Journal of NeuroEngineering and Rehabilitation

RESEARCH

Open Access

Achieving precision assessment of functional clinical scores for upper extremity using IMU-Based wearable devices and deep learning methods

Weinan Zhou[1,2], Diyang Fu[1,2], Zhiyu Duan[1,2], Jiping Wang[1,2], Linfu Zhou[3] and Liquan Guo[1,2*]

Abstract

Stroke is a serious cerebrovascular disease, and rehabilitation following the acute phase is particularly crucial. Not all rehabilitation outcomes are favorable, highlighting the necessity for personalized rehabilitation. Precision assessment is essential for tailored rehabilitation interventions. Wearable inertial measurement units (IMUs) and deep learning approaches have been effectively employed for motor function prediction. This study aims to use machine learning techniques and data collected from IMUs to assess the Fugl-Meyer upper extremity subscale for post-stroke patients with motor dysfunction. IMUs signals from 120 patients were collected during a clinical trial. These signals were fed into a gated recurrent unit network to complete the scoring of individual actions, which were then aggregated to obtain the total score. Simultaneously, on the basis of the internal correlation between the Fugl–Meyer assessment and the Brunnstrom scale, Brunnstrom stage prediction models of the arm and hand were established via the random forest and extremely randomized trees algorithm. The experimental results show that the proposed models can score Fugl-Meyer items with a high accuracy of 92.66%. The R^2 between the doctors' score and the model's score is 0.9838. The Brunnstrom stage prediction models can predict high-quality stages, achieving a Spearman correlation coefficient of 0.9709. The application of the proposed method enables precision assessment of patients' upper extremity motor function, thereby facilitating more personalized rehabilitation programs to achieve optimal recovery outcomes.

Trial registration: Clinical trial of telerehabilitation training and intelligent evaluation system, ChiCTR2200061310, Registered 20 June 2022-Retrospective registration.

Keywords Stroke, Fugl–meyer assessment, Brunnstrom stage, Deep learning

*Correspondence:
Liquan Guo
guolq@sibet.ac.cn
[1]School of Biomedical Engineering (Suzhou), Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui 230026, China
[2]Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, Suzhou, Jiangsu 215163, China
[3]Department of Respiratory and Critical Care Medicine, The First Affiliated Hospital, Nanjing Medical University, Nanjing, China

Introduction

Stroke is the third leading cause of disability worldwide [1]. Each year, ≈795,000 people experience a new or recurrent stroke [2]. Hemiplegia is the most common post-stroke symptom in most cases, and the upper extremities (UE) of hemiplegic patients are more seriously affected than their lower extremities (LE) are [3, 4]. To restore motor function, stroke survivors need to participate in rehabilitation programs in hospitals, outpatient clinics, or nursing facilities that provide professional guidance and equipment resources, which means that stroke rehabilitation requires a sustained and coordinated effort from a large team and expensive costs [5, 6]. Rehabilitation interventions are beneficial across a number of neurological conditions because they result in a decrease in the severity of disability [7] However, not all interventions report effective outcomes [8]. The need for developing patient-specific interventions is paramount, and subject-specific interventions are based on precise assessment of the patient [9–11]. Importantly, rehabilitation specialists should be provided with tools to monitor the motor recovery process, assess whether the ongoing intervention is leading to the anticipated clinical results, and adjust the intervention if needed.

The Fugl-Meyer assessment (FMA) [12], Brunnstrom stage [13], action research arm test (ARAT) [14], and wolf motor function test (WMFT) [15] are commonly used assessment tools. In particular, the FMA is considered one of the most comprehensive quantitative measures of motor impairment [16]. Unfortunately, regularly conducting these assessments throughout the entire intervention period is both time-consuming and impractical [17]. Traditional motor function assessments, such as the FMA, often require skilled clinicians to perform detailed evaluations during each session, which may last 30 min [4, 18]. Moreover, the results of these assessments are often susceptible to subjective bias, as they rely on clinicians' judgment in scoring patients' performance [19]. Factors such as clinician experience, interpretation of movement quality, and even mood or fatigue levels can influence scoring, leading to variability and inconsistencies in the evaluation process. To address these issues, an increasing number of researchers have collected and recorded motion data from patients with motor dysfunction via wearable devices to carry out assessments [20–22]. Wearable device technology offers the advantages of real-time, noninvasive continuous monitoring of patient activity, providing objective data support, and enabling personalized, remote rehabilitation possibilities [23–25]. In motion monitoring, wearable devices based on IMUs are extensively utilized in research and clinical settings [26, 27], which enables longitudinal assessment of UE motor function [28], thus facilitating personalized rehabilitation interventions [29, 30].

Zhang et al. [31] developed a desktop robot that collected motion signals from patients and used these signals to build three machine learning models to assess the WMFT. Adans Dester et al. [17] employed IMUs to collect motor signals from stroke and traumatic brain injury patients and developed machine learning models to assess the Functional Ability Scale (FAS). They then used the predicted FAS scores along with motion signals to build a model via balanced random forests to complete the FMA. The coefficient of determination between the predicted and actual scores reached 0.86. While the above studies have made significant progress, they often face limitations in practical clinical application for two reasons. First, the limited number of patients in their study reduces the robustness and stability of the model. Second, they employed traditional machine learning methods for assessment. Traditional machine learning models rely heavily on human expertise for feature selection, potentially missing complex and subtle patterns in the data that could improve predictive performance.

Recurrent neural networks (RNNs) are a class of deep learning algorithms designed to handle sequential signals [32]. They can directly classify or predict signals without feature extraction, ensuring an end-to-end processing flow and improved accuracy. Long short-term memory (LSTM) [33] is among the most commonly used RNN algorithms. Lee et al. [34] applied LSTM networks to process IMU-based gait signals and assess gait fatigue. Li et al. [35] proposed a multimodal evaluation framework using LSTM to quantitatively assess hand motor function in post-stroke hemiplegia patients. Compared with machine learning algorithms, deep learning algorithms like LSTM have a larger number of parameters, requiring more samples to ensure the creation of an optimal model. However, collecting large amounts of clinical data is extremely challenging. Therefore, the application of deep learning models in motion signal-based motor function assessment, especially for stroke rehabilitation, remains underexplored.

To establish a novel system that assists doctors in conducting precise rehabilitation assessment, this study introduces a sensor network and accompanying deep learning-based motor function assessment models. First, we conducted a clinical trial involving 120 stroke patients in which wearable devices were used to collect motion signals during the clinical assessment process, ensuring an adequate sample size for subsequent model development. Then, on the basis of the data, we developed a deep learning model based on gated recurrent units (GRUs) to assess UE motor function. The GRUs is a variant of the RNN, but it requires fewer parameters than the LSTM does, thus reducing the risk of overfitting. We totally achieved the following two objectives: (1) precise scoring of individual actions in the FMA and comprehensive

assessment of UE function and (2) classification of the Brunnstrom stage for arms and hands in stroke patients. Compared with previous studies, this study has the following advantages:

(1) Large Sample Size for Robustness and Clinical Applicability: With a large sample size, this study enhances the model's robustness and generalizability, overcoming the limitations of previous studies with smaller cohorts. A larger dataset improves the model's adaptability to diverse patient conditions, increasing its applicability in real-world clinical settings.
(2) Deep Learning for Enhanced Model Performance: By utilizing deep learning techniques, this study removes the need for manual feature extraction, enabling the model to automatically learn complex patterns from raw motion signals. This method captures nonlinear relationships and enhances the model's predictive performance, leading to more accurate motor function assessment across various patient profiles.

Materials and methods

Participants

The data used for model development and validation were obtained from a clinical trial conducted at Tangdu Hospital and Xi'an Gaoxin Hospital, which was approved by the hospital's Ethics Committee. The clinical trial registration number is ChiCTR2200061310. A total of 120 patients were enrolled in the trial, and all participants signed the informed consent form. The patient criteria are as follows, and their details are presented in Table 1.

Table 1 Basic information of the 120 patients

Characteristics	Mean ± Standard Deviation	Min–Max	95% Confidence Interval
Age (year)	56.47 ± 9.70	33.00–74.00	54.71–58.22
Height (cm)	167.80 ± 6.95	150.00–180.00	166.50–169.00
Weight (kg)	69.68 ± 10.95	48.00–114.00	67.69–71.66
Pulse (/min)	77.59 ± 7.95	60.00–110.00	76.16–79.03
FMA Score (point)	43.28 ± 25.09	13.00–97.00	33.68–48.54
UE FMA Score (point)	36.76 ± 18.65	7.00–66.00	31.67–41.85
Characteristics		Vaule	
Gender (male/female)		86/34	
Stroke type (cerebral infarction/hemorrhage)		52/68	
Disease stage (subacute/chronic)		0/120	

The inclusion criteria for patients were as follows: (1) stroke confirmed by CT or MRI; (2) aged between 30 and 75 years; (3) stable recovery with motor dysfunction caused by stroke, 15–180 days after onset (convalescent period), with Brunnstrom upper extremity and/or lower extremity motor function grades II–VI; (4) ability to follow the research protocol; and (5) ability to understand the study's purpose, adhere to the protocol, and provide informed consent.

The exclusion criteria for patients were as follows: (1) patients with significant cognitive or consciousness disorders that would prevent completion of the FMA; (2) patients with other major limb injuries, such as fractures, severe arthritis, amputations, etc.; (3) patients with joint contractures; (4) patients with disabilities as defined by law (e.g., blindness, deafness, mutism, intellectual disabilities, mental disorders, and physical disabilities); and (5) patients with severe comorbidities that were deemed unsuitable for participation by the researcher.

Data collection protocol

Data collection devices

The collection system consists of wearable devices and a computer, as shown in Fig. 1. The wearable devices include two rehabilitation armbands and one rehabilitation glove. The armbands were equipped with IMUs. The IMUs were constructed via the MPU9250 chip (InvenSense, "America"). The IMU accelerometer featured a measurement range of ± 2 g, the IMU gyroscope offered 16-bit resolution with a range of ± 250°/s, and the IMU magnetometer had 14-bit resolution with a range of ± 4800 μT. Through chip computation, the output signals are **Pitch**, **Yaw**, and **Roll**, which represent the sensor orientation as follows: **Yaw** indicates rotation about the vertical axis, **Pitch** indicates rotation about the horizontal transverse axis, and **Roll** indicates rotation about the longitudinal axis. Therefore, they can reflect the rotational state of the body in three-dimensional space and are commonly used to describe the range of motion and angles of joints and arms. The glove was equipped with piezoresistive bending sensors and IMUs positioned at the back of the five fingers and the dorsum of the hand, respectively. The piezoresistive sensors monitor finger flexion, and the IMUs track wrist movements. The piezoresistive sensors exhibit a resistance of 25 kΩ at 180° in a stationary state and 125 kΩ at 90° at full bending. The output signals for the glove are the flexion signals of five fingers (**F1–F5**) and the **pitch** of the hand. The sampling rate of the aforementioned sensors is 50 Hz.

The rehabilitation armbands are worn on the upper arm (S1) and forearm (S2) on the hemiplegic side. The rehabilitation glove is worn directly on the hemiplegic hand (S3). The ZigBee protocol was chosen for wireless transmission, which meets the requirements of high fault

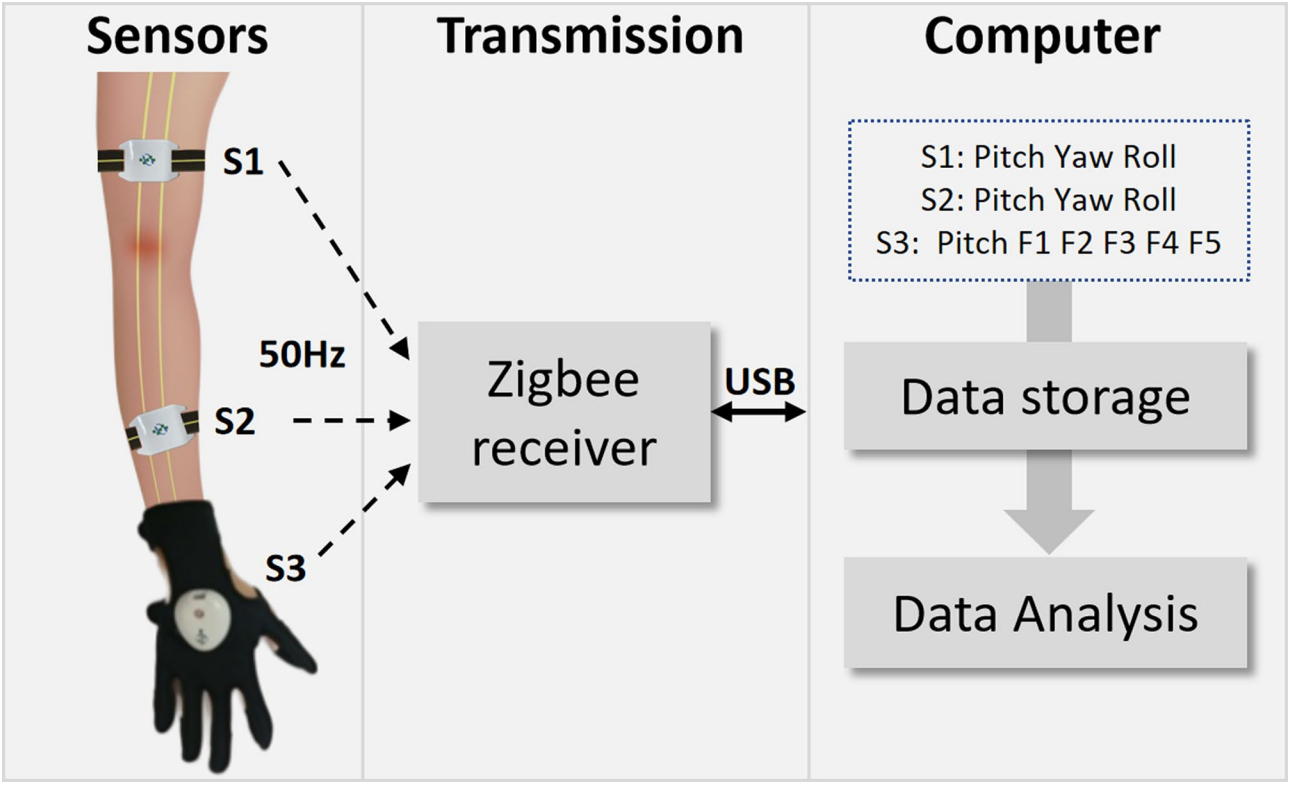

Fig. 1 The data collection devices for the experiment

tolerance and low cost. The ZigBee receiver can receive signals from the sensors in real time and store them on a personal computer, which handles data storage and performs signal analysis.

Actions for data collection

The upper extremity Fugl-Meyer assessment (UE-FMA) subscale includes 33 actions for assessment (items 1 to 33), with scores of 0, 1, or 2, where 0 indicates the inability to perform the action, 1 indicates partial completion, and 2 indicates smooth execution. Since all patients included in our study were at Brunnstrom Stage II-VI (Patients in the stage I are unable to complete the clinical trial due to physical reasons), they exhibited reflex abilities and thus scored full marks for reflex-related actions during data collection. Therefore, we excluded three reflex-related actions (Original items 1, 2, and 18). The original item 26 is a grip strength test that requires a force sensor for data collection; therefore, this item was excluded. Items 31 to 33 assess different aspects of the same action. Thus, we retained item 31 and removed items 32 and 33. In total, after excluding the six items mentioned above, we retained 27 items with 27 different actions, for which motion data were collected during the execution of these 27 assessment tasks. A description of these 27 actions is provided in Table 2.

Data Collection Procedure

Once the patient was ready, the physical therapist guided the patient to wear the wearable devices and powered them on. Each action was performed 3–5 times. The therapist then instructed the patient to perform the 27 actions sequentially. The physical therapist used a timer to record the start and end times for each stable action performed by the patient, ensuring a timing error of no more than 0.5 s. During this process, the physical therapist scored each action on the basis of the standardized assessment criteria. Moreover, the wearable devices captured the patient's motion signals in real time, transmitted them to a wireless receiver via the ZigBee protocol, and saved them on a computer.

The enrolled subjects participated in two data collection sessions, the first at baseline and the second at discharge, which occurred 3 weeks after the baseline assessment. All 120 patients participated in the first data collection session, resulting in 120 samples. For the second data collection session, 3 weeks later, a few patients were missing, resulting in 102 samples, resulting in 222 total collected samples. Each sample consists of data from 27 actions. The data for each action can be represented as $X \in R^{12 \times N}$, where 12 represents the number of signals and N represents the signal length.

Table 2 The description of the 27 actions

Action ID	Action Description
A1	Shoulder elevation
A2	Shoulder retraction
A3	Shoulder abduction (at least 90°)
A4	Shoulder external rotation
A5	Elbow flexion
A6	Forearm supination
A7	Shoulder adduction and internal rotation
A8	Elbow extension
A9	Forearm pronation
A10	Hand touching the lumbar spine
A11	Shoulder flexion to 90°(with elbow joint at 0°)
A12	Forearm pronation or supination (with shoulder joint at 0° and elbow joint at 90°)
A13	Forearm pronation (with shoulder abduction at 90°and elbow joint at 0°)
A14	Forearm pronation and supination (with shoulder flexion ranging from 90 to 180°and elbow joint at 0°)
A15	Forearm pronation or supination (with shoulder flexion between 30 to 90°and elbow joint at 0°)
A16	Wrist dorsiflexion (with elbow joint at 90°and shoulder joint at 0°)
A17	Wrist flexion and extension (with elbow joint at 90°and shoulder joint at 0°)
A18	Wrist dorsiflexion (with elbow joint at 0°and shoulder joint at 30°)
A19	Wrist flexion and extension (with elbow joint at 0° and shoulder joint at 30°)
A20	Wrist circumduction
A21	Finger flexion (flexion of all fingers together)
A22	Finger extension (extension of all fingers together)
A23	Thumb adduction with all joints at 0 position
A24	Thumb pinch (holding a pencil with the thumb)
A25	Grasp a cylindrical object
A26	Grasp a spherical object
A27	Finger-to-nose test (performed five times consecutively)

Signal processing

Preprocessing consists of the following steps, as shown in Fig. 2. The first step is the segmentation of valid signals, which aims to identify and isolate the relevant signal segments. During the data collection, the therapist used a timer to record the time points of the patient's stable actions. The valid signals are then segmented by matching the recorded starting point and end point with the sampling points collected by the wearable device.

The second step is filtering and sampling. A digital Butterworth bandstop filter was applied to eliminate power frequency interference, with a cutoff frequency of 50 Hz and a stopband gain of -40 dB. A moving average filter was used to remove random noise, with a filter width set to 10. Owing to the varying speeds of movements across patients, the length of valid signals also varies. Linear sampling was applied to resample the segmented data, and the signal length was fixed to 300 to meet the requirements of subsequent deep learning algorithms. The choice of a fixed length of 300 was based on two factors: first, the deep learning algorithm (RNN) requires inputs of uniform length, and second, the average signal length was approximately 300. Then, the resampled signals undergo zero-mean normalization, where the mean of the data is subtracted, centering the data around zero.

The third step addresses handling missing signals. Data collection was performed entirely by the therapist, with no involvement of engineering personnel. Occasionally, the therapist made operational errors, leading to the absence of certain signals for some patients (approximately 2% missing, with at most one missing action for one patient). The procedure for handling missing signals is as follows: for the i-th sample where the j-th movement signal is missing and its movement score label is s (where s = 0,1,2), one-tenth of the samples with the same label s and no missing signals are randomly selected. The mean of the missing signal is then computed from the selected samples, and the original missing value is replaced by this mean.

Upper extremity assessment algorithm

The process of constructing the UE rehabilitation assessment algorithm on the basis of motion signals is shown in Fig. 3. First, 27 GRU networks were developed, one for individual actions, to generate individual scores. These scores were then summed to calculate the total upper score. Subsequently, arm-related scores were extracted and used with a tree-based model for arm Brunnstrom stage prediction and hand-related scores were utilized for hand Brunnstrom stage prediction.

Scoring models for the UE-FMA subscale

Each item in the UE-FMA subscale represents specific motor functions, with motion signals differing in components. We utilized the GRU network, a variant of the traditional RNN, to build individual scoring models for each item. The GRU network comprises multiple GRU units, which offer a simpler architecture than conventional LSTM units do while effectively capturing long-term dependencies in sequential data. The input to the network is a preprocessed signal sample $X \in R^{12\times300}=\{x_1,x_2 \ldots x_{300}\}$, where each timestamp x_i represents a feature vector. The hidden state at the final timestamp is passed through a fully connected layer and a Softmax layer to produce the output—the item score. We constructed 27 GRU networks, one for each item, resulting in 27 outputs. The sum of these outputs provides the total score for the UE-FMA subscale.

The GRU network employs a bidirectional approach, processing signals in both forward and backward directions to comprehensively capture temporal dependencies. Owing to the uneven distribution of stages among

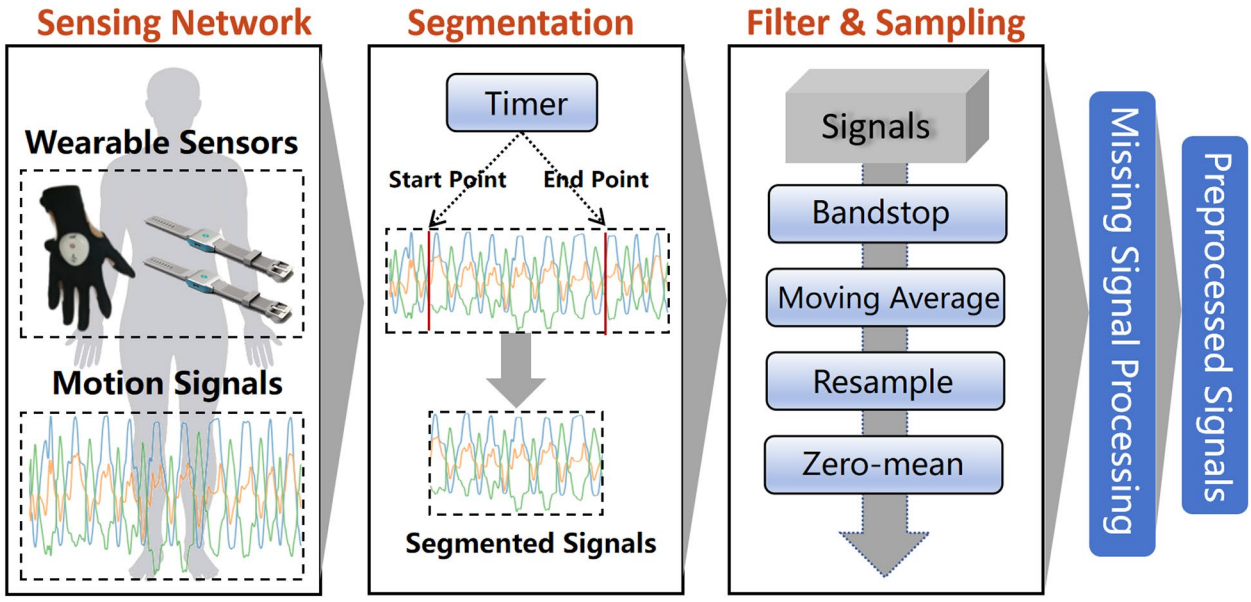

Fig. 2 The pipeline of data preprocessing

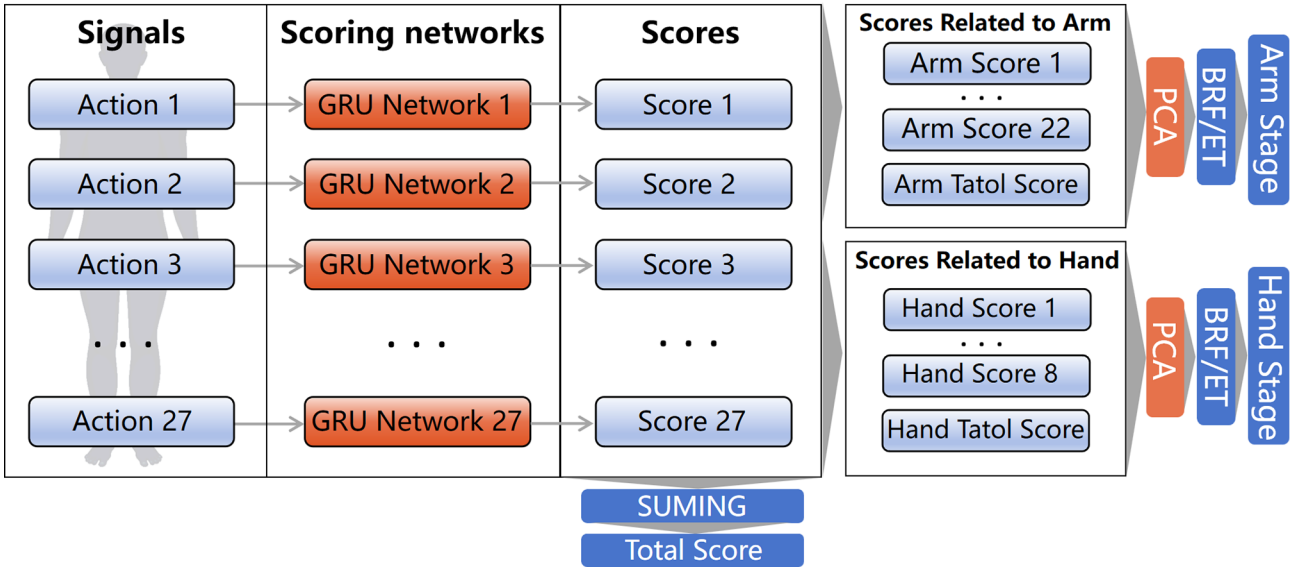

Fig. 3 The pipeline of building the upper extremity rehabilitation evaluation models

the 120 patients, a large proportion of scores were clustered around 0 and 1, resulting in data imbalance. To address this issue, we implemented a data balancing strategy during model training. Specifically, we used upsampling of the minority classes in the training set to achieve a more balanced data distribution across all categories. In this process, we replicated samples from the underrepresented classes to increase their frequency, ensuring that each class contributed more equally to the training process. Additionally, we applied a random sampling technique where the samples in the minority classes were randomly selected with replacement until the class distribution approached that of the majority class. To avoid overfitting, we also monitored the performance on the validation set and adjusted the upsampling strategy as necessary. By doing so, we aimed to improve the model's ability to generalize and reduce the bias toward the majority classes.

Prediction models for the arm and hand brunnstrom stage

Intrinsic correlations exist between various motor function scales [36]. Considering the intrinsic correlations between the FMA and Brunnstrom stage, we used the

scores from the 27 items in the UE-FMA to predict the Brunnstrom stage of the arm and hand.

22 arm-related items and 8 hand-related items were identified out of the 27 actions. For arm Brunnstrom stage prediction models, we used the scores of these 22 items and their total score, totaling 23 features, to build the prediction model. For hand Brunnstrom stage prediction models, we used the scores of the 8 items and their total score, totaling 9 features, to build the prediction model. Considering the large number of features, we applied principal component analysis (PCA) for dimensionality reduction. The final feature dimensions were reduced to six.

The random forest (RF) and extremely randomized tree (ERT) algorithms were employed to build the prediction models. Both algorithms use multiple decision trees to form the model and combine their outputs for final predictions. They can be used for classification and regression tasks and perform well on nonlinear problems. The RF performs sampling with replacement and feature selection randomization, selecting the best split point at each node, resulting in lower bias but higher computational cost. In contrast, ERT do not sample the data, but use the entire dataset and randomly select split points. They exhibit slightly higher bias but lower variance, faster training speed, and better resistance to overfitting. The model uses the above features as inputs and outputs the Brunnstrom stage.

Model establishment and validation

The data were split into training and testing datasets at a 3:1 ratio. The GRU networks were built and trained via PyTorch 2.1.2. The RF and ERT models were built via Scikit-learn. The model hyperparameters are shown in Table 3.

The accuracy, recall, precision, and F1 score were employed to evaluate the performance of the individual movement scoring model; the coefficient of determination (R^2) and root mean square error (RMSE) were used to evaluate the performance of the scoring models for the UE-FMA subscale; and the accuracy, recall, precision, F1 score, and Spearman correlation were used to evaluate the performance of the prediction model for the arm & hand Brunnstrom stage.

Table 3 The hyperparameters when building the models

Model	Hyperparameter	Value
GRU	Learning Rate	0.01
	Batch Size	16
	Epochs	50
	Optimizer	Adam
RF	Number of Trees	30
	Max_depth	6
ERT	Number of Trees	30
	Max_depth	6

Accuracy (Ac) refers to the proportion of correctly predicted samples to the total number of samples. Recall (Re) refers to the proportion of samples correctly predicted as positive cases to the total number of actual positive cases. Precision (Pr) refers to the proportion of samples correctly predicted as positive cases to the total number of predicted positive cases. The F1 score is the harmonic mean of precision and recall and is used to provide a balanced evaluation of the performance of classification models. The coefficient of determination (R^2) indicates the degree of fit of the model to the observed data. The root mean square error (RMSE) represents the average deviation between the observed values and the values predicted by the model. Spearman correlation is a nonparametric measure of the strength and direction of association between two ranked variables. Pearson correlation requires data to be continuous, linear, and normally distributed, whereas Spearman correlation is based on ranks and is suitable for nonnormally distributed data, particularly when the relationship is nonlinear or when the data are ordinal. Since the Brunnstrom stage is noncontinuous, have ranking features, and is typically ordinal, Spearman correlation is more appropriate for measuring the strength and direction of the relationship between the predicted and true Brunnstrom stage values.

Results

This study constructed 27 GRU networks to score the 27 items in the UE-FMA. The performance of these models is illustrated in Fig. 4. Figures 4(A), 4(B), 4(C), and 4(D) represent the accuracy, recall, precision, and F1 score of the 27 models, respectively (detailed values are provided in the appendix). The blue bars represent the results without applying the data balancing strategy, whereas the orange bars represent the results after applying the strategy. With the exception of a few items, the model performance improved after the data balancing strategy was implemented.

The total UE-FMA score was obtained by summing the scores of the 27 models. Figure 5 shows the RMSE and R^2 values comparing the total scores derived from the proposed models with those assessed by the therapists. Figure 5(A) shows the results without applying the data balancing strategy, with an R^2 of 0.9774 ($p < 0.001$) and an RMSE of 2.8347, whereas Fig. 5(B) illustrates the results with the strategy applied, with an improved R^2 of 0.9838 ($p < 0.001$) and a reduced RMSE of 2.4016.

The scores of the 27 items were used as features to construct RF and ERT models, capturing the relationship between FMA and Brunnstrom stage for both the arm and hand. Table 4 presents the performance of the prediction model using RF, while Table 5 presents the results

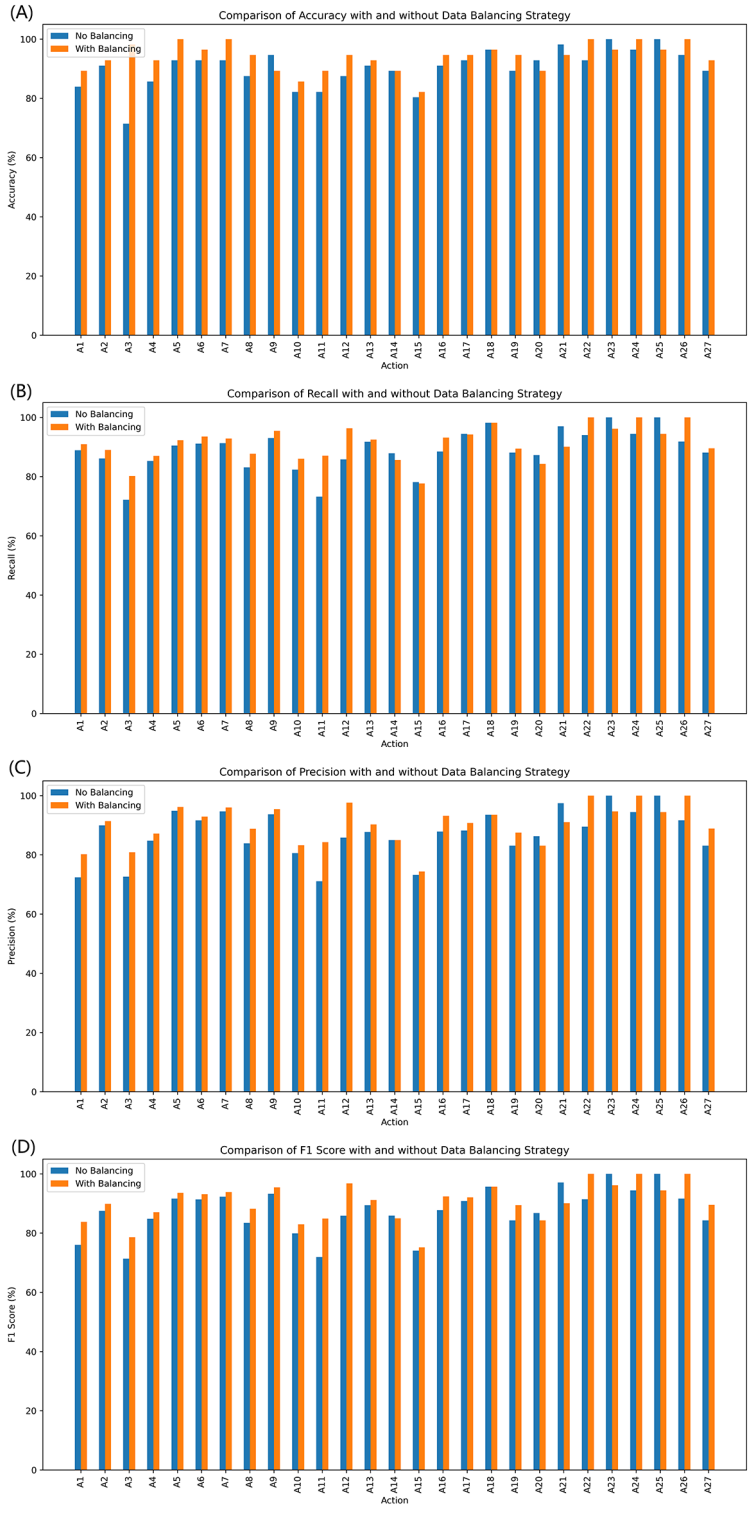

Fig. 4 The performance of the 27 GRU networks

with ERT. The ERT model outperformed the other models for both arm and hand stage prediction, achieving accuracies of 76.72% and 82.09%, with Spearman correlation coefficients of 0.9475 and 0.9709, respectively.

Discussion

In summary, through wearable devices and intelligent algorithms, scoring models for the UE-FMA and Brunnstrom stage prediction models for the arm and

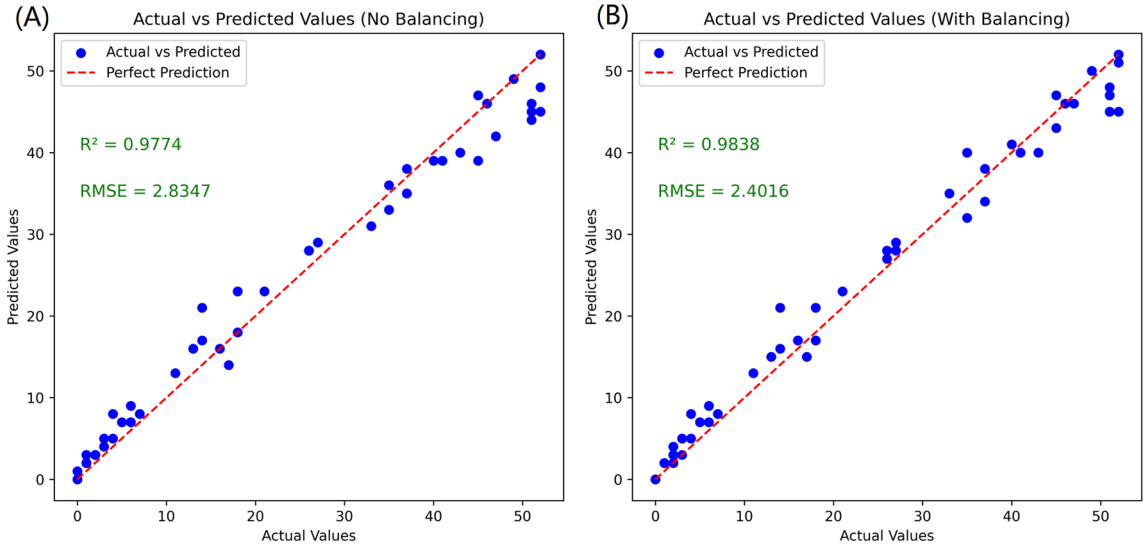

Fig. 5 The RMSE and the R2 between the total upper extremity FMA scores

Table 4 The accuracy, recall, precision, F1-score, and Spearman correlation coefficient of the RF model

	Ac	Re	Pr	F1	Spearman
Arms	73.20%	64.97%	64.12%	62.01%	0.9368(p<0.001)
Hands	79.49%	64.77%	67.26%	61.77%	0.9633(p<0.001)

Table 5 The accuracy, recall, precision, F1-score, and Spearman correlation coefficient of the ERT model

	Ac	Re	Pr	F1	Spearman
Arms	76.72%	60.31%	60.23%	58.07%	0.9475(p<0.001)
Hands	82.09%	68.63%	69.48%	65.02%	0.9709(p<0.001)

hand were developed. These models demonstrate promising performance in evaluating motor function in stroke patients, offering the potential to be used in clinical practice to assess UE motor function. By leveraging wearable devices, these models can be deployed in various clinical settings or even at home, making it easier for patients to undergo continuous monitoring and rehabilitation. This could significantly reduce the burden on clinicians, allowing them to focus more on personalized treatment plans on the basis of accurate, real-time data.

Upon reconsidering the scoring criteria for each item, we found that the boundaries between 0 and 1 point, as well as between 1 and 2 points, can be ambiguous in clinical scoring. For example, for the criterion of elbow extension, a score of 0 indicates an inability to perform the movement, a score of 1 indicates that the elbow can bend 90 degrees, and a score of 2 indicates that it can bend 180 degrees. However, according to the therapists' descriptions, the actual scores are often within ranges. For example, one therapist might define 0–45 degrees as 0 points, 45–135 degrees as 1 point, and more than 135 degrees as 2 points. Moreover, the defined ranges vary across therapists, and subjective factors can lead to different final evaluations. This phenomenon occurs because the FMA uses a discrete three-point scale for each item, while the movement performance is continuous, leading to a ceiling effect [37, 38]. To address this issue, refining the scoring system for each item to a decimal level could provide a solution. Finer scoring not only addresses the ceiling effect but also enables a more precise evaluation of the patient's motor function. Moreover, the ceiling effect may mitigate the reduction in patients' motivation for rehabilitation. However, finer grading significantly increases the difficulty of clinical assessment, and its clinical applicability remains to be evaluated. Nonetheless, the model we proposed achieved relatively good performance at the three-point scale, providing a foundation for future efforts to refine the scoring by engineering methods.

On the basis of the scores for the 27 items, this study established Brunnstrom stage prediction models for the arms and hands. Although the prediction accuracy is not very high, the discrepancies between the misclassified samples and their true labels are small, with at most a 1-stage difference. This is reflected by the Spearman correlation coefficient between the model's predicted stages and the true labels. The performance of the models established via ERT outperforms that of RF. This is primarily due to data imbalance. The ERT model performs better at handling imbalanced datasets because it introduces more randomness in selecting features and split points during construction, resulting in greater diversity among the trees. This diversity reduces model bias and helps to better address minority class samples in imbalanced datasets. One of the most common challenges in developing clinically relevant deep learning models is data imbalance. While collecting more data from minority groups seems to be the most straightforward approach,

it is difficult to implement clinically, as patient distributions are inherently imbalanced in practice. Therefore, in engineering, employing balancing strategies or using algorithms specifically designed for imbalanced datasets becomes the optimal solution.

Importantly, each item of the FMA reflects the synergistic relationships between different movements. Movements within these synergies are characterized by coordinated patterns of muscle activity, such as flexion synergy or extension synergy, which involve multiple joints working together in a specific pattern. When applying PCA to reduce the dimensionality of features, it is essential to recognize that the specific components resulting from dimensionality reduction do not directly correspond to any specific synergy or movement pattern. This is a limitation of PCA, as the interpretability of the features becomes less transparent, and the direct association between synergies and the reduced components becomes unclear. While PCA is effective in reducing dimensionality and preventing overfitting, it can complicate the precise interpretation of specific synergies or movement patterns in terms of their contribution to the final results. In summary, if we have sufficiently balanced raw data and a large enough sample size, we do not need any additional engineering remedies, as such methods often come with side effects. However, even though the inclusion of 120 patients in our study far exceeds that of other studies, we still had to employ such methods to address issues arising from imbalanced sample distribution and insufficient sample size.

Another limitation is that the current research lacks a fully integrated software platform that could facilitate clinical application. This study has only completed the design of the hardware system and the development of the evaluation algorithm, without completing the software platform development. This significantly limits its clinical application. In future work, we plan to develop a fully integrated software platform to address this gap. The platform seamlessly integrates the hardware system with the evaluation algorithm, providing a user-friendly interface for the complete process of data collection, analysis, and output of evaluation results. This will greatly reduce the workload of clinicians and minimize subjective influence in the evaluation. Moreover, we will further optimize the system functions on the basis of clinical needs, such as adding personalized evaluation modules, automatically generating reports, and integrating remote monitoring and rehabilitation guidance modules, thereby increasing the system's practicality and clinical applicability.

Conclusion

In this study, we developed an upper extremity motor function assessment system. Rehabilitation wearable devices were employed to collect motion signals from stroke patients. GRU networks were employed to develop scoring models for individual items in UE-FMA, thus enabling the calculation of the total score. Furthermore, RF and ERT algorithms were implemented to develop Brunnstrom stage prediction models for assessing the functions of the arm and hand. The scoring models achieved an average accuracy exceeding 92.66% for individual items. The R^2 value between the model-generated scores and those given by clinicians was 0.9838, with an RMSE of 2.4016. The Spearman correlation coefficients between the model-predicted Brunnstrom stage and the actual stage were 0.9475 for the arm and 0.9709 for the hand. These results highlight the potential of combining motion sensors with deep learning to accurately assess upper motor functions in stroke patients, which could facilitate the creation of personalized rehabilitation programs and promote better recovery outcomes for patients. In the future, we plan to expand the dataset to develop more granular, decimal-level evaluation models, improving precision and enhancing generalizability. Additionally, we aim to develop a comprehensive system framework that integrates data collection, assessment algorithms, and an evaluation platform, ensuring ease of use for clinicians and patients.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12984-025-01625-9.

Supplementary Material 1

Acknowledgements
Not applicable.

Author contributions
W.N.Z. and L.Q.G. contributed to the conception and design of the study. L.F.Z., J.P.W., and Z.Y.D. completed clinical trials and collected the motion data. D.Y.F. completed the categorization and preprocessing of data. W.N.Z. wrote the code for the models in the study and completed the validation of the model. W.N.Z. wrote the first draft of the manuscript. All authors contributed to the manuscript revision and approved the submitted version.

Funding
This research was funded by grants from the National Key Research and Development Program of China (Grant No. 2022YFC0710800) and the National Key Research and Development Program of China (Grant No. 2018YFC1313602).

Data availability
To protect the privacy of clinical trial participants, the data will be available upon request.

Declarations

Ethics approval and consent to participate
The clinical trial was conducted at Tangdu Hospital and Xi'an High-Tech Hospital and was approved by the medical Ethics Committee of Tangdu Hospital Airforce Medicine University.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 13 August 2024 / Accepted: 7 April 2025
Published online: 16 April 2025

References

1. Katan M, Luft A. Global burden of stroke. Semin Neurol. 2018;38(2):208–11.
2. Martin SS, Aday AW, Almarzooq ZI, Anderson CAM, Arora P, Avery CL, Baker-Smith CM, Barone Gibbs B, Beaton AZ, Boehme AK, et al. 2024 heart disease and stroke statistics: A report of US and global data from the American heart association. Circulation. 2024;149(8):e347–913.
3. Li H, Zhao G, Zhou Y, Chen X, Ji Z, Wang L. Relationship of EMG/SMG features and muscle strength level: an exploratory study on tibialis anterior muscles during plantar-flexion among hemiplegia patients. Biomed Eng Online. 2014;13:5.
4. Wang C, Peng L, Hou ZG, Li J, Zhang T, Zhao J. Quantitative assessment of Upper-Limb motor function for Post-Stroke rehabilitation based on motor synergy analysis and Multi-Modality fusion. IEEE Trans Neural Syst Rehabil Eng. 2020;28(4):943–52.
5. Winstein CJ, Stein J, Arena R, Bates B, Cherney LR, Cramer SC, Deruyter F, Eng JJ, Fisher B, Harvey RL, et al. Guidelines for adult stroke rehabilitation and recovery: A guideline for healthcare professionals from the American heart association/american stroke association. Stroke. 2016;47(6):e98–169.
6. Bo W, Cavuoto L, Langan J, Subryan H, Bhattacharjya S, Huang M-C, Xu W. A progressive prediction model towards home-based stroke rehabilitation programs. Smart Health. 2022;23:100239.
7. Walker WC, Pickett TC. Motor impairment after severe traumatic brain injury: A longitudinal multicenter study. J Rehabil Res Dev. 2007;44(7):975–82.
8. Teasell RW, Murie Fernandez M, McIntyre A, Mehta S. Rethinking the continuum of stroke rehabilitation. Arch Phys Med Rehabil. 2014;95(4):595–6.
9. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med. 2015;372(9):793–5.
10. Hulsen T, Jamuar SS, Moody AR, Karnes JH, Varga O, Hedensted S, Spreafico R, Hafler DA. McKinney EF: from big data to precision medicine. Front Med (Lausanne). 2019;6:34.
11. Niederberger E, Parnham MJ, Maas J, Geisslinger G. 4 Ds in health research-working together toward rapid precision medicine. EMBO Mol Med. 2019;11(11):e10917.
12. Fugl-Meyer AR, Jääskö L, Leyman I, Olsson S, Steglind S. The post-stroke hemiplegic patient. 1. a method for evaluation of physical performance. Scand J Rehabil Med. 1975;7(1):13–31.
13. Brunnstrom S. Motor testing procedures in hemiplegia: based on sequential recovery stages. Phys Ther. 1966;46(4):357–75.
14. Lyle RC. A performance test for assessment of upper limb function in physical rehabilitation treatment and research. Int J Rehabil Res. 1981;4(4):483–92.
15. Wolf SL, Catlin PA, Ellis M, Archer AL, Morgan B, Piacentino A. Assessing Wolf motor function test as outcome measure for research in patients after stroke. Stroke. 2001;32(7):1635–9.
16. Gor-García-Fogeda MD, Molina-Rueda F, Cuesta-Gómez A, Carratalá-Tejada M, Alguacil-Diego IM, Miangolarra-Page JC. Scales to assess gross motor function in stroke patients: A systematic review. Arch Phys Med Rehabil. 2014;95(6):1174–83.
17. Adans-Dester C, Hankov N, O'Brien A, Vergara-Diaz G, Black-Schaffer R, Zafonte R, Dy J, Lee SI, Bonato P. Enabling precision rehabilitation interventions using wearable sensors and machine learning to track motor recovery. NPJ Digit Med. 2020;3:121.
18. Gladstone DJ, Danells CJ, Black SE. The fugl-meyer assessment of motor recovery after stroke: a critical review of its measurement properties. Neurorehabil Neural Repair. 2002;16(3):232–40.
19. Lee HH, Kim DY, Sohn MK, Shin YI, Oh GJ, Lee YS, Joo MC, Lee SY, Han J, Ahn J, et al. Revisiting the proportional recovery model in view of the ceiling effect of Fugl-Meyer assessment. Stroke. 2021;52(10):3167–75.
20. Yu L, Xiong D, Guo L, Wang J. A remote quantitative Fugl-Meyer assessment framework for stroke patients based on wearable sensor networks. Comput Methods Programs Biomed. 2016;128:100–10.
21. 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.
22. Li Y, Zhang X, Gong Y, Cheng Y, Gao X, Chen X. Motor function evaluation of hemiplegic Upper-Extremities using data fusion from wearable inertial and surface EMG sensors. Sens (Basel) 2017, 17(3).
23. Patel S, Park H, Bonato P, Chan L, Rodgers M. A review of wearable sensors and systems with application in rehabilitation. J Neuroeng Rehabil. 2012;9:21.
24. Wang Q, Markopoulos P, Yu B, Chen W, Timmermans A. Interactive wearable systems for upper body rehabilitation: a systematic review. J Neuroeng Rehabil. 2017;14(1):20.
25. Lee SI, Adans-Dester CP, Grimaldi M, Dowling AV, Horak PC, Black-Schaffer RM, Bonato P, Gwin JT. Enabling stroke rehabilitation in home and community settings: A wearable Sensor-Based approach for Upper-Limb motor training. IEEE J Transl Eng Health Med. 2018;6:2100411.
26. Formstone L, Huo W, Wilson S, McGregor A, Bentley P, Vaidyanathan R. Quantification of motor function Post-Stroke using novel combination of wearable inertial and mechanomyographic sensors. IEEE Trans Neural Syst Rehabil Eng. 2021;29:1158–67.
27. Park YS, An CS, Lim CG. Effects of a rehabilitation program using a wearable device on the upper limb function, performance of activities of daily living, and rehabilitation participation in patients with acute stroke. Int J Environ Res Public Health 2021, 18(11).
28. Waddell KJ, Strube MJ, Tabak RG, Haire-Joshu D, Lang CE. Upper limb performance in daily life improves over the first 12 weeks poststroke. Neurorehabil Neural Repair. 2019;33(10):836–47.
29. Dobkin BH, Dorsch A. The promise of mHealth: daily activity monitoring and outcome assessments by wearable sensors. Neurorehabil Neural Repair. 2011;25(9):788–98.
30. Miao S, Shen C, Feng X, Zhu Q, Shorfuzzaman M, Lv Z. Upper limb rehabilitation system for stroke survivors based on Multi-Modal sensors and machine learning. IEEE Access. 2021;9:30283–91.
31. Zhang M, Chen J, Ling Z, Zhang B, Yan Y, Xiong D, Guo L. Quantitative evaluation system of upper limb motor function of stroke patients based on desktop rehabilitation robot. Sens (Basel) 2022, 22(3).
32. Zaremba W, Sutskever I, Vinyals O. Recurrent Neural Network Regularization. *ArXiv* 2014, abs/1409.2329.
33. Hochreiter S, Schmidhuber J. Long Short-Term memory. Neural Comput. 1997;9(8):1735–80.
34. Lee YJ, Wei MY, Chen YJ. Multiple inertial measurement unit combination and location for recognizing general, fatigue, and simulated-fatigue gait. Gait Posture. 2022;96:330–7.
35. Li C, Yang H, Cheng L, Huang F, Zhao S, Li D, Yan R. Quantitative assessment of hand motor function for Post-Stroke rehabilitation based on HAGCN and multimodality fusion. IEEE Trans Neural Syst Rehabil Eng. 2022;30:2032–41.
36. de Oliveira R, Cacho EW, Borges G. Post-stroke motor and functional evaluations: a clinical correlation using Fugl-Meyer assessment scale, Berg balance scale and Barthel index. Arq Neuropsiquiatr. 2006;64(3b):731–5.
37. De Weerdt WJG, Harrison MA. Measuring recovery of arm-hand function in stroke patients: A comparison of the Brunnstrom-Fugl-Meyer test and the action research arm test. Physiotherapy Can. 1985;37(2):65–70.
38. Rabadi MH, Rabadi FM. Comparison of the action research arm test and the Fugl-Meyer assessment as measures of upper-extremity motor weakness after stroke. Arch Phys Med Rehabil. 2006;87(7):962–6.

Publisher's note