, Elvrum et al., 2018) as a tool for measuring upper limb movement for children with bilateral CP. Once again, the hypothesis of a relationship between the BoHA clinical score and the technological Asymmetry Index has been tested. In Hoyt et al. (2020) an index reflecting the intensity and frequency of movement was analytically calculated and compared with each participant's MA-2 scores to assess real-world activity. These metrics offered a first step towards objective quantification of motor behavior and were relatively easy to compute and interpret. However, they often lacked sensitivity in capturing the complexity of movement quality, and their predictive capacity compared to functional outcomes was limited. Moreover, such features typically relied on arbitrary thresholds or simplified assumptions, thus restricting their generalizability across different populations and clinical conditions. Furthermore, the large amounts of raw data produced by these devices are not directly interpretable by clinicians, highlighting the need for advanced computational techniques to extract clinically meaningful features. In this scenario, data analysis becomes a significant challenge. While wearables offer a key advantage for medium- and long-term monitoring, they also generate large volumes of data that require advanced processing techniques—beyond the classical statistical methods used in previously cited studies. Indeed, traditional data analysis methods do not scale well when datasets grow in size and complexity; in addition, difficulties can occur when dealing with multidimensional time-series data, namely continuous streams of signals from X, Y, and Z axes. These limitations paved the way for more sophisticated analytical methods based on ML.

2.3 Machine Learning Approaches for Movement Analysis in Children

In recent years, ML and deep learning algorithms have been increasingly applied to accelerometer and IMU data to automatically identify relevant movement patterns and classify motor and physical activity. Physical activity is a broad umbrella term that includes measuring how much, how often, and how intense a child's physical activity is. Works in this area focused on developing classification models to distinguish between types of physical activities in children or to predict the related energy expenditure (Ahmadi et al., 2020a, 2020b, 2020c, 2020d, 2020e; de Vries et al., 2011; Ruch et al., 2011; Trost et al., 2012). The growing interest in estimating energy expenditure is seen in works such as Zheng et al. (2013), Ahmadi et al. (2020e), Coyle-Asbil et al. (2024), which aimed to model metabolic intensity using accelerometry and supervised algorithms. Other studies addressed free-living activity recognition and modelling under more ecologically valid conditions: for instance, Chowdhury et al. (2018) investigated deep learning and ensemble methods to classify diverse real-world movements in young children. Compared to rule-based approaches, ML methods can capture complex, non-linear relationships in the data and thus provide richer representations of motor function. Closest to our study, Ahmadi et al. (2020a) and Trost et al. (2016) designed and tested classification models that account for atypical movement profiles, while Mathew et al. (2024) and Letts et al. (2024) emphasized diverse cohorts and the importance of inclusive dataset development for training robust models. In this context, the researchers have explored a wide variety of computational models, Random Forest (RF), Support Vector Machine (SVM), Feed-Forward Networks classifiers (Ahmadi et al., 2020a, 2020b, 2020c, 2020d; Letts et al., 2024; Trost et al., 2016) or ensemble learning algorithms (Chowdhury et al., 2017) to improve physical activity recognition accuracy compared to the single classifier algorithms. Pattern recognition techniques on upper-arm data were also employed to identify activity categories (Ruch et al., 2011). Artificial Neural Networks (ANNs) were also used to model nonlinear patterns in children's movements (de Vries et al., 2011; Trost et al., 2012). Overall, these studies span a broad range of pediatric populations, from toddlers to adolescents, as well as children with motor impairments. Many early and mid-stage works, including de Vries et al. (2011), Trost et al. (2012), and Zheng et al. (2013), recruited school-aged children, typically between 5 and 15 years. These cohorts often participated in structured activity protocols such as walking, running, and stair climbing. A growing number of studies target preschool-aged children (3–5 years), recognizing that movement patterns in this group are highly variable and unstructured. Notable examples include (Ahmadi et al., 2020b, 2020c, 2020e; Chowdhury et al., 2018; Coyle-Asbil et al., 2024). These studies often feature play-based or semi-free-living tasks to better simulate real-world behaviors.

From this overview, it emerges that wearable sensors are a powerful means to objectively characterize motor function beyond what can be observed in clinical assessments, but very few studies in the literature focus in the monitoring of a children population at home, and that the development of predictive digital biomarkers, derived through advanced ML techniques, correlating with established clinical scales, such as the AHA, remains an underexplored area. Despite advancements in wearable sensing and ML-based movement analysis, critical gaps remain. To the best of our knowledge, no prior work has specifically investigated the spontaneous movement of children in their natural daily environments, for more than 2 days, using wearable sensors and ML techniques. The development of such a digital biomarker could serve as a powerful tool to quantify the spontaneous behaviour

continuously and in a naturalistic environment. Based on the previous experiences on Actigraph validation compared with clinical scores, our study aims to develop an ML-based framework to analyze long-term movement data and derive a digital biomarker, the DAB, that provides information on children's spontaneous movement and predicts clinical scores.

3 Methodology

The overall goal of this study is to define the DAB, a new digital biomarker that correlates the clinical assessment of the upper limb use of a subject with his/her behavior in a daily life environment, thus monitoring the children also in a non-clinical environment, providing a score that can be compared to then subject AHA score. For this, we collected Actigraph time series data from 80 children, in both clinical and home settings. At the time of clinical data collection, the children were evaluated with the AHA test (see Sect. 3.1). Using the data collected in the clinical setting, we trained a series of ML models suitable for time series data that can classify children into TD or UCP groups, based on fixed-sized windows from their actigraph data. The models with the highest performance were then applied to all the data collected in the home setting, computing, for each child and each model, the fraction of fixed-size windows classified as TD. The fractions from all models were then employed to estimate the AHA score of each child, using linear regression. The prediction of the regressor constitutes the novel DAB score, which can be applied to windows of different sizes of the time series data. We apply this to the data collected in the home setting, using overlapping 6-hour windows, to monitor children in their naturalistic environment.[1] Figure 1 shows a graphical representation of the entire methodology, while the rest of this section provides detailed descriptions for each step.

3.1 Participants and Data Acquisition Setup

The study involved 80 children, including 26 children with TD and 54 with UCP, with mean ages and standard deviation of 10.28 ± 4.3 years and 10.86 ± 4.1 years, respectively. Children with UCP were recruited among the patients of the Stella Maris Clinic while children with TD participated on a voluntary basis .[2] Data were collected using actigraph accelerometers (wGT3X-BT Monitor, ActiGraph, Florida, FL, model 7164; 4.6 cm × 3.3 cm × 1.5 cm, 19 g, Firmware v1.8.0) worn on both wrists. All subjects were first evaluated during the AHA clinical evaluation session, after which they were instructed to wear the actigraph sensors as much as possible for the next week. The single activity monitor consists of a tri-axis accelerometer with a dynamic range of 8 g, able to accurately detect the accelerations associated with upper arm movement. The sampling rate of the acceleration signals was set at 80 Hz and the data were stored locally into non-volatile flash memory, in gravitational units (i.e., 1 g is equal to the Earth standard gravitational unit). At the end of the acquisition, we had data recorded in a clinical setting during AHA, and in a home setting during the week.

[1] The entire analysis was implemented in Python and the code is available on GitHub (Filogna et al., https://github.com/giordanoscerra/AInCP-ML).

[2] Data are publicly available at https://doi.org/10.5281/zenodo.15019553 at request. All families filled in a written consent to participate in the study. Ethics approval was obtained from the Tuscany Paediatric Ethics Committee, Italy (78/2016). This study was subsequently registered on clinicaltrials.gov (NCT03054441).

Raw actigraph output consists of acceleration measured in the three dimensions. However, the ActiLife software also computes, for each dimension, an aggregated value based on the raw accelerometer data, called Activity Counts (AC), , which is the data we consider. We first transformed the AC data into a representable format: for each (aggregated) point in the series, we calculated the magnitude of the 3D activity vector ($|v| = \sqrt{(x^2 + y^2 + z^2)}$), thus obtaining two series for the Dominant (D) and Non-Dominant (ND) hand, respectively, which can be represented in a bivariate time series. Furthermore, we computed the pointwise Asymmetry Index (AI). This indicator represents the degree of asymmetry between the dominant and non-dominant hand of a subject. It can take discrete values between -100 (total limb asymmetry towards the non-dominant hand) and 100 (total limb asymmetry towards the dominant hand) (Beani et al., 2019). The AI is the only empirical indicator that can be computed directly from the actigraph data; in fact, it can be calculated from the data using this formula:

$$AI = (v_D - v_{ND})/(v_D + v_{ND}) \times 100$$

where v_D and v_{ND} are the magnitude of the 3D activity vectors for the D and ND upper limbs, respectively.

Besides accelerometer monitoring, subjects were also clinically described, i.e. tested using with AHA and classified according to MACS. Subjects with TD were assigned by convention with an AHA scores of 100 and MACS scores of 0.

3.2 Dataset

Our dataset is divided into two parts: one part is related to the AHA clinical test (called the *clinical* Time Series (TS) data), while the other concerns measurements taken over the following days of the subjects' daily life (called the *home* TS data). Although the recordings differ in duration, each consists of multivariate time series: at every time point, the three acceleration components (x, y, z) from each wrist are observed simultaneously. By computing the Euclidean norm of the three coordinates, we obtain one univariate time series per wrist. Taken together, these form a bivariate time series (one for each hand).

- **Clinical TS data**. Data related to the AHA session vary in length according to the subject's condition, ranging from a minimum of 11 min to a maximum of 27 min (averaging 18 min). The test is designed with age-appropriated playful activities in which objects and toys requiring bi-manual use are presented in a semi-structured context.
- **Home TS data**. Each subject was recorded outside of the clinical environment for one week. Subjects were deliberately not given clear instructions on how to behave during this period, in order to avoid stimulating non spontaneous or unnatural movements. Because of this, the at-home data are interesting when attempting to understand an individual's "true" motor activity. The immediate consequence of recording data in unstructured and free environments is, of course, the presence of high amount of noise within the measurements, which makes the analysis of asymmetry between the two limbs less significant and more challenging, especially among the many daily activities that do not involve the use of both limbs. A further consequence of the subjects' semi-independence during the week is the decision to wear or not wear the sensors during the night, an aspect that proved important during the development of our study.

Overall, at the end of the data acquisition process we have acquired, for each subject:

- the clinical AHA score (i.e., 0–100),
- the MACS level (i.e., 0–3),
- the clinical TS data from actigraphs,
- the home TS data from actigraphs.

3.3 Data Preprocessing

For preprocessing, we did not use a traditional approach focused on feature extraction from time series data through statistical theory using averages, standard deviations, and variances. Instead, we opted to use ML models specifically designed to work with time series (appropriately pre-processed) to maintain awareness of the environment in which we are working. For this, some pre-processing of the time series for each patient was needed.

First, when comparing different time series, it was necessary to standardize the length of the sessions, considering that they differ in duration. We thus divided the series into windows of fixed length, which we will call *samples*. In particular, we tested samples of 300, 600, and 900 s. This was applied to both clinical TS data, which are shorter, and home TS data. For each window, the sampling procedure checks the following: first, if a series is not long enough, points from the beginning of the series are concatenated until the desired length for the sample is reached. On the other hand, if the duration of a series exceeds the desired length, it is divided into more than one sample. The "excess" is removed equally from the beginning and the end of the series, as these moments are typically less significant. This procedure resulted in some data loss: the samplings of 300, 600, and 900 s lost 8.5%, 23.7%, and 26.9% of the total seconds for each series, respectively. Our results are shown with 300-second samples. Other window sizes provided similar performance, so we opted for the size that minimises data loss.

After splitting, for each patient, we have one or more samples in the clinical TS dataset and many samples in the home TS data. Each sample consists of two time series: one for the dominant limb and one for the non-dominant limb. To avoid working with multivariate time series, we then represented each sample with a single time series, resulting from a composition of the two upper limbs. In particular, we tested three *composition* techniques:

- **Concatenation**: the two series are concatenated. Specifically, the non-dominant series follows the dominant one. This results in a time series that is twice as long.
- **Difference**: here, we subtracted the time series of the non-dominant hand from that of the dominant hand, point by point. This results in a time series that represents the difference between the magnitudes of the two hands, which can therefore also take on negative values.
- **Point-to-point AI**: the AI formula has been applied to each point of the two series. As a result, a kind of "normalized" version of the difference is obtained, which does not take into account the individual intensities but oscillates between − 100 and 100.

Before proceeding with the ML analysis, the data was divided into *train* and *test* datasets, to enable out-of-sample model evaluation. The split was done at the participant level, with data from 8 children placed in the test dataset and 72 children in the training dataset. The

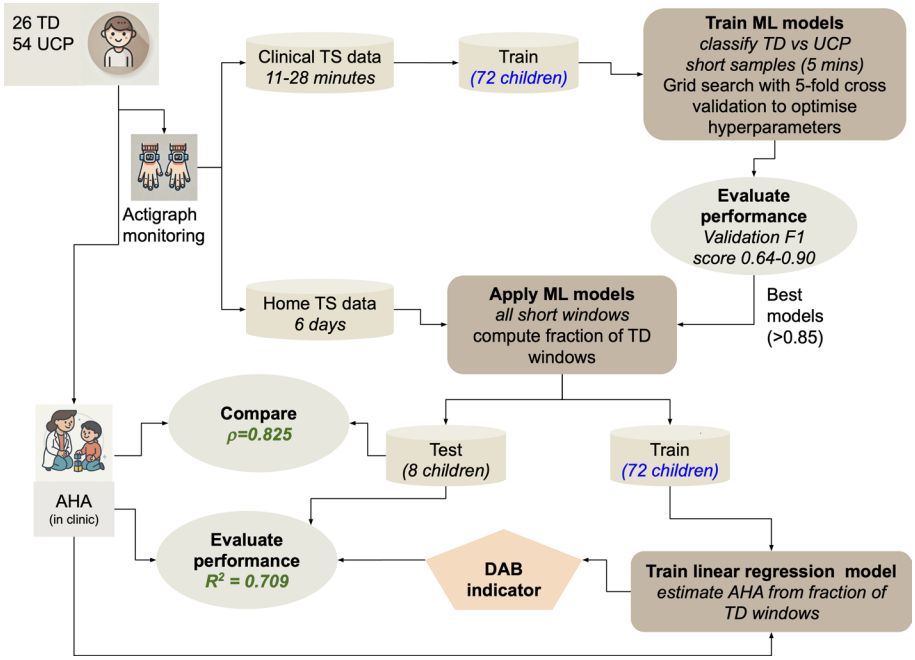

Fig. 1 Methodology. Graphical representation of the steps taken in our study, to obtain the final DAB. The methodology employed a 10-fold cross validation procedure, the figure only shows one fold. However, the evaluation results shown in green are averages over the 10 folds

split was stratified, so that the fraction of TD children in test and train data followed the same distribution. We acknowledge that CP children could be different among themselves, however further stratification based on MACS/AHA values was not attempted, due to the limited number of subjects in our data. The procedure was repeated 10 times, resulting thus in a 10-fold cross validation. Model performance will be reported as averages over all splits, overcoming thus possible issues with differences between different test subsets. The same train dataset was used to (1) build the classifier models and (2) to fit the regression model that computes the DAB. The test data instead was used to evaluate (1) the intermediate fraction of TD windows and (2) the regression performance. Please note that, as described in the next section, building the classification models from training data consisted of a further internal 5-fold cross-validation procedure, to tune training hyperparameters. Therefore, we are in a nested cross-validation setting, with 10 external folds and 5 internal folds.

3.4 ML Analysis

The ML analysis consisted of two consecutive steps: we first explored various **models to classify** children with TD or UCP based on their samples extracted from clinical TS data, after which we trained a **linear regressor** to use the classification output to predict AHA scores for our subjects. The sktime[3] (Löning et al., 2019) and scikit-learn (Pedregosa et al., 2011) Python libraries were used for the implementation.

[3] https://zenodo.org/records/15151312.

3.4.1 Classification Models

We used ML models that take time series data as input, rather than needing to manually compute a set of numerical features. This has the advantage that the results do not depend on the features computed, but the models themselves extract the meaningful information from the time series. The only necessary preprocessing is to obtain time series windows of uniform size, as described in the previous section. We decided to concentrate on ML models for time series data since preliminary experience with classical ML on this problem did not yield satisfactory results.

Among possible models, we empirically selected two classification and two clustering methods. For clustering models, the output clusters were transformed into classes by computing the majority label in each cluster, using the data labels, which allowed us to treat clustering methods as if they were classifiers. The classification methods included one distance-based method (ShapeDTW—Shape Discrete Wavelet Transform, Zhao & Itti, 2018), and one dictionary-based method (BOSSEnsemble—Ensemble of Bag of Symbolic Fourier Approximation Symbols, Schäfer, 2015). Other models in these categories, such as for instance K Nearest Neighbours methods, were tested, but performance was low so they were removed from the analysis. For clustering, we selected two partitioning-based methods: Time Series K-Means and Time Series K-Medoids (Kaufman & Rousseeuw, 1990; Lloyd, 1982). Further details on the four models employed are provided in Supplementary material.

We tested all the composition techniques described above on each type of model. We used a 5-fold internal cross-validation grid search to optimize hyperparameter values. The Supplementary Material section includes a table that indicates the hyperparameters optimised for each model type (Table 6 in Sect. B). The best hyperparameter configurations were selected based on the weighted F1 score. As a final evaluation, we computed the Mean Validation Score (MVS) from the 5 scores obtained from the internal 5-fold cross-validation.

The model with the best MVS was applied to the home TS data. The model predicted whether each 300 s sample in the home TS data belonged to a subject with TD or UCP. For each subject, we aggregated all classifications and computed a percentage TD score, i.e. the fraction of samples classified as TD. We evaluated the ability of this measure to estimate whether a subject belongs to the TD group or not, by computing **Pearson correlations** (ρ) with the AHA scores assigned by clinicians, only on home test data. We computed one ρ for each fold, and we report the average values in the following section. This further validated our ML models, demonstrating that their overall output correlates with clinician observations.

3.4.2 Regression Model

The fraction of samples classified as TD computed for our best model can be a good indicator of upper limb activity, but it has no connection to existing clinical scores, making it difficult to interpret. For this, we added a further step to our methodology: using the fraction of TD windows over the entire home TS data computed by multiple best models (those with MVS > 0.85) to predict the AHA score for each child. Given that we observed high Pearson correlations between this fraction and AHA, we used a linear regressor for this task. Specifically, the regressor takes as input features a list of TD fractions (the percentage

TD score above), as computed by various classification models, and computes an estimated AHA score. It was trained on TD fractions computed on home TS training data and tested on TD fractions computed on home TS test data. We report average R^2 values over the 10 external folds of our analysis.

3.4.3 The Daily AHA Biomarker

The regression model can be applied to any actigraph time series to predict an AHA-like value based on the output of several classifiers on short fixed-size windows. We call this value the DAB. In our case, we apply it to study the activity of children during the monitored week. We divided the week into *overlapping 6-hours* windows (72 samples of 300 s, 5 min each). Note that we do not consider a window to be valid unless it contains at least a percentage of valid samples: a sample is valid if it contains at least one significant (intensity greater than zero) movement. Good results were obtained by setting this threshold to 75%. Each 72-sample window overlaps with the previous one by 71 samples (one sample right shift). This solution would clearly help with the continuous monitoring of the DAB during the week's activities.

4 Results

4.1 TD and UCP Classification

We first inspect the performance of the ML models in the classification task. Table 1 shows, for each fold, the best validation F1 scores (MVS) and the number of models with MVS > 0.85, out of a total of 12 possibilities—4 model types (i.e., BOSSEnsemble, ShapeDTW, TimeSeriesMedoids, TimeSeriesKMeans, see Sect. 3.4.1) times 3 composition techniques (i.e., Concatenation, Difference, Point-to-Point AI, see Sect. 3.3). We note that many models achieve very good performance, with the top MVS reaching 0.90. In the supplementary material, we include a series of tables that show, for each fold, the actual parameter configurations for all models that crossed the 0.85 threshold. The most frequent configurations utilized TimeSeriesKMeans and TimeSeriesKMedoids with specific hyperparameters optimized for each iteration, showing that in this case unsupervised models tend to outperform supervised models.

To further evaluate the classification performance, we report the relation between the fraction of TD samples in home time series test data and AHA and MACS scores, for each patient. Figure 2 shows this relation for all models with MVS > 0.85 (a) and also for the best performing model alone (b). The plots combine results on test data for all folds, thus showing all children in one plot. The average correlation for the top models over all folds is 0.826 (corresponding to the results shown in Fig. 2b), confirming a good link between model predictions and AHA scores assigned by clinicians. We also note a good correlation with MACS level: children with low MACS generally have a higher fraction of TD samples, as expected. We do observe lower performance on children with low AHA/high MACS, which could be partially due to the fact that very acute symptoms include involuntary movements, which are recorded by actigraphs and probably interpreted by classifiers as increased movement symmetry (Table 5).

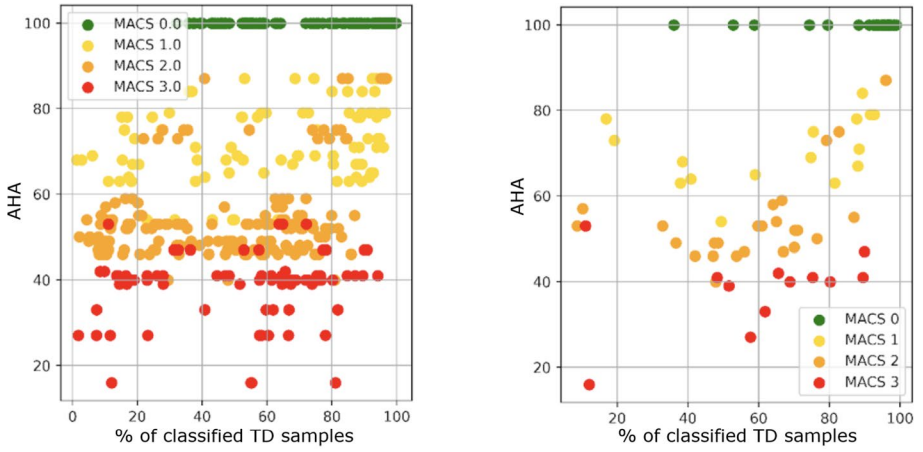

Fig. 2 **a** AHA versus the fraction of TD samples computed with predictions by multiple best models for each of the 10 folds. **b** AHA versus the fraction of TD samples, computed with predictions by the best model of each of the 10 folds

Table 5 Classification performance for each fold

Fold	Number of top classifiers	Max MVS
0	6 out of 12	0.8771
1	4 out of 12	0.8859
2	6 out of 12	0.8915
3	9 out of 12	0.904
4	8 out of 12	0.881
5	3 out of 12	0.8756
6	7 out of 12	0.8768
7	6 out of 12	0.9027
8	4 out of 12	0.8696
9	7 out of 12	0.875

4.2 Regression and the Daily AHA Biomarker

The performance of the regression model on test data, over the all 10 folds, is shown in Fig. 3. The output of the regression model is our final indicator, the DAB. We note very good prediction abilities, with an average R^2 of 0.709 over the 10 folds. DAB values also correspond well to MACS levels, again with subjects with high MACS generally showing lower DAB values. Of course, there is some variability, but we note that a similar variability is also present when comparing the two clinical scores: children in the same MACS category will have a range of AHA values, with the exception of TD children where all AHA values were evaluated as 100. We also note that the children with MACS 3 are better placed after regression compared to what we saw for the fraction of TD samples before (Fig. 2b). This means that by adopting the ensemble approach and training the regressor, our methodology managed to correct the probable involuntary movements in these children.

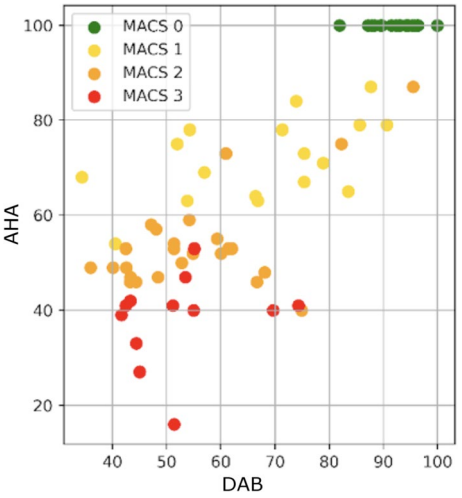

Fig. 3 Real versus predicted AHA, over 10 folds

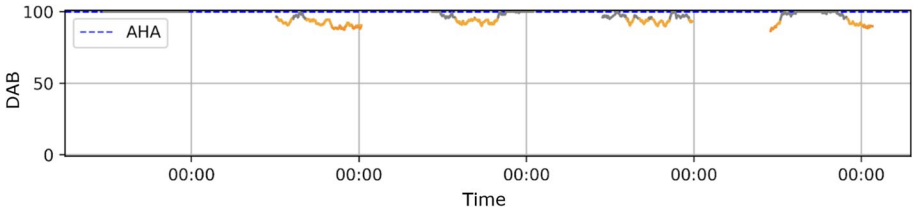

Fig. 4 This plot represents weekly Home-AHA predictions, compared to the real AHA value. Grey represents prediction close to the clinical AHA value, while green and yellow are over and under estimation respectively. This subject belongs to the control group, therefore the related AHA is 100. Home-AHA exceeding 100 are clipped

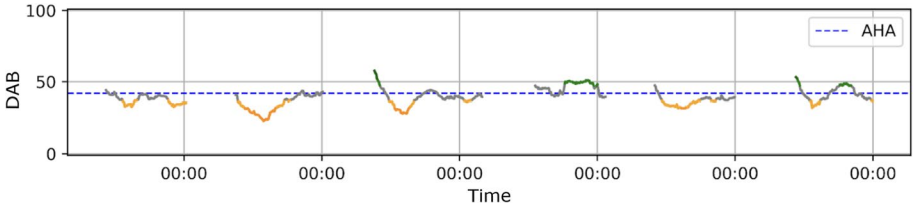

Fig. 5 This plot represents the Home-AHA predictions in a weekly period of time, in relation to the real AHA value. Grey represents values close to the clinical AHA (± 5), green and yellow are over and under estimation respectively (± 10)

We employ the DAB, measured above on all home TS data, to perform continuous monitoring of the children's activity outside the clinical environment, using overlapping 6-hour windows. Figures 4 and 5 show two examples for two children (test dataset) with very different AHA values: one with TD and one with UCP. For the child with TD, we observe that our indicator maintains high values despite temporary drops. This is to be expected, considering that some real-life activities imply some upper-limb asymmetry (e.g. eating, writing).

For the child with UCP, instead, DAB values are lower and vary around their AHA score. Sometimes, the child appears to maintain a higher level of activity compared to the clinical setting, while most of the time it remains below the AHA level. The results validate our indicator and demonstrate that the DAB can be useful in monitoring children with UCP at home.

5 Discussion

This study contributes to the emerging field of digital health monitoring by introducing a novel digital biomarker—the Daily AHA Biomarker—capable of predicting upper limb functional use in children with UCP from accelerometer data collected in unstructured environments of real-life. Traditional approaches for validating wearable sensors have primarily focused on comparing sensor-derived metrics with clinical scales both within and outside controlled environments. The AHA has been widely used as a reference standard for evaluating upper limb function in children with UCP, providing reliable and well-established clinical scores. Indeed, prior works have first established the feasibility of using accelerometers to characterize upper limb asymmetry in UCP; for example, Beani et al. (2019) introduced the Asymmetry Index based on structured clinical sessions and showed its alignment with AHA scores. While this method ensures robust assessments in controlled contexts, it does not fully capture how motor function translates into real-life spontaneous activity. Moreover, such studies primarily rely on descriptive statistics or correlation analyses, limiting their ability to analyse long-term, multivariate time-series data. Some studies, including Hoyt et al. (2020) and Beani et al. (2025), extended the analysis to longer, real-life recordings using standard statistical techniques.

In contrast, our work introduces a novel methodology that leverages a ML pipeline to compute a biomarker derived directly from wearable sensor data and clinical scores. Instead of simply validating sensor-based movement data against AHA scores, our models directly predict AHA scores and reflect daily motor activity from multivariate time series recorded by bilateral wrist-worn actigraphs, using clustering, ensemble methods, and regression. This approach produces a clinically interpretable and ecologically valid digital biomarker (DAB), supporting continuous assessment in naturalistic settings and offering a level of personalization and scalability not addressed in previous studies. Compared to studies relying solely on standard statistical analysis or limited structured settings (Burgess et al., 2024; Konrad et al., 2022), the DAB enables continuous, interpretable, and objective monitoring outside clinical environment, using the same 0–100 scale as the AHA, thereby facilitating its integration into clinical workflows. This represents a crucial step forward, as the goal is not merely to estimate AHA values (although the AHA scores serve as a foundation) but to establish a clinically grounded and controlled index that can be used for continuous monitoring beyond the clinical setting.

Concerning sensor placement, previous studies mainly employed triaxial accelerometers, in varying contexts. Hip-mounted sensors were used in structured or semi-structured settings by studies such as de Vries et al. (e2011), Trost et al. (2012, 2016), Zheng et al. (2013), Ahmadi et al. (2020c, 2020d), providing stable measurements for lower-body movement and it has long been the standard for accelerometry in children. In contrast, wrist-worn sensors have gained popularity, for younger children and free-living simulations as in Ahmadi et al. (2020b, 2020c), Chowdhury et al. (2018). In fact, these studies show that children

better tolerate wrist placement, and it may also capture expressive arm movement during play. For these reasons, we adopted the wrist placement. However, multi-position comparisons are reported in works like (Ahmadi et al., 2020a), which evaluated wrist, hip, and thigh placement in children with CP, whereas Letts et al. (2024) and Goodlich et al. (2020) explored hybrid configurations, and Mathew et al. (2024) emphasized wearable system design for maximal coverage with minimal discomfort. Variability is also evident in data collection protocols, ranging from highly structured trials (Ahmadi et al. 2020d; Trost et al. 2012, 2016) to semi-structured play and free-living observations (Ahmadi et al., 2020e; Al-Sowi et al., 2023; Chowdhury et al., 2018) reflecting a trade-off between ecological validity and modelling complexity. Unlike most prior studies that employ ML supervised classification (not in children) to recognize predefined activity types (Ahmadi et al., 2020b, 2020c, 2020d, 2020e; de Vries et al., 2011; Trost et al., 2012), or to estimate energy expenditure from accelerometer data (Ahmadi et al., 2020b; Chowdhury et al., 2018; Zheng et al., 2013), our approach builds an outcome-driven model that predicts a clinically validated functional scale (the AHA) for children with CP.

Furthermore, while traditional models often operate on features extracted from structured tasks, our method handles noisy, long-term multivariate time-series recorded in free-living conditions. For instance, Trost et al. (2016) developed and tested decision tree models to classify physical activity intensity on structured 7 sequences of activities in children with CP, whereas Mathew et al. (2024) studied semi-structured upper limb activities for only 5 min of free play of 10 children with CP. Similarly, Chowdhury et al. (2017) demonstrated that ensemble learning improves activity recognition from wrist acceleration; in contrast, our study uses ensemble learning as part of a stacked regression architecture for clinical score estimation, emphasizing interpretability and functional relevance rather than classification robustness alone. In contrast to the relevant literature, by combining unsupervised time-series clustering with a regression-based ensemble models, the DAB offers a personalized, interpretable, and scalable alternative for continuous remote assessment which is something rarely addressed in ML studies for pediatric neurorehabilitation, including CP-focused like the above mentioned Ahmadi et al. (2020a) or Trost et al. (2016), where the focus was to classify children according to 3 different macro-categories of movements.

Overall, our models demonstrated strong discriminative power between children with TD and UCP, achieving top F1 scores above 0.90 and a significant correlation with AHA scores (up to $\rho = 0.826$). This aligns with prior observations suggesting that activity asymmetries captured through wrist-worn sensors correlate with unilateral motor impairment severity (Beani et al., 2019; Burgess et al., 2024). Furthermore, the regression step, which transforms regression outputs into DAB scores, yielded an average $R^2 = 0.709$, validating the clinical relevance of the predicted values. Capturing variation in spontaneous bimanual use, the DAB serves as a proxy for functional capability, thus showing a capacity of reflecting real-life motor behaviour and enabling longitudinal and context-aware assessment, highlighting patterns, variations, and changes that traditional assessments might miss. Our novel biomarker offers several advantages over existing UCP clinical scores: unlike the AHA score, which requires clinic visits and certified and experienced evaluators, the DAB can be measured anywhere; it supports pointwise or continuous monitoring over time windows of any size. It is objective, in contrast to the MACS level based on clinician-parent agreement; it is more easily interpretable, using the same scale as the AHA score, rather than a derived asymmetry index.

Nonetheless, certain limitations must be acknowledged. Our approach requires wearing actigraphs for long periods of time by reducing compliance in real-world deployments and reliance on wrist-mounted accelerometers that may miss subtle hand or finger movements. In addition, a potential source of variability in our measurements arises from the placement and attachment tightness of the sensors, which were applied by children or caregivers at home. Although instructions were provided to standardize sensor positioning, some variability due to sensor placement is unavoidable in unsupervised daily-life conditions. Thus, in this study, we collapsed the tri-axial accelerometer data into a single magnitude to obtain a robust, orientation-invariant measure of movement intensity per wrist. While this approach is well-suited to our goal of quantifying asymmetry, it inevitably discards directional information. As future work, analysing the full three-axis time series could provide richer characterisations of movement patterns, potentially enhancing asymmetry detection and enabling the extraction of more fine-grained features. Additionally, processing raw accelerometer and gyroscope signals—rather than relying solely on aggregated data—could offer deeper insight into the indicator's robustness and generalizability across different sensing conditions. Lastly, the proposed DAB should be further validated by using a wider range of IMU-based hardware to assess its cross-platform reliability. Another important aspect to improve is to increase the discrimination power of the DAB between subject with TD (AHA evaluated at 100) and UCP with high scores of AHA (very close to 100): an improvement on this aspect would increase even more the impact of such a digital biomarker in the children neurodevelopmental clinical practice. Moreover, validation on broader populations and evaluation of responsiveness to rehabilitation are needed. In summary, the DAB represents a significant step toward integrating wearable sensor data and ML into pediatric neurorehabilitation. By enabling fine-grained, real-world monitoring of functional use, it has the potential to support clinicians in personalized treatment planning, facilitate earlier detection of progress or regress, and ultimately improve the quality of care for children with UCP.

6 Conclusions and Future Work

The presented results successfully demonstrated the feasibility of using ML techniques to analyze time series data from wearable sensors in children with UCP in a daily life environment. The developed pipeline showed high accuracy in differentiating between children with UCP and TD, and the Daily AHA Biomarker we introduced showed a very good relation to AHA scores, meaning that it is also applicable to continuous monitoring of children. These findings highlight the potential of wearable technology and ML to enhance clinical assessments and provide continuous monitoring, ultimately improving the management and care of children with UCP.

The method we presented is by all means an ensemble method: we use multiple ML models to classify time series samples, and their output becomes the features of the regression model. This approach combines the advantages of each model type, reducing possible bias.

The indicator introduced in this paper is intended to be used to provide clinical support, and not independently by patients. It is an additional instrument for the clinician to evaluate children and should not substitute the clinical staff who continues to make clinically relevant decisions. However, its utility must be proved in clinical tests before it can be used in general practice. In preparation for this, we have included it in a Dashboard that was built

Fig. 6 Integration of DAB into the developed visualization Dashboard

to collect patient data, which is available to doctors at the IRCCS Stella Maris Foundation in Italy (see snapshot in Fig. 6).

Beyond application to the clinical setting, future work will investigate further models, to be used either in ensemble with the current ones, or separately. In particular, we plan to explore the use of deep learning models such as LSTMs to extract a representation of the time series data. For this, however, further data is necessary, as the current one is quite limited in size, and would result in model underfitting when using models with a large number of parameters. We plan to collect data from additional patients and ideally combine it with public actigraph data. Public data could be used either in a self-supervised training phase or by using transfer learning.

In conclusion, the proposed approach enables remote, interpretable, and scalable monitoring of upper limb function and has the potential to support more personalized rehabilitation strategies. Our findings address the need for a more ecological and dynamic assessment, enriching the clinical assessment beyond the traditional clinical environments and moving further to the quite exclusive use of parent-reported questionnaires for daily life assessment. This methodology will increase family awareness, helping them become active participants in the assessment, monitoring, and rehabilitation process, thus fostering a collaborative approach between clinicians and caregivers. DAB will also provide clinicians with valuable information for the detection of spontaneous upper limb use during daily life, crucial for tailoring interventions. Overall, the DAB represents a promising step towards integrating wearable technology and machine learning into routine pediatric neurorehabilitation.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10994-025-06950-7.

Acknowledgements This work was partially supported by the project Tuscany Health Ecosystem (THE), ECS00000017, Spoke 3, CUP: B83C22003920001 (European Union - Next Generation EU, in the context of The National Recovery and Resilience Plan, Investment 1.5 Ecosystems of Innovation); by the AInCP Project that has received funding from the European Commission, Horizon Europe Research and Innovation Action under GA n. 101057309; by the Università di Pisa under the "PRA—Progetti di Ricerca di Ateneo" - Project no. PRA_2022_81 and through the SPARK project; by the project PRIN PNRR Favorite, P20228T3EE; by the Italian Ministry of Health, RC2025; by the Italian National Group for Scientific Computation GNCS-INdAM; by the European Union - Next Generation EU, in the context of The National Recovery and Resilience Plan, Mission 4 Component 2, Investment 1.1, Call PRIN 2022 D.D. 104 02-02-

2022—MEDICA Project, CUP I53D23003720006; by the European Union—Horizon 2020 Program under the scheme "INFRAIA-01-2018-2019—Integrating Activities for Advanced Communities", project "SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics", Grant Agreement n. 871042.

Author Contributions All authors designed the study; S.F., G.P., A.S., D.M., G.Sc. developed the methodology of the study; S.F., G.P., A.S. prepared the first draft of the manuscript; S.F.,E.B.,G.Sg. were responsible for the data acquisition; S.F., G.P., A.S., D.M., G.Sc. were responsible for the data curation and analysis; E.B.,G.Sg. were responsible for the clinical assessment scoring; G.P.,G.Sg. were project administrator. All the authors have reviewed and agreed to the current version of the manuscript.

Data Availability The analysis pipeline is available at: https://github.com/giordanoscerra/AInCP-ML. Ethi cs approval was obtained from the Tuscany Paediatric Ethics Committee, Italy (78/2016). This study was subsequently registered on clinicaltrials.gov (NCT03054441). Data are publicly available at https://doi.org/10.5281/zenodo.15019553 at request.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Ahmadi, M. N., Chowdhury, A., Pavey, T., & Trost, S. G. (2020a). Laboratory-based and free-living algorithms for energy expenditure estimation in preschool children: A free-living evaluation. *PLOS ONE, 15*(5), 1–14. https://doi.org/10.1371/journal.pone.0233229

Ahmadi, M. N., Brookes, D., Chowdhury, A., Pavey, T., & Trost, S. G. (2020b). Free-living evaluation of laboratory-based activity classifiers in preschoolers. *Medicine & Science in Sports & Exercise, 52*(5), 1227–1234.

Ahmadi, M. N., O'Neil, M. E., Baque, E., Boyd, R. N., & Trost, S. G. (2020c). Machine learning to quantify physical activity in children with cerebral palsy: Comparison of group, group-personalized, and fully-personalized activity classification models. *Sensors, 20*(14), 3976. https://doi.org/10.3390/s20143976

Ahmadi, M., O'Neil, M., Fragala-Pinkham, M., Lennon, N., & Trost, S. (2018). Machine learning algorithms for activity recognition in ambulant children and adolescents with cerebral palsy. *Journal of NeuroEngineering and Rehabilitation, 15*(1), 105. https://doi.org/10.1186/s12984-018-0456-x

Ahmadi, M. N., Pavey, T. G., & Trost, S. G. (2020d). Machine learning models for classifying physical activity in free-living preschool children. *Sensors, 20*(16), 4364. https://doi.org/10.3390/s20164364

Ahmadi, M. N., Pfeiffer, K. A., & Trost, S. G. (2020e). Physical activity classification in youth using raw accelerometer data from the hip. *Measurement in Physical Education and Exercise Science, 24*(2), 129–136. https://doi.org/10.1080/1091367X.2020.1716768

Ahmadi, M. N., & Trost, S. G. (2022). Device-based measurement of physical activity in pre-schoolers: Comparison of machine learning and cut point methods. *PLoS One, 17*(4), Article e0266970. https://doi.org/10.1371/journal.pone.0266970

Al-Sowi, A. M., AlMasri, N., Hammo, B., & Al-Qwaqzeh, F.A.-Z. (2023). Cerebral palsy classification based on multi-feature analysis using machine learning. *Informatics in Medicine Unlocked, 37*, Article 101197. https://doi.org/10.1016/j.imu.2023.101197

Au, W. W., Recchia, F., Fong, D. Y., Wong, S. H. S., Chan, D. K. C., Capio, C. M., Yu, C. C. W., Wong, S. W. S., Sit, C. H. P., Ip, P., Chen, Y.-J., Thompson, W. R., & Siu, P. M. (2024). Effect of wearable activity trackers on physical activity in children and adolescents: a systematic review and meta-analysis. *The Lancet Digital Health, 6*(9), 625–639. https://doi.org/10.1016/S2589-7500(24)00139-0

Beani, E., Cavalieri, M. F., Filogna, S., Barzacchi, V., Cianchetti, M., Maselli, M., Martini, G., Menici, V., Prencipe, G., Sicola, E., Cioni, G., & Sgandurra, G. (2025). Wearable sensors for measuring spontaneous upper limb use in children with unilateral cerebral palsy and typical development. *Journal of NeuroEngineering and Rehabilitation, 22*(1), 71. https://doi.org/10.1186/s12984-025-01601-3

Beani, E., Maselli, M., Sicola, E., Perazza, S., Cecchi, F., Dario, P., Braito, I., Boyd, R., Cioni, G., & Sgandurra, G. (2019). Actigraph assessment for measuring upper limb activity in unilateral cerebral palsy. *Journal of NeuroEngineering and Rehabilitation, 16*(1), 30. https://doi.org/10.1186/s12984-019-0499-7

Braito, I., Maselli, M., Sgandurra, G., Inguaggiato, E., Beani, E., Cecchi, F., Cioni, G., & Boyd, R. (2018). Assessment of upper limb use in children with typical development and neurodevelopmental disorders by inertial sensors: a systematic review. *Journal of NeuroEngineering and Rehabilitation, 15*(1), 94. https://doi.org/10.1186/s12984-018-0447-y

Burgess, A., Oftedal, S., Boyd, R. N., Reedman, S., Trost, S. G., Ware, R. S., & Sakzewski, L. (2024). Construct validity of the both hands assessment using wrist-worn accelerometers. *Physical & Occupational Therapy In Pediatrics, 44*(1), 42–55. https://doi.org/10.1080/01942638.2023.2207635

Chowdhury, A. K., Tjondronegoro, D., Chandran, V., & Trost, S. G. (2017). Ensemble methods for classification of physical activities from wrist accelerometry. *Medicine & Science in Sports & Exercise, 49*(9), 1965–1973. https://doi.org/10.1249/MSS.0000000000001291

Chowdhury, A. K., Tjondronegoro, D., Zhang, J., Hagenbuchner, M., Cliff, D., & Trost, S. G. (2018). Deep learning for energy expenditure prediction in pre-school children. In *IEEE conference on biomedical and health informatics (BHI2018)*. https://eprints.qut.edu.au/116767/

Coyle-Asbil, H. J., Burk, L., Brandes, M., Brandes, B., Buck, C., Wright, M. N., & Vallis, L. A. (2024). Energy expenditure prediction in preschool children: A machine learning approach using accelerometry and external validation. *Physiological Measurement, 45*(9), Article 095015.

de Vries, S. I., Engels, M., & Garre, F. G. (2011). Identification of children's activity type with accelerometer-based neural networks. *Medicine & Science in Sports & Exercise, 43*(10), 1994–199.

Eliasson, A.-C., Krumlinde-Sundholm, L., Rösblad, B., Beckung, E., Arner, M., Ohrvall, A.-M., & Rosenbaum, P. (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), 549–554. https://doi.org/10.1017/s0012162206001162

Elvrum, A.-K.G., Zethræus, B.-M., Vik, T., & Krumlinde-Sundholm, L. (2018). Development and validation of the both hands assessment for children with bilateral cerebral palsy. *Physical & Occupational Therapy In Pediatrics, 38*(2), 113–126. https://doi.org/10.1080/01942638.2017.1318431

Filogna, S., Marchi, D., Prencipe, G., Scerra, G., & Sirbu, A. DAB: A new digital biomarker to monitor everyday upper limb use in children with unilateral cerebral palsy. https://github.com/giordanoscerra/AInCP-ML

González Barral, C., & Servais, L. (2025). Wearable sensors in paediatric neurology. *Developmental Medicine & Child Neurology*. https://doi.org/10.1111/dmcn.16239

Goodlich, B. I., Armstrong, E. L., Horan, S. A., Baque, E., Carty, C. P., Ahmadi, M. N., & Trost, S. G. (2020). Machine learning to quantify habitual physical activity in children with cerebral palsy. *Developmental Medicine & Child Neurology, 62*(9), 1054–1060. https://doi.org/10.1111/dmcn.14560

Himmelmann, K., McIntyre, S., Goldsmith, S., Smithers-Sheedy, H., & Watson, L. (2018). Epidemiology of cerebral palsy. In F. Miller, S. Bachrach, N. Lennon, & M. O'Neil (Eds.), *Cerebral palsy* (pp. 1–16). Springer. https://doi.org/10.1007/978-3-319-50592-3_9-1

Hoyt, C. R., Brown, S. K., Sherman, S. K., Wood-Smith, M., Van, A. N., Ortega, M., Nguyen, A. L., Lang, C. E., Schlaggar, B. L., & Dosenbach, N. U. F. (2020). Using accelerometry for measurement of motor behavior in children: Relationship of real-world movement to standardized evaluation. *Research in Developmental Disabilities, 96*, Article 103546. https://doi.org/10.1016/j.ridd.2019.103546

Idri, A., Benhar, H., Fernández-Alemán, J. L., & Kadi, I. (2018). A systematic map of medical data preprocessing in knowledge discovery. *Computer Methods and Programs in Biomedicine, 162*, 69–85. https://doi.org/10.1016/j.cmpb.2018.05.007

Kaufman, L., & Rousseeuw, P. J. (1990). *Partitioning around medoids (Program PAM)* (pp. 68–125). Wiley. https://doi.org/10.1002/9780470316801.ch2

Konrad, J., Marrus, N., & Lang, C. E. (2022). A feasibility study of bilateral wrist sensors for measuring motor traits in children with autism. *Perceptual and Motor Skills, 129*(6), 1709–1735. https://doi.org/10.1177/00315125221125275. PMID: 36065830.

Krumlinde-sundholm, L., & Eliasson, A.-C. (2003). Development of the assisting hand assessment: A rasch-built measure intended for children with unilateral upper limb impairments. *Scandinavian Journal of Occupational Therapy, 10*(1), 16–26.

Letts, E., King-Dowling, S., Kwan, M. Y. W., Obeid, J., Cairney, J., & Trost, S. G. (2024). Machine learning derived physical activity in preschool children with developmental coordination disorder. *Developmental Medicine & Child Neurology*. https://doi.org/10.1111/dmcn.16186

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory, 28*(2), 129–137. https://doi.org/10.1109/TIT.1982.1056489

Löning, M., Bagnall, A. J., Ganesh, S., Kazakov, V., Lines, J. & Király, F.J. (2019). sktime: a unified interface for machine learning with time series. CoRR. arXiv preprint arXiv:1909.07872

Mathew, S. P., Dawe, J., Musselman, K. E., Petrevska, M., Zariffa, J., Andrysek, J., & Biddiss, E. (2024). Measuring functional hand use in children with unilateral cerebral palsy using accelerometry and machine learning. *Developmental Medicine and Child Neurology, 66*(10), 1380–1389.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Randall, M. (1999). *Melbourne assessment of unilateral upper limb function. The Test Administration Manual*. Royal Children's Hospital. https://books.google.it/books?id=DINlAAAACAAJ

Ruch, N., Rumo, M., & Mäder, U. (2011). Recognition of activities in children by two uniaxial accelerometers in free-living conditions. *European Journal of Applied Physiology, 111*(8), 1917–1927. https://doi.org/10.1007/s00421-011-1828-0

Sabry, F., Eltaras, T., Labda, W., Alzoubi, K., & Malluhi, Q. (2022). Machine learning for healthcare wearable devices: The big picture. *Journal of Healthcare Engineering, 2022*(1), 4653923. https://doi.org/10.1155/2022/4653923

Schäfer, P. (2015). The boss is concerned with time series classification in the presence of noise. *Data Mining and Knowledge Discovery, 29*(6), 1505–1530. https://doi.org/10.1007/s10618-014-0377-7

Silva, L. V. T. T., Vegas, M., Ricci, N. A., Sá, C. S. C., & Alouche, S. R. (2022). Selecting assessment tools to characterize upper limb function of children with cerebral palsy: A mega-review of systematic reviews. *Developmental Neurorehabilitation, 25*(6), 378–391. https://doi.org/10.1080/17518423.2022.2046656

Sousa, A. C., Ferrinho, S. N., & Travassos, B. (2023). The use of wearable technologies in the assessment of physical activity in preschool- and school-age youth: Systematic review. *International Journal of Environmental Research and Public Health, 20*(4), 3402. https://doi.org/10.3390/ijerph20043402

Trost, S. G., Fragala-Pinkham, M., Lennon, N., & O'Neil, M. E. (2016). Decision trees for detection of activity intensity in youth with cerebral palsy. *Medicine & Science in Sports & Exercise, 48*(5), 958.

Trost, S. G., Wong, W.-K., Pfeiffer, K. A., & Zheng, Y. (2012). Artificial neural networks to predict activity type and energy expenditure in youth. *Medicine & Science in Sports & Exercise, 44*(9), 1801.

wGT3X-BT, A.: ActiGraph wGT3X-BT. https://theactigraph.com/actigraph-wgt3x-bt

Zhao, J., & Itti, L. (2018). shapedtw: Shape dynamic time warping. *Pattern Recognition, 74*, 171–184. https://doi.org/10.1016/j.patcog.2017.09.020

Zheng, Y., Won, W.-K., Guan, X., & Trost, S. (2013). Physical activity recognition from accelerometer data using a multi-scale ensemble method. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 27(2), pp. 1575–1581). https://doi.org/10.1609/aaai.v27i2.18997

Authors and Affiliations

Silvia Filogna[1] · Giuseppe Prencipe[2] · Alina Sîrbu[4] · Elena Beani[1,3] · Davide Marchi[2] · Giordano Scerra[2] · Giuseppina Sgandurra[1,3]

✉ Giuseppe Prencipe
giuseppe.prencipe@unipi.it

✉ Alina Sîrbu
alina.sirbu@unibo.it

Silvia Filogna
silvia.filogna@fsm.unipi.it

Elena Beani
elena.beani@unipi.it; elena.beani@fsm.unipi.it

Davide Marchi
d.marchi5@studenti.unipi.it

Giordano Scerra
g.scerra1@studenti.unipi.it

Giuseppina Sgandurra
giuseppina.sgandurra@unipi.it; giuseppina.sgandurra@fsm.unipi.it

[1] Department of Developmental Neuroscience, IRCCS Fondazione Stella Maris, Calambrone, Italy

[2] Department of Computer Science, University of Pisa, Pisa, Italy

[3] Department of Clinical and Experimental Medicine, University of Pisa, Pisa, Italy

[4] Department of Computer Science and Engineering, University of Bologna, Bologna, Italy