From Pilot Data to Protocol: Sample-Size Guidance for Multimodal Intent Detection in Assistive Wearable Robotics

Mohamad Reza Shahabian Alashti[1], Shadiya Alingal Meethal[1], Patrick Holthaus[1], Gabriella Lakatos[1], and Farshid Amirabdollahian[1]

Robotics Research Group, School of Physics, Engineering and Computer Science, University of Hertfordshire, Hatfield, AL10 9AB, United Kingdom
{m.r.shahabian, s.alingal-meethal4, p.holthaus, g.lakatos, f.amirabdollahian2}@herts.ac.uk

Abstract. Designing a sensible data-collection protocol for intent detection in assistive exoskeletons and wearable robots is challenging: too little data yields unstable models, while collecting "as much data as possible" is costly and often unnecessary. This challenge is particularly relevant in social robotics and human-robot interaction, where HAR from biosignals such as EMG is used to infer user intent and physical state for safer, more adaptive assistance. We propose a generic, model-agnostic procedure that turns a small pilot recording into concrete sample-size guidance using learning curves. Using a 32-channel high-density EMG (HD-EMG) grid on the thigh and a co-located IMU, we record eight repetitions of six lower-limb activities relevant to locomotion assistance and extract 100 ms windows. We compare RF, SVM, LDA, and a ResNet-18 CNN under leave-one-trial-out (LOTO) evaluation, estimating the training fraction needed to reach 90 % of each model's peak accuracy and the plateau where adding 10 % more data yields < 1 percentage-point gain. On this pilot dataset, RF/SVM/LDA typically meet both criteria after 20–30 % of available windows, whereas the CNN continues to improve up to approximately 70–90 %. IMU features outperform HD-EMG alone, and EMG+IMU fusion achieves the highest accuracy. Overall, the protocol provides per-model sample targets and a principled stopping rule to reduce recording and recalibration burden in data-efficient exoskeleton intent detection for human augmentation and wellbeing.

Keywords: assistive robotics, wearable robotics, intent detection, HD-EMG, IMU, sample size, data collection protocol.

1 Introduction

Wearable robotic systems such as exoskeletons, prostheses and assistive orthoses are increasingly deployed for rehabilitation, assistance and human augmentation [20]. A central technical challenge in these systems is *intent detection*: recognising the user's intended activity or transition early enough to trigger appropriate

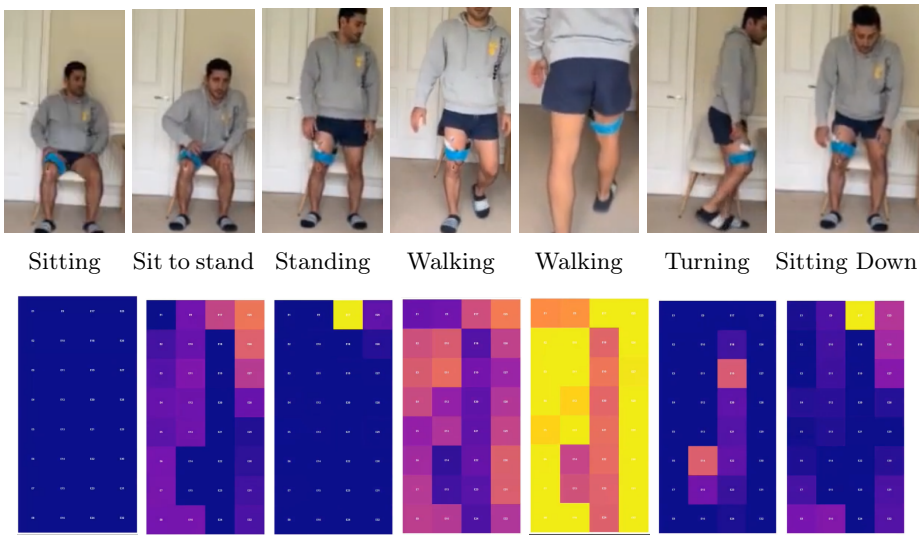

Fig. 1: Example activity sequence and corresponding sensor heatmaps. The top row shows representative frames of the subject performing each activity, and the bottom row shows the corresponding spatial heatmaps of HD-EMG (same ordering).

assistance [12,11]. Reliable intent detection enables smooth, proactive control and largely determines whether a device feels responsive and intuitive to its user.

Besides model choice, one key practical barrier to deploying wearable HAR systems is the lack of clear guidance on how much data is sufficient for robust training under realistic variability [5]. In many applications, data collection is constrained by time, user fatigue, and limited access to participants, and may begin with a single-user pilot recorded in a home environment [10,14]. This motivates a generic framework that can turn a small pilot dataset into quantitative sample-size guidance, enabling evidence-based planning of data collection and iteration without committing to a specific robot platform [3].

High-density electromyography (HD-EMG) is a promising modality for intent detection because it captures spatial patterns of muscle activation with much higher resolution than traditional bipolar EMG (BP-EMG) [18,7,16]. Dense electrode grids (e.g., 4×8 or 8×8 arrays) provide detailed information about motor unit recruitment and muscle synergies[7], and have been successfully used for upper-limb gesture recognition, prosthetic control and torque estimation, and are increasingly being explored for lower-limb locomotion and exosuit control [17,19,8]. Figure 1 illustrates a representative sample from the dataset collected for this paper, showing a participant performing the activity across frames alongside the corresponding 4×8 (32-channel) HD-EMG heatmap across all channels.

Despite extensive work on classification algorithms for EMG- and IMU-based HAR, from linear discriminants and support vector machines to convolutional and recurrent neural networks [16,13,9], a more basic practical question remains under-addressed in the wearable robotics literature:

> *How much labelled data is needed to train reliable intent detection models for a given sensor configuration and activity set?*

In many applications, data collection is constrained by time, participant burden, and limited access to users; therefore, studies often begin with a small-scale pilot dataset collected in-the-wild or in-lab settings [5,10,14]. Collecting too little data risks brittle models that fail under realistic variability, whereas collecting too much data wastes laboratory time and participant effort. Although learning-curve analysis is well established in machine learning [15], concrete, modality-specific guidance for HD-EMG and IMU intent detection remains scarce [3].

With our approach, we address this problem from a practical, data-driven perspective. We consider an early-stage scenario typical of many assistive robotics projects: a custom modular data-acquisition board has been developed for a lower-limb soft exosuit, the candidate activities have been defined, and an initial single-subject pilot dataset is available, but the final protocol (number of trials, recording time per activity, and choice of classifier) is still open. We specifically present a novel approach aimed to support the development of assistive wearable robotics, such as exoskeletons and soft exosuits, where intent recognition affects not only functional performance but also interaction quality, trust, and willingness to use a device repeatedly in everyday settings. Reducing the data-collection and recalibration burden is therefore a human-centred requirement as much as a technical one. The main contributions of this work are:

1. **Pilot-to-protocol sample-size estimation:** a simple, model-agnostic framework that uses learning curves from a small pilot dataset to estimate sample-size requirements for wearable intent detection (using a relative 90% of each model's own peak criterion and a plateau rule).
2. **Realistic multimodal dataset instantiation:** an instantiation of the protocol on a lower-limb HD-EMG and IMU dataset collected with a modular wireless acquisition board in a domestic home environment, including common activities and transitions.
3. **Classifier and modality analysis:** a comparative evaluation of RF, SVM, LDA, and a ResNet-18 CNN, and an empirical comparison of EMG-only, IMU-only, and fusion models, highlighting how modality choice interacts with data efficiency and achievable accuracy.
4. **Actionable data-collection targets:** per-model and per-activity guidelines for minimum and recommended training windows, reported both as sample counts and as equivalent recording times, enabling practical planning of repetitions and session duration.

Overall, while the numerical thresholds depend on the specific sensor setup and activity set considered here, the methodology is general and can be reused as a design tool when planning new datasets for assistive and wearable robotics.

2 Background and Related Work

This section motivates our problem setting from three angles: (i) intent detection as a core requirement for usable wearable/assistive robots, (ii) HD-EMG and IMU sensing as complementary modalities for wearable HAR, and (iii) the lack of practical guidance on how much labelled data is required to train robust models under realistic variability.

Lower-limb exoskeletons and assistive exosuits have matured from laboratory prototypes into commercially available devices for gait rehabilitation, industrial assistance and mobility support [20]. For these systems to be effective and acceptable in daily life, assistance must be delivered in synchrony with the user's movement intentions rather than reacting with a long delay. A wide range of control interfaces have been explored for intent detection, including mechanical interaction forces, foot switches, joint encoders, IMUs and various bio-signals such as EMG [12]. Non-invasive, EMG-based interfaces are particularly attractive because they provide a direct window into the neural drive to muscles while remaining compatible with wearable form factors [17].

In practice, intent detection for wearable robots often reduces to an HAR problem on a constrained activity set. Examples include recognising locomotion modes (level walking, stairs, ramps and transitions) for powered prostheses and exoskeletons, or distinguishing between postures such as sitting, standing and walking for assist-as-needed support. The focus of this work is on such locomotion-related activities, with an emphasis on how much training data is needed to reach robust performance for different model classes.

We adopt this intent-detection-as-HAR framing for a lower-limb activity set and focus on a practical design question that is usually left implicit: how much training data is needed to achieve robust performance for different model classes.

Conventional EMG-based intent detection typically relies on a small number of bipolar electrodes placed over specific muscles. While effective in controlled laboratory settings, such sparse configurations are sensitive to electrode placement and often require time-consuming manual adjustment [16]. HD-EMG addresses these limitations by using dense grids of electrodes that cover a larger muscle region. This not only provides richer spatial information about muscle activation patterns but also allows for spatial filtering and channel selection strategies that improve robustness to electrode shift and movement artefacts [18,7]. HD-EMG has been successfully used for upper-limb gesture recognition, prosthetic control and torque estimation, and is increasingly being explored for lower-limb locomotion and exosuit control [17,19,8].

IMUs are a standard choice for wearable HAR, capturing segment orientation and acceleration with low cost and power consumption [11,4]. When EMG and IMU are combined, the two modalities offer complementary information: EMG reflects the neural command, often preceding visible movement, while IMU captures the executed kinematics. Multi-modal fusion of EMG and kinematic sensors has been shown to improve recognition accuracy and robustness in both prosthetic control and activity recognition scenarios [2,13,1].

Rather than proposing a new sensing modality or fusion architecture, we use an HD-EMG grid with a co-located IMU to quantify how modality choice (EMG, IMU, and fusion) affects data efficiency and the sample size required to reach performance saturation.

In machine learning, learning curves, plots of model performance as a function of training set size, are a standard tool for diagnosing underfitting and overfitting and for projecting the benefits of collecting more data [15]. From a statistical perspective, sample-size determination is typically framed in terms of desired confidence intervals, effect sizes or generalisation error bounds, and learning curves are widely used to characterise the benefit of additional data [15,9]. However, such analyses are rarely carried out explicitly in the HAR literature, especially for high-dimensional physiological signals such as HD-EMG.

Existing EMG and multimodal datasets for intent detection and HAR span a wide range of sizes, from a few thousand to hundreds of thousands of labelled windows. Most studies justify their chosen recording durations pragmatically (e.g., "five repetitions of each activity"), without investigating whether this amount of data is over- or under-provisioned for the models they use. Moreover, data requirements are often model-dependent: shallow linear classifiers may saturate quickly, whereas deep networks with millions of parameters typically require much larger datasets to reach their full potential [13,9].

We address this lack of practical guidance by explicitly constructing learning curves for four representative model families on an HD-EMG+IMU pilot dataset and translating them into concrete per-model and per-activity sample-size targets. The outcome is a reusable pilot-to-protocol procedure that others can apply to their own sensors, activities and model choices, rather than universal fixed numbers.

3 Pilot Dataset and Recording Protocol

This section describes the pilot dataset that underpins our *pilot-to-protocol* methodology. In this paper, the pilot is not intended as a benchmark dataset; instead, it is a deliberately small, early-stage recording used to (i) construct learning curves under a fixed sensor configuration and activity set and (ii) convert those curves into quantitative sample-size targets and equivalent recording times. Accordingly, the dataset is recorded as a short, repeatable, cue-driven routine to reduce label noise and isolate the effect of training-set size, while still reflecting the locomotion-related activities relevant to lower-limb assistance. We first describe the activity routine and trial structure, then the sensing hardware and placement, followed by the windowing/label assignment procedure, and finally the offline signal-processing pipeline used to generate model inputs.

3.1 Activity routine and trials

The pilot study was designed around a short routine that captures common activities of daily living and relevant transitions for lower-limb assistance:

Table 1: Per-activity sample and duration distribution of the pilot dataset (all eight trials combined).

Activity	# Windows	Duration [s]	Proportion [%]
Sitting	1,311	65.3	20.5
Sitting Down	602	30.0	9.4
Standing	1,781	88.7	27.8
Standing Up	586	29.2	9.1
Turn Left	917	45.7	14.3
Walking	1,210	60.3	18.9
Total	6,407	319.1	100.0

sitting → standing up → standing → walking → standing → turn left → standing → sitting down.

This sequence is executed twice per trial, yielding sixteen labelled segments per trial. We recorded eight trials, each lasting approximately 50 s, in a controlled environment representative of the intended in-home use with the participant seated on a chair, walking a short straight segment, and turning around in the same direction each time.

During each trial, the participant followed on-screen and audio instructions provided by the data-acquisition (DAQ) software. The current activity label (e.g. *Sitting*, *Standing Up*, *Walking*) was displayed as large text on the screen, and a short beep sound indicated when to transition to the next activity in the predefined sequence. This cueing scheme ensured that the subject performed all activities in the intended order, with consistent timing across repetitions, without requiring additional verbal guidance.

Figure 2 illustrates one complete trial from the pilot dataset. The top panel shows the HD-EMG activity (root-mean-square envelope summed across all 32 channels), while the bottom panel shows the four IMU quaternion components. Coloured background bands mark the manually annotated activity segments, highlighting how the sit-to-walk routine unfolds over time and how EMG bursts and IMU orientation changes co-vary with the labels.

The continuous sensor streams were segmented into fixed-length windows (Section 3.3), producing a total of 6,407 labelled windows across six activity classes. Table 1 summarises the class distribution in terms of number of windows, total duration and relative proportion.

The repeated, scripted trials provide a minimal but structured dataset for constructing learning curves, and the trial-wise repetition supports leave-one-trial-out (LOTO) evaluation to study data efficiency.

3.2 Sensors and acquisition board

Signals were acquired using the modular wireless biosignal platform of Doukakis *et al.* [6]. The system consists of a Leader board and up to three Follower boards,

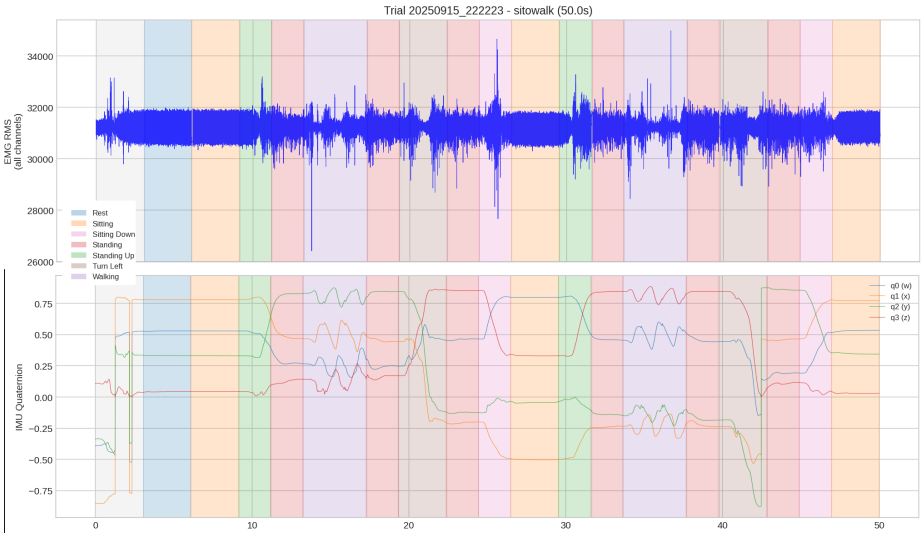

Fig. 2: Example 50 s trial from the pilot dataset. Top: HD-EMG activity (RMS envelope summed across all 32 channels). Bottom: IMU quaternion components (q_0, q_1, q_2, q_3). Coloured background bands indicate the annotated activity segments (rest, sitting, sitting down, standing, standing up, turn left, walking).

each hosting a 16-channel EMG analogue front end (AFE) and an STM32WB55 microcontroller with an integrated 2.4 GHz radio. The Leader carries a 9-axis BNO086 IMU and aggregates data from Followers over high-speed SPI before streaming it wirelessly to a USB dongle. The AFE bandwidth is approximately 9–286 Hz and the effective sampling rate for EMG channels is around 1.4 kHz, with an input-referred noise on the order of 0.8 μV_{RMS} [6].

In this study, we used one Leader and one Follower in monopolar mode, yielding 32 HD-EMG channels arranged as a 4×8 grid on the right anterior–lateral thigh over the quadriceps (primarily rectus femoris and vastus lateralis), with the integrated IMU on the Leader board mounted at the same site. The dataset analysed here corresponds to an effective sampling rate of:

$$f_s \approx 1446 \text{ Hz},$$

as determined by the firmware timer configuration on the acquisition board. This configuration matches realistic conditions for lower-limb exosuit experiments while keeping the hardware overhead modest.

3.3 Windowing and labelling

To transform continuous streams into fixed-size samples suitable for supervised learning, we use sliding windows of 100 ms with 50% overlap. Given a recording

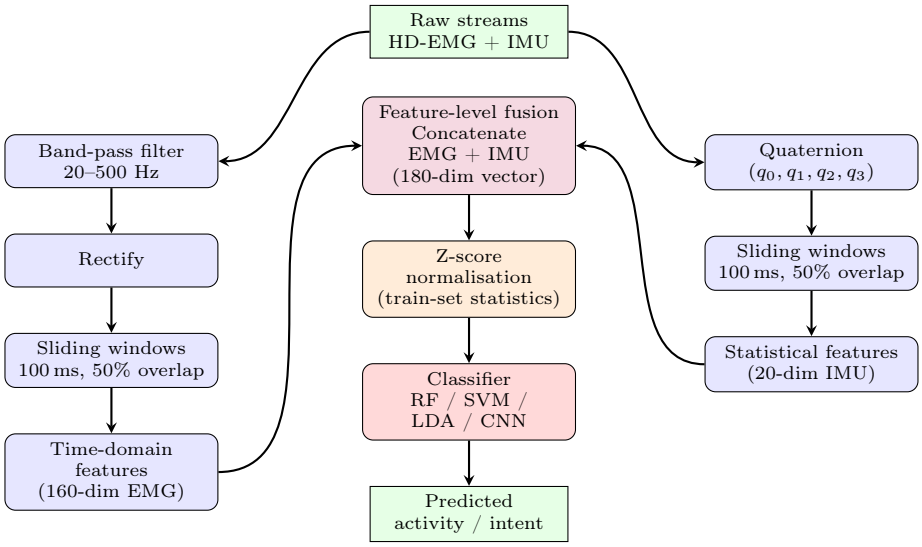

Fig. 3: Signal-processing pipeline for HD-EMG and IMU data. EMG and IMU streams are processed in parallel and fused at the feature level.

of duration t (in seconds), the number of windows N_{win} is

$$N_{\text{win}} = \left\lfloor \frac{t f_s - N_w}{N_s} \right\rfloor + 1, \tag{1}$$

where $N_w = 144$ is the number of samples per window and $N_s = 72$ is the step size. Each window is assigned the activity label that occupies the majority of its samples. Windows spanning activity transitions are rare because transitions themselves are short, and they are labelled according to the dominant activity.

3.4 Signal-processing pipeline

The complete offline pipeline used in this study is summarised in Fig. 3. HD-EMG and IMU streams are first pre-processed and segmented into overlapping windows. For the classical models (RF, SVM, LDA) we extract hand-crafted features from each window, optionally fuse EMG and IMU feature vectors, normalise them, and train a classifier. For the CNN baseline, the EMG windows follow the same pre-processing up to the windowing stage but are then fed directly to the ResNet-18 architecture without feature engineering.

The pipeline standardises pre-processing and input representations across classical and deep models, ensuring that observed differences in learning curves reflect data requirements rather than inconsistent signal preparation.

4 Feature Extraction and Models

This section defines the input representations and model families used to construct learning curves and derive sample-size targets. We compare classical classifiers (RF, SVM, LDA) trained on hand-crafted EMG/IMU features, CNN baselines trained directly on windowed signals, and three multimodal fusion strategies. We then specify the LOTO evaluation and the 90%-of-peak and plateau criteria used to translate learning curves into data-collection guidance.

4.1 Hand-crafted features for classical models

For RF, SVM and LDA, we follow a standard feature-engineering pipeline. From each HD-EMG channel, we extract five time-domain descriptors—mean absolute value (MAV), root mean square (RMS), waveform length (WL), zero crossings (ZC) and slope sign changes (SSC)—yielding $32 \times 5 = 160$ EMG features per window. From the IMU quaternion components, we compute mean, standard deviation, minimum, maximum and range, producing $4 \times 5 = 20$ IMU features per window. All features are z-score normalised using training-set statistics. We consider three feature sets: (1) EMG-only (160 features), (2) IMU-only (20 features), and (3) early-fusion EMG+IMU (180 features obtained by concatenation).

4.2 CNN baselines for EMG and IMU

For the deep learning baseline, we use convolutional architectures tailored to each modality.

HD-EMG CNN (ResNet-18). For HD-EMG, we adopt a modified ResNet-18 backbone that treats each window as a 2D "image" of size $32 \times N_w$ (channels $\times$ time). The first convolutional layer is adapted to single-channel input, the initial max-pooling layer is removed to preserve temporal resolution, and the final fully connected (FC) layer is replaced with a 6-way classifier. Batch normalisation and dropout are applied after each residual block, and an adaptive average pooling layer aggregates the last feature map into a fixed-length 512-dimensional embedding, which is fed to the final FC layer. This architecture is used for the EMG-only CNN results reported in the paper and as the EMG branch in the fusion models.

IMU CNN (temporal Conv1D). For IMU-only experiments, we do not reuse ResNet-18. Instead, we employ a lightweight 1D temporal CNN that operates directly on the quaternion time series. Each input window is represented as a tensor of shape $[N_w \times 4]$ (time $\times$ quaternion components $q_0, \ldots, q_3$). The network consists of three Conv1D–BatchNorm–ReLU blocks with kernel size 5 and padding 2, interleaved with max-pooling layers to reduce the temporal dimension progressively. An adaptive average pooling layer collapses the final feature map to a 128-dimensional embedding, followed by a small classifier head (two fully connected layers with dropout) that outputs class logits. This IMU

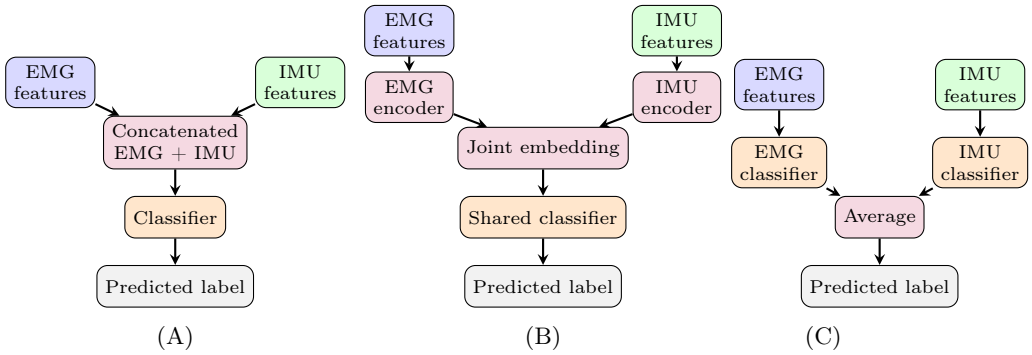

Fig. 4: Fusion architectures used in this work: (A) early feature-level fusion (EMG and IMU features concatenated before a single classifier), (B) intermediate dual-branch fusion (separate EMG/IMU encoders feeding a shared classifier), and (C) late decision-level fusion (separate EMG and IMU classifiers combined at the output level). Blue boxes: EMG branch; green boxes: IMU branch; purple boxes: fusion/embedding layers; orange boxes: classifiers; grey boxes: final prediction.

CNN is used for IMU-only results and as the IMU branch in the intermediate- and late-fusion configurations.

In the intermediate-fusion model, the 512-dimensional EMG embedding from the ResNet-18 branch and the 128-dimensional IMU embedding from the temporal CNN are concatenated into a joint feature vector and passed through a shared FC classifier. In the late-fusion model, independent EMG and IMU CNN classifiers are trained, and their posterior probabilities are averaged at inference time.

4.3 Fusion strategies

To explore the benefits of combining modalities, we consider three fusion strategies when both EMG and IMU are available:

- **Early fusion**: feature concatenation followed by a single classifier.
- **Intermediate fusion**: a dual-branch CNN with separate EMG and IMU streams merged in a shared embedding layer.
- **Late fusion**: independent EMG and IMU classifiers whose posterior probabilities are averaged.

4.4 Training and evaluation protocol

We use a leave-one-trial-out (LOTO) cross-validation scheme over the eight repetitions of the activity routine. At each fold, one trial is held out for testing, and the remaining seven are used for training. Learning curves are constructed by randomly subsampling the available training windows at fractions

$\{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}\%$ and averaging test accuracy across all folds and three random seeds.

Let $A_{\max}$ denote the mean test accuracy achieved by a given model when trained on 100% of the available training data. We define the *minimum acceptable performance* as

$$A_{\text{thr}} = 0.9\, A_{\max}, \tag{2}$$

i.e., 90% of the model's own peak accuracy. For each fold, let N_{tot} denote the number of labelled windows available in the training split (seven trials). For each mode,l we find the smallest training fraction $p^\star$ at which the learning curve crosses A_{thr}. The corresponding minimum number of training windows is

$$N_{\min} = p^\star N_{\text{tot}}. \tag{3}$$

To account for variability across future users or recording conditions, we also define a *recommended* sample size as

$$N_{\text{rec}} = 1.2\, N_{\min}, \tag{4}$$

i.e., 20% more than the minimum. Both $N_{\min}$ and N_{rec} are converted into recording time by inverting (1) for a single continuous recording.

Defining a plateau region on the learning curve In addition to the 90%-of-peak criterion, we also characterise the *plateau* region of each learning curve. Let $A(p)$ denote the mean test accuracy when training on a fraction $p \in \{10, 20, \ldots, 100\}\%$ of the available training windows. We define the discrete increment

$$\Delta A(p) = A(p + 10) - A(p).$$

Intuitively, the learning curve has plateaued once additional data yields only marginal gains in accuracy. Operationally, we declare that a model has reached its plateau at the smallest training fraction p_{plat} such that all subsequent increments are below a tolerance δ:

$$p_{\text{plat}} = \min \{p \ : \ |\Delta A(p')| \leq \delta \ \text{ for all } \ p' \geq p\},$$

with $\delta = 0.01$ (one percentage point). Training fractions beyond p_{plat} are thus in a regime of diminishing returns, where collecting more data increases accuracy by less than one absolute percentage point per additional 10% of training data. In the results, we use this plateau analysis as a complementary sanity check for the sample-size thresholds derived from the 90%-of-peak rule.

5 Results

This section reports the empirical outcomes of the pilot-to-protocol framework. We first compare overall performance across model families and sensing modalities, then analyse learning curves to derive model-specific sample-size thresholds and plateau regions, and finally translate these global requirements into per-activity quotas and error patterns via per-class allocations and confusion matrices.

Table 2: Accuracy and macro-F1 for all models and fusion strategies with leave-one-trial-out evaluation. Unimodal rows use EMG-only or IMU-only features.

Model	Fusion	Modality	Accuracy	F1-Macro
Late_Fusion_RF	Late	EMG+IMU	0.822	0.795
RF_fusion	Early	EMG+IMU	0.805	0.785
RF_Intermediate	Intermediate	EMG+IMU	0.783	0.760
SVM_fusion	Early	EMG+IMU	0.781	0.736
Late_Fusion_SVM	Late	EMG+IMU	0.779	0.737
RF_imu	—	IMU	0.766	0.735
LDA_Intermediate	Intermediate	EMG+IMU	0.755	0.715
SVM_Intermediate	Intermediate	EMG+IMU	0.752	0.709
LDA_fusion	Early	EMG+IMU	0.749	0.711
CNN_Late_Fusion	Late	EMG+IMU	0.741	0.708
Late_Fusion_LDA	Late	EMG+IMU	0.739	0.682
CNN_Intermediate_Fusion	Intermediate	EMG+IMU	0.732	0.682
CNN_emg	—	EMG	0.720	0.669
SVM_imu	—	IMU	0.718	0.649
CNN_Early_Fusion	Early	EMG+IMU	0.717	0.664
SVM_emg	—	EMG	0.715	0.658
LDA_emg	—	EMG	0.708	0.658
RF_emg	—	EMG	0.707	0.654
CNN_imu	—	IMU	0.678	0.619
LDA_imu	—	IMU	0.640	0.519

5.1 Overall model performance and modalities

Table 2 reports the accuracy and macro-F1 score for all models and fusion strategies when trained on 100% of the training data (LOTO). The best overall performance (82.2% accuracy, 79.5% macro-F1) is obtained by a late-fusion RF that combines EMG and IMU posteriors. Among EMG-only models, the CNN reaches 72.0–72.6% accuracy, slightly outperforming RF, SVM and LDA, which cluster around 70–71%.

Fig. 5(b) compares modalities at the 100% training point over models. IMU-only consistently outperforms EMG-only, and combining EMG and IMU improves performance further; the best result on this pilot dataset is achieved by late-fusion RF. In the remainder of this section, we use EMG-only models as a conservative baseline to isolate how model family affects data requirements.

5.2 Learning curves and model-specific sample size

Figure 5a shows the learning curves for the four EMG-only models as the fraction of training data increases. All models exhibit the expected diminishing returns: the largest gains occur between 10% and 30% of the training windows, after which additional data yields only incremental improvements. To obtain a model-specific notion of "enough" data, we define the minimum acceptable performance as 90% of each model's own peak accuracy and take $p^\star$ to be the smallest training fraction at which this threshold is reached.

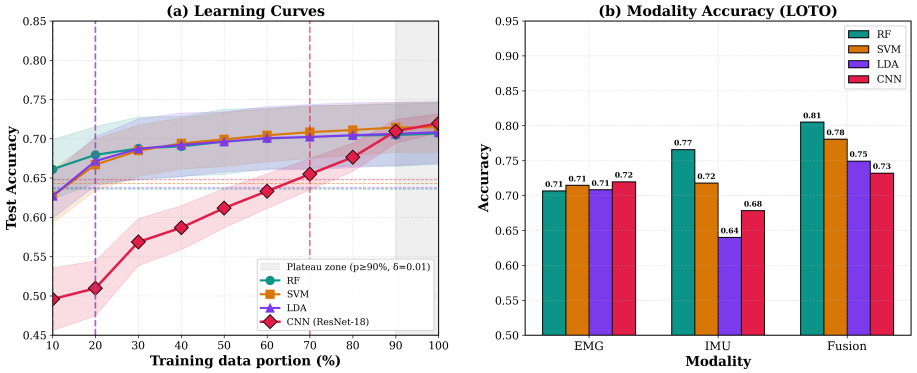

Fig. 5: Learning curves and modality comparison. (a) LOTO test accuracy vs. training fraction (10–100%) for RF, SVM, LDA, and ResNet-18 using EMG-only features; Horizontal dashed lines indicate 90% of each model's peak accuracy; vertical dashed lines mark the smallest training fraction at which each curve first reaches that threshold (p^*).The grey shaded area denotes the plateau region (δ=0.01). (b) Peak LOTO accuracy for EMG, IMU, and early-fusion (EMG+IMU), averaged over models.

On this dataset, RF reaches its 90%-of-peak threshold at 10% of the training windows, SVM and LDA at around 20%, and the CNN at roughly 70%. These values translate into the minimum and recommended sample sizes in Table 3, obtained by scaling $p^\star$ to the total number of available windows and adding a 20% safety margin. From a practical standpoint, this means that, for the scripted routine considered here, traditional models already operate near their best EMG-only performance after only a few pilot trials, whereas the CNN continues to benefit from substantially more data.

Figure 6 makes this observation more explicit by analysing the plateau of each learning curve. We compute the absolute increment in accuracy between successive training fractions, $|\Delta A(p)| = |A(p{+}10) - A(p)|$, and define the plateau onset p_{plat} as the smallest fraction such that $|\Delta A(p')| \leq 0.01$ for all $p' \geq p_{\text{plat}}$. For RF, SVM and LDA, the plateau begins between 20% and 30%; beyond this point, adding 10% of extra data improves accuracy by less than one percentage point. In contrast, the CNN exhibits a much later plateau, with noticeable gains up to about 90% of the data and only very small changes thereafter. Together, the 90%-of-peak rule and the plateau analysis show that more data does not automatically translate into better performance: once a model's learning curve has flattened, additional repetitions of the same routine deliver sharply diminishing returns.

Here, N_{min} and N_{rec} are reported as equivalent windows (and time) for a continuous per-subject recording; learning-curve fractions are estimated within LOTO folds but translated into recording-duration targets for protocol planning.

From a practical perspective, N_{min} marks the point where additional data mainly improves confidence rather than changing performance, while N_{rec} adds

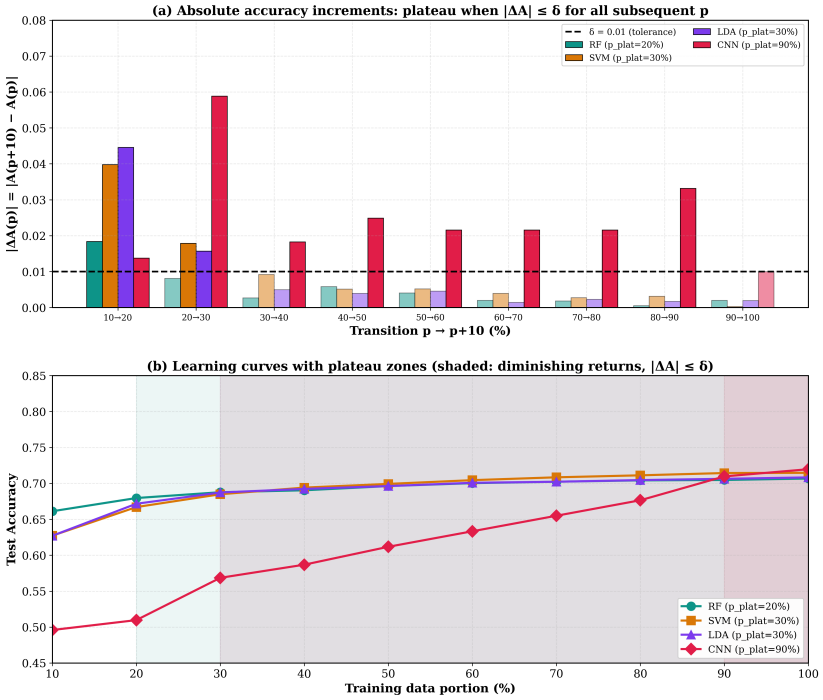

Fig. 6: Plateau analysis on the learning curves. (a) Absolute accuracy increments $|\Delta A(p)| = |A(p+10) - A(p)|$ for RF, SVM, LDA, and CNN over each 10% step (e.g. 10→20, 20→30). The horizontal dashed line is the tolerance $\delta = 0.01$. Bars with $p \geq p_{\text{plat}}$ are shown with reduced opacity (plateau region), where p_{plat} is the smallest p such that $|\Delta A(p')| \leq \delta$ for all $p' \geq p$. (b) Learning curves with the plateau zone shaded for each model. Beyond p_{plat}, additional data yields less than one percentage point of accuracy gain per extra 10% of training data.

a conservative 20% margin. Together with the plateau thresholds in Fig. 6, the results indicate that traditional EMG-only models reach near-saturation with well under two minutes of labelled data for this routine, whereas the CNN benefits from several minutes before diminishing returns dominate. These thresholds inform target recording durations for a larger-scale acquisition.

5.3 Per-activity sample requirements

The minimum sample sizes in Table 3 specify how many windows are needed in total for each model, but they do not by themselves determine how those windows should be distributed across activities. To preserve the empirical class distribution of the pilot dataset when scaling up data collection, we allocate

Table 3: Minimum and recommended number of training windows for each classifier, based on the 90%-of-peak accuracy threshold $A_{\mathrm{thr}} = 0.9A_{\max}$ from Figure 5a. $N_{\min}$ is the smallest number of windows whose accuracy exceeds A_{thr}; $N_{\mathrm{rec}} = 1.2N_{\min}$ adds a 20% safety margin. $T_{\min}$ and T_{rec} are the corresponding continuous-recording durations obtained by inverting (1).

Model	$N_{\min}$	N_{rec}	$T_{\min}$ [min]	T_{rec} [min]
RF	641	769	0.5	0.6
SVM	1,281	1,537	1.1	1.3
LDA	1,281	1,537	1.1	1.3
CNN	4,485	5,382	3.7	4.5

per-activity quotas by simple proportional rescaling:

$$N_{\min}^{(a)} = \mathrm{round}\left(N_{\min} \frac{n_a}{N_{\mathrm{tot}}} \right), \tag{5}$$

where n_a is the number of windows of activity a in the full pilot and N_{tot} is the total number of windows. Figure 7 visualises the resulting minimum per-activity counts for RF, SVM, LDA and the CNN, assuming an EMG-only configuration.

Because the scripted routine naturally contains more repetitions of the static postures than of the short transitions, sitting and standing receive the largest quotas, followed by walking, with sitting down, standing up and turning left represented by fewer windows. It is important to emphasise that these per-activity numbers are not derived from separate class-wise learning curves; they simply maintain the same class proportions as the pilot while honouring each model's global $N_{\min}$. In other words, the pilot tells us how many windows are needed in total for a given model to reach stable EMG-only performance, and the per-activity quotas provide a convenient way of reproducing the same class balance in a larger recording.

Viewed together with the plateau analysis, this has two practical implications. First, for RF, SVM and LDA, the learning curves flatten after roughly 20–30% of the available data, so repeatedly executing the same routine beyond this point will mostly add more examples of already frequent classes without substantially improving overall accuracy. Second, if confusion matrices reveal that particular activities remain difficult (for example, specific transitions or turning), then it may be more effective to deliberately oversample those activities or to enrich the protocol (e.g. with additional movement variations or multi-day recordings) rather than uniformly increasing the number of repetitions of the entire sequence.

5.4 Confusion patterns

To understand which activities drive the remaining errors once the models have reached their plateau, we inspect the EMG-only confusion matrices. Figure 8 reports the leave-one-trial-out confusion matrix for the RF classifier; SVM, LDA

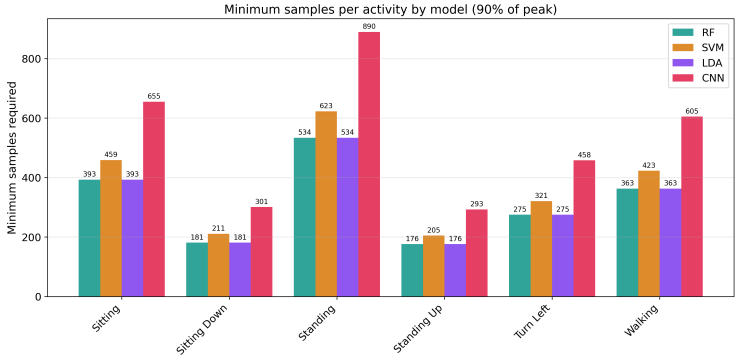

Fig. 7: Minimum per-activity training windows for each EMG-only model, obtained by scaling the global $N_{\min}$ values from Table 3 with the empirical class proportions of Table 1.

and CNN exhibit very similar patterns. Across all four EMG-only models, the two static postures are recognised most reliably: sitting and standing typically achieve recalls in the high 88–93% range. Walking is moderately robust, with recalls around 66%, whereas the three dynamic and transitional activities—sitting down, standing up and turning left—remain the most challenging, with recalls mostly between 40% and 50%. Errors cluster in intuitive ways: sitting down is often confused with standing and walking, standing up with sitting and walking, and turning left with both standing and walking. These confusions reflect the fact that the underlying muscle activation patterns are brief, overlap in time with neighbouring postures, and are recorded from a single muscle group on one leg.

Importantly, the plateau analysis indicates that, for RF, SVM and LDA, adding more repetitions of the same routine beyond roughly 20–30% of the maximum dataset size will only slightly change these confusion structures and will not remove the dominant error modes: overall accuracy improves by less than one percentage point per additional 10% of data. In practical terms, this means that simply making static postures even more frequent by recording many more trials is unlikely to eliminate the dominant misclassifications. Instead, the pilot results suggest two complementary strategies for designing a larger dataset: (i) explicitly rebalancing the protocol to allocate more recording time to the hardest activities identified in the confusion matrices, and (ii) increasing informational richness through additional sensing modalities or more diverse recording conditions (e.g. multi-day sessions or varied walking paths).

6 Discussion

This paper addresses a practical gap in wearable HAR/intent detection research: despite extensive work on sensing modalities and classifiers, recording duration

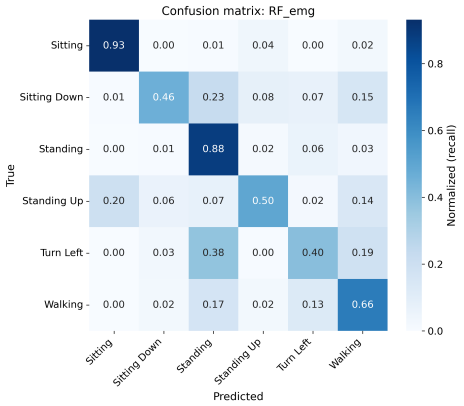

Fig. 8: EMG-only confusion matrix for the Random Forest classifier.

and repetitions are still often chosen pragmatically, and dataset size is rarely analysed in terms of sufficiency for the intended models and variability [15,5]. In assistive wearable robotics, this is also an HRI issue: unreliable intent detection can cause delayed or inappropriate assistance, reducing perceived responsiveness, comfort, and trust, and lowering willingness to use a device repeatedly. By converting learning curves into sample-size and recording-time targets, we contribute a protocol-centric tool that supports reliable control with reduced user burden.

Link to the paper objectives. The results establish a pilot-to-protocol method for estimating data needs, demonstrate it on a lower-limb HD-EMG+IMU pilot, and translate outcomes into practical per-model and per-activity recording targets. The central HRI implication is that data-efficient protocol design can shorten calibration and recalibration while maintaining interaction reliability, supporting trust and long-term adoption.

Learning curves as a stopping principle. Learning curves are widely used in machine learning [15] but are seldom used to decide how much data to collect in wearable HAR. Here, they show strong model dependence: RF/SVM/LDA reach near-saturation quickly on this routine, whereas the CNN benefits from substantially more data before diminishing returns. The plateau analysis operationalises the point at which additional repetitions provide marginal gains. Practically, the same workflow can be applied during data collection: record a small pilot, fit learning curves for candidate model families, and continue recording only until performance reliably exceeds the 90%-of-peak threshold and enters the plateau region. This reduces participant fatigue and setup time, and limits over-collecting repetitive data that may not improve real-world robustness; conversely, if curves have not stabilised, additional data is justified to avoid inconsistent assistance that undermines user confidence and perceived safety.

Why a scripted pilot and what it enables. The pilot in this paper is not intended as a benchmark dataset; it is a measurement tool for protocol planning. A short, cue-driven routine reduces label noise and isolates the effect of training-set size, enabling a direct conversion from learning-curve thresholds to window counts and recording time. Mapping these totals into per-activity quotas provides actionable guidance for planning repetitions under common constraints on participant availability and fatigue [5], and supports shorter, more sustainable calibration routines that improve acceptability in HRI.

Modality implications. The modality comparison is consistent with prior multimodal HAR work showing that inertial sensing is highly informative and fusion can improve robustness [11,4,2,13]. In our pilot, IMU-only outperforms EMG-only, and fusion performs best overall, suggesting that modest sensor diversity may yield larger benefits than substantially increasing EMG recording time alone. For HRI, improved robustness translates into smoother assistance and fewer unexpected responses, supporting trust and comfort, while EMG remains valuable for muscle-level intent interfaces [17,7]. The proposed procedure is modality-agnostic and can be rerun for IMU-only and fusion pipelines to quantify their data requirements under the intended protocol.

Limitations and future work. The thresholds reported here are specific to a single participant, a short structured routine, and the selected model set, and should be interpreted as pilot-derived planning values. The key next step is to quantify how sample-size targets change under deployment variability, particularly across subjects and days, where EMG drift and electrode placement shifts can degrade stability and thus impact perceived reliability in HRI. Future work will extend the protocol to multi-subject, multi-day recordings and explore additional model families and alternative sufficiency criteria beyond 90% of peak (e.g., uncertainty, calibration, or task-utility measures) to better capture trade-offs between accuracy, user burden, and perceived safety and responsiveness.

7 Conclusion

This paper presented a practical, data-driven protocol for estimating how many labelled windows are sufficient for lower-limb intent detection framed as wearable-sensor activity recognition. Using a small pilot dataset recorded with a modular wireless HD-EMG+IMU platform, we combined learning-curve analysis, a 90%-of-peak criterion and an explicit plateau definition to derive global and per-activity sample-size targets for representative classical and deep models.

On the pilot routine, RF reached the 90%-of-peak threshold at about 10% of the available windows, SVM and LDA at around 20%, and the ResNet-18 CNN at roughly 70%, with classical models plateauing by approximately 20–30%. IMU-only models outperformed EMG-only, and multimodal fusion achieved the best overall performance (late-fusion RF performing best). Confusions persisted mainly in dynamic and transitional activities, suggesting that targeted protocol

adjustments (e.g. oversampling difficult classes, adding movement variation, or richer sensing) are more effective than uniformly scaling repetitions once learning curves flatten.

The proposed pilot-to-protocol workflow is general and can be reused for other sensors, activities and model families, helping reduce unnecessary recording and recalibration effort while supporting responsive, trustworthy assistance. Future work will extend the analysis to multi-subject, multi-day recordings and integrate the resulting guidance into real-time intent-detection pipelines for assistive exosuits.

Acknowledgment

This research was funded by the European Union's Horizon Europe programme via the SWAG Project (Grant Agreement No. 101120408) and by Horizon Europe UK Research and Innovation (UKRI) through the UK government's Horizon Europe funding guarantee (Grant No. 10079504).

References

1. Alashti, M.R.S., Abadi, M.H.B., Holthaus, P., Menon, C., Amirabdollahian, F.: Efficient skeleton-based human activity recognition in ambient assisted living scenarios with multi-view cnn. In: 2024 10th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob). pp. 979–984. IEEE (2024)
2. Atzori, M., Gijsberts, A., Castellini, C., Caputo, B., Hager, A.G.M., Elsig, S., Giatsidis, G., Bassetto, F., Müller, H.: Electromyography data for non-invasive naturally-controlled robotic hand prostheses. Scientific data **1**(1), 140053 (2014)
3. Beleites, C., Neugebauer, U., Bocklitz, T., Krafft, C., Popp, J.: Sample size planning for classification models. Analytica chimica acta **760**, 25–33 (2013). https://doi.org/10.1016/j.aca.2012.11.007
4. Bulling, A., Blanke, U., Schiele, B.: A tutorial on human activity recognition using body-worn inertial sensors. ACM Computing Surveys (CSUR) **46**(3), 1–33 (2014)
5. Chen, W., Lin, S., Thompson, E., Stankovic, J.: Sensecollect: We need efficient ways to collect on-body sensor-based human activity data! Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **5**(3), 1–27 (2021). https://doi.org/10.1145/3478119
6. Doukakis, A., Smyrli, A., Livadas, M., Ribeiro, H.D.M., Meethal, S.A., Alashti, M.R.S., Lakatos, G., Holthaus, P., Amirabdollahian, F.: A modular, wireless and wearable biosignal acquisition platform. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) (2026), to appear
7. Farina, D., Holobar, A., Merletti, R., Enoka, R.M.: Decoding the neural drive to muscles from the surface electromyogram. Clinical neurophysiology **121**(10), 1616–1623 (2010)
8. Gopura, R., Bandara, D., Kiguchi, K., Mann, G.K.: Developments in hardware systems of active upper-limb exoskeleton robots: A review. Robotics and Autonomous Systems **75**, 203–220 (2016)

9. Hammerla, N.Y., Halloran, S., Plötz, T.: Deep, convolutional, and recurrent models for human activity recognition using wearables. arXiv preprint arXiv:1604.08880 (2016)
10. Hoelzemann, A., Van Laerhoven, K.: A matter of annotation: an empirical study on in situ and self-recall activity annotations from wearable sensors. Frontiers in Computer Science **6**, 1379788 (2024). https://doi.org/10.3389/fcomp.2024.1379788
11. Lara, O.D., Labrador, M.A.: A survey on human activity recognition using wearable sensors. IEEE communications surveys & tutorials **15**(3), 1192–1209 (2012)
12. Lobo-Prat, J., Kooren, P.N., Stienen, A.H., Herder, J.L., Koopman, B.F., Veltink, P.H.: Non-invasive control interfaces for intention detection in active movement-assistive devices. Journal of neuroengineering and rehabilitation **11**(1), 168 (2014)
13. Ordóñez, F.J., Roggen, D.: Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. Sensors **16**(1), 115 (2016)
14. Paraschiakos, S., Cachucho, R., Moed, M., van Heemst, D., Mooijaart, S., Slagboom, E.P., Knobbe, A., Beekman, M.: Activity recognition using wearable sensors for tracking the elderly. User Modeling and User-Adapted Interaction **30**(3), 567–605 (2020). https://doi.org/10.1007/s11257-020-09268-2
15. Perlich, C., et al.: Learning curves in machine learning. (2010)
16. Phinyomark, A., N. Khushaba, R., Scheme, E.: Feature extraction and selection for myoelectric control based on wearable emg sensors. Sensors **18**(5), 1615 (2018)
17. Scheme, E., Englehart, K.: Electromyogram pattern recognition for control of powered upper-limb prostheses: state of the art and challenges for clinical use. Journal of Rehabilitation Research & Development **48**(6), 643–660 (2011)
18. Staudenmann, D., Roeleveld, K., Stegeman, D.F., Van Dieën, J.H.: Methodological aspects of semg recordings for force estimation–a tutorial and review. Journal of electromyography and kinesiology **20**(3), 375–387 (2010)
19. Xia, P., Hu, J., Peng, Y.: Emg-based estimation of limb movement using deep learning with recurrent convolutional neural networks. Artificial organs **42**(5), E67–E77 (2018)
20. Young, A.J., Ferris, D.P.: State of the art and future directions for lower limb robotic exoskeletons. IEEE Transactions on Neural Systems and Rehabilitation Engineering **25**(2), 171–182 (2016)