PLOS DIGITAL HEALTH

RESEARCH ARTICLE

PrimSeq: A deep learning-based pipeline to quantitate rehabilitation training

Avinash Parnandi[1☯], Aakash Kaku[2☯], Anita Venkatesan[1], Natasha Pandit[1], Audre Wirtanen[1], Haresh Rajamohan[2], Kannan Venkataramanan[2], Dawn Nilsen[3], Carlos Fernandez-Granda[2,4‡*], Heidi Schambra[1,5,6‡*]

1 Department of Neurology, New York University Langone Health, New York, United States of America, **2** Center for Data Science, New York University, New York, United States of America, **3** Department of Rehabilitation and Regenerative Medicine, Columbia University, New York, United States of America, **4** Courant Institute of Mathematical Sciences, New York University, New York, United States of America, **5** Department of Rehabilitation Medicine, New York University Langone Health, New York, United States of America, **6** Neuroscience Institute, New York University Langone Health, New York, United States of America

☯ These authors contributed equally to this work.
‡ These authors jointly supervised this work.
* cfgranda@cims.nyu.edu (CF-G); heidi.schambra@nyulangone.org (HS)

Citation: Parnandi A, Kaku A, Venkatesan A, Pandit N, Wirtanen A, Rajamohan H, et al. (2022) PrimSeq: A deep learning-based pipeline to quantitate rehabilitation training. PLOS Digit Health 1(6): e0000044. https://doi.org/10.1371/journal.pdig.0000044

Editor: Matthew Chua Chin Heng, National University of Singapore, SINGAPORE

Received: January 31, 2022

Accepted: April 12, 2022

Published: June 16, 2022

Data Availability Statement: The data and the code are now publicly available at SimTK (https://simtk.org/projects/primseq) and Github (https://github.com/aakashrkaku/seq2seq_hrar), respectively.

Funding: This work was funded by the American Heart Association/Amazon Web Service postdoctoral fellowship 19AMTG35210398 (A.P.), NIH R01 LM013316 (C.F.G., H.S.), NIH K02 NS104207 (H.S.), NIH NCATS UL1TR001445 (H. S.), and NSF NRT-HDR 1922658 (A.K., C.F.G.). The

Abstract

Stroke rehabilitation seeks to accelerate motor recovery by training functional activities, but may have minimal impact because of insufficient training doses. In animals, training hundreds of functional motions in the first weeks after stroke can substantially boost upper extremity recovery. The optimal quantity of functional motions to boost recovery in humans is currently unknown, however, because no practical tools exist to measure them during rehabilitation training. Here, we present PrimSeq, a pipeline to classify and count functional motions trained in stroke rehabilitation. Our approach integrates wearable sensors to capture upper-body motion, a deep learning model to predict motion sequences, and an algorithm to tally motions. The trained model accurately decomposes rehabilitation activities into elemental functional motions, outperforming competitive machine learning methods. PrimSeq furthermore quantifies these motions at a fraction of the time and labor costs of human experts. We demonstrate the capabilities of PrimSeq in previously unseen stroke patients with a range of upper extremity motor impairment. We expect that our methodological advances will support the rigorous measurement required for quantitative dosing trials in stroke rehabilitation.

Author summary

Stroke commonly damages motor function in the upper extremity (UE), leading to long-term disability and loss of independence in a majority of individuals. Rehabilitation seeks to restore function by training daily activities, which deliver repeated UE functional motions. The optimal number of functional motions necessary to boost recovery is unknown. This gap stems from the lack of measurement tools to feasibly count functional

funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

motions. We thus developed the PrimSeq pipeline to enable the accurate and rapid counting of building-block functional motions, called primitives. PrimSeq uses wearable sensors to capture rich motion information from the upper body, and custom-built algorithms to detect and count functional primitives in this motion data. We showed that our deep learning algorithm precisely counts functional primitives performed by stroke patients and outperformed other benchmark algorithms. We also showed patients tolerated the wearable sensors and that the approach is 366 times faster at counting primitives than humans. PrimSeq thus provides a precise and practical means of quantifying functional primitives, which promises to advance stroke research and clinical care and to improve the outcomes of individuals with stroke.

Introduction

Most individuals with stroke have persistent motor deficits in an upper extremity (UE) [1], resulting in a significant loss of function and independence in activities of daily living (ADLs) [1–3]. Stroke rehabilitation seeks to promote UE motor recovery and restore UE function. To this end, a major focus of rehabilitation is the repeated practice of ADLs, which are comprised of functional motions made by the UE [4]. Functional motions are goal-directed and purposeful, typically targeting objects in the context of executing an ADL [4,5].

In animal models of rehabilitation, when hundreds of UE functional motions are trained per day in the early weeks after stroke, recovery can be potentiated [6,7]. In humans, the number of UE functional motions needed to boost recovery is less clear, primarily because a feasible method to count UE functional motions does not currently exist. To date, the quantitation of UE functional motions—a critical parameter for establishing the effective rehabilitation dose [8]—has been challenged by the trade-off between pragmatism and precision.

Most rehabilitation researchers and clinicians opt for pragmatism, using time-in-therapy to estimate how much training has occurred [9–15]. Problematically, time metrics do not directly read out motion content or quantity, as the delivery of therapy is not standardized [11]. For a given rehabilitation session, the number of functional motions can vary considerably across subjects and institutions, or training may focus on nonfunctional motions unrelated to ADL execution, such as stretching, weight-bearing, or strengthening exercises [16,17]. Even if a time-based intervention is successful [9,13], its lack of content detail—precisely *what* and *how much* was trained—hinders replication by other researchers or clinicians. Some investigators have sought to automatically detect functional versus nonfunctional UE motion using inertial measurement units (IMUs) and machine learning [18–20], but this approach still outputs time spent in functional motion, rather than the quantity of functional motions trained.

A different pragmatic approach has been to focus on UE joint motions (e.g., forearm rotation, elbow flexion, shoulder flexion) made by stroke patients. This approach uses wearable sensors and deep learning to identify or grade UE joint motions, with the goal of remotely monitoring rehabilitation training or ADL execution at home [21,22]. However, identifying joint displacement does not disentangle functional from nonfunctional UE motions. The identification of functional motions is important, because their repeated practice is key for inducing activity-dependent neuroplasticity, engaging stroke-induced neuroplasticity, and promoting behavioral recovery [6,7,23–26].

Some rehabilitation researchers have instead opted for precision by manually counting functional motions [17,27–30]. This approach meticulously identifies training content and quantity, but has practical challenges. Functional motions are fluid, fast, and difficult to

disambiguate in real time. Using treating therapists to count these motions would be a distraction from clinical remediation, and using independent observers would likely outstrip the financial and labor resources of most institutions. Even with offline video analysis, the identification and counting of functional motions is prohibitively time- and personnel-intensive [5]. The impracticality of manual tallying is thus a major obstacle to its widespread adoption in research and clinical settings.

To overcome these limitations, we seek to directly identify and measure individual functional motions in rehabilitation. We previously reported that rehabilitation activities can be entirely decomposed into elemental function motions, which we call functional primitives [5,31–33]. There are five main classes of functional primitives: *reach* (UE motion to make contact with a target object), *reposition* (UE motion to move into proximity of a target object), *transport* (UE motion to convey a target object in space), *stabilization* (minimal UE motion to hold a target object still), and *idle* (minimal UE motion to stand at the ready near a target object). Much like words in a paragraph, primitives are strung together to execute an activity. Primitives typically have a short duration and one goal, resulting in simple motion phenotypes. These phenotypes are surprisingly consistent even across species, activities, and motor impairment [5,34,35], indicating that it is possible to identify primitives regardless of individual or context. We thus use primitives as units of motion that are readily identifiable, quantifiable, and replicable.

Here, we present the Primitive Sequencing pipeline (PrimSeq), a deep learning-based framework to automatically identify and count functional primitives in rehabilitation training. Inspired by deep learning methods for speech recognition, our approach uses a sequence-to-sequence model to generate sequences of functional primitives, which are then counted [36]. PrimSeq encompasses three main steps: (1) capture of upper body motion during rehabilitation with wearable inertial measurement units (IMUs), (2) generation of primitive sequences from IMU data with the trained deep learning model, and (3) tallying of primitives with a counting algorithm. We developed PrimSeq in chronic stroke patients performing a battery of rehabilitation activities. We show that in previously unseen stroke patients with a range of motor impairment, PrimSeq robustly identifies primitives in various rehabilitation activities and outperforms other activity-recognition methods. IMU-based motion capture is also well tolerated by patients, and PrimSeq considerably diminishes the time and burden of quantifying functional motions.

Results

To collect a variety of functional primitives to train the deep learning model, we recorded and labeled functional motion from 41 chronic stroke patients with UE paresis (Fig 1 and Table 1). While patients performed common rehabilitation activities (Fig 1A and S1 Table), we captured upper-body motion with an array of IMUs and video cameras (Fig 1B). To generate ground truth labels, trained human coders viewed the videotaped activities and segmented them into functional primitives (Fig 1C). The coders annotated the beginning and end of each primitive on the video, which applied a primitive label to the corresponding segment of IMU data. Inter-rater reliability of primitive labeling was high between the coders and an expert (Cohen's $K \geq 0.96$). We split the labeled IMU data into a training set (n = 33 patients; 51,616 primitives; see primitive class distribution in Materials and methods) and an independent test set (n = 8 patients; 12,545 primitives). Data splits were balanced for impairment level and paretic side (Table 1). We note that the number and variety of primitives, not the number of subjects, makes this dataset robust for the development of deep learning approaches. We used the labeled training set for model fitting, parameter adjustment, and hyperparameter tuning. We

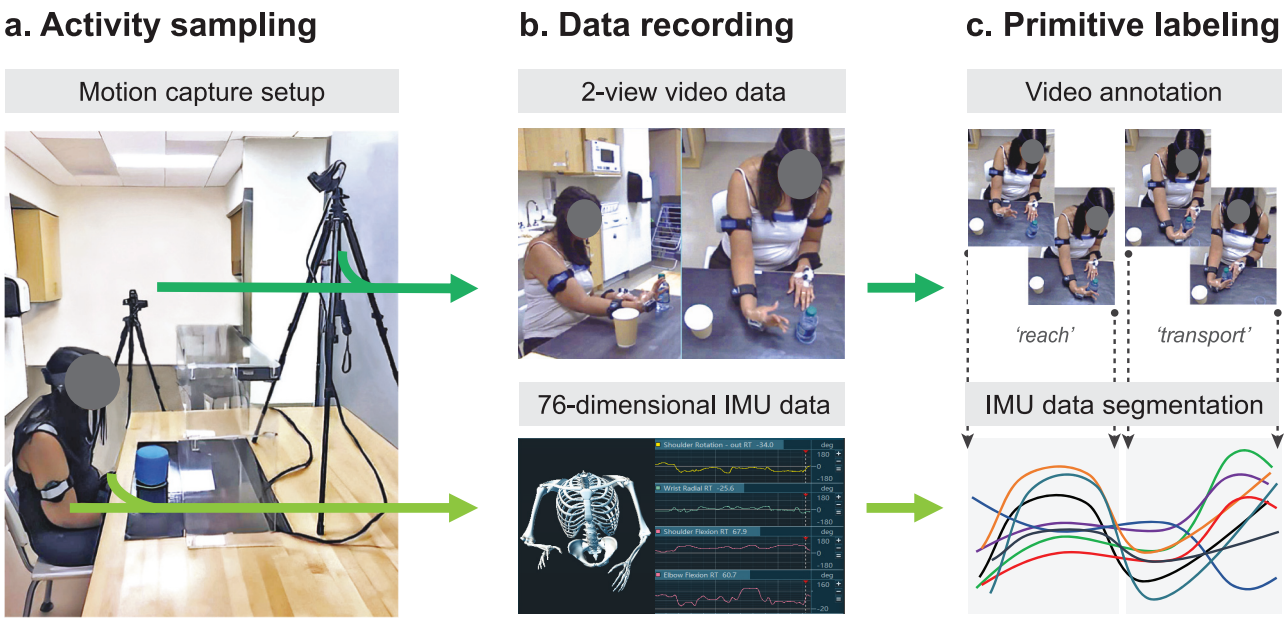

Fig 1. Functional motion capture and labeling. (A) Activity sampling. As patients performed rehabilitation activities, functional motion was synchronously captured with two video cameras (dark green arrow) placed orthogonal to the workspace and nine inertial measurement units (IMUs, light green arrow) affixed to the upper body. **(B) Data recording**. The video cameras generated 2-view, high-resolution data. The IMU system generated 76-dimensional kinematic data (accelerations, quaternions, and joint angles). A skeletal avatar of patient motion and joint angle offsets were monitored for electromagnetic sensor drift. **(C) Primitive labeling**. Trained coders used the video recordings to identify and annotate functional primitives (dotted vertical lines). These annotations labeled and segmented the corresponding IMU data. Interrater reliability was high between the coders and expert (Cohen's K for reach, 0.96; reposition, 0.97; transport, 0.97; stabilization, 0.98; idle, 0.96).

https://doi.org/10.1371/journal.pdig.0000044.g001

used the previously unseen test set for an unbiased model evaluation, employing ground truth labels to assess primitive counting and classification performance.

We designed a sequence-to-sequence (Seq2Seq) encoder-decoder deep learning model that predicts primitives from IMU data patterns. Its encoder module is a three-layer bi-directional Gated Recurrent Unit (GRU) with 3,072 hidden representations. Its decoder module is a single-layer GRU with 6,144 hidden representations. Using this architecture, Seq2Seq maps a window of motion data to a sequence of primitives (see Materials and methods for architecture design and training details).

Table 1. Demographics and clinical characteristics of patients.

	Training set	Test set
Patient n	33	8
Primitive n	51,616	12,545
Age (Years)	56.3 (21.3–84.3)	60.9 (42.6–84.3)
Gender n	18 female: 15 male	4 female: 4 male
Stroke type	30 ischemic: 3 hemorrhagic	8 ischemic
Time since stroke (years)	6.5 (0.3–38.4)	3.1 (0.4–5.7)
Paretic side	18 left: 15 right	4 left: 4 right
UE-FMA score (points)	48.1 (26–65)	49.4 (27–63)

The patient cohort (n = 41) was divided into a training set and test set, with no overlap. Impairment level was measured using the upper extremity Fugl-Meyer Assessment (UE-FMA) score, with a maximum normal score of 66. Total n or average values with ranges are shown.

https://doi.org/10.1371/journal.pdig.0000044.t001

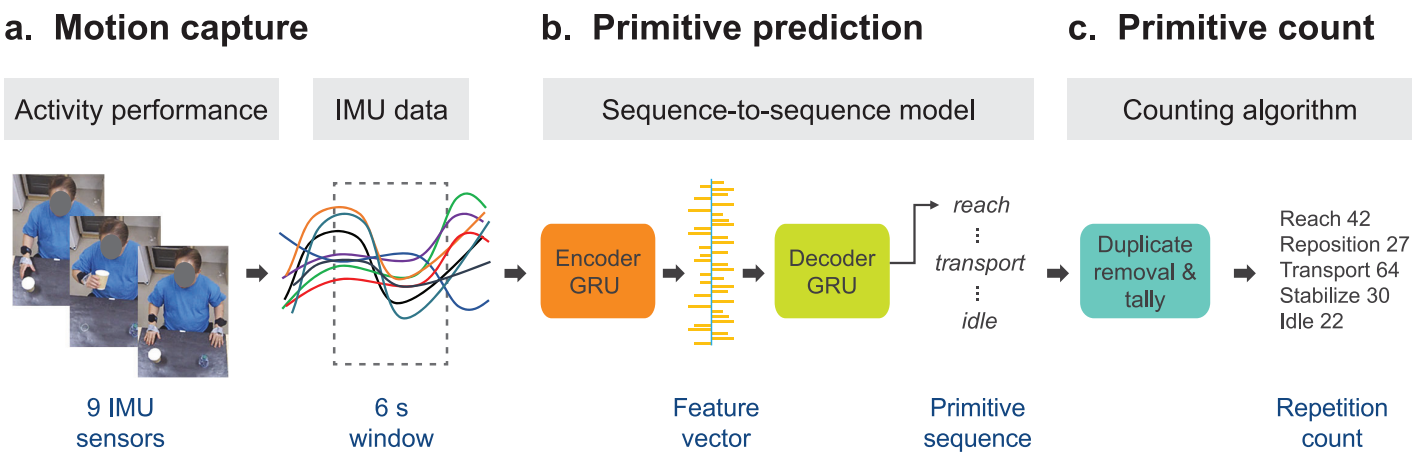

Fig 2. The PrimSeq pipeline. (A) Motion capture. Upper-body motion data are captured with IMUs during performance of rehabilitation activities. IMU data are divided into six-second windows. **(B) Primitive prediction**. The windowed IMU data are fed into the sequence-to sequence (Seq2Seq) deep learning model. Seq2Seq uses a Gated Recurrent Unit (GRU) to sequentially encode IMU data into a feature vector, which provides a condensed representation of relevant motion information. A second GRU then sequentially decodes the feature vector to generate the primitive sequence. **(C) Primitive count**. A counting algorithm then removes primitive duplicates at window boundaries and tallies the predicted primitives.

https://doi.org/10.1371/journal.pdig.0000044.g002

The PrimSeq pipeline has three main steps (Fig 2). First, IMU data from rehabilitation activities are recorded and divided into 6-second windows (Fig 2A). Second, this windowed data is fed to the Seq2Seq model. The encoder GRU generates a single feature vector capturing relevant motion information, which is processed by the decoder GRU to estimate a sequence of primitives (Fig 2B). Third, a counting algorithm removes any duplicates at window boundaries and tallies primitives from the sequences (Fig 2C).

PrimSeq has high counting performance across primitive classes and activities

We first examined the counting performance of PrimSeq (Fig 3). In the separate primitive classes, the approach counted on average 282 reaches, 206 repositions, 408 transports, 291 stabilizations, and 258 idles across combined activities (Fig 3A). In the separate rehabilitation activities, the approach counted on average 40–308 primitives across combined classes (Fig 3B). To assess the similarity of PrimSeq counts to the actual number of primitives performed, we compared predicted versus ground truth counts per primitive class and activity. PrimSeq generated primitive counts that were 86.1–99.6% of true counts for the separate primitive classes (Fig 3A) and 79.1–109.1% of true counts for the separate activities (Fig 3B). We also examined counting errors at the single-subject level, finding that they were consistently low for primitive classes (reach 7.2 ± 9.9%, reposition 13.3 ± 9.2%, transport 0.4 ± 17.7%, stabilization 8.0 ± 42.7%, and idle 6.6 ± 10.5%) and for activities (shelf task -10.7 ± 26.0%, tabletop task -2.3 ± 13.4%, feeding 5.9 ± 16.7%, drinking 5.5 ± 25.0%, combing 11.8 ± 7.8%, donning glasses 14.5 ± 23.7%, applying deodorant 11.5 ± 26.2%, face washing 12.2 ± 30.1%, and tooth-brushing 6.3 ± 31.2%). In most cases, the variance in predicted counts could be attributed to inter-individual variance in true counts. However, for stabilizations, an excess variance in predicted counts could be explained by increased prediction errors, which we further discuss below.

Seq2Seq error examination

PrimSeq generated primitive counts that closely approximate true counts, but gross tallies do not reveal if the Seq2Seq model identified primitives that were actually performed. For

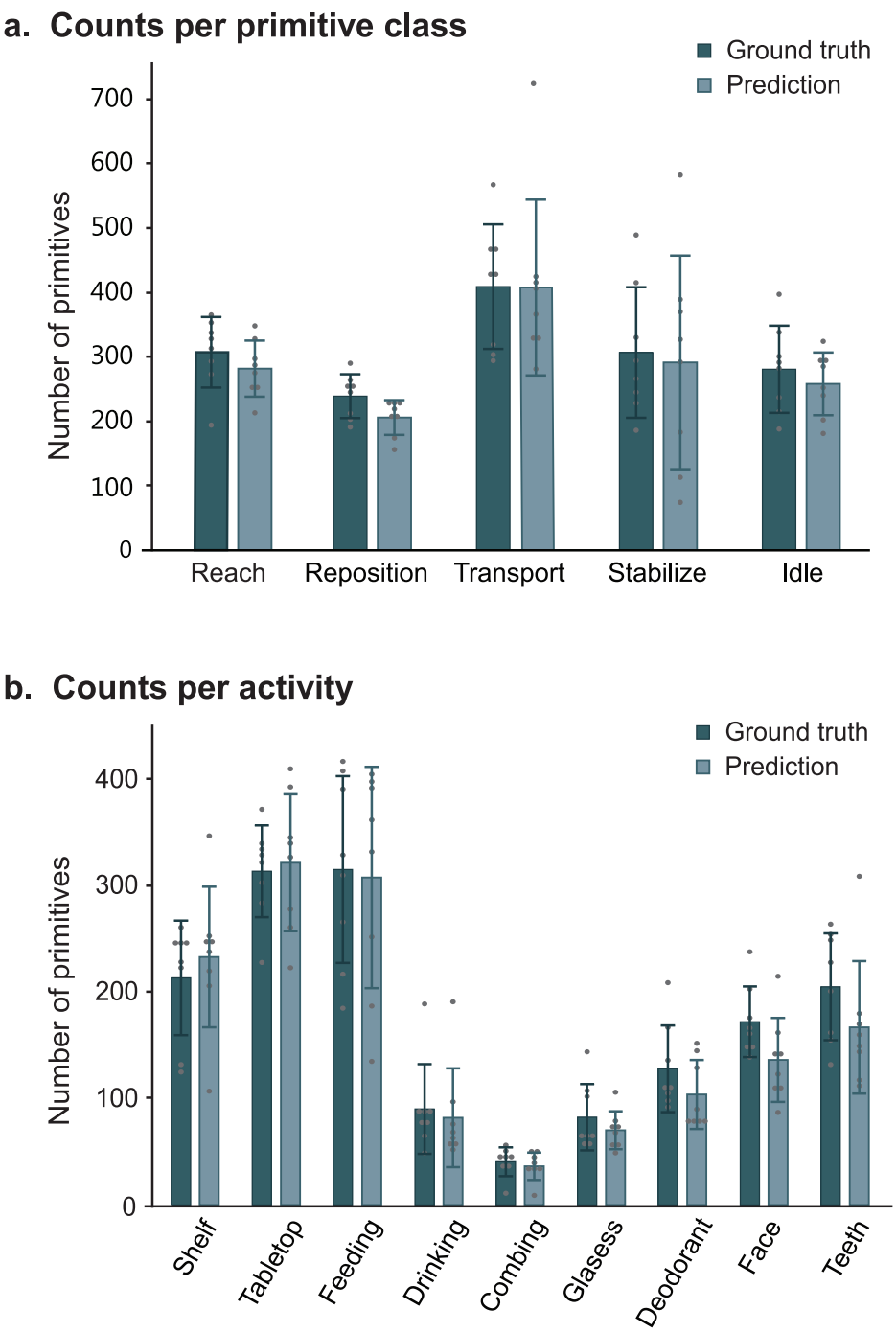

Fig 3. Counting performance of PrimSeq. Shown are the mean ± standard deviation of the ground truth and predicted counts for patients in the test set; each dot represents a single subject. **(A) Counts per primitive class**. Activities were combined. The pipeline generated counts that were similar to true counts for each primitive class (mean percent of true counts: reach, 92.2%; reposition, 86.6%; transport, 99.5%; stabilization, 91.1%; and idle, 93.3%). **(B) Counts per activity**. Primitive classes were combined. The pipeline generated counts that were similar to true counts for each activity (mean percent of true counts: shelf task, 109.1%; tabletop task, 102.5%; feeding, 97.6%; drinking, 93.7%; combing hair, 89.8%; donning glasses, 85.2%; applying deodorant, 81.1%; washing face, 79.1%; and brushing teeth, 81.4%).

https://doi.org/10.1371/journal.pdig.0000044.g003

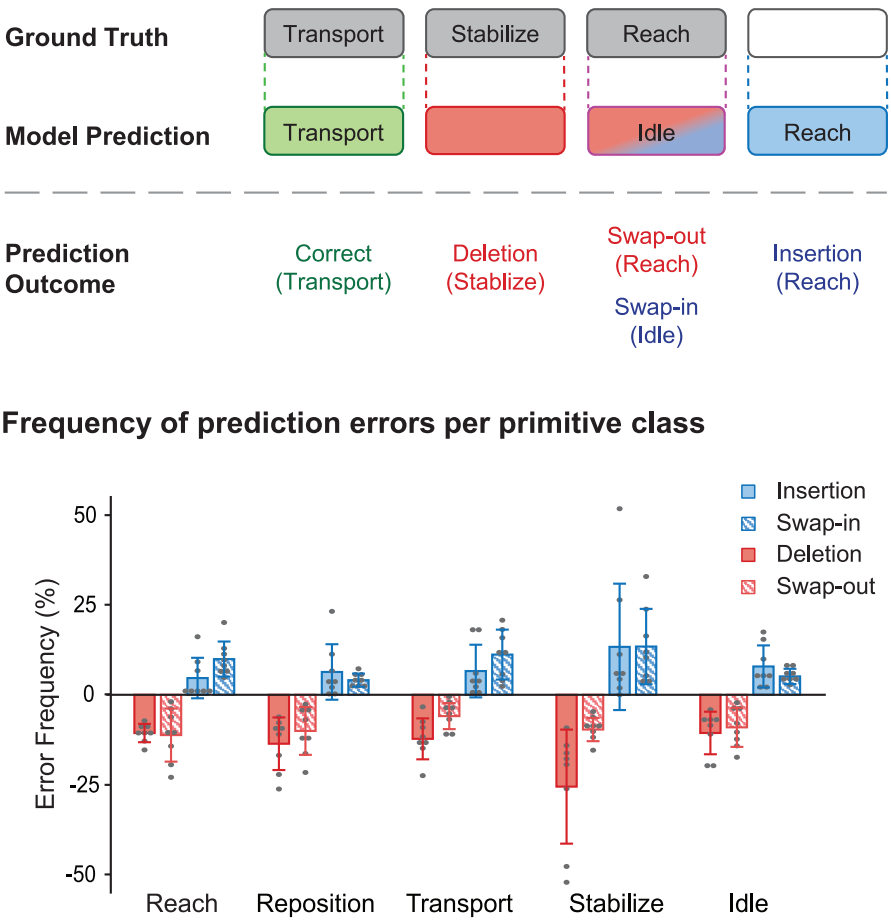

Fig 4. Prediction errors by Seq2Seq. (A) Sequence matching. Shown is a schematic depicting the different types of prediction outcomes, comparing the ground truth sequence (top row) against the predicted sequence (bottom row). Seq2Seq could produce a true positive (correct prediction; green), false negative (deletion or swap-out error; pink), or false positive (insertion or swap-in error; blue). In the example shown, transport was correctly predicted; stabilize was incorrectly deleted; reach was incorrectly swapped-out while idle was incorrectly swapped-in; and reach was incorrectly inserted. **(B) Frequency of prediction errors per primitive class**. Shown are the mean frequency ± standard deviation of prediction errors for patients in the test set; each dot represents a single subject. Activities were combined, and erroneous counts were normalized to ground truth counts in each primitive class. Deletion errors happened when primitives were incorrectly removed from the prediction, and occurred with modestly low frequency, except for stabilizations. Swap-out errors happened when primitives were incorrectly removed from the prediction and instead predicted as another class, and occurred with modestly low frequency. Insertion errors happened when primitives were incorrectly added to the prediction, and occurred with low frequency, except for stabilizations. Swap-in errors happened when primitives were incorrectly predicted instead of the actual primitive class, and occurred with modestly low frequency.

https://doi.org/10.1371/journal.pdig.0000044.g004

example, the model may fail to predict a reach that happened but may later predict a reach that did not happen; the net result of these two errors is that a reach is spuriously credited to the count. We thus examined the nature of predictions made by Seq2Seq (Fig 4). We compared the predicted and ground truth sequences using the Levenshtein algorithm [37], identifying two types of prediction errors: false negatives and false positives (Fig 4A). A false negative occurred when Seq2Seq did not predict a primitive that actually happened because it

erroneously missed the primitive (deletion error) or erroneously predicted another primitive class (swap-out error). A false positive occurred when Seq2Seq predicted a primitive that did not actually happen because it erroneously added the primitive (insertion error) or erroneously predicted this primitive class (swap-in error).

We examined the frequency of these error types with respect to true counts in each primitive class, which adjusts for differences in number of primitives performed (Fig 4B). Seq2Seq had a modest frequency of deletion errors for most primitives (10.6–13.6%) except stabilizations (25.5%) and a modest frequency of swap-out errors (5.9–11.2%) for all primitives. Seq2Seq also had a modest frequency of insertion errors (4.6–13.3%) and swap-in errors (4.0–13.4%) for all primitives.

Overall, Seq2Seq had a tolerable error rate, but had the most difficulty classifying stabilizations. The motion phenotype of stabilizations, which allows for some minimal motion in the UE [5], may account for different classification errors. Seq2Seq could blend the minimal motion of stabilizations into the beginning or end of an adjacent motion-based primitive (i.e., reach, reposition, transport), leading to its deletion. Conversely, the model could mistake periods of diminished motion in adjacent primitives as a stabilization, leading to an insertion or swapping-in. In addition, the model had an increased frequency of swapping-in stabilizations for idles (S1 Fig). This error could be attributed to the lack of IMU finger data necessary to identify grasp, which is a major phenotypic distinction between these two minimal-motion primitives [5].

Seq2Seq classification performance

To assess the overall classification performance of Seq2Seq to predict primitives, we computed sensitivity and false discovery rate (FDR). Sensitivity represents the proportion of true primitives that were correctly predicted and included in the count. The FDR, a measure of overcounting, represents the proportion of predicted primitives that were incorrectly predicted and included in the count. We assessed Seq2Seq classification performance for separate primitive classes, activities, and patient impairment levels (Fig 5).

Seq2Seq classifies most primitives well

We first examined Seq2Seq classification performance for each primitive class (Fig 5A). The model had high mean sensitivities for most primitives (0.76–0.81) except stabilizations (0.64). The model also had low mean FDRs for most primitives (0.11–0.16) except stabilizations (0.23). For stabilizations, distinct prediction errors drove the modest classification performance: their spurious removal (false negatives) decreased sensitivity, whereas their spurious addition (false positives) increased overcounting.

Seq2Seq classifies primitives best in structured functional activities

We next examined Seq2Seq classification performance for each activity (Fig 5B). The model had excellent performance with structured activities, such as moving an object to fixed locations on a shelf (mean sensitivity 0.92, FDR 0.07) and tabletop (mean sensitivity 0.91, FDR 0.06). Its performance declined with more naturalistic activities, such as drinking (mean sensitivity 0.74, FDR 0.16) and feeding (mean sensitivity 0.74, FDR 0.19). Seq2Seq had its lowest performance with the tooth-brushing activity (mean sensitivity 0.62, FDR 0.21).

Seq2Seq performs well for patients with mild to moderate UE impairment

We also examined if Seq2Seq performance was affected by impairment level (Fig 5C). Seq2Seq had a stable sensitivity (0.71–0.82) that did not vary with UE-FMA score (Spearman's

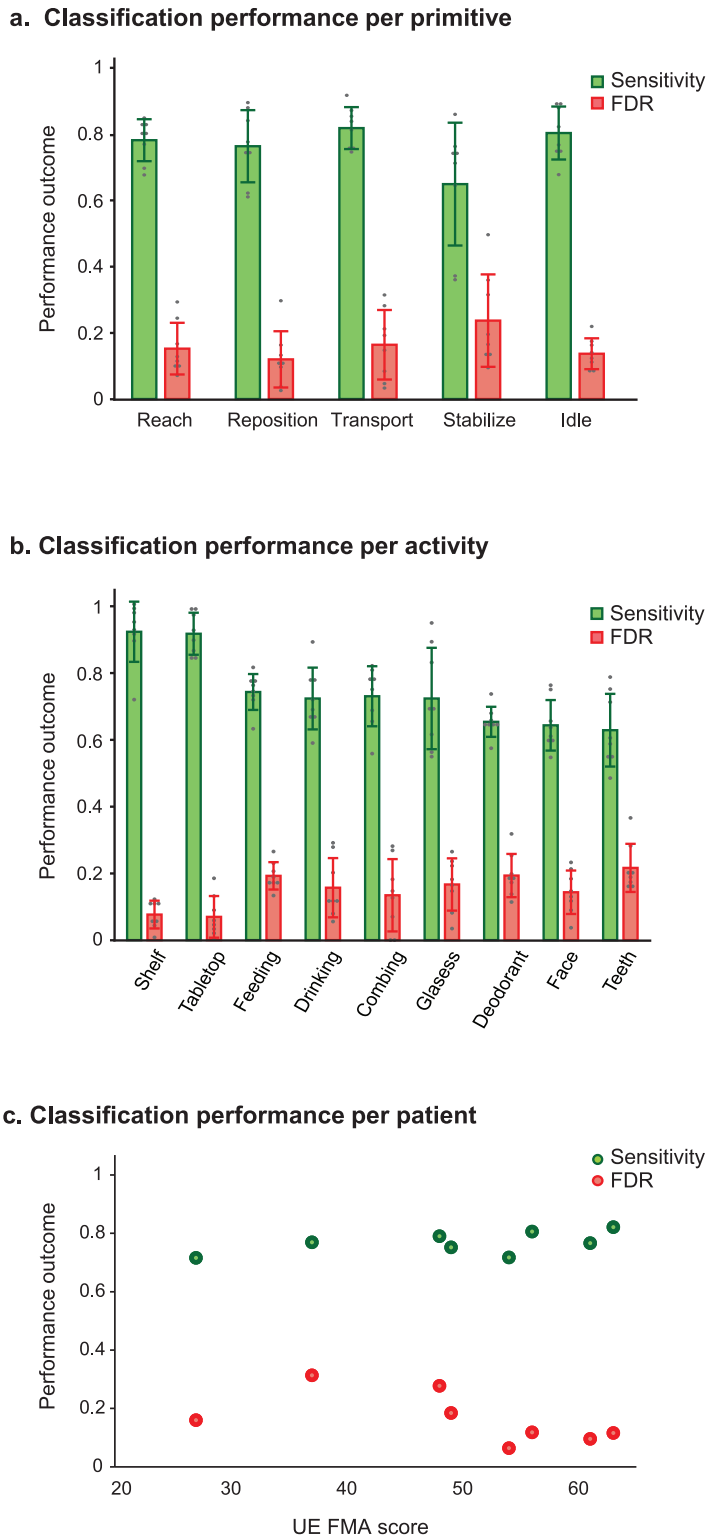

Fig 5. Classification performance of Seq2Seq. (A) Classification performance per primitive class. Prediction outcomes (FP, FN, TP) for activities were combined. Shown are mean ± standard deviation of Seq2Seq sensitivity and FDR for patients in the test set; each dot represents a single subject. Mean sensitivity was high and mean FDR was low for most primitives except stabilizations. **(B) Classification performance per activity**. Prediction outcomes for primitive classes were combined. Shown are mean ± standard deviation of Seq2Seq sensitivity and FDR for patients in the test set; each dot represents a single subject. Mean sensitivity was high and mean FDR was low for structured

activities such as the shelf and tabletop tasks, but were more modest for more complex activities. **(C) Classification performance per patient**. Prediction outcomes for activities and primitives were combined. Shown are sensitivity and FDR values per patient with respect to their upper extremity Fugl-Meyer Assessment (UE-FMA) score. Seq2Seq sensitivity was not affected by impairment level (p = 0.171), but there was a trend for reduced FDRs with higher UE-FMA scores (p = 0.069), driven by one patient.

https://doi.org/10.1371/journal.pdig.0000044.g005

correlation (ρ(6) = 0.54, p = 0.171, 95% confidence interval (CI) [-0.12, 0.87]). The FDR ranged more widely (0.31–0.64) and showed a trend for decreasing as UE-FMA scores increased (ρ(6) = -0.69, p = 0.069, 95% CI [-0.11, -0.91]). This trend was driven by one patient (UE-FMA 37), whose stabilizations and transports were excessively overcounted by Seq2Seq.

Seq2Seq outperforms benchmarks

Finally, we benchmarked Seq2Seq against competitive models used in human action recognition: convolutional neural network (CNN), action segment refinement framework (ASRF), and random forest (RF; Fig 6 and S2 Fig) [31,38,39]. The CNN and RF made predictions at each 10-ms time point, which were smoothed to generate primitive sequences [40]. ASRF is a state-of-the-art action recognition method that directly generates primitive sequences. We aggregated patients, primitive classes, and activities to examine the overall sensitivity and FDR of each model. We also examined the F_1 score, the harmonic mean between sensitivity and precision (1-FDR), which reflects global classification performance. We used bootstrapping and unpaired, two-tailed t-tests to statistically compare the classification performance of Seq2Seq against the other models.

Seq2Seq had a significantly higher sensitivity than the other models (Seq2Seq, 0.767; CNN, 0.727; ASRF, 0.720; RF, 0.497; all $t_{498} > 20.7$, all $p < 0.0001$). Seq2Seq had a significantly lower FDR than CNN and RF (Seq2Seq, 0.166; CNN, 0.196; RF, 0.213; all $t_{498} > 22.8$, all $p < 0.0001$), but was outperformed by ASRF (0.160; $t_{498} = 7.5$, $p < 0.0001$). Seq2Seq also had a significantly higher F_1 score than the other models (Seq2Seq, 0.799; CNN, 0.763; ASRF, 0.775; RF, 0.609; all

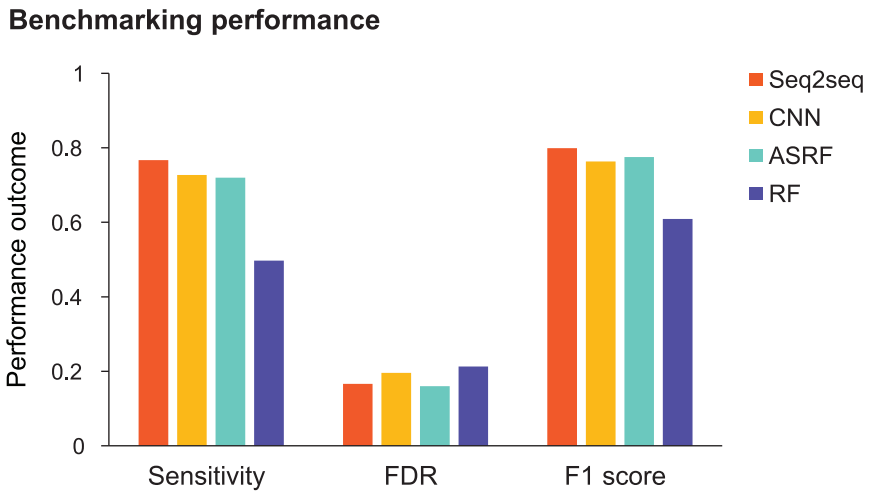

Fig 6. Benchmarking Seq2Seq performance. Prediction outcomes were aggregated from primitives, activities, and patients. Shown are the sensitivity, false discovery rate (FDR), and F1 score for Seq2Seq, convolutional neural network (CNN), action segment refinement framework (ASRF), and random forest (RF). Seq2Seq had the highest sensitivity (Seq2Seq, 0.767; CNN, 0.727; ASRF, 0.720; RF, 0.497). Seq2Seq FDR was lower than CNN and RF but marginally higher than ASRF (Seq2Seq, 0.166; CNN, 0.196; ASRF, 0.160; RF, 0.213). Seq2Seq had the highest F1 score (Seq2Seq, 0.799; CNN, 0.763; ASRF, 0.775; RF, 0.609).

https://doi.org/10.1371/journal.pdig.0000044.g006

$t_{498} > 18.4$, all $p < 0.0001$) indicating its best overall classification performance of the models tested.

PrimSeq provides a practical solution to count functional motion repetitions

Finally, we examined the practicality of using the PrimSeq pipeline. From the motion capture standpoint, most patients reported minimal difficulty with donning and calibrating the IMU array and minimal discomfort with wearing them (median scores 0 and 1, respectively, on an 11-point visual analog scale; S3 Fig). IMU setup time was minimal, requiring on average 12.9 ± 4.1 minutes. Electromagnetic sensor drift was also late to emerge during recording, occurring after 63.7 ± 28.1 minutes, with recalibration taking 1–2 minutes. Collectively, these observations indicate that an IMU system could support motion capture that is well tolerated, efficient, and stable.

From the primitive identification standpoint, the trained Seq2Seq model and counting algorithm identified and tallied primitives 370 times faster than human coders. For example, to process 6.4 hours of recorded activities, Seq2Seq had an execution time of 1.4 hours (13 s per minute of recording) whereas trained human coders required 513.6 hours or 12.8 work-weeks (4,815 s per minute of recording). Furthermore, if inference is made on a high-performing computer with an advanced graphical processing unit, processing lags between incoming IMU data and primitive prediction are estimated at ~15 seconds. These observations indicate that PrimSeq could provide near-immediate feedback about primitive counts, which would help patients and therapists meet training goals during rehabilitation sessions [41].

Discussion

To date, the measurement of functional motions in UE stroke rehabilitation has been an elusive technical challenge, hampered by imprecision or impracticality. To address these obstacles, we developed PrimSeq, an approach that integrates wearable IMU data, a trained deep learning model, and a counting algorithm to quantify functional primitives performed during rehabilitation. We found that PrimSeq generated accurate primitive counts across primitive classes and activities. We also found that the Seq2Seq model also had a moderate-to-high classification performance across primitive classes, rehabilitation activities, and impairment levels, outperforming state-of-the-art action recognition models. Finally, we found that PrimSeq was a practical approach for UE motion capture and processing. These results indicate that PrimSeq provides a pragmatic solution for the accurate identification of motion content and quantity during rehabilitation.

The generalizability of our approach depends in part on the adoption of functional primitives as units of measure. As with any classification in machine learning, users must confirm that the predicted classes are relevant to their question or application. We reason that, for the purpose of objectively measuring rehabilitation training, primitives are reliable and effective units of measure. As singular motion and minimal-motion events, primitives directly transcribe the series of actions taken by individuals to execute more complex activities [5]. Because primitives are circumscribed and mutually exclusive, they avoid lumping together a variable quantity or series of motions under one class designation, as can occur with functional motion detection [18–20] or human activity recognition [42–44]. We have consistently found that rehabilitation activities can be entirely broken down into constituent functional primitives [5, 31, 32]. Primitives thus provide a metric by which rehabilitation training can be consistently measured, enabling the critical appraisal and replication of dosed rehabilitation interventions.

Our approach has some limitations to consider. PrimSeq identifies functional motions but does not measure how normally they are performed. This information is important for tracking recovery and tailoring rehabilitation. Future work could characterize and reference the normative kinematics of the primitive classes, which could generate continuous measurements of abnormal primitive performance. In addition, PrimSeq was trained on motion from chronic stroke patients performing a circumscribed battery of rehabilitation activities. To increase clinical utility, additional model training and refinement could be undertaken "in the wild" with subacute stroke patients undergoing inpatient rehabilitation. A more extensive sampling of primitives could be expected to boost classification performance on unstructured activities, and make PrimSeq robust for application in different recovery stages. Finally, the classification performance of Seq2Seq was limited in some cases (e.g. stabilizations, tooth-brushing activity). Future work could employ alternative deep learning models with explainable artificial intelligence to identify sources of confusion, which could then be targeted to improve classification performance [45, 46]. Alternatively, the motion capture setup could be expanded to generate information that is lacking but may be critical for more precise classification. For example, the grasp of an object is a major phenotypic feature that distinguishes stabilizations (grasp present) from idles (grasp absent) [5], but our current IMU array does not capture finger motion. The incorporation of videography data and computer vision [47–49] or the addition of an instrumented glove [50] could provide grasp data, delivering key motion details that are necessary for classification.

In conclusion, we present a novel pipeline that measures functional motion repetitions in UE rehabilitation activities. PrimSeq is a foundational step toward the precise and pragmatic quantification of rehabilitation dose, and overcomes considerable time, personnel, and financial barriers. Our approach has the potential to support rigorous rehabilitation research and quantitative clinical delivery, which are vitally needed to improve stroke outcomes.

Materials and methods

Subjects

We studied 41 chronic stroke patients with upper extremity (UE) paresis. Patients gave written informed consent to participate. This study was approved by the Institutional Review Board at New York University Langone Health, in accordance with the Declaration of Helsinki. Patient demographics and clinical characteristics are reported in Table 1.

Enrollment criteria

Eligibility was determined by electronic medical records, patient self-report, and physical examination. Patients were included if they were $\geq$ 18 years old, premorbidly right-handed, able to give informed consent, and had unilateral motor stroke with contralateral UE weakness scoring $<$ 5/5 in any major muscle group [51]. Patients were excluded if they had: hemorrhagic stroke with mass effect, or subarachnoid or intraventricular hemorrhage; traumatic brain injury; musculoskeletal, major medical, or non-stroke neurological condition that interferes motor function; contracture at shoulder, elbow, or wrist; moderate UE dysmetria or truncal ataxia; apraxia; visuospatial neglect; global inattention; or legal blindness. Stroke was confirmed by radiographic report. Lesions in non-motor areas or the opposite hemisphere were allowed barring bilateral weakness. Both ischemic and hemorrhagic stroke were included, as motor deficits do not substantially differ between the two types [52]. Stroke patients were chronic ($>$ 6 months post-stroke) except for two patients (3.1 and 4.6 months post-stroke).

Primitive dataset generation

Patients participated in two to three sessions lasting ~2.5 hours. Sessions were typically one to three days apart (average 2.6 days). At the first session, we recorded patient height and measured UE impairment level with the UE Fugl-Meyer Assessment (FMA), where a higher score (maximum 66) indicates less impairment [53]. Patients then performed five trials of nine rehabilitation activities while their upper body motion was recorded (S1 Table and Fig 1A). We identified activities using a standardized manual of occupational therapy (OT) [54]. From these, we identified activities commonly practiced during inpatient stroke rehabilitation through survey of seven OTs with expertise in stroke rehabilitation. Patients were seated in front of a workspace (table or sink counter) at a distance that allowed the nonparetic UE to reach, without trunk flexion, to the furthest target object. Workspaces were adjusted to standard heights (table, 76 cm; counter, 91 cm). We placed the target objects at fixed locations using marked, laminated cardboard mats (table) or measured distances (counter). We used standardized instructions that outlined the major goals of the activity. Because most activities in the battery are bimanual, we instructed patients to use their paretic UE to the best of their ability.

Motion capture

To record patient motion, we affixed nine inertial measurement units (IMUs; Noraxon, USA) to the C7 and T10 spine, pelvis, and both hands using Tegaderm tape (3M, USA) and to both arms and forearms using Velcro straps (Fig 1A). Each IMU is small (length: 3.8 cm; width: 5.2 cm; height: 1.8 cm), lightweight (34 g), and captures 3D linear acceleration, 3D angular velocity, and 3D magnetic heading at 100 Hz. The motion capture software (myoMotion, Noraxon, USA) applies a Kalman filter to the linear accelerations, angular velocities, and magnetic heading to generate 3D unit quaternions for each sensor [55]. We used coordinate transformation matrices to transform the generated quaternions to a sensor-centric framework, which represents the rotation of each sensor around its own axes. The motion capture software also applies a proprietary height-scaled skeletal model to the IMU data to generate 22 anatomical angles of the upper body (S2 Table). The motion capture system thus generates a 76-dimensional dataset every 10 ms consisting of the following: 27 dimensions of accelerations (9 IMUs × 3D accelerations per IMU), 27 dimensions of quaternions (9 IMUs × 3D quaternions per IMU), and 22 joint angles. These data are displayed on a software interface alongside an avatar of the patient (Fig 1B). As an additional feature, we added the side of the patient's paretic UE (left or right) to each 10 ms time step, resulting in a 77-dimensional dataset.

The motion capture system records patient motion with high precision (accelerometry accuracy ± 0.001 g; gyroscopic accuracy ± 1.25˚; anatomical angle accuracy ± 2˚), performing as well as the gold-standard optical system [56]. We monitored online for electromagnetic sensor drift by visually inspecting the joint angle data for baseline shifts and ensuring that avatar motions matched those of the patient. Patients were immediately recalibrated if drift was observed. Recalibration required standing with UEs straight at the sides (arms, forearms, and wrists in neutral position with elbows extended) and took less than two minutes.

We recorded UE motion with two high-speed (60 Hz), high-definition (1088 × 704 resolution) cameras (Ninox, Noraxon) positioned orthogonally less than two meters from the patient (Fig 1A). The cameras have a focal length of f4.0 mm and a large viewing window (length: 2.5 m, width: 2.5 m; Fig 1B). The cameras ran on the same clock as the IMUs and video and IMU recordings were synchronized.

Data labeling

Human coders identified the functional primitives performed in the rehabilitation activities. The five classes of functional primitives are *reach* (UE motion to make contact with a target object), *reposition* (UE motion to move proximate to a target object), *transport* (UE motion to convey a target object in space), *stabilization* (minimal UE motion to hold a target object still), and *idle* (minimal UE motion to stand at the ready near a target object). Coders were trained on a functional motion taxonomy that operationalizes primitive identification [5]. The coders used the video recordings to identify and label the start and end of each primitive, which simultaneously segmented and labeled the synchronously recorded IMU data. To ensure the reliable labeling of primitives, an expert (A.P.) inspected one-third of all coded videos. Interrater reliability between the coders and expert was high, with Cohen's K coefficients ≥ 0.96. Coders took on average 79.8 minutes to annotate one minute of recording.

We split the resulting ground truth dataset of into a training set (n = 33 patients; 51,616 primitives: 9840 reaches, 8028 repositions, 12471 transports, 11445 stabilizations, and 9832 idles) and test set (n = 8 patients; 12,545 primitives: 2510 reaches, 1948 repositions, 3331 transports, 2475 stabilizations, and 2281 idles) to independently train and test the deep learning model. Patient selection was random but constrained to balance impairment level and paretic side. The IMU dataset, including its data splits, are available on https://simtk.org/projects/primseq.

Deep learning model development

Inspired by speech recognition models [57], we used a sequence-to-sequence (Seq2Seq) deep learning model to perform the task of predicting primitive sequences. To handle the higher dimensionality of the IMU data, we increased the model's input nodes to 77 (from 40 for speech) and increased the hidden dimensionality to 3,072 (from 512 for speech). To provide sufficient context of the time series given lower sampling rates of 100 Hz for motion data (versus 16,000 Hz for speech), we expanded the window size of the input data to 6 s (from 10 ms for speech).

The architecture of Seq2Seq has two modules: a feature encoder consisting of a three-layer, bi-directional Gated Recurrent Unit (GRU) with 3,072 hidden representations, and a feature decoder consisting of a single-layer GRU with 6,144 hidden representations.

Seq2Seq performs primitive identification in two steps (Fig 2). The encoder GRU first encodes the data window to generate a 6,144-element feature vector. This step reduces the dimensionality of the high-dimensional IMU data, which enables it to learn relevant features from the IMU data for the downstream task of sequence prediction. The decoder GRU then decodes this feature vector to generate the sequence of primitives. The generated sequence of primitives is then passed through a counting algorithm that tallies the functional motion repetitions while removing duplicate primitives at the window boundaries. Additional model details are presented in recent work [36].

We trained Seq2Seq by minimizing a loss function based on the cross-entropy between the predicted and ground truth primitive sequences using the Adam optimizer [58]. We used a learning rate of $5x10^{-4}$. Because primitive overcounting may lead to accidental under-training of patients in rehabilitation, we prioritized keeping the average false discovery rate (FDR) $<$ 20% while maximizing the average sensitivity during model training. We ensured this balance by stopping the model training early based on the Action Error Rate (AER), computed as the total number of changes needed on the predicted sequence to match the ground truth, normalized to the length of the ground truth sequence.

We used a window size of 6 s for primitive prediction with Seq2Seq. During model training, the middle 4 s of the window was predicted, with the flanking 1 s of data providing the model additional temporal context for prediction. We further maximized the training data by adding a window slide of 0.5 s, which also helped the model learn primitive boundaries. During model testing, the window size was 6 s and middle 4 s of the window was predicted. To enable the flanking during model testing, we set the slide to 4 s. A preliminary experiment was performed with window sizes of 2 s, 4 s, and 8 s. The window size of 6 s resulted in the lowest validation AER.

We selected and cross-validated the hyperparameters for Seq2Seq with four different splits of the training set. In each split, 24 or 25 patients were used for training and 9 or 8 patients were used for validation. We selected the hyperparameters for each of the four models based on their validation AERs. Each split yields a separate model that generates independent prediction probabilities per primitive. The prediction probabilities from the four models were averaged, or ensembled, and the primitive with the highest probability was taken as the Seq2Seq prediction. The ensembled prediction was also fed back into the four models to inform the next prediction in the data window.

After Seq2Seq training and hyperparameter estimations were done on the training set, we applied the trained Seq2Seq model to the test set to assess its counting and classification performance in data from previously unseen patients. The test set was not used for feature selection, preprocessing steps, or parameter tuning. Code implementing the model, including instructions for training and hyperparameter selection and comparisons with other action-recognition methods on benchmark datasets, are available on https://github.com/aakashrkaku/seq2seq_hrar. An overlay of model predictions with respect to ground truth primitives is demonstrated on a patient video (S1 Video).

Analysis of counting performance

To visualize the ability of PrimSeq to correctly count primitives, we tallied the predicted and ground truth primitive counts per subject in the test set. These counts are displayed as means and standard deviations in Fig 3A and 3B. To analyze PrimSeq counting performance for each primitive class, we combined all activities and normalized the predicted counts to ground truth counts. To analyze PrimSeq counting performance for each activity, we combined primitive classes and normalized the predicted counts to ground truth counts. Counting performance is reported as the mean percent and standard deviation of true counts. We also examined counting errors at the single-subject level, because mean tallies may obscure erroneous counting (e.g. an average of under-counts and over-counts would wash out errors). We calculated counting error per subject as the difference between true and predicted counts, normalized to true count. Single-subject counting errors are reported as mean percent and standard deviation.

Analysis of prediction outcomes and error frequency

To examine the nature of predictions that Seq2Seq made per primitive class, we combined activities and compared predicted against ground truth sequences for each patient in the test set. We used the Levenshtein sequence-comparison algorithm to match the predicted and ground truth primitive sequences [37]. This step generated the prediction outcomes of false negative (FN), false positive (FP), or true positive (TP; Fig 4A).

False negatives, or primitives spuriously removed from the prediction, could arise from a deletion error (the model did not predict a primitive that actually happened) or a swap-out error (the model did not predict the actual primitive but instead predicted an incorrect class).

False positives, or primitives spuriously added to the prediction, could arise from an insertion error (the model predicted a primitive that did not actually happen) or a swap-in error (the model predicted an incorrect primitive class instead of the actual primitive). True positives were primitives that were correctly predicted.

We examined the frequency and type of prediction errors (FN, FP) made by Seq2Seq, normalizing prediction errors to ground truth counts to adjust for different quantities of primitives. The frequencies of prediction errors are presented as mean and standard deviation for the test set patients. To further assess which classes of primitives were mistaken for each other by Seq2Seq, we generated a confusion matrix to examine swap-out and swap-in errors (S1 Fig).

Analysis of classification performance

To examine the classification performance of Seq2Seq, we computed the classification performance metrics of sensitivity and false discovery rate (FDR) with respect to primitive class, activity, and impairment level.

Sensitivity, also known as true positive rate or recall, represents the proportion of ground truth primitives that were correctly predicted. It is calculated as:

$$Sensitivity = \frac{TP}{TP + FN}$$

FDR, a type of overcount, represents the proportion of predicted primitives that were incorrectly predicted. It is calculated as:

$$FDR = \frac{FP}{TP + FP}$$

To calculate sensitivity and FDR per primitive class, we combined prediction outcomes (i.e., TP, FN, and FP) from all activities for each test set patient (Fig 5A). To calculate sensitivity and FDR per activity, we combined prediction outcomes from all primitives for each test set patient (Fig 5B). Sensitivity and FDR are reported as means and standard deviations across test set patients.

Finally, to assess classification performance at different levels of UE impairment, we combined prediction outcomes from all primitives and activities for each test set patient and calculated sensitivity and FDR (Fig 5C). We examined if these performance metrics varied with ordinal UE-FMA scores using Spearman's correlation (ρ).

Model benchmarking

We compared Seq2Seq against three benchmark models used in human action recognition: convolutional neural network (CNN), random forest (RF), and action segment refinement framework (ASRF).

We examined the classification performance of a CNN that we previously developed to predict primitives from IMU data [31]. Each layer in the CNN computes linear combinations of outputs of the previous layer, weighted by the coefficients of convolutional filters. The model includes an initial module that helps to map different physical quantities captured by IMU system (e.g., accelerations, joint angles, and quaternions) to a common representation space. The model also uses adaptive feature-normalization to increase the robustness of the model to shifts in the distribution of the data, which can occur when the model is applied to new patients.

We also examined the classification performance of RF, a conventional machine learning model, which has previously been used for human activity recognition [38], including distinguishing functional from nonfunctional motion using wrist-worn sensors [18–20]. RF uses a number of decision trees on randomly selected sub-samples of the dataset to make predictions. We input into the model a set of statistical features for each data dimension, including its mean, maximum, minimum, standard deviation, and root mean square. These features capture useful information for motion identification, such as the energy and variance of the motion.

CNN and RF generate primitive predictions at each 10-ms time point. To generate primitive sequences, we smoothed the pointwise predictions of these models using a weighted running average approach. To perform the smoothing, we used a Kaiser window [40] whose parameters (window size and relative sidelobe attenuation) were selected using the best validation performance.

We also examined the classification performance of ASRF, a state-of-the-art deep learning model for action recognition [39]. ASRF is composed of two CNN modules: a segmentation module and a boundary detection module. The segmentation module performs the pointwise predictions of the primitives, and the boundary detection module detects the boundaries of the primitives. These pointwise primitive predictions are combined with the detected boundaries for smoothing and final sequence generation. During smoothing, the model takes the most frequent pointwise prediction between two detected boundaries as the final prediction for that segment.

To benchmark Seq2Seq against these alternative models, we combined prediction outcomes from all patients, primitives, and activities (confusion matrices are shown in S2 Fig). In addition to calculating each model's overall sensitivity and FDR, we also calculated its F_1 score. The F_1 score, a balance between sensitivity and FDR, captures the global classification performance of a model. The F_1 score ranges between 0 and 1, and a value of 1 indicates perfect classification. It is calculated as:

$$F_1 = 2\left(\frac{sensitivity\,(1-FDR)}{sensitivity+(1-FDR)}\right) = 2\left(\frac{TP}{TP+0.5(FN+FP)}\right)$$

Statistical examination of model performance

We report the sensitivities, FDRs, and F_1 scores for each model on the test set. To examine if the models significantly differed in their classification performance, we bootstrapped the test set, which consisted of 324 single ADL trials aggregated across test subjects. From this test set, we randomly subsampled 81 trials with replacement to create a bootstrap set. We created 250 such bootstrapped sets, which were independently fed into each model to generate a distribution of performance metrics (sensitivities, FDRs, and F_1 scores) for each model. We used unpaired, two-tailed Student's t-tests to compare the performance metrics of Seq2Seq against each model, and used Bonferroni correction for multiple comparisons. We performed all statistical analyses in Python. Significance was set at $\alpha = 0.05$.

Practicality assessment of PrimSeq

To assess the practicality of PrimSeq, we first examined whether patients found the IMUs challenging to wear. We used a visual analogue scale (VAS) ranging from 0 (least) to 10 (most) to examine if donning IMUs (application and calibration) was difficult or wearing IMUs was uncomfortable. VAS scores were obtained at the end of each session and are reported as median and range across patients (S3 Fig). We recorded the time to don and calibrate the

IMUs in a subset of 10 patients, reported as mean and standard deviation. We also recorded the onset time of electromagnetic sensor drift in all patients, reported as mean and standard deviation, and the time needed for recalibration, reported as a range.

Finally, to compare the labeling speed of trained human coders against the trained Seq2Seq model, we recorded how long humans and the model took to label all activities from a subset of 10 patients (6.4 h of recordings). Seq2Seq processed the IMU data on a high-performing computer with an advanced graphical processing unit (GPU, 10 trillion floating-point operations per second, memory bandwidth of 900 GB/s). Total processing times are reported. We estimated processing lags between incoming IMU data and model predictions by summing the time of each interstitial operation: transferring data between the IMUs and myoMotion receiver, calling the API, preprocessing the data (quaternion transformation, z-score normalization), sizing the data windows, predicting the primitives on the GPU, and displaying and storing the data.

Supporting information

S1 Table. Activity battery. These representative rehabilitation activities were used to generate an abundant sample of functional primitives for model training. Activity parameters include the workspace setup, target objects, and instructions to complete each task. The table and counter edges are their anterior edges closest to the patient. Patients could perform the actions within the activity in their preferred order.
(DOCX)

S2 Table. Anatomical upper body angles. The motion capture system (myomotion, Noraxon, USA) used 9 IMUs and a proprietary height-scaled model to generate 22 upper body angles, shown in relation to their joint of origin. ‡Shoulder total flexion is a combination of shoulder flexion/extension and shoulder ad-/abduction. *Thoracic angles are computed between the cervical (C7) and thoracic (T10) vertebrae. †Lumbar angles are computed between the thoracic vertebra and pelvis.
(DOCX)

S1 Fig. Classification performance of Seq2Seq. Shown is a confusion matrix, with values normalized to the ground truth primitive count. The diagonal values represent the sensitivity per primitive, or how often the model correct predicted a primitive that was actually performed. The non-diagonal values represent the identification errors made by Seq2Seq. Rows reflect swap-out errors for ground truth primitives and indicate how often a ground truth primitive was incorrectly predicted as another primitive class. Seq2Seq made modest swap-out errors for all primitives (reach, 0.9–6.2%; reposition, 0.4–4.7%; transport, 0.7–2.0%; stabilizations, 1.1–4.1%; idle, 0.3–4.6%). Columns reflect swap-in errors for predicted primitives and indicate how often an incorrect primitive was predicted instead of the ground truth primitive. Seq2Seq made modest swap-in errors for all primitives (reach, 0.9–4.1%; reposition, 0.3–1.1%; transport, 1.7–6.2%; stabilizations, 1.6–4.7%; idle, 0.4–2.2%). We note that some of the errors made by the model could be explained by the lack of finger information from the IMU setup (e.g. confusion between reaches and transports, idles and stabilizations). These primitives have similar motion phenotypes and are distinguished by grasp onset/amount.
(EPS)

S2 Fig. Classification performance of Seq2Seq, convolutional neural network (CNN), action state representation framework (ASRF), and random forest. Shown are confusion matrices for each model with values normalized to the ground truth primitive count. The diagonal values represent the sensitivity per primitive, or how often the model correctly predicted

a primitive that was actually performed. The non-diagonal values represent the identification errors made by the models. Rows reflect swap-out errors for ground truth primitives and indicate how often a ground truth primitive was incorrectly predicted as another primitive class. Columns reflect swap-in errors for predicted primitives and indicate how often an incorrect primitive was predicted instead of the ground truth primitive. Comparing sensitivities (diagonal values), CNN outperformed Seq2Seq in classifying repositions (80.7% versus 75.4%) but underperformed in classifying the remaining primitives. ASRF outperformed Seq2Seq in classifying reaches (81.6% versus 77.6%) and repositions (79.0% versus 75.4%), but underperformed in classifying the remaining primitives. Random forest underperformed Seq2Seq in classifying all primitives.
(EPS)

S3 Fig. Patient tolerance of IMUs for motion capture. Shown are box plots of the visual analogue scale (VAS) ratings by the patients. We asked two questions of each patient (n = 41) at the end of each session: 'how difficult was the setup (IMU application and calibration)?' and 'how uncomfortable were the IMUs?' Patients scored their responses on an ordinal scale ranging from 0 (not at all) to 10 (most). Most patients reported minimal difficulty with the setup during data collection (median score 0, range 0–8) and minimal discomfort with wearing the IMUs (median score 1, range 0–10), highlighting the unobtrusiveness of IMUs for motion capture. The 25-75th interquartile range (IQR) are shown as the lower and upper limits of the box plots, median values and 1.5*IQR are shown as the green dotted line and error bars respectively, and outliers are shown as black dots.
(EPS)

S1 Video. Visualization of model predictions with respect to ground truth primitives. Shown is a mildly impaired patient performing a combing activity. Human coders used the videotaped activity to identify and label primitives performed by the impaired side (circled); these ground truth labels are shown on the upper right. The trained sequence-to-sequence model used the IMU data to predict primitives performed by the impaired side; these predictions are shown on the upper left.
(AVI)

Acknowledgments

We wish to thank Emily Fokas for assistance with preparing figures, Huizhi Li for software development to visualize model predictions, and the following for their assistance with annotating videos: Ronak Trivedi, Sanya Rastogi, Adisa Velovic, Vivian Zhang, Candace Cameron, Nicole Rezak, Sindhu Avuthu, Chris Yoon, Sirajul Islam, Caprianna Pappalardo, Alexandra Alvarez, Bria Barstch, Tiffany Rivera, and Courtney Nilson. We thank Drs. Jose Torres and Cen Zhang for assistance with identifying stroke patients.

Author Contributions

Conceptualization: Dawn Nilsen, Carlos Fernandez-Granda, Heidi Schambra.

Data curation: Avinash Parnandi, Anita Venkatesan, Natasha Pandit, Audre Wirtanen, Heidi Schambra.

Formal analysis: Avinash Parnandi, Carlos Fernandez-Granda, Heidi Schambra.

Funding acquisition: Carlos Fernandez-Granda, Heidi Schambra.

Investigation: Avinash Parnandi, Aakash Kaku, Haresh Rajamohan, Kannan Venkataramanan, Carlos Fernandez-Granda, Heidi Schambra.

Methodology: Avinash Parnandi, Aakash Kaku, Dawn Nilsen, Carlos Fernandez-Granda, Heidi Schambra.

Project administration: Carlos Fernandez-Granda, Heidi Schambra.

Resources: Carlos Fernandez-Granda, Heidi Schambra.

Software: Avinash Parnandi, Aakash Kaku, Carlos Fernandez-Granda.

Supervision: Carlos Fernandez-Granda, Heidi Schambra.

Validation: Avinash Parnandi, Aakash Kaku, Carlos Fernandez-Granda, Heidi Schambra.

Visualization: Avinash Parnandi, Heidi Schambra.

Writing – original draft: Avinash Parnandi, Heidi Schambra.

Writing – review & editing: Avinash Parnandi, Dawn Nilsen, Carlos Fernandez-Granda, Heidi Schambra.

References

1. Virani SS, Alonso A, Aparicio HJ, Benjamin EJ, Bittencourt MS, Callaway CW, et al. Heart disease and stroke statistics—2021 update: a report from the American Heart Association. Circulation. 2021; 143 (8):e254–e743. https://doi.org/10.1161/CIR.0000000000000950 PMID: 33501848
2. Lawrence ES, Coshall C, Dundas R, Stewart J, Rudd AG, Howard R, et al. Estimates of the prevalence of acute stroke impairments and disability in a multiethnic population. Stroke. 2001; 32(6):1279–84. https://doi.org/10.1161/01.str.32.6.1279 PMID: 11387487.
3. Kwakkel G, Veerbeek JM, van Wegen EE, Wolf SL. Constraint-induced movement therapy after stroke. Lancet Neurol. 2015; 14(2):224–34. https://doi.org/10.1016/S1474-4422(14)70160-7 PMID: 25772900.
4. Gillen G. Upper extremity function and management. In: Gillen G, editor. Stroke rehabilitation: a function-based approach: Elsevier; 2011. p. 218–20.
5. Schambra HM, Parnandi A, Pandit NG, Uddin J, Wirtanen A, Nilsen DM. A Taxonomy of Functional Upper Extremity Motion. Frontiers in Neurology—Neurorehabilitation. 2019; 10:857. Epub 2019/09/05. https://doi.org/10.3389/fneur.2019.00857 PMID: 31481922.
6. Murata Y, Higo N, Oishi T, Yamashita A, Matsuda K, Hayashi M, et al. Effects of motor training on the recovery of manual dexterity after primary motor cortex lesion in macaque monkeys. Journal of Neurophysiology. 2008; 99(2):773–86. Epub 2007/12/21. https://doi.org/10.1152/jn.01001.2007 PMID: 18094104.
7. Jeffers MS, Karthikeyan S, Gomez-Smith M, Gasinzigwa S, Achenbach J, Feiten A, et al. Does Stroke Rehabilitation Really Matter? Part B: An Algorithm for Prescribing an Effective Intensity of Rehabilitation. Neurorehabilitation Neural Repair. 2018; 32(1):73–83. Epub 2018/01/18. https://doi.org/10.1177/1545968317753074 PMID: 29334831.
8. Hayward KS, Churilov L, Dalton EJ, Brodtmann A, Campbell BC, Copland D, et al. Advancing stroke recovery through improved articulation of nonpharmacological intervention dose. Stroke. 2021; 52 (2):761–9. https://doi.org/10.1161/STROKEAHA.120.032496 PMID: 33430635
9. Ward NS, Brander F, Kelly K. Intensive upper limb neurorehabilitation in chronic stroke: outcomes from the Queen Square programme. J Neurol Neurosurg Psychiatry. 2019; 90(5):498–506. Epub 2019/02/17. https://doi.org/10.1136/jnnp-2018-319954 PMID: 30770457.
10. Dromerick A, Lang C, Birkenmeier R, Wagner J, Miller J, Videen T, et al. Very early constraint-induced movement during stroke rehabilitation (VECTORS): a single-center RCT. J Neurology. 2009; 73 (3):195–201. https://doi.org/10.1212/WNL.0b013e3181ab2b27 PMID: 19458319
11. Lohse KR, Pathania A, Wegman R, Boyd LA, Lang CE. On the Reporting of Experimental and Control Therapies in Stroke Rehabilitation Trials: A Systematic Review. Arch Phys Med Rehabil. 2018; 99 (7):1424–32. Epub 2018/02/08. https://doi.org/10.1016/j.apmr.2017.12.024 PMID: 29412168.
12. Lohse KR, Lang CE, Boyd LA. Is more better? Using metadata to explore dose-response relationships in stroke rehabilitation. Stroke. 2014; 45(7):2053–8. https://doi.org/10.1161/STROKEAHA.114.004695 PMID: 24867924.

13. Dromerick AW, Geed S, Barth J, Brady K, Giannetti ML, Mitchell A, et al. Critical Period After Stroke Study (CPASS): A phase II clinical trial testing an optimal time for motor recovery after stroke in humans. Proc Natl Acad Sci U S A. 2021; 118(39). Epub 2021/09/22. https://doi.org/10.1073/pnas.2026676118 PMID: 34544853.

14. Winstein CJ, Wolf SL, Dromerick AW, Lane CJ, Nelsen MA, Lewthwaite R, et al. Effect of a Task-Oriented Rehabilitation Program on Upper Extremity Recovery Following Motor Stroke: The ICARE Randomized Clinical Trial. JAMA. 2016; 315(6):571–81. Epub 2016/02/13. https://doi.org/10.1001/jama.2016.0276 PMID: 26864411.

15. Dawson J, Liu CY, Francisco GE, Cramer SC, Wolf SL, Dixit A, et al. Vagus nerve stimulation paired with rehabilitation for upper limb motor function after ischaemic stroke (VNS-REHAB): a randomised, blinded, pivotal, device trial. The Lancet. 2021; 397(10284):1545–53. https://doi.org/10.1016/S0140-6736(21)00475-X PMID: 33894832

16. Kimberley TJ, Samargia S, Moore LG, Shakya JK, Lang CE. Comparison of amounts and types of practice during rehabilitation for traumatic brain injury and stroke. J Rehabil Res Dev. 2010; 47(9):851–62. https://doi.org/10.1682/jrrd.2010.02.0019 PMID: 21174250

17. Lang CE, Macdonald JR, Reisman DS, Boyd L, Jacobson Kimberley T, Schindler-Ivens SM, et al. Observation of amounts of movement practice provided during stroke rehabilitation. Arch Phys Med Rehabil. 2009; 90(10):1692–8. Epub 2009/10/06. https://doi.org/10.1016/j.apmr.2009.04.005 PMID: 19801058.

18. Lum PS, Shu L, Bochniewicz EM, Tran T, Chang L-C, Barth J, et al. Improving accelerometry-based measurement of functional use of the upper extremity after stroke: Machine learning versus counts threshold method. J Neurorehabilitation neural repair. 2020; 34(12):1078–87. https://doi.org/10.1177/1545968320962483 PMID: 33150830

19. Tran T, Chang L-C, Almubark I, Bochniewicz EM, Shu L, Lum PS, et al., editors. Robust Classification of Functional and Nonfunctional Arm Movement after Stroke Using a Single Wrist-Worn Sensor Device. 2018 IEEE International Conference on Big Data (Big Data); 2018: IEEE.

20. Bochniewicz EM, Emmer G, McLeod A, Barth J, Dromerick AW, Lum P. Measuring Functional Arm Movement after Stroke Using a Single Wrist-Worn Sensor and Machine Learning. J Stroke Cerebrovasc Dis. 2017; 26(12):2880–7. Epub 2017/08/07. https://doi.org/10.1016/j.jstrokecerebrovasdis.2017.07.004 PMID: 28781056.

21. Panwar M, Biswas D, Bajaj H, Jöbges M, Turk R, Maharatna K, et al. Rehab-net: Deep learning framework for arm movement classification using wearable sensors for stroke rehabilitation. IEEE Transactions on Biomedical Engineering. 2019; 66(11):3026–37. https://doi.org/10.1109/TBME.2019.2899927 PMID: 30794162

22. Shen C, Ning X, Zhu Q, Miao S, Lv H. Application and comparison of deep learning approaches for upper limb functionality evaluation based on multi-modal inertial data. Sustainable Computing: Informatics Systems. 2022; 33:100624.

23. Nudo RJ, Milliken GW, Jenkins WM, Merzenich MM. Use-dependent alterations of movement representations in primary motor cortex of adult squirrel monkeys. J Neurosci. 1996; 16(2):785–807. https://doi.org/10.1523/JNEUROSCI.16-02-00785.1996 PMID: 8551360.

24. Nudo RJ. Recovery after brain injury: mechanisms and principles. Front Hum Neurosci. 2013; 7:887. Epub 2014/01/09. https://doi.org/10.3389/fnhum.2013.00887 PMID: 24399951.

25. Overman JJ, Clarkson AN, Wanner IB, Overman WT, Eckstein I, Maguire JL, et al. A role for ephrin-A5 in axonal sprouting, recovery, and activity-dependent plasticity after stroke. Proc Natl Acad Sci U S A. 2012; 109(33):E2230–9. https://doi.org/10.1073/pnas.1204386109 PMID: 22837401.

26. Kim SY, Hsu JE, Husbands LC, Kleim JA, Jones TA. Coordinated Plasticity of Synapses and Astrocytes Underlies Practice-Driven Functional Vicariation in Peri-Infarct Motor Cortex. J Neurosci. 2018; 38(1):93–107. https://doi.org/10.1523/JNEUROSCI.1295-17.2017 PMID: 29133435.

27. Birkenmeier RL, Prager EM, Lang CE. Translating animal doses of task-specific training to people with chronic stroke in 1-hour therapy sessions: a proof-of-concept study. Neurorehabilitation Neural Repair. 2010; 24(7):620–35. Epub 2010/04/29. https://doi.org/10.1177/1545968310361957 PMID: 20424192.

28. Lang CE, MacDonald JR, Gnip C. Counting Repetitions: An Observational Study of Outpatient Therapy for People with Hemiparesis Post-Stroke. J Neurol Phys Ther. 2007; 31(1):3–10. 213736065; https://doi.org/10.1097/01.npt.0000260568.31746.34 PMID: 17419883.

29. Lang CE, Strube MJ, Bland MD, Waddell KJ, Cherry-Allen KM, Nudo RJ, et al. Dose response of task-specific upper limb training in people at least 6 months poststroke: A phase II, single-blind, randomized, controlled trial. Annals of neurology. 2016; 80(3):342–54. https://doi.org/10.1002/ana.24734 PMID: 27447365

30. Waddell KJ, Birkenmeier RL, Moore JL, Hornby TG, Lang CE. Feasibility of high-repetition, task-specific training for individuals with upper-extremity paresis. Am J Occup Ther. 2014; 68(4):444–53. Epub 2014/07/10. https://doi.org/10.5014/ajot.2014.011619 PMID: 25005508

31. Kaku A, Parnandi AR, Venkatesan A, Pandit N, Schambra HM, Fernandez-Granda C. Towards data-driven stroke rehabilitation via wearable sensors and deep learning. Proceedings of machine learning research. 2020; 126:143–71. PMID: 34337420

32. Guerra J, Uddin J, Nilsen D, McLnerney J, Fadoo A, Omofuma IB, et al. Capture, learning, and classification of upper extremity movement primitives in healthy controls and stroke patients. IEEE International Conference on Rehabilitation Robotics. 2017; 2017:547–54. Epub 2017/08/18. https://doi.org/10.1109/ICORR.2017.8009305 PMID: 28813877.

33. Parnandi A, Uddin J, Nilsen D, Schambra H. The pragmatic classification of upper extremity motion in neurological patients: a primer. Frontiers in Neurology—Stroke. 2019; 10:996. https://doi.org/10.3389/fneur.2019.00996 PMID: 31620070

34. Fanti C. Towards automatic discovery of human movemes: California Institute of Technology. http://resolver.caltech.edu/CaltechETD:etd-02262008-172531; 2008.

35. Sumbre G, Fiorito G, Flash T, Hochner B. Neurobiology: motor control of flexible octopus arms. Nature. 2005; 433(7026):595–6. Epub 2005/02/11. https://doi.org/10.1038/433595a PMID: 15703737.

36. Kaku A, Liu K, Parnandi AR, Rajamohan HR, Venkataramanan K, Venkatesan A, et al. Sequence-to-Sequence Modeling for Action Identification at High Temporal Resolution. ArXiv. 2021;arXiv:2111.02521.

37. Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. Soviet physics doklady. 1966; 10(8):707–10.

38. Kwapisz JR, Weiss GM, Moore SA. Activity recognition using cell phone accelerometers. ACM SigKDD Explorations Newsletter. 2011; 12(2):74–82. https://doi.org/10.1145/1964897.1964918

39. Ishikawa Y, Kasai S, Aoki Y, Kataoka H. Alleviating Over-segmentation Errors by Detecting Action Boundaries. 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). 2021:2321–30.

40. Kaiser J, Schafer R. On the use of the I 0-sinh window for spectrum analysis. IEEE Trans Acoust. 1980; 28(1):105–7.

41. Klassen TD, Dukelow SP, Bayley MT, Benavente O, Hill MD, Krassioukov A, et al. Higher doses improve walking recovery during stroke inpatient rehabilitation. Stroke. 2020; 51(9):2639–48. https://doi.org/10.1161/STROKEAHA.120.029245 PMID: 32811378

42. Lemmens RJ, Janssen-Potten YJ, Timmermans AA, Smeets RJ, Seelen HA. Recognizing complex upper extremity activities using body worn sensors. PLoS One. 2015; 10(3):e0118642. https://doi.org/10.1371/journal.pone.0118642 PMID: 25734641.

43. Hassan MM, Huda S, Uddin MZ, Almogren A, Alrubaian M. Human Activity Recognition from Body Sensor Data using Deep Learning. J Med Syst. 2018; 42(6):99. Epub 2018/04/18. https://doi.org/10.1007/s10916-018-0948-z PMID: 29663090.

44. Attal F, Mohammed S, Dedabrishvili M, Chamroukhi F, Oukhellou L, Amirat Y. Physical Human Activity Recognition Using Wearable Sensors. Sensors (Basel). 2015; 15(12):31314–38. Epub 2015/12/23. https://doi.org/10.3390/s151229858 PMID: 26690450.

45. Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang G-Z. XAI—Explainable artificial intelligence. Science Robotics. 2019; 4(37):eaay7120. https://doi.org/10.1126/scirobotics.aay7120 PMID: 33137719

46. Samek W, Wiegand T, Müller K-R. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv 2017; arXiv:1708.08296.

47. Ma M, Marturi N, Li Y, Leonardis A, Stolkin R. Region-sequence based six-stream CNN features for general and fine-grained human action recognition in videos. Pattern Recognition. 2018; 76:506–21. https://doi.org/10.1016/j.patcog.2017.11.026

48. Cao Z. ST, Wei S., Sheikh Y. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017;2017:7291–9.

49. Cao Z. HG, Simon T., Wei S., Sheikh Y. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. arXiv. 2018;arXiv:1812.08008v1

50. Lin B-S, Lee IJ, Yang S-Y, Lo Y-C, Lee J, Chen J-L. Design of an Inertial-Sensor-Based Data Glove for Hand Function Evaluation. Sensors (Basel, Switzerland). 2018; 18(5):1545. https://doi.org/10.3390/s18051545 PMID: 29757261.

51. Medical Research Council of the United Kingdom Aids to Examination of the Peripheral Nervous System. Palo Alto, CA: Pendragon House; 1978.

52. Saulle MF, Schambra HM. Recovery and Rehabilitation after Intracerebral Hemorrhage. Semin Neurol. 2016; 36(3):306–12. https://doi.org/10.1055/s-0036-1581995 PMID: 27214706.

53. Fugl-Meyer AR, Jaasko L, Leyman I, Olsson S, Steglind S. The post-stroke hemiplegic patient. 1. a method for evaluation of physical performance. Scandinavian Journal of Rehabilitation Medicine. 1975; 7(1):13–31. Epub 1975/01/01. PMID: 1135616.

54. Lang CE, Birkenmeier RL. Upper-extremity task-specific training after stroke or disability: A manual for occupational therapy and physical therapy: AOTA Press; 2014.

55. Sabatini AM. Quaternion-based extended Kalman filter for determining orientation by inertial and magnetic sensing. IEEE Trans Biomed Eng. 2006; 53(7):1346–56. https://doi.org/10.1109/TBME.2006.875664 PMID: 16830938

56. Balasubramanian S. Comparison of angle measurements between Vicon and Myomotion systems: Arizona State University; 2013.

57. Chan W, Jaitly N, Le QV, Vinyals O. Listen, attend and spell. arXiv 2015; arXiv:1508.01211v2.

58. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv. 2014;arXiv:1412.6980.