Evaluating generalization of arm movement identification using machine learning
From structured to semi-structured environments

Akbari, Sahel; Horemans, Herwin L.D.; Bussmann, Johannes B.J.; Zgonnikov, Arkady

DOI
10.1016/j.compbiomed.2025.111167

Publication date
2025

Document Version
Final published version

Published in
Computers in biology and medicine

Citation (APA)
Akbari, S., Horemans, H. L. D., Bussmann, J. B. J., & Zgonnikov, A. (2025). Evaluating generalization of arm movement identification using machine learning: From structured to semi-structured environments. *Computers in biology and medicine*, *198*(Pt A), 111167. Article 111167. https://doi.org/10.1016/j.compbiomed.2025.111167

Computers in Biology and Medicine 198 (2025) 111167

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Evaluating generalization of arm movement identification using machine learning: From structured to semi-structured environments

Sahel Akbari [a,b,*], Herwin L.D. Horemans [a,c], Johannes B.J. Bussmann [a,c], Arkady Zgonnikov [b]

[a] Dept. Rehabilitation Medicine, Erasmus MC University Medical Center, The Netherlands
[b] Dept. Cognitive Robotics, Faculty of Mechanical Engineering, Delft University of Technology, The Netherlands
[c] Rijndam Rehabilitation, The Netherlands

ARTICLE INFO

Keywords:
Activities of daily life
Arm movement identification
Home-based rehabilitation
Inertial measurement units
Machine learning

ABSTRACT

Home-based rehabilitation is essential for stroke survivors, facilitating motor recovery and improving activities-of-daily-life performance. Recent advances in wearable technologies and machine learning promise to revolutionize home-based arm rehabilitation by providing detailed movement analysis. However, machine learning algorithms for arm movement identification are predominantly trained and tested in the same environments. Their ability to generalize to novel environments remains largely unknown, hindering practical applications. This paper investigates the ability of two established machine learning models to generalize a structured, lab-based environment to a more realistic, semi-structured kitchen environment. Twelve healthy participants performed various arm activities, involving three arm movement types (reaching, lifting, and pronation/supination). In addition to evaluating the generalization of movement identification, we compared algorithm performance for two different sensor configurations: four Inertial Measurement Units (IMUs) on the arm versus a single IMU on the wrist. We employed a Random Forest (RF) classifier and a hybrid deep learning model combining convolutional and recurrent neural networks, evaluating both subject-specific and group approaches. Trained in the structured environment, the RF classifier predicted activities in the semi-structured environment with 86.54% (subject-specific) and 77.37% (group) balanced accuracy, based on the four-sensor configuration, while the hybrid model reached 87.96% and 82.96% accuracy. The accuracy was lower with a single wrist IMU; the RF classifier showed a smaller decrease than the hybrid model. Our findings demonstrate that the investigated arm movement identification algorithms generalize well across environments even with the minimal sensor configuration, indicating the potential for future applications in home-based stroke rehabilitation.

1. Introduction

Stroke is a leading cause of disability worldwide, with upper limb impairment being a common consequence that significantly impacts the performance of daily activities [1,2]. Rehabilitation plays a crucial role in treating and facilitating motor recovery, thereby improving stroke survivors' ability to perform activities of daily life (ADL) [3]. However, the complexity and duration of rehabilitation, along with the high cost of specialized therapies and limited access to facilities, often make clinical rehabilitation prolonged, inconvenient, costly and not sustainable [4]. This necessitates the exploration of remote or home-based rehabilitation techniques [5,6].

Moreover, to optimally improve motor recovery, increasing the intensity of rehabilitation is essential, as higher intensity is associated with better functional outcomes in motor recovery [7]. However, current clinical practices often involve low-intensity approaches, which fail to promote the optimal neuroplastic changes necessary for meaningful recovery, especially in the upper limbs [8]. In this context, continuous monitoring of ADL offers clinically relevant insights for tailoring post-stroke therapy in home settings, and has the potential to provide more intensive rehabilitation outcomes compared to traditional clinical practices [9–11].

Research in this area has highlighted various approaches to categorize and monitor arm movements more accurately [12]. For instance, Bochniewicz's study [13] categorized arm movements during ADL into functional and non-functional, offering a more targeted framework for continuous monitoring of daily arm use. Similarly, Rand's study [14]

[*] Corresponding author at: Dept. Rehabilitation Medicine, Erasmus MC University Medical Center, The Netherlands.
E-mail address: s.akbari@erasmusmc.nl (S. Akbari).

https://doi.org/10.1016/j.compbiomed.2025.111167
Received 3 January 2025; Received in revised form 7 March 2025; Accepted 28 September 2025
Available online 11 October 2025

demonstrated the potential of wrist accelerometer data, combined with self-reported information, to predict the level of affected upper extremity use in daily life, specifically 12 months post-stroke. However, simply measuring affected arm use may not reflect the appropriate dosage of specific movements required to tailor rehabilitation efforts effectively [15]. Therefore, it is crucial to go beyond just assessing the time spent on functional tasks and instead focus on a more precise monitoring of the motion content and quality of arm movements.

Human motions are systematically placed into a hierarchical structure based on their time scales and goals [16], where functional primitives, or isolated movements, serve as the foundational elements of activities. This hierarchical understanding of motion offers profound insights into neurological disorders at a granular level. If the motion content is not fully comprehended, it becomes challenging to establish any relationship between the prescribed repetitions of movement and the efficacy of rehabilitation interventions [17].

A promising approach to accurately measure motion content in daily life involves motion sensing technologies. Wearable sensors, characterized by their lightweight, portability, and small size, are being increasingly integrated stroke tele-rehabilitation, post-stroke recovery at home, and motion analysis applications such as gait analysis [18–20]. Among the wearable sensors, inertial measurement units (IMUs) are particularly well-suited for measuring upper extremity motions. IMUs offer extraordinarily rich data by capturing both linear and rotational movements [21].

In recent years, arm activity identification using IMU sensors has been enhanced by machine learning (ML) techniques. These include supervised and unsupervised methods, ranging from support vector machines (SVM) [22] and linear discriminant analysis (LDA) [22], to ensemble models like random forest (RF) [23], k-means clustering [24], and convolutional neural networks (CNN) [25,26]. Deep learning architectures, such as CNNs, have been reported to achieve higher accuracy in arm movement identification compared to conventional machine learning models [25–29]. Moreover, Murad and Pyun [30] found that deep recurrent neural networks can outperform CNN models for certain tasks.

Despite these advancements, much remains unknown about the potential of ML-based arm movement identification for real-life applications. In particular, existing studies mainly test and train models on similar datasets from the same settings and activities [25–27,29], and often in fully constrained settings which limits their relevance for real-life applications [22,31]. Recent studies have begun to address this challenge; for instance, Gomez.et al. [23] investigated the ability of their models to identify reach versus non-reach actions using an activities-of-daily-life (pizza-making) dataset. However, in general, there is currently little clarity regarding the ability of arm movement identification models to generalize across separate environments and different individuals, largely due to the lack of inclusion of varied datasets used for evaluation.

To address this gap, here we evaluated the generalization of ML-based arm movement identification models by focusing on three arm movement types, including different activities performed in different settings for training and testing phases. We trained ML-based models on healthy participants' data obtained in a structured environment resembling clinical settings; we then tested models on data captured in a semi-structured setting that mimicked real-world home environments. Unlike prior work that relied primarily on within-dataset validation, our approach allows us to evaluate the generalization of models to more dynamic and varied conditions.

Furthermore, we also considered generalization across random individuals to evaluate the model's ability to adapt to new, previously unseen subject's data. Therefore, we followed two distinct approaches: (1) a subject-specific approach, where the training (structured trials) and testing (semi-structured trial) datasets came from the same subject, and (2) a group approach, where structured data from all participants is used for training, while a random subject's semi-structured trial is used for testing.

Building on the findings of previous studies, we selected two models for our evaluation: The first is a model based on Random Forest (an established classification technique) and the second is a new variant of a state-of-the-art hybrid deep learning model which was previously proposed for ankle movement classification [32].

Our primary aim in this paper was to evaluate to what extent ML-based models can generalize from structured environments to more challenging, semi-structured environments that more closely resemble patients' home environments. We followed both subject-specific and group approaches to evaluate our models, and we focused on three main arm movements: reaching (extension/flexion of forearm), lifting (rotation of forearm around the elbow joint), and supination/pronation (rotation of the wrist around the forearm axis). We evaluated and compared the performance of the two selected models, particularly examining whether the hybrid model's architecture provides an edge in classifying sequential data into three movement types, based on its ability to capture both spatial and temporal features.

Our second aim was to explore the feasibility of arm movement identification using minimal sensor configuration, which is crucial for clinical applications. We used two alternative sensor configurations: the first one including four IMU sensors (on participants' dominant hand, lower arm, upper arm, and shoulder) and the second one incorporating only one IMU sensor on the dominant wrist, similar to a conventional smart watch. To our knowledge, this study is the first attempt to evaluate the use of minimal sensor configuration to generalize arm movement identification from one setting to another (clinic-home scenario).

2. Methodology

2.1. Participants

This study initially recruited fifteen healthy participants. However, due to data quality issues, three participants were excluded: two due to a large share of missing data, likely caused by IMU data transmission issues, and one due to an abnormally fast task completion time that impeded accurate arm movement measurement. The final dataset included twelve participants (eight women, four men) with a mean age of 29.3 ± 5.9 years Among these, eleven participants were right-handed and one participant was left-handed. All participants were healthy adults with no history of brain injury or limb paralysis. Informed consents were received from all participants. This study received ethical approval from the Human Research Ethics Committee of Delft University of Technology, application number 4189.

2.2. Data acquisition

We employed five QSense 9DOF motion tracking sensors (2M Engineering, [33]), each consisting of a 3-axis accelerometer, a 3-axis gyroscope, and a 3-axis magnetometer. These inertial measurement units (IMUs) were positioned on the dominant arm in a distal-to-proximal arrangement, including placements on the hand, wrist, lower arm (one-third of the forearm, close to the wrist), upper arm (one-third of the upper arm, close to elbow joint), and the shoulder near the acromion. Sensors on the hand, lower arm, upper arm, and shoulder constituted the primary sensor configuration and were affixed directly to the body using adhesive stickers. In contrast, the wrist sensor was configured separately to explore the feasibility of movement identification in a watch-like setup. This sensor was encased in a rubber wristband and worn like a conventional watch.

For the purpose of arm movement identification, only the raw linear acceleration and raw rotational rate data channels were utilized, as the magnetometer data were significantly influenced by environmental factors, such as magnetic field disturbances. The IMUs sampled data at a frequency of 50 Hz and transmitted data via Bluetooth to a computer running the QSense Motion Application.

Each trial was individually recorded using a video camera (720p HD, 30 frames per second), for subsequent manual data labeling, and stored separately alongside the corresponding IMU recordings.

2.3. Experiment protocols

For each participant, data were collected using both a structured and a semi-structured protocol. The primary distinction between these two lies in their level of standardization and control, with the structured protocol providing uniformity and the semi-structured protocol introducing variability by allowing participants to perform tasks naturally. The structured protocol involved performing three isolated arm movements: reaching (extension/flexion of the forearm), lifting (rotation of the forearm around the elbow joint), and supination/pronation (pro/supination; rotation of the wrist around the forearm axis). During this protocol, the participant sat in front of a table and performed these movements. Researchers provided explicit instructions on how to execute the movements while maintaining proper posture. To maintain uniformity, each movement started and ended at a designated position in front of the participant. Throughout the recording, participants remained seated with minimal body movement. Each movement was repeated ten times, with a three-second pause between repetitions. To distinguish between sets of movements, subjects were asked to clap twice before starting the next set of ten repetitions. The structured protocol was designed to include five trials in total, with each trial consisting of thirty arm movements (3 movements × 10 repetitions).

The selection of these arm movements was based on recommendations from [24,26], as well as consultations with therapists involved in this project, who emphasized the significance of these movements as indicators of rehabilitation progress in performing ADL. Reaching was assumed as the extension/flexion of the forearm in the x-y plane, lifting was described as a combined extension/flexion of the elbow and abduction/adduction of the shoulder, moving towards the mouth and back (similar to drinking task), and pro/supination was introduced as the inward and outward rotation of the wrist along the forearm axis. This controlled setup ensured consistent and objective data collection and it created a standardized training dataset.

The semi-structured dataset was collected using a kitchen activity protocol, where participants completed a set of predefined activities, related to meal preparation (Table 1). It incorporated different variations of the three previously mentioned arm movements within a mock kitchen environment, at Rijndam rehabilitation center. Each participant was required to complete all activities on the list, starting in front of a designated cabinet. However, during the task, they were free to adjust their body position relative to objects and surfaces and to perform movements in their own natural way. Variation in arm movement execution was intentional. For example: 1- Reaching included both reaching for high objects and placing objects on a flat surface, all labeled under the same movement category. 2- Lifting was defined as bringing the arm towards the mouth, which could occur while holding an object (e.g., drinking from a cup, or biting into bread) or without an object (e.g., washing the face). 3- Pronation/Supination was performed in different ways, such as rotating the forearm while spreading butter on bread or turning a water tap. Unlike the structured dataset, participants were not given specific instructions on where to sit, stand, how to perform the activity or to maintain their body position. This dataset allowed for more natural, functional movements that closely resemble real-life activities. While the sequence of tasks was predefined, individual execution varied, resulting in a more diverse and ecologically test dataset.

Moreover, a few additional activities, such as walking, standing, and sitting down, were included to create a logical flow and help participants feel more natural in their environment. Participants were asked to repeat the entire set of activities twice, in two separate trials. However, only one trial was primarily used for analysis, with the second trial reserved as a backup in case of any errors in activity performance.

Table 1
Activities performed in the semi-structured protocol.

#No	Activities	Category
1	Reach a piece of bread from the cabinet	Reaching high-A
–	*Walk to the table	Random
2	Place the bread on the table	Reaching low-A
–	*Walk back to the kitchen	Random
3	Open the water tap	Pro/Supination-C
4	Wash your face in the meantime	Lifting-B
5	Close the water tap	Pro/supination-C
6	Reach the cup from the dryer	Reaching-A
7	Take a sip	Lifting-B
–	*Walk to the table	Random
8	Put back the cup on the table	Reaching-A
–	* Take a seat	Random
9	Reach the knife from the table	Reaching-A
10	Spread butter on the bread using a Knife	Pro/supination-C
11	Eat the piece of a bread	Lifting-B
12	Turn the page of the book, look underneath	Pro/supination-C
13	Reach the cup from the table and hold it	Reaching-A
–	*Stand from the chair	Random
–	*Walk around the kitchen	Random
14	Put back the cup to the dryer	Reaching-A

2.4. Data labeling

To label the data, all movements were classified as either reaching (A), lifting (B), or pro/supination (C). Each movement was considered a forward–backward motion comprising two phases. For example, reaching was followed by a retraction, with the entire sequence classified as reaching.

Two researchers manually labeled all movements in both the structured and semi-structured trials using video recordings as the ground truth. The videos were synchronized with the IMU recordings based on the claps at the start and end of each trial. Some kitchen activities in the semi-structured trials involved composite movements. For instance, opening and closing a water tap involved both a reaching movement towards the tap and a subsequent pro/supination. In such cases, the start and end of the activity was annotated based on the intended movement, while the non-targeted movement was disregarded. For example, in the tap activity, only the pro/supination was labeled, and the reaching component was excluded.

The labels from structured trials were served to train and validate ML models, while those from semi-structured trials assessed test accuracy.

2.5. Data pre-processing

Before entering the data into a classifier, general pre-processing steps were performed.

Filtering. The raw sensor data were band-pass filtered using a 3rd-order Butterworth filter with low and high cutoff frequencies set at 0.1 Hz and 12 Hz, respectively [23,24]. This band-pass filter was designed to eliminate low-frequency artifacts and high-frequency noise introduced by physical effects, such as arm movement drifts.

Equal length. In this experiment, five IMU sensors simultaneously measured different arm segments and all IMUs transmitted data in unique-sized packets. However, desynchronization in the Bluetooth transmission rate of individual sensors occasionally resulted in some sensors losing a few data packets. To address this, we aligned the sensor data by truncating the longest sensor output to match the length of the shortest one.

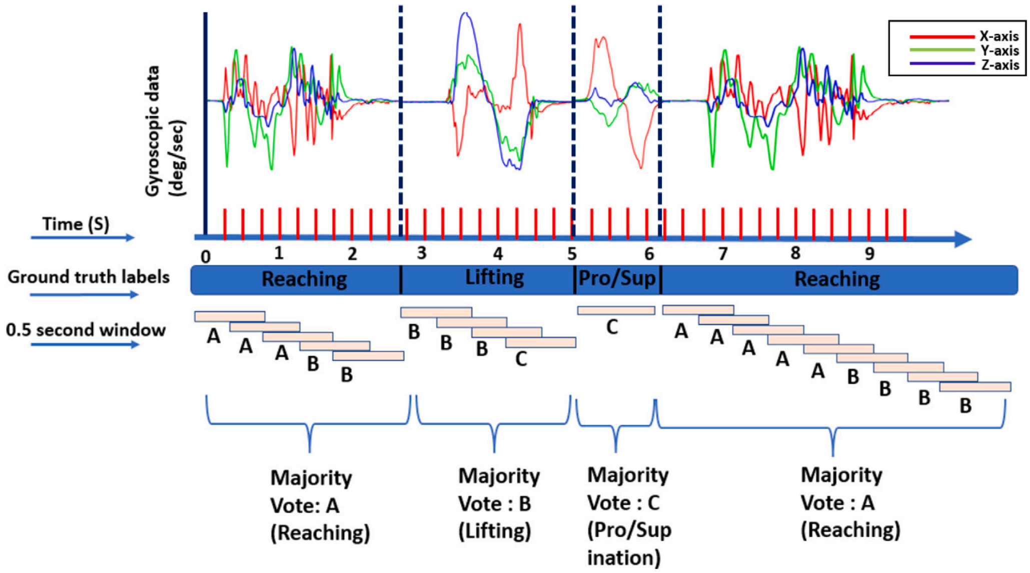
Gyroscopic data (deg/sec)
X-axis
Y-axis
Z-axis
Time (S)
0
1
2
3
4
5
6
7
8
9
Ground truth labels
Reaching
Lifting
Pro/Sup
Reaching
0.5 second window
A A A B B
B B B C
C
A A A A A B B B B
Majority Vote: A (Reaching)
Majority Vote : B (Lifting)
Majority Vote : C (Pro/Sup ination)
Majority Vote : A (Reaching)

Fig. 1. Segmentation of data by class (ground truth labels) and window size. Each data fragment consists of equal-length windows. The gyroscopic data is plotted to visually illustrate the segmentation process applied across the entire dataset.

Data segmentation. The data were originally segmented with the labels (movement classes) by aligning the IMU signals with the video recordings. However, this class-segmented data contained examples of the three movements with varying lengths, reflecting the individual differences in the speed of performing these motions. To generate uniform-length data segments for use by the ML models, every single movement was segmented into 0.5 s windows with 50% overlap between consecutive segments (Fig. 1). Segmenting data into windows of consistent length is generally recommended before applying many machine learning models [23–26]. The 0.5-s window was preferred over other tested window sizes (1 and 2 s) because it resulted in better ability of the selected models to distinguish between the three movement types.

2.6. Machine learning models for movement identification

In this study, we approached arm movement identification as a classification problem. A classification model aims to assign items to discrete groups (in this case, three arm movement types) based on a specific set of features. We evaluated two classification models, one based on Random Forest and one hybrid deep learning model. The rationale behind proposing these two models was to compare the performance of a more traditional ML model, in our case Random Forest (based on extracting relevant features manually from data), with a more sophisticated neural network architecture (in our case the hybrid model) which does not depend on manual feature extraction and optimization process. Both models were implemented in Python 3.11, using Scikit-learn (Random Forest) and Keras (for the hybrid model) libraries.

2.6.1. Random forest

Random Forest is an ensemble machine learning algorithm widely used for movement identification applications and for time series data classification [13,23,25,34–36]. Notably, results from [36] demonstrated that Random Forest outperformed other machine learning models in intra-subject classification of arm movements for both control and stroke groups (similar to our case with having subject-specific models).

Random Forest builds multiple decision trees by splitting data based on feature values, then aggregates their predictions to enhance accuracy and determine the final output [37].

Hyper-parameter optimization. The key hyper-parameters of the RF model include the number of decision trees in the forest ("n-estimators"), the maximum depth of each decision tree ("max-depth"), and the number of features to consider when determining the best split at each node ("max-features"). We employed a pragmatic grid search technique to optimize our RF model, testing a limited set of values for each hyper-parameter during the training and validation phases.

- rf-param-grid
  - 'n-estimators': [100, 200, 500]
  - 'max-depth': [10, 20]
  - 'max-features': ['sqrt', 5, 10]

For hyper-parameter optimization, we utilized the leave-one-out approach [38] to allocate the structured trials for training and validation. During each iteration, the model was trained using four trials of data from structured protocol while performing a grid search over various combinations of hyper-parameters. The model was then validated on the fifth structured trial. This was repeated five times, iterating over each trial assigned to the validation set. A flowchart illustrating the trial allocation and hyper-parameter tuning process is shown in Fig. 2.

Feature extraction. For the RF model, we selected 8 different types of commonly used signal-based features. These features were (1) the mean of the signal [9,13], (2) the standard deviation of signal [9,13,23], (3) the root mean square [10,23,35]: RMS = $\sqrt{\frac{1}{T}\int_0^T x^2(t)\,dt}$, (4) the minimum value of signal, (5) the maximum value of the signal, (6) the slope of the signal: [9], Slope = $\frac{dx(t)}{dt}$, (7) the skewness of the signal [23], Skewness = $\frac{\sum_{i=1}^{n}(x_i-\bar{x})^3/N}{\sigma^3}$, and (8) the kurtosis of the signal [23], Kurtosis = $\frac{\sum_{i=1}^{n}(x_i-\bar{x})^4/N}{\sigma^4}$.

Each feature was computed over each channel (X,Y,Z) of accelerometer and gyroscope, across four sensors located on hand, lower arm, upper arm and shoulder. This resulted in a total of eight(features)* three(axes)* two(sensor types)* four(sensor number) = 192 features. The feature extraction process was performed for each 0.5 s (25 samples in time) window segments contained within a specific data fragment, labeled as A, B or C. The same computation was performed for the secondary sensor configuration, involving only a wrist sensor and resulted in eight(features)*three(axes)*two(sensor types) = 48 features.

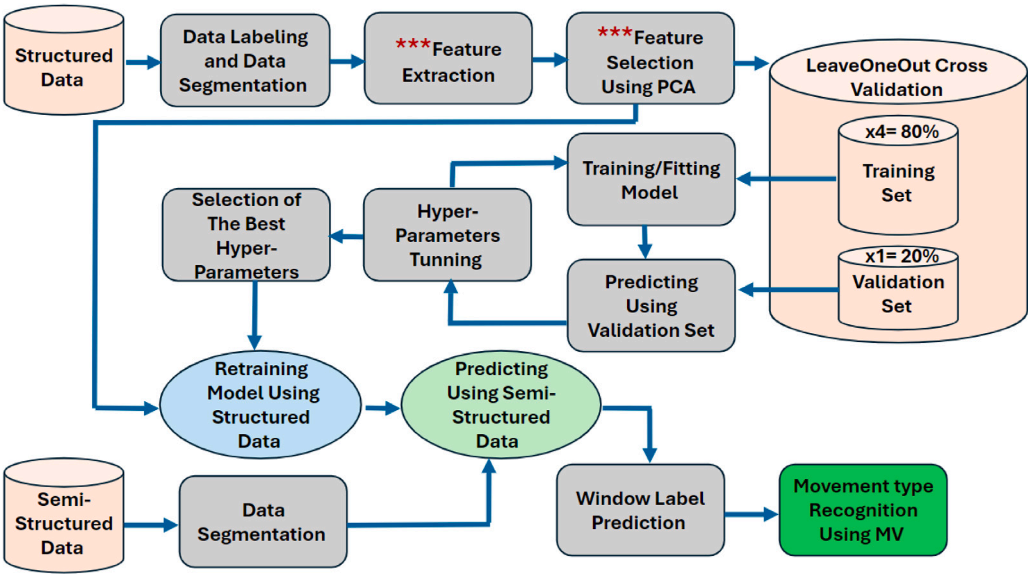
Structured Data
Data Labeling and Data Segmentation
***Feature Extraction
***Feature Selection Using PCA
LeaveOneOut Cross Validation
x4= 80%
Training Set
x1= 20%
Validation Set
Training/Fitting Model
Predicting Using Validation Set
Hyper-Parameters Tunning
Selection of The Best Hyper-Parameters
Retraining Model Using Structured Data
Predicting Using Semi-Structured Data
Semi-Structured Data
Data Segmentation
Window Label Prediction
Movement type Recognition Using MV

Fig. 2. Overview of the main steps for training and validating the machine learning models for three arm movement identification using leave-one-out cross-validation on structured trials and testing on semi-structured trial. Note that blocks with *** were applied only for running the RF model and not for the hybrid model.

Feature scaling. We used "preprocessing.StandardScaler" class from the "Scikit-learn" library to standardize features, ensuring they contribute equally to the model by centering them around the mean and scaling by the standard deviation.

Feature optimization. For feature optimization, we used Principal Component Analysis (PCA), a common feature reduction technique that projects a multi-dimensional feature matrix into a lower-dimensional feature matrix, while preserving the significant characteristics of the data [39]. Among all the extracted principal components (PC), the two first PC1 and PC2, have the most significant contribution to translate data into a lower-dimension. Each feature (each data point in the scatter plot) is plotted according to its loadings on PC1 and PC2, in a 2D plane. Features are considered more important if they are farther from the origin in the PC1-PC2 plane, as indicated by their Euclidean distance. We selected the twenty features most distant from the origin (Fig. 3). While PC1 and PC2 can be used directly to transform data into a lower dimension, this approach may obscure the relationship between the original features and their corresponding sensors [40].

Following the feature optimization process, the same twenty most important features identified during training and validation were also extracted from the semi-structured trial. This approach was chosen to establish a logical consistency between the two different settings, ensuring that the features used for classification were comparable across both structured and semi-structured settings.

2.6.2. Hybrid model (CNN+LSTM)

The second model we used was a variant of a previously proposed hybrid model [32,41], which combines convolutional neural network (CNN) and a long short-term memory (LSTM) network architectures (Fig. 4). Unlike Random Forest, the CNN+LSTM model does not require pre-defined features for classification, as it automatically learns internal representations that are useful for the classification task [23]. The rationale behind choosing this model was the success of CNNs in identifying arm movement activities [25–29], as well as the assumption that extracting both temporal dependencies and spatial features could enhance identification accuracy for sequential time series data. Moreover, Liang's study [42] demonstrated superior performance in predicting lower-limb joint moments during locomotive activities using their proposed BiLSTM+CNN model compared to other neural networks.

The CNN component extracts useful implicit features by convolving filters across the spatial dimension (over channels). However, CNNs are not designed to capture temporal dependencies across time-steps. To address this, the model includes an LSTM layer to learn temporal dependencies in the features extracted by CNN. The input to the hybrid model was a 0.5 s window (25 samples in time) with 24 data channels (The acceleration and gyroscope data from all the four sensors in X, Y, Z directions).

The model includes two 1D-convolution layers, each followed by a batch normalization layer to normalize the output of the previous layer, a max-pooling layer to reduce the input dimensions, and a ReLU activation function to introduce non-linearity to the model. The first convolution operation was done by a 1D-Conv layer with 32 filters, kernel size of 3 and a stride rate of 1. The second convolution was done by a 1D-Conv layer with 64 filters and the same kernel size and stride rate. The max-pooling layers had a pool size of 2. The model included an LSTM layer with 64 memory units included to capture temporal dependencies. The output module included two dense layers: the first with 256 neurons, and the second with three neurons, representing the three types of arm movements in this study. The final dense layer used a softmax activation function to handle the multi-class classification problem.

The hybrid model used Adam optimizer [43] with learning rate of 0.001 and a batch size of 32 trains over 10 epochs and it used categorical cross-entropy as a loss function.

The training and validation were performed over the structured trials with iteratively allocating one trial for validation and the other four trials for training. This results in 20% splitting data for cross-validation. After tuning the hyper-parameters, the model was re-trained once again on the entire structured trials. The general steps for training and validation are presented in Fig. 2.

2.7. Model evaluation

Each model was evaluated by training on structured trials and using the leave-one-out approach to fine tune the hyper-parameters during the validation phase (still on structured trials). For testing the models' performance, the semi-structured trial was used. This strategy for training on more constrained setting and test on less constrained setting allowed us to evaluate how well the models generalize to more challenging scenarios. This was done to gauge the potential of model

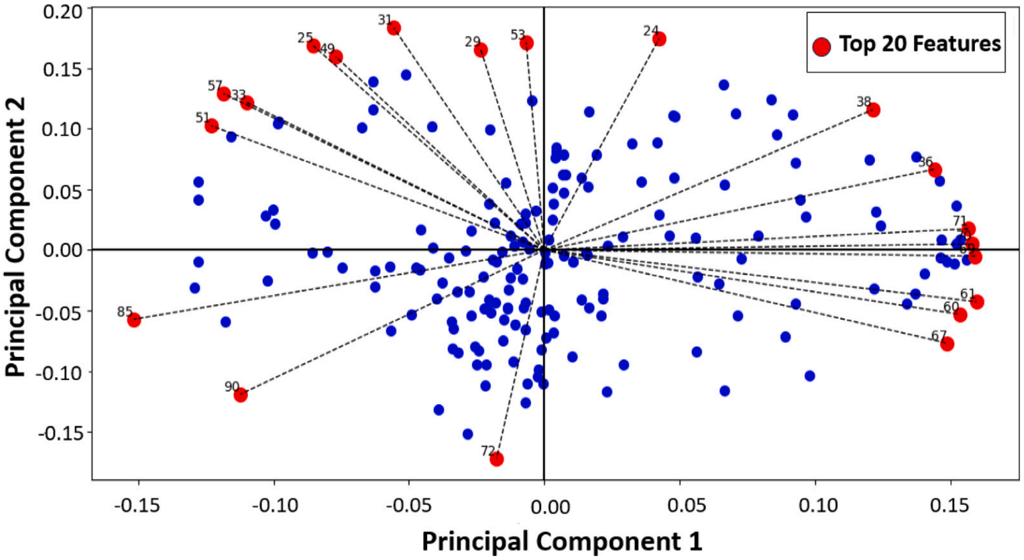

Fig. 3. Selection of the 20 most important features based on PC1 and PC2 loadings and their distance from the origin. The red dots indicate the selected features and their corresponding numbers out of the total 192 features.

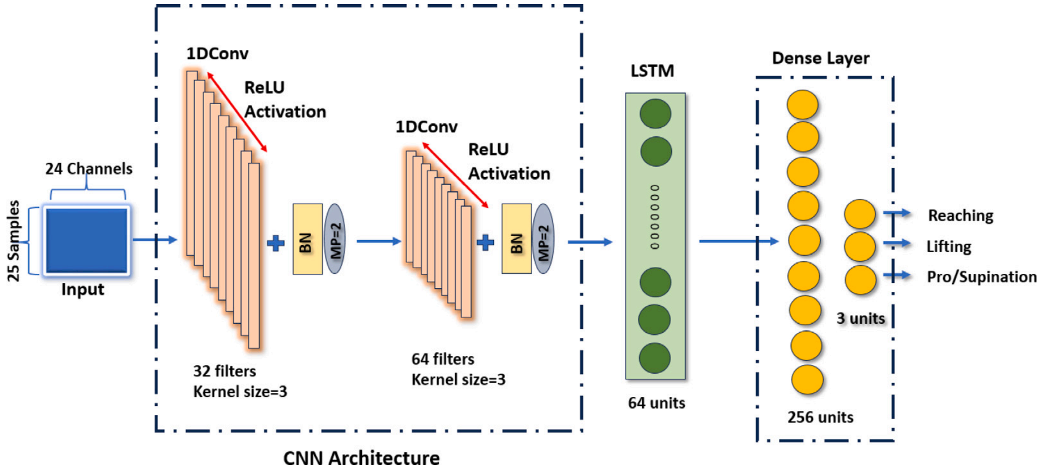

Fig. 4. Hybrid model architecture. The input to the model is a window of (25Samples*24Channels). Blocks BN and MP refer to Batch Normalization and MaxPooling layers, respectively. This architecture is a new variant of previously proposed model by Li M, and colleagues.et al. [32].

applicability in more realistic settings, where the movements performed during training and testing can differ significantly.

Along with the general idea of using separate datasets for training and testing, we evaluated our models using two distinct approaches:

Subject-specific approach:. This is defined as using all structured trials from one subject for training/validation, while the corresponding semi-structured trial from the same subject is used for testing.

Group approach:. This approach is defined as using the first trial of structured data from all participants together for training/validation, while a randomly selected participant's semi-structured trial is used for testing. This setup ensures that the model can adapt to data from different individuals. The approach is further divided into two variations: either the target subject, whose semi-structured trial is used for testing, is included in the training process or excluded.

2.7.1. Evaluation metrics

The overall model performance was evaluated with the *accuracy* metric that measures the proportion of correctly predicted instances relative to the total number of instances. However, for the multi-class classification problem with possible class imbalances and potential differences in accuracies between different classes, it is crucial to evaluate the classification performance of each class separately. For this, we used sensitivity (also known as recall), which measures the proportion of correctly predicted movements in a specific class with respect to the total number of actual movements in that class [44]:

$$S = \frac{TP}{TP + FN} \times 100\%,$$

where TP is the number of true positive and FN is the number of false negative classification for each class.

Furthermore, we used balanced accuracy which provides a holistic view of model performance for multi-class classification problems:

$$\text{Balanced Accuracy} = \frac{(S_A + S_B + S_C)}{3} \times 100\%,$$

where $S_{A,B,C}$ are sensitivities for each class.

We evaluated model performance in two different scenarios:

1. **Predicting class labels of each 0.5 s window**. Here we used balanced accuracy for (a) validation/training (on structured trials), and (b) testing (on the semi-structured trial).

2. **Predicting the type of entire arm movements (activity prediction).** Since we performed the movement identification at a high time granularity (window size of 0.5 s with 50% overlap), the window labels prediction cannot be directly translated into movement identification. Therefore, we averaged out the labels for all the windows included within a single movement, using the majority of the votes procedure [45] (Fig. 1). For each movement, label predictions based of all 0.5 s time windows fully contained within that movement were considered, and the label with the highest occurrence (most votes) was assigned to the entire movement. In cases where there was an equal number of votes for different classes, the label of the middle window of that data fragment was selected as the final classification label. The majority of votes approach was employed to predict the full set of isolated arm movement labels during structured trials as well as the labels for kitchen activities during the semi-structured trial. Additionally, after predicting the kitchen activity labels in semi-structured trial, the sensitivities per each class of movement were calculated.

3. Results

3.1. Generalization from structured to semi-structured environments

3.1.1. Subject-specific models

We found that both models (Tables 2, 3), achieved near perfect balanced accuracy in predicting window labels during structured trials. However, their accuracy decreased notably to predict window labels in semi-structured trial (99.15% vs. 69.08% for RF and 96.51% vs. 58% for hybrid model).

Despite the substantial drop in balanced accuracy for window labels prediction in semi-structured trial, employing the majority of votes generalized activity predictions throughout distinct settings, effectively (86.54% for RF and 87.96% for hybrid model). These improvements might be due to the fact that models may have shown a higher error in predicting the labels of a shorter fragment of data, but taking the average of these sequential data predictions (majority of votes) can help mitigating these errors.

We found no evidence of major differences between the models in activity prediction balanced accuracy, considering the subject-specific models (Tables 2, 3) (87.96% for the hybrid model vs. 86.54% for RF). However, RF demonstrated higher balanced accuracy for window predictions in semi-structured trial than the hybrid model (69.08% vs. 58%). This suggests that, although the hybrid model is less accurate than RF in predicting individual windows, it may slightly outperform RF when the task shifts to predicting labels for sequential data. Additionally, the hybrid model significantly improved sensitivity in detecting reaching movements compared to RF, with an increase of 11.9%. This movement was designed in our experiment to have the most variation (high-reach, low-reach in kitchen activities) among all the other movements. However, this improvement was accompanied by a 5.56% decrease in sensitivity for predicting lifting movements.

3.1.2. Group models

For both variations of group models, (1-including the target subject in the training/validation process and 2-excluding the target subject from training/validation), the generalization from structured to semi-structured trial performed similarly to subject-specific models, exhibiting higher training/validation accuracy but lower test accuracy (Table 4).

Furthermore, we found that both RF and hybrid group models performed worse than subject-specific models (86.54% vs. 77.37% for RF and 87.96% vs. 82.96% for the hybrid model) (Table 4) in predicting the activity labels. However, the performance reduction was smaller for the hybrid model compared to RF, indicating a better ability of the hybrid model to generalize across random individuals.

Furthermore, similar to the hybrid subject-specific model, the hybrid group model improved the sensitivity in detecting reaching movements compared to the RF group model, showing a 45.24% increase.

3.2. Comparison of sensor configurations

Our analyses revealed that the wrist-only sensor configuration performed worse than the four-sensor configuration for both RF and hybrid subject-specific and group models, but not by a large margin (Figs. 5, 6). The average accuracy in predicting activities decreased by 10.57% and 26.22% when using a single wrist IMU sensor, compared to using four IMUs, as evaluated by the RF and hybrid subject-specific models, respectively. This was followed by a 3.24% and 14.32% decrease when considering the RF and hybrid group models, respectively. Compared to their subject-specific counterparts, both models – especially the RF model – are less sensitive to a reduction in the number of sensors.

We also found that the variability in accuracy across subjects was more pronounced for the hybrid model (regardless of whether a subject-specific or group approach was used) than for the RF model (Figs. 5, 6). This suggests that the hybrid model's performance is more sensitive to individual differences when only one sensor is used. Despite the decrease in classification accuracy with a single sensor, the RF subject-specific (75.97%) and RF group models (74.135%) continued to produce reasonably sufficient results.

4. Discussion

In this paper, we investigated the generalization of machine-learning-based models for identification of various arm activities, including three arm movement types – reaching, lifting, and pro/supination – in healthy adults from a structured clinic-like setting to a semi-structured home-like environment. We trained two established machine learning models: a Random Forest (RF) model and a state-of-the-art hybrid deep learning model [32]. We evaluated these models following two distinct approaches: 1-Subject-Specific approach and 2- Group approach. These machine learning models analyzed data from two sensor configurations: a single IMU on the dominant wrist and four IMUs placed along the dominant arm.

Our results demonstrated that both RF and hybrid models, following both evaluation approaches (subject-specific and group approaches) were remarkably accurate when predicting movements using structured trials, with balanced accuracy close to 100%. However, when applied to the data from a more natural, semi-structured protocol without prior exposure to that data, the prediction accuracy declined, reflecting the challenge of identifying arm movements in less controlled environments. At the same time, the performance was only reduced by 12 to 14 percentage points for subject-specific models and 13 to 18 percentage points for group models, resulting in average balanced accuracies around 86% to 88%, and 77% to 82%, respectively. This suggests that using the majority of votes method to predict entire activity labels significantly improved balanced accuracy, particularly within the semi-structured trial. Building on these promising results, future studies could benefit from incorporating more realistic simulations of real-life scenarios, allowing participants to engage in activities of their interest.

In general, both RF and hybrid models performed better with the subject-specific approach than with the group approach. This suggests that while the group approach increases the amount of training data, it does not necessarily improve performance, as the increased variability among subjects introduces additional challenges.

In comparing the performance of the two selected models (RF and Hybrid), there were two instances where the hybrid model slightly outperformed the RF model. First, the hybrid model performed better when combined with majority voting to predict activity labels across sequences. This suggests that the hybrid model may be better suited for tasks that require capturing the temporal dynamics of arm movements, particularly for continuous activity identification. Second, the group hybrid model experienced a smaller decrease in accuracy compared to the subject-specific hybrid model, whereas the accuracy drop was more pronounced in the RF model. This suggests that the hybrid model is a better candidate for scenarios where insufficient data is available from

Table 2
Evaluation results for the Random Forest Subject-Specific model. The balanced accuracy for predicting movement types in structured trials approached 100% for all subjects, and hence is not shown.

Subject	Windows prediction balanced accuracy (%)		Activity prediction balanced accuracy (%)			
	Training/Validation (structured trials)	Test (semi-structured trial)	Test (semi-structured trial)	Sensitivity (%), Test (semi-structured trial)		
				Reaching (A)	Lifting (B)	Pro/supination (C)
S1	99.15	74.82	88.89	100	66.67	100
S2	99.63	61.20	77.38	57.14	100	75
S3	98.73	76.98	95.23	85.71	100	100
S4	98.31	68.45	84.12	85.71	66.67	100
S5	99.43	67.70	85.71	57.14	100	100
S6	98.64	64.36	84.12	85.17	66.67	100
S7	98.84	83.62	100	100	100	100
S8	99.17	64.28	79.36	71.43	66.67	100
S9	100	65.71	84.12	85.71	66.67	100
S10	99.81	61.05	85.71	57.14	100	100
S11	99.70	63.37	82.14	71.43	100	75
S12	98.39	77.47	91.66	100	100	75
Total	99.15 ± 0.57	69.08 ± 7.35	86.54 ± 6.48	79.76 ± 16.61	86.11 ± 17.16	93.75 ± 11.30

Table 3
Evaluation results for the hybrid deep learning subject-specific model. The balanced accuracy for predicting movement types in structured trials approached 100% for all subjects, and hence is not shown.

Subject	Windows prediction balanced accuracy (%)		Activity prediction balanced accuracy (%)			
	Training/Validation (structured trials)	Test (semi-structured trial)	Test (semi-structured trial)	Sensitivity (%), Test (semi-structured trial)		
				Reaching (A)	Lifting (B)	Pro/supination (C)
S1	96.06	57.59	100	100	100	100
S2	98.68	61.52	88.89	100	66.67	100
S3	98.41	53.01	83.33	100	100	50
S4	96.09	61.22	72.22	100	66.67	50
S5	96.00	62.30	88.89	100	66.67	100
S6	97.58	63.75	88.89	100	66.67	100
S7	93.20	60.24	90.47	71.43	100	100
S8	96.40	51.07	90.47	71.43	100	100
S9	96.17	54.17	73.01	85.71	33.33	100
S10	94.74	46.70	79.36	71.43	66.67	100
S11	97.14	62.90	100	100	100	100
S12	97.72	61.60	100	100	100	100
Total	96.51 ± 1.54	58.00 ± 5.49	87.96 ± 9.61	91.66 ± 12.86	80.55 ± 22.28	91.66 ± 19.46

Table 4
Evaluation results for comparison between subject-specific models and group models.

Performance	RF balanced accuracy (%)					Hybrid balanced accuracy (%)				
	Training/ Validation (structured)	Activity prediction test (semi-structured)	Sensitivity			Training/ Validation (structured)	Activity prediction test (semi-structured)	Sensitivity		
			A	B	C			A	B	C
Subject- Specific models	99.15	86.54	79.76	86.11	93.75	96.51	87.96	91.66	80.55	91.66
Group models target-Subject excluded in training	95.38	77.37	48.80	91.66	91.66	95.67	82.96	94.04	86.10	68.75
Group models target-Subject included in training	95.15	77.27	46.42	91.66	93.75	95.65	82.6	96.42	80.55	70.83

the same subject, as it can generate reasonably accurate predictions using movement data from other subjects.

We acknowledge that more advanced models could potentially perform better than the ones we considered. For instance, transformer-based models [46] have shown promise in time-series classification by capturing long-range dependencies through a self-attention mechanism. However, since little is known about their use in upper arm movement classification, we chose models that have been previously reported for this task to mainly focus on generalization across different setups. Moreover, our findings suggest that simpler models can still achieve strong generalization performance.

Regarding sensor configuration, our results showed that four measuring sensors (IMU), and therefore more comprehensive movement data from different arm segments, improved the movement identification accuracy compared to a single wrist sensor only. However, increasing the number of sensors may increase the computational load due to the larger volume of data to process. According to the work of Parnandi [22], the best classification performance during a rehabilitation-like activity was achieved using seven IMUs placed on the pelvis, sternum, head, active shoulder, upper arm, forearm, and hand, out of a total of eleven possible sensors. This aligns with our findings, as we excluded sensors on the pelvis, sternum, and head to focus specifically on arm movements, which was the objective of our study.

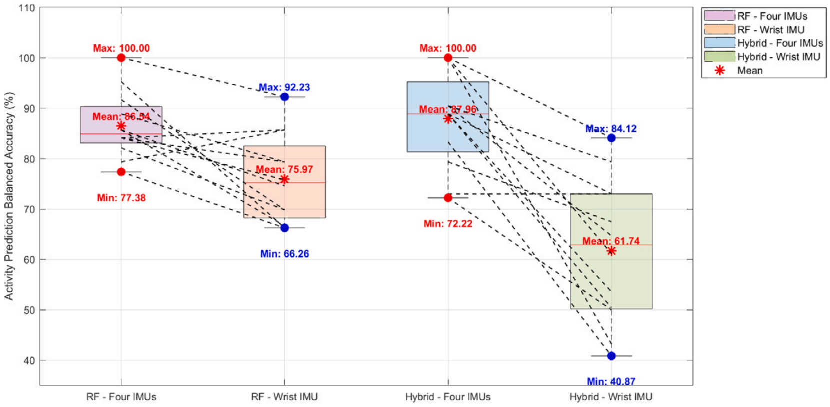

Fig. 5. Subject-Specific models: Comparison of activity prediction balanced accuracies between two sensor configurations using both RF and Hybrid models. Dashed lines connect values corresponding to each participant.

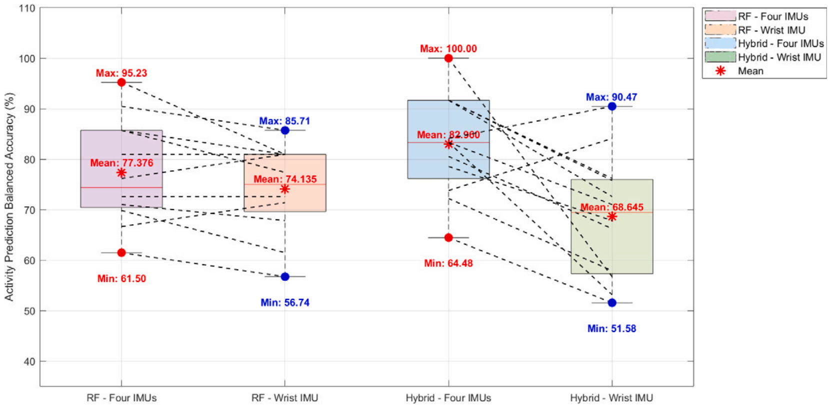

Fig. 6. Group models: Comparison of activity prediction balanced accuracies between two sensor configurations using both RF and Hybrid models. Dashed lines connect values corresponding to each participant.

Still, our findings suggest that even one wrist sensor can provide data to distinguish between three considered arm movements with balanced accuracy approx. 75% (for the RF model), with the chance level for ternary classification being 33%. This highlights the strength of our study, as it adds insights into the trade-off between complexity and accuracy for practical arm monitoring when using distinct datasets.

Our results bring an important contribution to this field of research, as few studies have evaluated the generalization of arm movement identification, using distinct datasets for training and testing phases. Gomez et al. [23] distinguished reach from non-reach movements of stroke patients (binary classification), with models tested on an activities-of-daily-life (pizza-making) dataset, and they reported 74.8% to 76.5% test accuracies.

As the key contribution of our work, we expanded the number of arm movements to include three distinct types (Reaching, Lifting, and Pronation/Supination). Moreover, we expanded the scenario to generalize across random subjects while ensuring that still the training and testing environments are entirely separated (structured vs. semi-structured protocols). This forced the ML models to handle greater variability among different settings and individuals, making movement identification more robust in real-life scenarios (e.g., kitchen activities). Considering that we had three classes (reaching, lifting, pro/supination) rather than two, our results add possible interpretations of differences in accuracy. Furthermore, the sensitivity for detecting reaching movements in our study, following the subject-specific approach (79.76% for the RF model and 91.66% for the hybrid model), was notably higher than the sensitivities reported in Gomez study (43.0% to 46.9%). This suggests a significant improvement in the sensitivity of reaching movement detection in our approach.

5. Limitations

This study was conducted on healthy participants as a proof of concept, making it challenging to directly translate these results to stroke patients. Since stroke patients often experience motor control impairments and muscle weakness, these factors could impact model performance. However, this study establishes a framework for generalizing movement identification across different settings by evaluating both subject-specific models (trained and tested on individual participant) and group models (trained on all participants and tested on a random subject). By demonstrating strong results from both strategies and carefully considering the use of separate datasets, we aimed to enhance the robustness of our methodology in identifying arm movements across various conditions (generalized across settings) and among different subjects (generalized across individuals or unseen subject). We believe that our contribution is an important step towards real-world application of arm activity identification in daily life for tele-rehabilitation by highlighting that such application can be possible without the need for prior exposure to similar movement data or a substantial amount of data from a single participant.

Additionally, this study relied on manual annotation by two researchers for all structured and semi-structured trials, which presented a few challenges. The annotation process was time-intensive, requiring individual labeling for each trial, and thus making it less practical for large datasets and real-time applications. At the same time, this manual annotation was only required for evaluating the generalization of the algorithms for research purposes, as it would allow us to evaluate the model performances more reliably; for the practical application of our results, we will not envision the need for manual annotation for daily life. Specifically, we envision that the model will need to be trained on previously collected and annotated patient data but then can potentially operate in a real-life setting on raw data without any annotations or prior knowledge of movement duration.

6. Conclusions

Overall, our study demonstrates the ability of machine learning models to accurately generalize the identification of arm activities based on IMU data when trained and tested in distinct settings. Our results contribute to addressing the challenge of identifying unseen arm activities in home-based rehabilitation. We believe that at a later stage, the clinical highlights of this study can be formulated in two ways: First, determining whether it is possible to identify specific arm activities in daily life by training models on simple, straightforward arm movements in constrained settings, without requiring the patient to follow extensive and burdensome practices. Second, evaluating the trade-off between high identification accuracy (while still maintaining distinct settings for training and testing phases) and the minimal sensor configuration.

Although both models' performance declined when using a single wrist sensor, the Random Forest (RF) model maintained reasonable accuracy within its range. In contrast, the hybrid model was more successful in capturing temporal dependencies due to its memory units, making it more effective when used with the majority vote approach. Additionally, the hybrid model showed greater promise for scenarios when limited data is available from the same subject, as it generalized better across different subjects [47].

CRediT authorship contribution statement

Sahel Akbari: Project administration, Visualization, Writing – original draft, Validation, Methodology, Software, Investigation, Resources, Formal analysis, Conceptualization. **Herwin L.D. Horemans:** Writing – review & editing, Investigation. **Johannes B.J. Bussmann:** Investigation, Writing – review & editing, Funding acquisition, Resources. **Arkady Zgonnikov:** Formal analysis, Conceptualization, Validation, Supervision, Writing – review & editing, Visualization, Investigation.

Ethical approval

This work involved human subjects in its research. All the ethical procedures and experimental protocols were approved by TU Delft Human Research Ethics Committee (HREC) under application number 4189.

Funding

The study was funded by Erasmus MC TKI-LSH (project number: EMCLSHM22015), with financial support from Rijndam Rehabilitation, The Netherlands.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

We would like to thank Ruben Regterschot for his role in the initial planning of the project, Rijndam Rehabilitation, and all participants for their cooperation in our measurements. We would also like to thank Christian Reinprecht and Alba Riera Cardona for their help in data labeling.

Data availability

Open dataset and analysis are available upon request.

References

[1] World Health Organization, World health organization, the top 10 causes of death, 2023, URL https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.

[2] E.J. Benjamin, M.J. Blaha, S.E. Chiuve, M. Cushman, S.R. Das, R. Deo, S.D. de Ferranti, J. Floyd, M. Fornage, C. Gillespie, C.R. Isasi, M.C. Jiménez, L.C. Jordan, S.E. Judd, D. Lackland, J.H. Lichtman, L. Lisabeth, S. Liu, C.T. Longenecker, R.H. Mackey, K. Matsushita, D. Mozaffarian, M.E. Mussolino, K. Nasir, R.W. Neumar, L. Palaniappan, D.K. Pandey, R.R. Thiagarajan, M.J. Reeves, M. Ritchey, C.J. Rodriguez, G.A. Roth, W.D. Rosamond, C. Sasson, A. Towfighi, C.W. Tsao, M.B. Turner, S.S. Virani, J.H. Voeks, J.Z. Willey, J.T. Wilkins, J.H.Y. Wu, H.M. Alger, S.S. Wong, P. Muntner, Heart disease and stroke statistics—2017 update: A report from the American heart association, Circulation 135 (10) (2017) e146–e603, http://dx.doi.org/10.1161/CIR.0000000000000485, URL https://www.ahajournals.org/doi/abs/10.1161/CIR.0000000000000485.

[3] G. Kwakkel, C. Stinear, B. Essers, M. Munoz-Novoa, M. Branscheidt, R. Cabanas-Valdés, S. Lakičević, S. Lampropoulou, A.R. Luft, P. Marque, S.A. Moore, J.M. Solomon, E. Swinnen, A. Turolla, M.A. Murphy, G. Verheyden, Motor rehabilitation after stroke: European stroke organisation (ESO) consensus-based definition and guiding framework, Eur. Stroke J. 8 (4) (2023) 880–894, http://dx.doi.org/10.1177/23969873231191304.

[4] G.B. Bonifacio, N.S. Ward, H.C.A. Emsley, J. Cooper, J. Bernhardt, Optimising rehabilitation and recovery after a stroke, Pr. Neurol. 22 (6) (2022) 478–485, http://dx.doi.org/10.1136/practneurol-2021-003004, URL https://pn.bmj.com/content/22/6/478.

[5] A.Y. Gelaw, B. Janakiraman, B.F. Gebremeskel, H. Ravichandran, Effectiveness of home-based rehabilitation in improving physical function of persons with stroke and other physical disability: A systematic review of randomized controlled trials, J. Stroke Cerebrovasc. Dis. 29 (6) (2020) 104800, http://dx.doi.org/10.1016/j.jstrokecerebrovasdis.2020.104800, URL https://www.sciencedirect.com/science/article/pii/S1052305720301841.

[6] J. Chen, M. Liu, D. Sun, Y. Jin, T. Wang, C. Ren, Effectiveness and neural mechanisms of home-based telerehabilitation in patients with stroke based on fMRI and DTI: A study protocol for a randomized controlled trial, Medicine 97 (3) (2018) URL https://journals.lww.com/md-journal/fulltext/2018/01190/effectiveness_and_neural_mechanisms_of_home_based.19.aspx.

[7] Veerbeek Janne Marieke, Erwin van Wegen, Roland van Peppen, van der Wees Philip Jan, Rietberg Marc, Hendriks Erik, Kwakkel Gert, What is the evidence for physical therapy poststroke? A systematic review and meta-analysis, PLoS One 9 (2) (2014) 1, http://dx.doi.org/10.1371/journal.pone.0087987.

[8] C.E. Lang, J.R. MacDonald, D.S. Reisman, L. Boyd, T. Jacobson Kimberley, S.M. Schindler-Ivens, T.G. Hornby, S.A. Ross, P.L. Scheets, Observation of amounts of movement practice provided during stroke rehabilitation, Arch. Phys. Med. Rehabil. 90 (10) (2009) 1692–1698, http://dx.doi.org/10.1016/j.apmr.2009.04.005, URL https://www.sciencedirect.com/science/article/pii/S0003999309003530.

[9] P.-W. Chen, N.A. Baune, I. Zwir, J. Wang, V. Swamidass, A.W.K. Wong, Measuring activities of daily living in stroke patients with motion machine learning algorithms: A pilot study, Int. J. Environ. Res. Public Heal. 18 (4) (2021) http://dx.doi.org/10.3390/ijerph18041634, URL https://www.mdpi.com/1660-4601/18/4/1634.

[10] S.H. Roy, M.S. Cheng, S.S. Chang, J. Moore, G.D. Luca, S.H. Nawab, C.J.D. Luca, A combined sEMG and accelerometer system for monitoring functional activity in stroke, IEEE Trans. Neural Syst. Rehabil. Eng. 17 (6) (2009) 585–594, http://dx.doi.org/10.1109/TNSRE.2009.2036615.

[11] R.R. Bailey, J.W. Klaesner, C.E. Lang, Quantifying real-world upper-limb activity in nondisabled adults and adults with chronic stroke, Neurorehabil. Neural Repair 29 (10) (2015) 969–978, http://dx.doi.org/10.1177/1545968315583720.

[12] G.R.H. Regterschot, J.B.J. Bussmann, M.H.J. Fanchamps, C.G.M. Meskers, G.M. Ribbers, R.W. Selles, Objectively measured arm use in daily life improves during the first 6 months poststroke: A longitudinal observational cohort study, J. NeuroEng. Rehabil. 18 (1) (2021) 51, http://dx.doi.org/10.1186/s12984-021-00847-x.

[13] E.M. Bochniewicz, G. Emmer, A. McLeod, J. Barth, A.W. Dromerick, P. Lum, Measuring functional arm movement after stroke using a single wrist-worn sensor and machine learning, J. Stroke Cerebrovasc. Dis. 26 (12) (2017) 2880–2887, http://dx.doi.org/10.1016/j.jstrokecerebrovasdis.2017.07.004, URL https://www.sciencedirect.com/science/article/pii/S1052305717303610.

[14] D. Rand, J.J. Eng, Predicting daily use of the affected upper extremity 1 year after stroke, J. Stroke Cerebrovasc. Dis. 24 (2) (2015) 274–283, http://dx.doi.org/10.1016/j.jstrokecerebrovasdis.2014.07.039, URL https://www.sciencedirect.com/science/article/pii/S1052305714003693.

[15] N. Balestra, G. Sharma, L.M. Riek, A. Busza, Automatic identification of upper extremity rehabilitation exercise type and dose using body-worn sensors and machine learning: A pilot study, Digit. Biomark. 5 (2) (2021) 158–166, http://dx.doi.org/10.1159/000516619.

[16] H.M. Schambra, A. Parnandi, N.G. Pandit, J. Uddin, A. Wirtanen, D.M. Nilsen, A taxonomy of functional upper extremity motion, Front. Neurol. 10 (2019) http://dx.doi.org/10.3389/fneur.2019.00857, URL https://www.frontiersin.org/journals/neurology/articles/10.3389/fneur.2019.00857.

[17] R.H. Da-Silva, S.A. Moore, C.I. Price, Self-directed therapy programmes for arm rehabilitation after stroke: A systematic review, Clin. Rehabil. 32 (8) (2018) 1022–1036, http://dx.doi.org/10.1177/0269215518775170.

[18] M. Mundt, A. Koeppe, S. David, T. Witter, F. Bamer, W. Potthast, B. Markert, Estimation of gait mechanics based on simulated and measured IMU data using an artificial neural network, Front. Bioeng. Biotechnol. 8 (2020) http://dx.doi.org/10.3389/fbioe.2020.00041, URL https://www.frontiersin.org/journals/bioengineering-and-biotechnology/articles/10.3389/fbioe.2020.00041.

[19] S.E. Grace J. Kim Avinash Parnandi, H. Schambra, The use of wearable sensors to assess and treat the upper extremity after stroke: A scoping review, Disabil. Rehabil. 44 (20) (2022) 6119–6138, http://dx.doi.org/10.1080/09638288.2021.1957027.

[20] S.P. Marco Iosa Pietro Picerno, G. Morone, Wearable inertial sensors for human movement analysis, Expert. Rev. Med. Devices 13 (7) (2016) 641–659, http://dx.doi.org/10.1080/17434440.2016.1198694.

[21] Q. Wang, P. Markopoulos, B. Yu, W. Chen, A. Timmermans, Interactive wearable systems for upper body rehabilitation: A systematic review, J. NeuroEng. Rehabil. 14 (1) (2017) 20, http://dx.doi.org/10.1186/s12984-017-0229-y.

[22] A. Parnandi, J. Uddin, D.M. Nilsen, H.M. Schambra, The pragmatic classification of upper extremity motion in neurological patients: A primer, Front. Neurol. 10 (2019) http://dx.doi.org/10.3389/fneur.2019.00996, URL https://www.frontiersin.org/journals/neurology/articles/10.3389/fneur.2019.00996.

[23] J.P. Gomez-Arrunategui, J.J. Eng, A.J. Hodgson, Monitoring arm movements post-stroke for applications in rehabilitation and home settings, IEEE Trans. Neural Syst. Rehabil. Eng. 30 (2022) 2312–2321, http://dx.doi.org/10.1109/TNSRE.2022.3197993.

[24] D. Biswas, A. Cranny, N. Gupta, K. Maharatna, J. Achner, J. Klemke, M. Jöbges, S. Ortmann, Recognizing upper limb movements with wrist worn inertial sensors using k-means clustering classification, Hum. Mov. Sci. 40 (2015) 59–76, http://dx.doi.org/10.1016/j.humov.2014.11.013, URL https://www.sciencedirect.com/science/article/pii/S0167945714002115.

[25] A. Kaku, A. Parnandi, A. Venkatesan, N. Pandit, H. Schambra, C. Fernandez-Granda, Towards data-driven stroke rehabilitation via wearable sensors and deep learning, in: Machine Learning for Healthcare Conference, PMLR, 2020, pp. 143–171.

[26] M. Panwar, D. Biswas, H. Bajaj, M. Jöbges, R. Turk, K. Maharatna, A. Acharyya, Rehab-net: Deep learning framework for arm movement classification using wearable sensors for stroke rehabilitation, IEEE Trans. Biomed. Eng. 66 (11) (2019) 3026–3037, http://dx.doi.org/10.1109/TBME.2019.2899927.

[27] S.H. Chae, Y. Kim, K.-S. Lee, H.-S. Park, Development and clinical evaluation of a web-based upper limb home rehabilitation system using a smartwatch and machine learning model for chronic stroke survivors: Prospective comparative study, JMIR Mhealth Uhealth 8 (7) (2020) e17216, http://dx.doi.org/10.2196/17216, URL http://www.ncbi.nlm.nih.gov/pubmed/32480361.

[28] Y. Chen, Y. Xue, A deep learning approach to human activity recognition based on single accelerometer, in: 2015 IEEE International Conference on Systems, Man, and Cybernetics, 2015, pp. 1488–1492, http://dx.doi.org/10.1109/SMC.2015.263.

[29] M. Panwar, S.R. Dyuthi, K.C. Prakash, D. Biswas, A. Acharyya, K. Maharatna, A. Gautam, G.R. Naik, CNN based approach for activity recognition using a wrist-worn accelerometer, in: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC, 2017, pp. 2438–2441, http://dx.doi.org/10.1109/EMBC.2017.8037349.

[30] A. Murad, J.-Y. Pyun, Deep recurrent neural networks for human activity recognition, Sensors 17 (11) (2017) http://dx.doi.org/10.3390/s17112556, URL https://www.mdpi.com/1424-8220/17/11/2556.

[31] R.J.M. Lemmens, Y.J.M. Janssen-Potten, A.A.A. Timmermans, R.J.E.M. Smeets, H.A.M. Seelen, Recognizing complex upper extremity activities using body worn sensors, PLoS One 10 (3) (2015) e0118642–, URL https://doi.org/10.1371/journal.pone.0118642.

[32] M. Li, J. Wang, S. Yang, J. Xie, G. Xu, S. Luo, A CNN-LSTM model for six human ankle movements classification on different loads, Front. Hum. Neurosci. 17 (2023) http://dx.doi.org/10.3389/fnhum.2023.1101938, URL https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2023.1101938.

[33] 2M Engineering, 9DOF motion tracking sensors, 2024, URL https://www.2mel.nl/wearable-product-development/motion-tracking-sensors.

[34] S. Kashi, R.F. Polak, B. Lerner, L. Rokach, S. Levy-Tzedek, A machine-learning model for automatic detection of movement compensations in stroke patients, IEEE Trans. Emerg. Top. Comput. 9 (3) (2021) 1234–1247, http://dx.doi.org/10.1109/TETC.2020.2988945.

[35] S.I. Lee, C.P. Adans-Dester, M. Grimaldi, A.V. Dowling, P.C. Horak, R.M. Black-Schaffer, P. Bonato, J.T. Gwin, Enabling stroke rehabilitation in home and community settings: A wearable sensor-based approach for upper-limb motor training, IEEE J. Transl. Eng. Heal. Med. 6 (2018) 1–11, http://dx.doi.org/10.1109/JTEHM.2018.2829208.

[36] P.S. Lum, L. Shu, E.M. Bochniewicz, T. Tran, L.-C. Chang, J. Barth, A.W. Dromerick, Improving accelerometry-based measurement of functional use of the upper extremity after stroke: Machine learning versus counts threshold method, Neurorehabil. Neural Repair 34 (12) (2020) 1078–1087, http://dx.doi.org/10.1177/1545968320962483.

[37] A. Criminisi, J. Shotton, E. Konukoglu, Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning, now, 2012, p. 1, http://dx.doi.org/10.1561/0600000035, URL http://ieeexplore.ieee.org/document/8187032.

[38] T.-T. Wong, Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation, Pattern Recognit. 48 (9) (2015) 2839–2846, http://dx.doi.org/10.1016/j.patcog.2015.03.009, URL https://www.sciencedirect.com/science/article/pii/S0031320315000989.

[39] S. Alavi, D. Arsenault, A. Whitehead, Quaternion-based gesture recognition using wireless wearable motion capture sensors, Sensors 16 (5) (2016) http://dx.doi.org/10.3390/s16050605, URL https://www.mdpi.com/1424-8220/16/5/605.

[40] I.T. Jolliffe, Graphical representation of data using principal components, Princ. Compon. Anal. (2002) 78–110.

[41] M.A. Khatun, M.A. Yousuf, S. Ahmed, M.Z. Uddin, S.A. Alyami, S. Al-Ashhab, H.F. Akhdar, A. Khan, A. Azad, M.A. Moni, Deep CNN-LSTM with self-attention model for human activity recognition using wearable sensor, IEEE J. Transl. Eng. Heal. Med. 10 (2022) 1–16, http://dx.doi.org/10.1109/JTEHM.2022.3177710.

[42] W. Liang, F. Wang, A. Fan, W. Zhao, W. Yao, P. Yang, Deep-learning model for the prediction of lower-limb joint moments using single inertial measurement unit during different locomotive activities, Biomed. Signal Process. Control. 86 (2023) 105372, http://dx.doi.org/10.1016/j.bspc.2023.105372, URL https://www.sciencedirect.com/science/article/pii/S1746809423008054.

[43] D.P. Kingma, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.

[44] J.C.F. Caballero, F.J. Martínez, C. Hervás, P.A. Gutiérrez, Sensitivity versus accuracy in multiclass problems using memetic pareto evolutionary neural networks, IEEE Trans. Neural Netw. 21 (5) (2010) 750–770.

[45] M. Dallel, V. Havard, Y. Dupuis, D. Baudry, A sliding window based approach with majority voting for online human action recognition using spatial temporal graph convolutional neural networks, in: Proceedings of the 2022 7th International Conference on Machine Learning Technologies, ICMLT '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 155–163, http://dx.doi.org/10.1145/3529399.3529425.

[46] S. Hossein Sadat Hosseini, N.N. Joojili, M. Ahmadi, LLMT: A transformer-based multi-modal lower limb human motion prediction model for assistive robotics applications, IEEE Access 12 (2024) 82730–82741, http://dx.doi.org/10.1109/ACCESS.2024.3413576.

[47] M. Atzori, M. Cognolato, H. Müller, Deep learning with convolutional neural networks applied to electromyography data: A resource for the classification of movements for prosthetic hands, Front. Neurorobotics 10 (2016) http://dx.doi.org/10.3389/fnbot.2016.00009, URL https://www.frontiersin.org/journals/neurorobotics/articles/10.3389/fnbot.2016.00009.