Comparing Teaching Strategies of a Machine Learning-based Prosthetic Arm

VAYNEE SUNGEELEE, Sorbonne Université, CNRS, Institut des Systèmes Intelligents et de Robotique, ISIR, Paris, France

TÉO SANCHEZ, Hochschule München, Munich Center for Digital Sciences and AI, MUC.DAI, Munich, Germany

NATHANAËL JARRASSÉ, Sorbonne Université, CNRS, Institut des Systèmes Intelligents et de Robotique, ISIR ,Paris, France

BAPTISTE CARAMIAUX, Sorbonne Université, CNRS, Institut des Systèmes Intelligents et de Robotique, ISIR, Paris, France

Pattern-recognition-based arm prostheses rely on recognizing muscle activation to trigger movements. The effectiveness of this approach depends not only on the performance of the machine learner but also on the user's understanding of its recognition capabilities, allowing them to adapt and work around recognition failures. We investigate how different model training strategies to select gesture classes and record respective muscle contractions impact model accuracy and user comprehension. We report on a lab experiment where participants performed hand gestures to train a classifier under three conditions: (1) the system cues gesture classes randomly (control), (2) the user selects gesture classes (teacher-led), (3) the system queries gesture classes based on their separability (learner-led). After training, we compare the models' accuracy and test participants' predictive understanding of the prosthesis' behavior. We found that teacher-led and learner-led strategies yield faster and greater performance increases, respectively. Combining two evaluation methods, we found that participants developed a more accurate mental model when the system queried the least separable gesture class (learner-led). Our results conclude that, in the context of machine learning-based myoelectric prosthesis control, guiding the user to focus on class separability during training can improve recognition performances and support users' mental models about the system's behavior. We discuss our results in light of several research fields : myoelectric prosthesis control, motor learning, human-robot interaction, and interactive machine teaching.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; Accessibility technologies; • **Computing methodologies → Learning from demonstrations**.

Additional Key Words and Phrases: Myoelectric prosthesis, Machine Learning, Interactive Machine Teaching, Training curriculum, Mental model

ACM Reference Format:
Vaynee Sungeelee, Téo Sanchez, Nathanaël Jarrassé, and Baptiste Caramiaux. 2024. Comparing Teaching Strategies of a Machine Learning-based Prosthetic Arm. In *29th International Conference on Intelligent User Interfaces (IUI '24), March 18–21, 2024, Greenville, SC, USA.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3640543.3645170

1 INTRODUCTION

Learning to control a myoelectric prosthesis is challenging and influenced by physiological differences, thus requiring tailored training for users. These prostheses capture users' intentions by mapping electromyographic (EMG) signals from muscular activations to categories associated with prosthesis output gestures. Myoelectric prostheses can use supervised

Manuscript submitted to ACM

machine learning (ML) algorithms to learn associations between EMG signal patterns and prosthesis gestures. With such devices, users can curate new examples for the prosthesis to update its mapping between muscular contractions and gestures.

However, pattern recognition-based prosthesis control has not been widely used in rehabilitation and commercial devices because strategies for user-mediated model training remain unclear, and therefore classification accuracy remains low in ecological settings. EMG signals can also change due to muscle tiredness or stump posture. As a result, user control deteriorates in ways that remain to be fully understood [15], and the classifier must be retrained to account for these changes.

Increasing our understanding of which model training strategies yield better performance and better system comprehension might fill the gap between in-lab and real-life performances for prosthesis control [15]. Previous works have addressed different strategies to train users to control their prostheses. It has been shown that training the user to provide more consistent and distinguishable muscle contraction patterns improves classification accuracy [16, 31]. Furthermore, research showed that real-time visual guidance improves the quality of EMG data [12, 13, 32]. Training the user to perform muscular contractions and train an ML model simultaneously can be seen as an HCI problem, where a co-adaptation occurs between the user and the classification model. This has been investigated particularly in the domains of Interactive Machine Learning (IML) [1] and Interactive Machine Teaching (IMT) [35]. The IML approach, at the intersection of HCI and Machine Learning, aims to include end-users, including non-ML expert stakeholders, in the development of ML models through dataset curation, model steering, or the choice of performance metrics. Interactive Machine Teaching specifically focuses on the inherent ability of humans to teach, i.e., to take the perspective of a learner and curate relevant training examples to address its weaknesses. IMT research calls for the design of interactions that leverage these human abilities, including the ability of users to develop an accurate mental model of the machine learner being taught [17, 38].

This paper aims to study different teaching strategies of ML-based myoelectric prostheses, with different degrees of freedom to organize the training sequence. We want to understand how certain factors affecting the gesture examples provided by users for training the ML-based prosthesis impact the system's performance and the user's understanding of the system's behaviour. In particular, this research aims to answer the following research question: *How does the teaching strategy affect recognition accuracy and user's understanding of an ML-based prosthesis when the training strategy is either random, directed by the user, or directed by the learner according to an optimisation criterion?* Two assumptions guide this research question:

Assumption 1. Allowing the user to take an active role ("teacher") by structuring their training session can help them test and improve their understanding of the system. This assumption is guided by a) the learning by design (LBD) approach [18], which argues that being engaged in the process of design enhances skill acquisition, and b) the interactive machine teaching approach [35], suggesting that users engaged in the role of a machine "teacher", rather than annotator, develop investigative and self-reflecting behaviors [37]. This is a *teacher-led* approach.

Assumption 2. Guiding users to select gesture classes that maximize the separability of the gesture classes might lead to better model accuracy. This assumption is led by the fact that in the context of prosthesis control, increasing the separability of EMG patterns in the feature space of the classifier is conceptually linked to higher accuracy [12]. Since the machine learner queries the human teacher, this is a *learner-led* approach.

To answer this research question, and test our assumptions, we conducted a controlled experiment investigating the effects of different training strategies on recognition accuracy and user's understanding of the recognition process. During training, gesture classes, i.e., the target movements to be triggered by the prosthesis are selected sequentially

according to three conditions: either at random (control condition), by the participants themselves, or by a gesture-separability algorithm. Then, participants must demonstrate the corresponding muscular contraction to create a training set and update the recognition model. They visualised a prosthetic hand performing the recognised gesture as feedback. Our contributions are twofold: 1) We empirically identified model training conditions that leverage recognition accuracy and user understanding of an ML-based myoelectric prosthesis, and 2) we provide directions for research and for the design of training sessions that could support users of myoelectric prostheses.

2 RELATED WORK

We first present how training has been designed for ML-based prosthesis control and then report findings from the Machine Teaching literature, which shed light on how humans teach concepts to machine learners.

2.1 User training for ML-based prosthesis control

2.1.1 User training strategies. The importance of user-training in the ecological context of prosthesis control through pattern recognition of muscle activity was first highlighted by Powell et al. [31]. In a ten-day experiment, they showed that amputees improved their control performance with the help of visual feedback, which showed the movement of a virtual prosthesis. Subjects were coached by the experimenter to produce more separable and consistent gestures. However, another study with able-bodied subjects [21] showed that subjects can improve with training alone, independently of feedback or with coaching during training. More recently, research has shown that feedback in the form of visualization of EMG data points in a Linear Discriminant Analysis classifier's (LDA) feature space can accelerate user-training [12, 13]. De Montivalet et al. [12] showed that using a continuous feedback based on the separability index of class means improved the accuracy of the retrained class without affecting overall accuracy. When compared to using the labels of the classifier as feedback, this effect was larger.

Since training alone can help humans provide better EMG patterns, it remains to be seen whether other training parameters can lead to further improvements. Only a few studies [2, 44] explored the effects of training curricula in a prosthesis context. However, this remains to be explored for ML-based prostheses. These studies applied principles derived from motor learning research [25, 33, 41] to determine effective training methods. For example, Bouwsema et al. [2] studied the order of practice tasks (based on functional uses of prosthesis) and their effect on movement time with a myoelectric simulator. Their findings suggest that performance in daily life is independent of training structure, but a blocked practice leads to faster learning than a random practice.

2.1.2 User training phases. Many training paradigms described in the literature involve iterating over one or several of the following phases: data collection, user training and testing [12, 20, 22, 31]. The data related to a gesture example is labelled automatically when a gesture is cued, and the procedure involves retraining the classifier on new data. This is sometimes called Supervised Recalibration. For example, De Montivalet et al. [12] retrain a specific gesture class after collecting data from all gesture classes. Users only get a chance to test the classifier at the testing phase, after a data collection phase to train the gesture classifier, and thus the user can only make adjustments to the classifier in the next training phase. This restrains human learning [4]. To address this problem, Fang et al. [13] adopt incremental training and users are provided with classifier feedback at each training trial. Nishikawa et al. [28] tackle this problem through an online learning mechanism. The user provides teaching signals when the virtual hand is moving unsatisfactorily, to generate new data. Undesirable data is eliminated based on different metrics. They show that this learning and data elimination mechanism simplifies decision boundaries. In both studies, gesture classes are cued randomly. Finally,

Szymaniak et al. [42] diverge from this 3-phase training protocol by proposing an active learning framework to address the cost of data labelling. Assuming that a dataset of EMG data can be collected prior to training, they simulate different sampling strategies and conclude that active learning surpasses random sampling.

The research in ML-based myoelectric control sought different ways to impart basic concepts of Machine Learning control to users during training, so that they can retrain their classifiers more efficiently. To understand how users can better steer model training in the desired way, we now look at how humans have taken the role of teachers to teach a machine.

2.2 Guiding humans to teach machines

Research in interactive machine learning (IML) and human-robot interaction (HRI) investigated how guidance can help human teachers convey concepts to a machine learner. Teaching guidance mainly takes two forms: *delegation of initiative* and *instructions*. Delegation of initiative mainly explored active learning (AL), i.e., the possibility for the machine learner to be curious and query novel and informative examples [40]. Cakmak et al. [5] investigated how humans teach a learning robot in a finite concept space (conjunction of nominal features). Their study compared self-directed teaching (no learner's query) with three active learning variants: full AL, mixed-initiative AL, or human-controlled AL. They found that active learning conditions, including human-controlled AL, resulted in significantly higher F1 scores than self-directed teaching. In another study, Cakmak et al. [8] explored instructional teaching guidance across more complex and realistic tasks, including binary classification of sketches: participants were primed with different heuristics, e.g. an efficient teaching strategy must sample examples close to the decision boundary. Their results found that instructional guidance influenced human teaching toward being more beneficial for the machine learner's accuracy. The authors pointed out participants' tendency to provide typical rather than borderline examples when unguided, to be more successful at curating positive rather than negative examples, and to forget what examples were already shown to the system. Several authors [8, 10, 37] concur about the complex entanglement between the interaction workflow, concept space (complexity of the concept), the learning algorithm, and the data domain. A large benchmark conducted by Pereira et al. [29] concerning Active Learners revealed that performance gains are uneven across models, domains, and to a lesser extent, sampling strategies. Preliminary findings by Sanchez et al. [37] comparing simulated active learning with real human teaching curricula in a multi-class recognition of images with a deep learning model also contradict Cakmak et al. [5], as self-directed human teachers seem to outperform simulated active learners.

Our work contributes to uncover the complex interaction between human teacher and machine learner by studying a real-world problem with infinite concept spaces and little-studied domains: the multi-class recognition of EMG signals. The proprioceptive nature of muscular contraction makes it challenging to design similar experiments to those in the visual domain. Active learning scenarios in which the learner queries output labels from input trajectories are not feasible, as EMG signals are hard to represent and too distinctive. Instead, our work investigates demo queries [7], i.e., the active learner chooses a class and queries a teacher's demonstration. Demo queries, also called active class selection [26], afford greater human control than label queries as users can still induce variations in the input data points.

2.3 Humans' mental models in interactive ML

The research literature involving a human teacher and a machine learner focuses primarily on model performance. However, an essential aspect in ML-based myoelectric control is users' comprehension of the model, also referred to as the user's mental model. Mental models about a machine learner can be *functional*—the comprehension of the

model's learning or final behavior— and *structural*—the understanding of the machine learner's underlying mechanisms (e.g., neural network, k-NN, etc.). In our case, we are interested in assessing participants' *functional* mental model, since the comprehension of the learner's behavior—its strengths and weaknesses—can help users adapt their muscular contractions and avoid errors in real-life scenarios.

The approaches employed to assess users' mental model of an ML system in the interactive ML literature often triangulate quantitative (objective understanding) and qualitative (self-reported understanding), including: 1) assessments of users' predictive accuracy (quantitative, objective), i.e., users' ability to predict the system's predictions [9, 39]; 2) verbalization analysis (qualitative, self-reported) from interviews [24, 39], think-aloud protocols [38, 39], or written responses to open-ended questions; 3) analysis of response from close-ended questions (qualitative, self-reported) during post-hoc surveys [5, 17, 19, 23, 24]. The research studies cited span the fields of interactive ML, human-robot interaction, and explainable AI and offer valuable insights for our work.

On the influence of learner's performance on users' comprehension, Hedlund et al. [17] found that the machine learner's (a reinforcement algorithm) performance can affect a teacher's mental model about the learner (capacity, reliability) and their own teaching capacity. Sanchez et al. [39] investigated how people use and understand two types of uncertainty feedback while teaching an image classifier. Although neither epistemic nor aleatoric uncertainty feedback improved users' predictive accuracy compared to each other, they also found a strong correlation between users' predictive accuracy (e.g., objective understanding) and the machine learner's final accuracy. This confirms the intuition that accurate models are easier to comprehend than inaccurate ones.

On the influence of curriculum on users' comprehension, Cakmak et al. [5] also conducted subjective ratings from a post-hoc survey, which revealed that people had a better mental model of the system's performance in the 3 active learning (AL) conditions rather than self-directed teaching. Participants, however, reported less engagement in the full AL condition. Using a think-aloud protocol, Sanchez et al. [38] explored humans' understanding when incrementally teaching a deep neural network to recognize sketches. Their results underscore the importance of the training sequence, both for the classifier performance and user understanding of the model behavior. In particular, curating examples which are imbalanced across classes at the beginning of the teaching was detrimental to learners' performances and users' comprehension. Researchers [5, 38, 43] converge on the usefulness of guidance prior or early in the teaching phase. Letting users interact with data trigger investigative behaviors [38] but requires engagement and time [9], suggesting a trade-off between efficiency and understanding.

Our work also combines methods to assess users' functional mental model: we investigate both objective (users' predictive abilities) and self-reported understanding (think-aloud protocol and survey). Our results might shed light on the relationship between the training curriculum and users' functional mental model in the specific use case of ML-based myoelectric prosthesis control. This use case is unique compared to the rest of the literature presented above, as it involves complex signals that results from users' implicit knowledge and fine-grained proprioception.

3 USER STUDY

We want to evaluate participants' ability to build accurate classification models of muscle activity based on different user-training strategies, and to investigate to what extent they foster the development of accurate mental models of the trained classifiers. We compare three model-training conditions under which participants incrementally train the gesture classifier: (1) Random-Condition (RC), which cued gesture classes randomly; (2) Teacher-led Condition (TLC), in which the user could choose the gesture classes for which to give examples; and (3) Learner-led Condition (LLC), which cued gesture classes based on their separability in the feature space. To assess the effects of the training condition on

participants' functional mental models, we combined objective and self-reported insights. In this section, we provide the details of the experimental protocol.

3.1 Participants

51 people participated in this study (25 males,25 females, and 1 non-binary, aged between 18 and 36, M = 24.8, SD = 4.2). We recruited able-bodied participants to ensure comparable prior knowledge and sensorimotor expertise with myoelectric control technology. They were required to have no neurological or upper-extremity musculoskeletal problems that might influence performance. They received 15 euros as compensation. Participants' experience in Machine Learning and Robotics in Healthcare varied from novice (no knowledge) to advanced (experienced in Machine Learning / Robotics in Healthcare) (see table 1). The University's Ethical Research Committee approved all experimental procedures.

	Knowledge in Healthcare Robotics	Knowledge in Machine Learning
(Expert)	0	0
	1	3
↓	3	7
	14	12
(No knowledge)	33	29

Table 1. Number of participants having knowledge in (1) Robotics and (2) Machine Learning. Ratings were provided on a likert scale of 0 (expert) to 5(no knowledge).

3.2 Task

We asked participants to teach the best possible recognition system for the following 8 gestures: (1) Palm Down, (2) Palm Up (3) Close Hand (4) Open Hand, (5) Close Pinch, (6) Open Pinch, (7) Rest Hand, (8) Point Index. These eight gestures are depicted as pictograms in Figure 1 and were specifically chosen because they are common actions to be performed by upper-limb amputated users fitted with current commercial myoelectric prostheses. They define the typical gesture set used in studies in rehabilitation engineering focusing on myoelectric prosthesis [12].

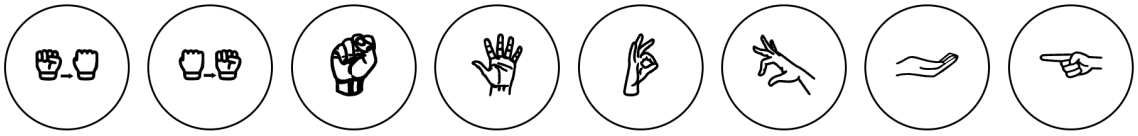


Fig. 1. The ML-based myoelectric prosthesis had to learn a set of gestures from examples provided by participants. We considered 8 gestures, from left to right: Palm Down, Palm Up, Close Hand, Open Hand, Close Pinch, Open Pinch, Rest Hand, and Point Index.

Training the best gesture recognition system implies that participants must consistently produce precise muscle contractions for the eight gesture categories, ensuring that the resulting model achieves minimal recognition errors. We implement an incremental training protocol, where the model is retrained for each new example demonstrated by participants.

As further detailed below, the interface shows gesture icons that indicate to the participants which gesture to perform now. The icon represents the gesture class used as a ground truth label (see Figure 1).

3.3 Apparatus and Setup

The experiment consists of four components: a muscle contraction sensor, a web-based interface, a Python-based software module, and a prosthetic arm. Figure 8a illustrates the setup with the various components we explain in this section. The surface myoelectric activity of the forearm is measured with a dedicated device, a Myo armband from Thalmic Labs. The Myo armband comprises 8 channels to record electro-physiological signals with an 8-bit resolution. No specific skin preparation was used before placing the armband on the participant's forearm, approximately 5 centimeters from the elbow joint.

EMG data are sent to the Python-based software module. Features are extracted from the Root Mean Square (RMS) of the raw EMG signal. It was computed from the surface EMG (sEMG) over a 128-ms-sliding analysis window, with a 32-ms overlap between successive windows. The dataset is available on Zenodo: https://doi.org/10.5281/zenodo.10528482. Among the wide variety of features that have been investigated in the literature [30], RMS features are known to be the most efficient and robust processing for classification of sEMG with LDA. In addition, they are standard features in the literature on myoelectric prosthesis arms, allowing the community to compare our results with other research such as Roche et al. [36] or De Montalivet et al. [12]. We recorded about 2 seconds of data and averaged the features to obtain a vector of 8 RMS values for the 8 channels of the Myo Armband. For classification, we used a Linear Discriminant Analysis (LDA) classifier (taken from the Scikit-learn python library [1]). We chose the LDA since it is the most used algorithm in an ecological setting to classify EMG signals [12, 45]. In addition, LDA, compared to more recent approaches with neural networks [27, 34], allows for fast training updates of the model, enabling incremental training. The Python-based software module sends classification predictions to the web-based interface through a web socket. The Python-based software module also sends the classification prediction to the prosthetic arm through WiFi to provide feedback on model performance to the participant.

The web-based interface was created using the Marcelle toolkit[2], a modular open-source toolkit for programming interactive machine learning applications. The interface shows the icons described above, and a capture button used by participants to record muscle contraction associated with gesture classes. Participants are told to hold the gesture and click on the 'Capture' button, which records EMG data for approximately 2 seconds.

To provide motion feedback, a TASKA, NZ prosthetic hand [3] along with a conventional electric wrist rotator is used. The prosthetic hand was programmed to enact the 8 gesture classes.

The training pipeline is as follows: a gesture class is selected (by the system or the participant, depending on the condition), and the participant provides an example of that gesture class through muscle contractions. After data windowing, the RMS is computed from the acquired EMG data. This feature is then used for classification with LDA. The predicted class is sent to the prosthetic hand to trigger the appropriate motor command.

3.4 Conditions and experimental design

We used a between-participant experimental design, where we compared three conditions, each with different ways of selecting gestures for training.

- `random condition (RC)`: at each trial during training, the system queries a gesture to be executed by the participant. This gesture is picked pseudo-randomly by shuffling the set of 8 gesture classes. Consequently,

[1] https://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.LinearDiscriminantAnalysis.html
[2] https://marcelle.dev [14]
[3] https://www.taskaprosthetics.com/

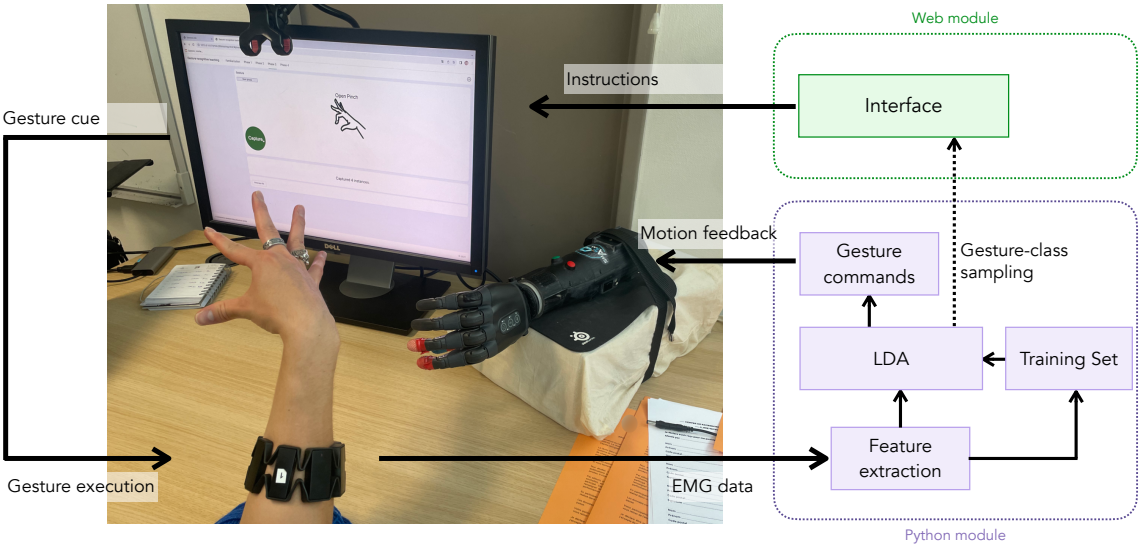


Fig. 2. Experiment setup: The participant wears the Myo armband on the forearm and performs a gesture. A screen in front of the participant displays the interface for capturing data. The interface for the random condition (RC) is shown here. The interfaces for the other conditions can be found in the appendix (see section 7.1). On the left, a prosthetic forearm (composed of a wrist rotator and a prosthetic hand) provides feedback during the training phase.

the number of gestures executed by a participant is balanced among gesture classes. There is no consecutive repetition of a gesture class, i.e., gesture classes always change from one trial to another.

- `teacher-led condition (TLC)`: at each trial during training, the participant selects the gesture class for which they wish to produce an example. The number of gestures per class executed by a participant depends on the sequence created, which can lead to an imbalanced dataset across classes. Participants (i.e. teachers) can also choose to provide several consecutive examples for a gesture class.
- `learner-led condition (LLC)`: the system (i.e. the learner) sorts gesture classes based on a separability index and queries the least separable gesture class at each trial during training. We compute the gesture class separability index as a ratio of the inter-class variance to the intra-class variance computed on the captured data (EMG features; see more details below). The intra-class variance for a class c is computed as the mean-variance of EMG data within c. The inter-class variance is computed as the mean of the variance of EMG data from class c and every other class c' different from c. A pilot study showed that, in practice, this strategy can create a highly imbalanced dataset depending on the initial data points. Therefore, to mitigate the curation of imbalanced training datasets, the gesture separability index is weighted by the inverse number of instances for this gesture (see Algorithm 1).

Two of the conditions, namely teacher-led and learner-led, can lead to training datasets which are not balanced. A common assumption in supervised machine learning is that training data points are independent and identically distributed (i.d.d) regarding the problem under scrutiny. This assumption is rarely met in simulated or human teaching, as well-explained in Zhu et al. [46]. Teaching sets are often non-independent and identically distributed (non-i.i.d.) as they might be designed, for instance, to mitigate the inherent ambiguity between two classes. For this reason, we do not enforce balanced training sets in the teacher-led and learner-led conditions in our experiment. Strictly forcing an

Algorithm 1: Gesture Class Separability Index

Input: A training dataset: $(X_{\text{train}}, y_{\text{train}})$
Output: Per-class separability index: SI
for c *in* $NumClasses$ **do**
$SI(c) \leftarrow \frac{Var_{\text{inter-class}}(c)}{Var_{\text{intra-class}}(c)}$;
$SI(c) \leftarrow SI(c) * \frac{ClassSize(c)}{\sum_{n=1}^{NumClasses} ClassSize(n)}$;
end

equal number of instances per class, besides the random condition, might undermine a crucial characteristic of teaching: planning and learner adaptivity.

3.5 Procedure

We welcomed participants and explained the task, the different phases, and the interface. Participants were fitted with a Myo armband. They were asked to place the elbow of the dominant arm on a ball of wool for comfort and to maintain this position throughout the experiment. To familiarize participants with the gestures, they were shown icons of each gesture and an image showing the corresponding gesture. Participants in all groups performed four phases (the procedure is also depicted in Figure 3).

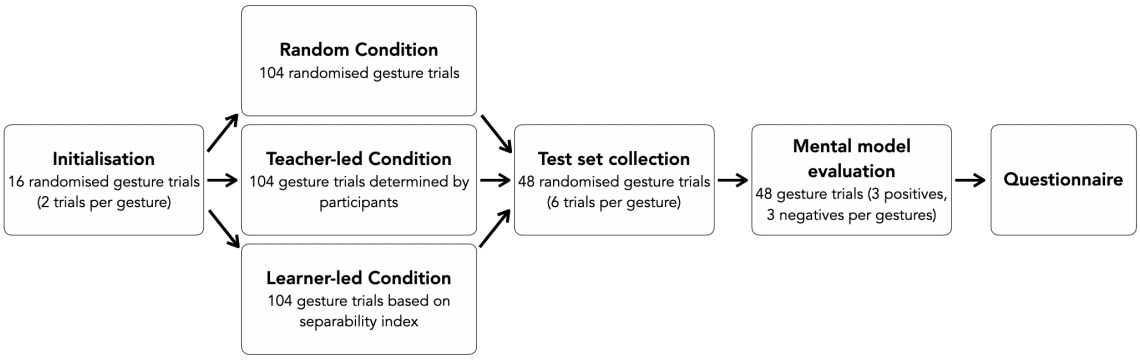


Fig. 3. The 5-phase experimental design. It involved a between-subject comparison of three conditions. Participants received feedback only during the training phase.

(1) Initialization. This phase aims to provide an initial dataset to train the gesture classification model before starting the training phase. Participants are asked to record two gesture examples per class (16 trials in total). Each gesture query is picked according to a pseudo-random strategy. There is no repetition of gestures.

(2) Training. During the training phase, participants are asked to give an example of a gesture (RC and LLC) or to choose a gesture for which to provide an example (TLC). Participants are briefed about the conditions: in the teacher-led condition (TLC), they are told to choose which gesture they want to train, and in the random condition (RC), they are told that the gesture classes are proposed in a random order. Finally, in the learner-led condition (LLC), participants are told that the most inconsistent gesture classes are prompted. In the RC and LLC conditions, gestures are presented as icons, and in the TLC condition, participants select gesture classes by clicking on the corresponding button illustrated by the icon. After each example, the classifier is trained on the

entire dataset collected until this point, and the participant receives gesture feedback from the prosthetic hand. For instance, if a participant performs the "close hand" gesture, and if the classifier correctly recognizes this example as "close hand", the prosthetic hand closes. Participants are told to pay attention to the feedback, which could be leveraged to reflect on future gestures. This phase includes 104 trials. The number of trials was decided based on a pilot study, which showed that the learning curve of the classification model plateaued after around 100 trials. In total, the aggregation of the examples from both the initialization and training phases leads to 120 gesture examples in the training set.

(3) Post-test. To assess the model's final accuracy after training, the participant completes a post-test. This post-test served as a data collection phase, comprising 48 trials to create a test set of 6 trials per gesture class. This phase was similar to the initialization phase in that the gestures were collected randomly and without any feedback.

(4) Positives-Negatives. To assess participants' mental model of the system, they were asked to complete a task that involved providing gestures while explaining their reasoning. In this task, participants were prompted to give three gesture examples that would be correctly recognized, called *positives*, and three examples that would incorrectly be recognized, called *negatives*. They were told that they could give different examples or repeat their examples. The examples could be drawn from observation during training or from their own understanding and interpretation. Participants were asked to verbalize (think aloud) their strategies to produce positive and negative examples. Therefore, this phase comprised 48 trials. Each participant, across conditions, experienced the same gesture order. The 3 positives were queried in a row, followed by the 3 negatives.

(5) Questionnaire. Finally, participants were asked to answer a questionnaire where they rated the perceived accuracy of gestures of the trained model. Perceived accuracy is measured on a scale of 0 to 100, where 0 means never correctly recognized and 100 means always correctly recognized.

3.6 Data Collection and analysis

We logged each example recorded by participants and the classifier trained by users after each trial to compute the model accuracy along trials. During the 'Positives-Negatives' phase, we collected additional EMG data and participants' verbalizations with audio recordings. Finally, we collected answers to the post-study questionnaire (cf. Appendix 7.2) to assess the participant's mental model. Based on these collected data, we computed the following measures:

Model Accuracy. We compute the model accuracy by selecting a trained model at a given trial, from trial 16 (after initialization) to trial 120 (at the end of training), and use data collected during the post-test phase as test set. The test set is balanced and contains 6 instances of each gesture. Thus, using Sklearn's API, we compute the accuracy score at each trial, computed as the fraction of correctly classified samples. Accuracy is computed per gesture class to take into account the variation between classes when calculating statistics between conditions.

Gesture separability. To understand how the separability of the EMG data evolves over time, we compute a measure of participants' ability to produce consistent and separable gestures over 3 phases: the start of training (16 gestures), end of training (120 gestures) and the post-test set (48 gestures). In the next section, these 3 phases will be designated as the variable Phase. Gesture separability is computed similarly to the separability index used in condition LLC: we compute the gesture class separability as a ratio of the inter-class variance to the intra-class variance computed on the captured data. This computation is done pairwise to ensure a comparable metric between the beginning of training (where only 2 instances per class are available) and the end of training.

True Positive and True Negative rates. To quantify differences in participants' understanding of two types of tasks, i.e., their capacity to identify positive and negative examples of gestures, we calculate the true positive rate, which is

the quantity of correctly classified gestures among the set of per-class positives. We then calculate the true negative rate for the negative examples, which is, in our case, the quantity of incorrectly classified gestures in the per-class negative examples. For both rates, high values mean an accurate model for positives and, respectively, a model that handles negatives correctly.

Participants' perception of class accuracy. We asked participants to complete a post-study questionnaire to assess their ability to keep track of training (cf Appendix 7). We compute the correlation between the perceived accuracy for each class and the actual accuracy provided by the model at the end of the training phase, using data collected during the post-test set as test set.

To compare each participant's answers in relation to all the other participants, we transform the raw accuracy values to standard scores (z-score).

4 RESULTS

In this section, we present the findings of the user study, starting with the analysis of the performance of the models built by participants, followed by the analysis of the mental models of the participants after having trained their gesture classifiers.

4.1 Machine learner's performance

In this section, we analyse the machine learner's performance and the extent to which the created training sets include separable gesture classes according to the training condition.

4.1.1 Model accuracy. We first analyzed whether the initial models built by participants (model comprising the initial 16 trials) are different between conditions, i.e. they result in significantly different accuracy values if tested on the post-test data. The mean accuracies for the initial models are as follows: 30.14 %(RC), 37.62 %(TLC) and 30.27 %(LLC). A Shapiro-Wilk test shows that the per-class accuracy values do not follow a normal distribution ($p < 0.05$). We conducted a non-parametric Kruskal-Wallis test, which shows no significant differences between initial models trained for each condition ($\chi^2 = 3.30, df = 2, p = 0.19$).

We then examined how the accuracy of the models changed based on the training conditions. We assessed model accuracy after each new gesture was demonstrated and added to the training set. Since the model is updated after each new gesture, the number of user trials equals the size of the training set, ranging from 1 to 104. We computed the model accuracy for each trial using post-test data as test set. The average learning curves for each condition and the 95% confidence intervals are shown in Figure 4a.

We evaluated the final model from the training phase, which is trained on all the examples demonstrated by users, using the post-test data as test set. The mean model accuracies for each training condition are as follows: 70.83% for RC, 65.81% for TLC, and 68.38% for LLC. To verify whether these means are significantly different, we conducted a Kruskal-Wallis test, taking ACCURACY as dependent variable and CONDITION as independent variable. We found no significant differences between training conditions on the per-class model accuracy ($\chi^2 = 0.74, df = 2, p = 0.69$). Therefore, the three conditions lead to equivalent model performance after the training phase.

Even though there were no statistical differences between the mean accuracies of the model at the beginning of training, we observed a variability between the initial mean values, which could suggest a certain level of heterogeneity between groups. Therefore, we carried out a complementary analysis in which we considered the average delta accuracy (difference between the final and initial accuracy). Figure 4b reports the results. A Kruskal-Wallis test, taking DELTA

ACCURACY as dependent variable and CONDITION as independent variable, shows a significant difference between conditions ($\chi^2 = 19.48, df = 2, p < 0.001$). A posthoc analysis with Dunn's test showed that LLC leads to a significantly higher improvement in accuracy compared to RC ($p < 0.05$) and TLC ($p < 0.001$), and there is a borderline difference between RC and TLC ($p = 0.053$). Consequently, although there is no difference between model accuracy created between the conditions, the performance improvement is larger when gestures are queried based on their separability and smaller when users choose the gestures to be trained.

Finally, we measured the learning rate to assess the evolution of skill development during training. We fit a linear regression model to the training set size (i.e., number of trials) and task performance. The logarithm of task performance plotted against the logarithm of the number of trials gives a straight line. To compare learning rates across conditions, we conducted a one-way ANOVA, taking LEARNING RATE as dependent variable and CONDITION as independent variable (a Shapiro-Wilk normality test shows that data are normally distributed ($p = 0.62$)). The ANOVA shows a significant effect of CONDITION (p<0.001). A post-hoc analysis shows that the learning rate is higher for TLC than LLC ($t = -4.655, df = 31.369, p < 0.001$). Furthermore, the learning rate is higher for RC than LLC ($t = 2.3891, df = 30.757, p = 0.023$).

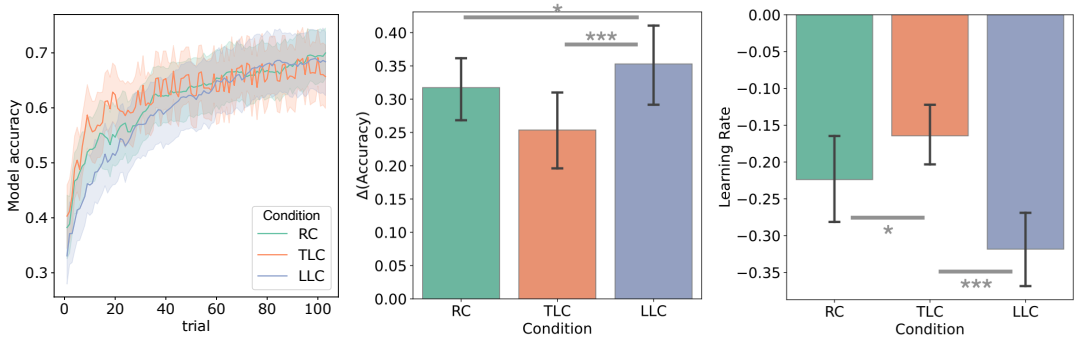


Fig. 4. (a) Model accuracy evolution during training. (b) The difference in model accuracy between the beginning and the end of the training phase for each condition. (c) Mean learning rates of model error reduction for each condition.

- **Finding 1.** Performance improvements of the ML-based myoelectric prosthesis are higher when gesture classes are queried to optimize gesture separability than in the random or Teacher-led curriculum.
- **Finding 2.** Performance improvements of the ML-based myoelectric prosthesis are faster when the user chooses gesture classes to demonstrate rather than queried to optimize gesture-separability or at random.

4.1.2 Gesture Class Separability. Incrementally training the gesture recognition model led to datasets that may differ across conditions. In this subsection, we analyze these demonstration datasets regarding gesture class separability. High gesture class separability implies concentrated gesture class clusters (low intra-class variance) far from the other (high inter-class variance). This index of gesture class separability is computed similarly to condition LLC (cf Section 3.6). Therefore, this measure indicates the consistency of the executed gestures, where more consistent gesture classes will lead to higher separability. Figure 5 depicts the mean separability values per condition at three different phases: at the beginning of the training session (after the first training trial), at the end of the training phase (after the last training trial) and at the post-test phase. First, we performed a Kruskal-Wallis test at each phase, considering CONDITION as independent variable and SEPARABILITY as dependent variable. We found a significant effect of CONDITION only at

the end of the training ($\chi^2 = 18.3, df = 2, p < .001$) where gesture class separability is higher for LLC than in the other conditions. We also inspected how the values of separability vary according to the phase. We performed Kruskal-Wallis non-parametric tests with Phase as an independent variable and Separability as dependent variable, for each Condition. We found that Phase has a significant effect on Separability for each condition ($\chi^2 = 13.5, df = 2, p < .01$ for RC, $\chi^2 = 30.2, df = 2, p < .001$ for LLC and $\chi^2 = 19.9, df = 2, p < .001$ for TLC). For each condition, gesture class separability values are higher at the end of training and at the post-test compared to the separability at the beginning of training. In addition, for LLC, separability is higher at the end of the training phase than post-test, as shown by Dunn's test , $p = 0.02$. It shows that overall, the gesture class separability increases, which can be related to the improvement in performing gestures during training, with a higher increase for LLC. However, at post-test, gesture class separability values are comparable across conditions.

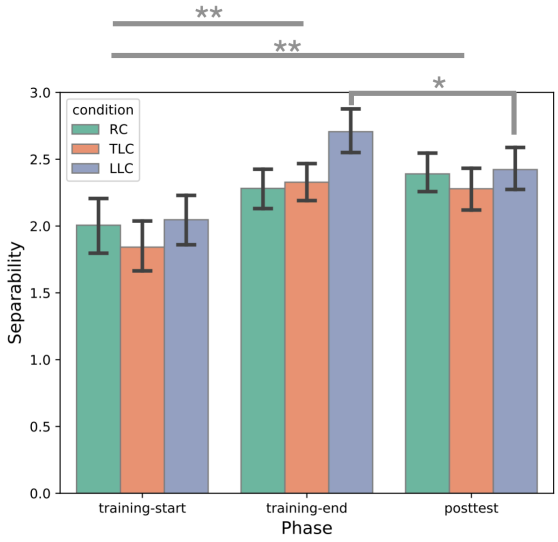


Fig. 5. Separability of gestures along different phases: at the start of training (after the first training trial) at the end of training (after the last training trial) and at post-test.

- **Finding 3.** The separability of gesture demonstrations increases during training, irrespective of the condition.
- **Finding 4.** Increase of gesture separability is significantly higher when gesture classes are queried to optimize gesture separability than in the Random or Teacher-led condition.

4.2 Human teachers' mental model

In this section, we assess the mental models of participants after they built their gesture classifier.

We assessed participants' mental model following two methods. First, we assessed participants' ability to predict where the machine learning model will make correct or incorrect predictions. We present results regarding participants' ability to execute correctly classified (positives) or incorrectly classified (negatives) examples for each gesture class. Second, we assess self-reported understanding through participants' answers to a post-training questionnaire on their perceived accuracy for each gesture class.

4.2.1 Participants' predictive abilities about the model. We want to test participants' ability to produce gestures of each class, which will be correctly recognized by the system (positives), and gestures that will not be correctly recognized (negatives). We evaluate the true positive and true negative rates of examples collected for both types (positive and negative), as well as participants' interpretation of these two types of examples. To compare the rates of true positive and true negative for each class and between conditions, we first conduct a Kruskal-Wallis test with True Positive Rate as dependent variable and Condition as independent variable. This yields a significant difference between conditions ($\chi^2 = 7.31, df = 2, p = 0.026$). Pairwise comparisons using Dunn's test show that the mean true positive rate is higher for condition LLC than condition TLC ($z = 2.49, p = 0.039$, see Figure 6). There is a borderline effect between RC and TLC ($z = 2.16, p = 0.061$) and no difference between RC and TLC ($z = 0.32, p = 0.75$). A similar test considering True Negative Rate as dependent variable and Condition as independent variable shows no differences between conditions for the negative examples. Finally, a Kruskal-Wallis test considering the type of test (True Positive Rate and True Negative Rate) as independent variable and the rate as dependent variable shows that True Negative Rates are significantly higher than True Positive Rates ($\chi^2 = 25.26, df = 1, p < 0.001$). Overall, participants' understanding of positive examples is worst for condition TLC, while participants' understanding of negative examples does not differ across conditions.

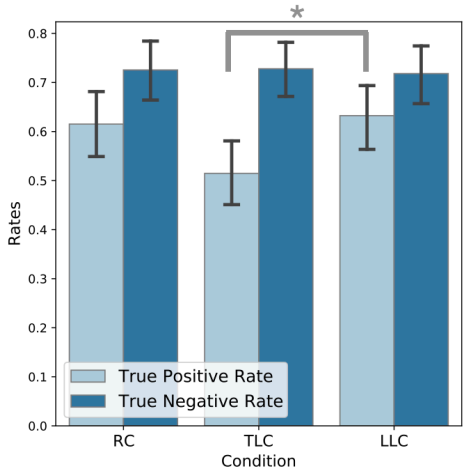


Fig. 6. True positive and negative rates from examples provided by participants after training. Participants have a less good mental model of what it means to produce positive examples of gestures when trained under condition TLC.

During the think-aloud phase, a number of participants (P4, P8, P19, P20, P41) explained that to create a positive, they exaggerated the gesture. For example, while giving a positive example of the gesture 'Palm Down', P41 said : *"[..]rotate the hand until I hurt myself"*. Some participants considered positive examples of gestures to be clear, slow, and precise : "*high up and poised, like this*" [P19], or "*I'm going to make a slow, detailed movement*" [P4]. On the other hand, negative examples were considered as fast, incomplete, and casual: "*A negative would be a bit of a jerky movement, like this. Maybe it's too fast for it*" [P19], or "*a lighter rotation*" [P1].

A few participants mentioned that it was difficult to imagine negative examples. P21 pointed out: "*In general, the robot recognizes the 'Close Hand' at all times, it'll be hard to create a negative for that*". Some participants who encountered errors during training easily identified examples that could confuse the system. While giving a negative example of

'Close Hand', P31 said: "*Earlier, it interpreted [Close Hand] as a downward rotational movement. I'm actually going to make a tiny downward movement too*". Besides recalling training errors, another strategy employed by participants to create negative examples was to combine gestures. For instance, while creating a negative example of 'Palm Down,' P23 said: "*It's a mixture of Close Hand and Palm Down*", which may have contributed in deceiving the model.

- **Finding 5.** Participants' predictive abilities about the trained model are above random. Surprisingly, participants can better reproduce negative than positive examples.
- **Finding 6.** Participants' ability to reproduce correctly identified examples is significantly lower when the user chooses gesture classes to demonstrate rather than queried to optimize gesture-separability or at random.

4.2.2 Participants' perception of class accuracy. As a second evaluation of mental models, participants were required to answer a questionnaire about their perception of how well the system learned to recognize each gesture. Participants' answers about their perceived accuracy of each gesture class had to be chosen among 11 discrete percentage values ranging from 0 (bad recognition) to 100 (excellent recognition). Since participants' answers depend on their rating style and the scale provided (e.g., a participant's perception of a good recognition can be 80% while another's perception can be 100%), we transform all values to z-scores. This allows us to consider participants' answers relative to each other. Then, we computed a Pearson correlation coefficient to assess the linear relationship between perceived accuracy for each gesture class at the end of training and the actual accuracy computed on the post-test data at the end of training. Figure 7 shows the results. The results indicate a positive relationship for LLC (Pearson's r score is 0.30, with $p < 0.001$, see Figure 7c) while there is no significant correlation for the other two conditions TLC and RC (see Figures 7a and 7b). This suggests that condition LLC allowed participants to better keep track of classification errors during training.

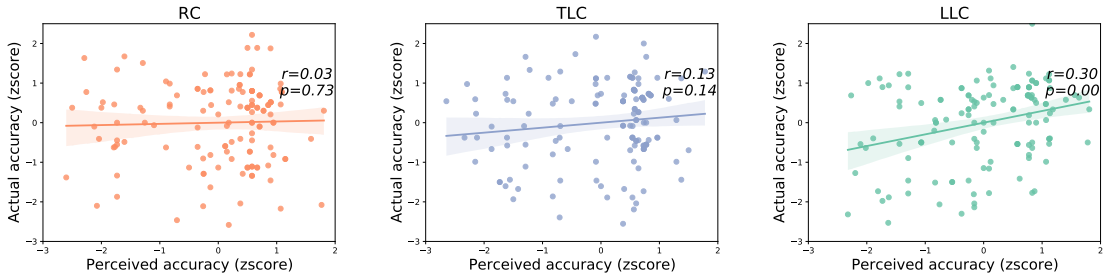


Fig. 7. Correlation of z-scores of actual accuracy values (y axis) and the corresponding z-scores of participants' perceived accuracy values (x axis) for each gesture class. The correlation is significant for condition LLC.

- **Finding 7.** Participants better perceive the ML-based myoelectric prosthesis accuracy when gesture classes are queried to optimize separability during training than for the Random or Teacher-led conditions.

5 DISCUSSION

In our study, we investigated user-driven teaching strategies for ML-based myoelectric prostheses and their influence on the machine learner's accuracy and the human teacher's mental model. From the machine learner perspective, we found that the model's performance maximally increases when gesture classes are queried to optimize gesture separability, outperforming both random and teacher-led strategies. That said, we also showed that training is faster when users

select the gesture class to be demonstrated. The three training strategies similarly led participants to increase the separability of examples across gesture classes, with a greater increase when gesture classes were queried to optimize gesture separability. From the human teacher perspective, participants' ability to predict the model's behaviors was more accurate, particularly in reproducing negative examples over positive ones. Participants' predictive abilities, though, diminish when they are left to select gesture classes during training, a discrepancy that is not observed with negative examples. Lastly, participants' perception of the model accuracy is better when gesture classes are queried to optimize separability during training. In the following subsections, we discuss the implications of the results, which can inform the design and further research on ML-based prosthesis control.

5.1 Influence of training strategy on model accuracy

Several aspects of a training strategy can contribute to its efficiency. We tested whether giving users the choice to select the gesture class to train (teacher-led condition) results in better recognition accuracy than imposing gesture classes to train (random condition and learner-led condition based on a gesture separability index). We found that the teacher-led strategy did not yield better accuracy than the other training strategies. This is in line with findings in the interactive machine teaching literature, where Cakmak et al. [5, 6] showed that self-directed human teaching is rarely optimal, including for tasks where the human teacher generates training data [8]. Conversely, we found that querying the least separable gesture class led to slower training despite a larger overall increase in accuracy. Previous works partially corroborate our results. Research showed that both instructional guidance [6, 8] and sharing initiative with an active learner [5] yield better accuracy than self-directed teaching. However, the same authors showed that teaching an active learner, even with partial initiative, is faster than self-directed teaching, i.e., active learners converge towards the maximal accuracy with fewer examples. The fact that our results do not align with Cakmak et al. [5] can be interpreted in the following way. First, the authors employed simplistic machine-teaching tasks using a finite concept space. Our work expands these results to a real-world problem with a complex and infinite concept space, and 8 output classes. Second, the implicit and proprioceptive nature of muscular contractions suggests different pathways, distinct from visual cues, to adapt to the machine learner. Third, unlike previous work, our study does not employ an active learner that queries labels but rather demonstrations. Such a scenario affords greater human control and engagement than label queries, and our results demonstrate its potential to support ML-based prosthesis control.

Teacher-led strategy resp. Learner-led strategy yield faster resp. greater performance increase. Looking at the problem solely from the point of view of learner performance, a direct implication would corroborate design guidelines outlined in previous works [5, 38], i.e., to share initiative at relevant moments of training. In our case, an optimal approach would be to let users direct the early stages of the training, and transition to the gesture-separability active learning later on. Alternatively, Powell et al. [31] highlighted the importance of coaching to incrementally incorporate gestures in a subject's training session based on the subject's capacity to create more separable EMG patterns. Looking at human-centered aspects, Dakpa et al. [11] showed that faster improvement of a classifier's performance can motivate users to pursue the training, hence leaning toward the use of teacher-led training in our use case. That being said, there is the risk that the averaged performance gaps obtained in section 4.1 might not be perceivable by a human teacher. Furthermore, it is likely that such results change with a different machine learner [29], or sensor device. Instead of scrutinizing averaged curricula performances, a promising research follow-up would be to investigate outliers, i.e., participants' curricula that led to extreme performances—either good or bad, compared to others. Second, we argue that human-oriented assessments, e.g., focused on engagement and comprehension, are more important than model

performances to involve users in the role of teacher. A valuable avenue for research would be to explore methods for users to effectively retrain their ML-based prosthesis during daily life activities.

5.2 Impact of training strategy on (human) skill acquisition

The gesture separability measure is used to assess the consistency of the demonstrations, i.e., to what extent gesture examples in the same class are close to each other, but far from examples in other classes. All conditions showed a trend towards higher separability across phases, with a larger increase from start to end when the least separable gesture is queried. Practice and repetition are necessary to build consistency in any motor movement [28, 32]. The first finding might indicate that participants learned how to perform gestures more consistently throughout the training session, i.e., participants improved their motor skills in the execution of gestures. In our case, the learner-led condition favored repetition between gestures of identical classes, suggesting that participants improved their motor skills better in this condition.

Conversely, increasing the variability of practice has been shown to improve the acquisition of motor skills [3]. This means that it is preferable to switch from one task to another rather than repeat the same task until it has been learned. This improvement in learning has been demonstrated in terms of retention and transfer. In particular, it has been shown that creating motor interference by switching task from one trial (or block of trials) to another decreases the rate of motor learning but increases retention [41]. An analogous observation in our case is that the randomized training curriculum prevented gesture class repetition by design. It is also likely that the teacher-led condition involved fewer repetitions than the learner-led strategy. Hence, investigating the learning effect, measured by gesture consistency, during the retention phase (typically after several days) presents a significant research opportunity. The insights gained could profoundly influence the application of such training protocols in real-world scenarios.

5.3 Users' understanding of gesture classification for prosthesis control

Our study uniquely investigated users' mental model of an ML-based myoelectric prosthesis. An accurate functional mental model, i.e., users' comprehension of the model's behavior—its strengths and weaknesses—can help users adapt their muscular contractions and avoid errors in real-life scenarios. Our evaluation method combines assessment of 1) users' predictive ability to execute gesture examples which will positively and negatively be classified, with 2) users' self-reported perception of the system's accuracy, i.e., questionnaire answers about their perceived accuracy of each gesture class. Both methods indicate the benefit of the learned-led strategy (based on gesture separability) on participants' mental model. On one hand, participants' ability to reproduce correctly identified examples is significantly higher in both the learner-led and the random condition than in the case where participants decided which gesture to train. No differences in demonstrating negative examples were found between conditions. On the other hand, users' perception of the system's accuracy was more accurate in the learner-led condition than in the teacher-led condition.

Two explanations can support these findings. First, the learner-led condition creates a curriculum that focuses on ambiguous classes, hence developing users' understanding of the most unstable gesture classes throughout training. A greater comprehension of borderline examples might translate to higher performance in guessing positive examples. This explanation corroborates with Cakmak et al. [5], who found that human teachers had a more accurate performance estimate of the learner in active learning modes, including human-controlled active learning, rather than self-directed teaching. The second explanation links with our previous results on the model's final accuracy: model behavior of an accurate model is easier to understand. Such an explanation would align with other empirical findings [17, 39] presented in the related-work section 2.3.

5.4 Limitations

Our study yielded one unexpected result: during the mental model assessment phase, participants were more able to produce negative than positive examples. This finding contradicts the verbalizations reported in section 4.2.1, where participants described demonstrating negative examples as more challenging. Participants' verbalizations also suggest that they approached positive examples differently than in earlier phases. Challenges in creating negative examples are reported in prior research [8]. This observation may be due to an experimental design artifact in the mental model assessment session. The think-aloud protocol, which required participants to demonstrate gestures while explaining their decisions aloud, might have imposed a higher cognitive load, resulting in changes in the execution of positive examples.

Asking participants to perform positive and negative examples is not a standard approach to assess participants' mental model, more precisely, their predictive abilities about the model's behavior. To assess participants' predictive accuracy, Sanchez et al. [39] sampled input data to be shown to participants, who had to guess if their model would correctly or incorrectly recognize the selected input. They also introduced an exploration phase after the teaching session in order for participants to familiarize with the final model's behavior. Cheng et al. [9] introduced variants of tests that assess participants' predictive ability about the model. These variants, called *unnamed attributes*, *alternative prediction*, and *decision prediction*, are also based on actual data points and suggested modifications. Querying positive and negative labels to participants from actual input data, as well as the variants introduced in Cheng et al. [9], are not feasible in the context of myoelectric prosthesis control as there is no way to cue a muscular contraction pattern: a signal visualization would not evocative. Furthermore, adding an exploration phase would have lengthened our experiment, potentially leading to participant fatigue, where prolonged engagement can reduce performance and increase errors due to mental exhaustion.

Finally, we assessed the participants on the basis of a signal which was the average EMG signal over two seconds of data collection. In doing so, we discarded the gestural strategies used by the participants to arrive at the final postures. Keeping the raw EMG signals would have allowed us to extend our assessment of users' ability to control a prosthesis by inspecting the extent to which they could achieve a given gesture. In particular, it would be interesting to examine the model predictions temporally during these two seconds of gesture stabilisation. Taking into account all the classification decisions made from the beginning to the end of the two seconds would have made it possible to measure the regularity with which a user was able to maintain the desired position once it had been reached and to devise more refined guidance strategies.

6 CONCLUSION

Modern myoelectric prostheses equipped with machine learning aim to provide a more personalized control than conventional methods. With ML-based prostheses, users become responsible for "teaching" the system with examples, i.e., demonstrating associations between their muscular contractions and the prosthesis' gestural response. In this article, we investigated model training strategies and tested their impact on the machine learner's performance and the human teacher's (i.e., the user's) comprehension of the system's behavior. Our lab experiment investigated three model training strategies: (1) the system cues gesture classes randomly (control), (2) the user selects gesture classes (teacher-led), (3) the system queries gesture classes based on their separability (learner-led).

Our findings indicate that both learner-led and teacher-led strategies have their merits, each contributing to important aspects of model and user training. The teacher-led strategy led to faster model accuracy increases early in the training.

The learner-led strategy resulted in larger increase in model accuracy, more consistent gestures, and more accurate mental models of users. These results highlight the potential of these teaching strategies in the context of prosthesis control, and suggest the benefit of demonstration queries to organize the curriculum, thus fostering accurate models and users' mental model. We discuss our results in the light of several bodies of research, namely myoelectric control, motor learning, human-robot interaction, and interactive machine teaching. Promising research directions include the design of training strategies with shared-initiative, experiments in ecological settings, as well as a thorough evaluation of users' engagement.

ACKNOWLEDGMENTS

This research was supported by the ARCOL project (ANR-19-CE33- 0001) – Interactive Reinforcement Co-Learning, from the French National Research Agency. The BYCEPS project (ANR-18-CE19-0004) funded the development of the prosthetic forearm. Data collection was made possible through INSEAD-Sorbonne Université's behavioral lab. Recruitment of participants was funded by Idex Sorbonne Université in the context of the program "Investissements d'Avenir". We wish to acknowledge and thank everyone involved in this research, and express our sincere gratitude to the anonymous reviewers for their constructive comments. We also wish to express special appreciation to Sébastien Mick for customizing the prosthetic forearm for this project.

REFERENCES

[1] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *Ai Magazine* 35, 4 (2014), 105–120.

[2] Hanneke Bouwsema, Corry K van der Sluis, and Raoul M Bongers. 2008. The role of order of practice in learning to handle an upper-limb prosthesis. *Archives of physical medicine and rehabilitation* 89, 9 (2008), 1759–1764.

[3] Daniel A Braun, Ad Aertsen, Daniel M Wolpert, and Carsten Mehring. 2009. Motor task variation induces structural learning. *Current Biology* 19, 4 (2009), 352–357.

[4] Nathan E Bunderson and Todd A Kuiken. 2012. Quantification of feature space changes with experience during electromyogram pattern recognition control. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 20, 3 (2012), 239–246.

[5] Maya Cakmak, Crystal Chao, and Andrea L Thomaz. 2010. Designing interactions for robot active learners. *IEEE Transactions on Autonomous Mental Development* 2, 2 (2010), 108–118.

[6] Maya Cakmak and Andrea L Thomaz. 2010. Optimality of human teachers for robot learners. In *2010 IEEE 9th International Conference on Development and Learning*. IEEE, 64–69.

[7] Maya Cakmak and Andrea L Thomaz. 2011. Active learning with mixed query types in learning from demonstration. In *Proc. of the ICML workshop on new developments in imitation learning*. Citeseer.

[8] Maya Cakmak and Andrea L Thomaz. 2014. Eliciting good teaching from humans for machine learners. *Artificial Intelligence* 217 (2014), 198–215.

[9] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[10] Yuchen Cui, Pallavi Koppol, Henny Admoni, Scott Niekum, Reid Simmons, Aaron Steinfeld, and Tesca Fitzgerald. 2021. Understanding the relationship between interactions and outcomes in human-in-the-loop machine learning. In *International Joint Conference on Artificial Intelligence*.

[11] R Dakpa and H Heger. 1997. Prosthetic management and training of adult upper limb amputees. *Current Orthopaedics* 11, 3 (1997), 193–202.

[12] Etienne de Montalivet, Kevin Bailly, Amélie Touillet, Noël Martinet, Jean Paysant, and Nathanael Jarrasse. 2020. Guiding the training of users with a pattern similarity biofeedback to improve the performance of myoelectric pattern recognition. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 28, 8 (2020), 1731–1741.

[13] Yinfeng Fang, Dalin Zhou, Kairu Li, and Honghai Liu. 2016. Interface prostheses with classifier-feedback-based user training. *IEEE transactions on biomedical engineering* 64, 11 (2016), 2575–2583.

[14] Jules Françoise, Baptiste Caramiaux, and Téo Sanchez. 2021. Marcelle: composing interactive machine learning workflows and interfaces. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. 39–53.

[15] Andreas Franzke. 2023. Machine-learning myoelectric prosthesis control: towards the advancement of assessing functional use and control skill. (2023).

[16] Jiayuan He, Dingguo Zhang, Ning Jiang, Xinjun Sheng, Dario Farina, and Xiangyang Zhu. 2015. User adaptation in long-term, open-loop myoelectric training: implications for EMG pattern recognition in prosthesis control. *Journal of neural engineering* 12, 4 (2015), 046005.

[17] Erin Hedlund, Michael Johnson, and Matthew Gombolay. 2021. The Effects of a Robot's Performance on Human Teachers for Learning from Demonstration Tasks. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 207–215.

[18] Janet L Kolodner, David Crismond, Jackie Gray, Jennifer Holbrook, and Sadhana Puntambekar. 1998. Learning by design from theory to practice. In *Proceedings of the international conference of the learning sciences*, Vol. 98. Atlanta, GA, 16–22.

[19] Samantha Krening and Karen M Feigh. 2018. Interaction algorithm effect on human experience with reinforcement learning. *ACM Transactions on Human-Robot Interaction (THRI)* 7, 2 (2018), 1–22.

[20] Morten B Kristoffersen, Andreas W Franzke, Raoul M Bongers, Michael Wand, Alessio Murgia, and Corry K van der Sluis. 2021. User training for machine learning controlled upper limb prostheses: a serious game approach. *Journal of neuroengineering and rehabilitation* 18, 1 (2021), 1–15.

[21] Morten B Kristoffersen, Andreas W Franzke, Corry K van der Sluis, Alessio Murgia, and Raoul M Bongers. 2019. The effect of feedback during training sessions on learning pattern-recognition-based prosthesis control. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 27, 10 (2019), 2087–2096.

[22] Morten B Kristoffersen, Andreas W Franzke, Corry K Van Der Sluis, Alessio Murgia, and Raoul M Bongers. 2020. Serious gaming to generate separated and consistent EMG patterns in pattern-recognition prosthesis control. *Biomedical Signal Processing and Control* 62 (2020), 102140.

[23] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more? The effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the sigchi conference on human factors in computing systems*. 1–10.

[24] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing*. IEEE, 3–10.

[25] Timothy D Lee and Richard A Magill. 1983. The locus of contextual interference in motor-skill acquisition. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 9, 4 (1983), 730.

[26] Rachel Lomasky, Carla E Brodley, Matthew Aernecke, David Walt, and Mark Friedl. 2007. Active class selection. In *Machine Learning: ECML 2007: 18th European Conference on Machine Learning, Warsaw, Poland, September 17-21, 2007. Proceedings 18*. Springer, 640–647.

[27] Nazmun Nahid, Arafat Rahman, and Md AR Ahad. 2020. Deep learning based surface EMG hand gesture classification for low-cost myoelectric prosthetic hand. In *2020 Joint 9th International Conference on Informatics, Electronics & Vision (ICIEV) and 2020 4th International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*. IEEE, 1–8.

[28] Daisuke Nishikawa, Wenwei Yu, Masaharu Maruishi, Ichiro Watanabe, Hiroshi Yokoi, Yukio Mano, and Yukinori Kakazu. 2000. On-line learning based electromyogram to forearm motion classifier with motor skill evaluation. *JSME International Journal Series C Mechanical Systems, Machine Elements and Manufacturing* 43, 4 (2000), 906–915.

[29] Davi Pereira-Santos, Ricardo Bastos Cavalcante Prudêncio, and André CPLF de Carvalho. 2019. Empirical investigation of active learning strategies. *Neurocomputing* 326 (2019), 15–27.

[30] Angkoon Phinyomark, Franck Quaine, Sylvie Charbonnier, Christine Serviere, Franck Tarpin-Bernard, and Yann Laurillau. 2013. EMG feature evaluation for improving myoelectric pattern recognition robustness. *Expert Systems with applications* 40, 12 (2013), 4832–4840.

[31] Michael A Powell, Rahul R Kaliki, and Nitish V Thakor. 2013. User training for pattern recognition-based myoelectric prostheses: Improving phantom limb movement consistency and distinguishability. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 22, 3 (2013), 522–532.

[32] Michael A Powell and Nitish V Thakor. 2013. A training strategy for learning pattern recognition control for myoelectric prostheses. *Journal of prosthetics and orthotics: JPO* 25, 1 (2013), 30.

[33] Luc Proteau, Yannick Blandin, Claude Alain, and André Dorion. 1994. The effects of the amount and variability of practice on the learning of a multi-segmented motor task. *Acta Psychologica* 85, 1 (1994), 61–74.

[34] R Rajapriya, K Rajeswari, and SJ Thiruvengadam. 2021. Deep learning and machine learning techniques to improve hand movement classification in myoelectric control system. *Biocybernetics and Biomedical Engineering* 41, 2 (2021), 554–571.

[35] Gonzalo Ramos, Christopher Meek, Patrice Simard, Jina Suh, and Soroush Ghorashi. 2020. Interactive machine teaching: a human-centered approach to building machine-learned models. *Human–Computer Interaction* 35, 5-6 (2020), 413–451.

[36] Aidan Dominic Roche, Ivan Vujaklija, Sebastian Amsüss, Agnes Sturma, Peter Göbel, Dario Farina, and Oskar C Aszmann. 2015. A structured rehabilitation protocol for improved multifunctional prosthetic control: a case study. *JoVE (Journal of Visualized Experiments)* 105 (2015), e52968.

[37] Téo Sanchez. 2022. *Interactive Machine Teaching with and for Novices*. Ph. D. Dissertation. Université Paris-Saclay.

[38] Téo Sanchez, Baptiste Caramiaux, Jules Françoise, Frédéric Bevilacqua, and Wendy E Mackay. 2021. How do people train a machine? Strategies and (Mis) Understandings. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–26.

[39] Téo Sanchez, Baptiste Caramiaux, Pierre Thiel, and Wendy E Mackay. 2022. Deep learning uncertainty in machine teaching. In *27th International Conference on Intelligent User Interfaces*. 173–190.

[40] Burr Settles. 2009. Active learning literature survey. (2009).

[41] John B Shea and Robyn L Morgan. 1979. Contextual interference effects on the acquisition, retention, and transfer of a motor skill. *Journal of Experimental psychology: Human Learning and memory* 5, 2 (1979), 179.

[42] Katarzyna Szymaniak, Agamemnon Krasoulis, and Kianoush Nazarpour. 2022. Recalibration of myoelectric control with active learning. *Frontiers in Neurorobotics* 16 (2022), 277.

[43] Emily Wall, Soroush Ghorashi, and Gonzalo Ramos. 2019. Using expert patterns in assisted interactive machine learning: A study in machine teaching. In *Human-Computer Interaction–INTERACT 2019: 17th IFIP TC 13 International Conference, Paphos, Cyprus, September 2–6, 2019, Proceedings, Part III 17*. Springer, 578–599.

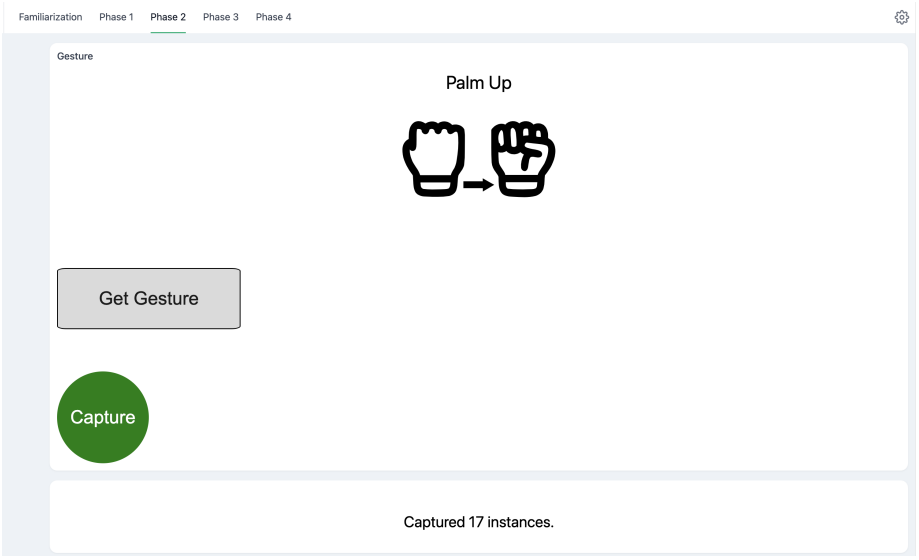
[44] Douglas L Weeks, David I Anderson, and Stephen A Wallace. 2003. The role of variability in practice structure when learning to use an upper-extremity prosthesis. *JPO: Journal of Prosthetics and Orthotics* 15, 3 (2003), 84–92.

[45] Aaron J Young, Levi J Hargrove, and Todd A Kuiken. 2011. The effects of electrode size and orientation on the sensitivity of myoelectric pattern recognition systems to electrode shift. *IEEE transactions on biomedical engineering* 58, 9 (2011), 2537–2544.

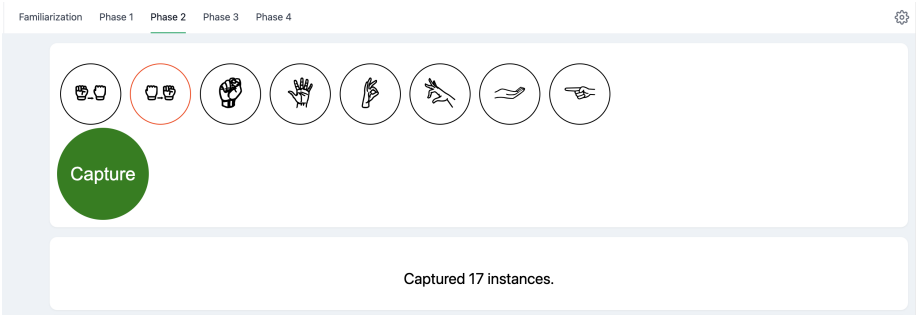
[46] Xiaojin Zhu, Adish Singla, Sandra Zilles, and Anna N Rafferty. 2018. An overview of machine teaching. *arXiv preprint arXiv:1801.05927* (2018).

7 APPENDIX

7.1 Interface for the Learner -Led and Teacher -Led conditions



(a) Interface for the Learner-Led condition. Clicking on the 'Get Gesture' button queries the next gesture.



(b) Screenshot of interface for the Teacher-Led Condition. It consists of 8 gesture icons and a 'Capture' button. Participants select gestures by clicking on icon buttons. The button border turns red to indicate that a gesture has been selected.

Fig. 8

7.2 Post-Training Questionnaire

According to your experience, after having taught the system to recognize the 8 gestures, how accurately will the system recognize each one? (0% - the system never recognizes the gesture to 100% - the system always recognizes the gesture)

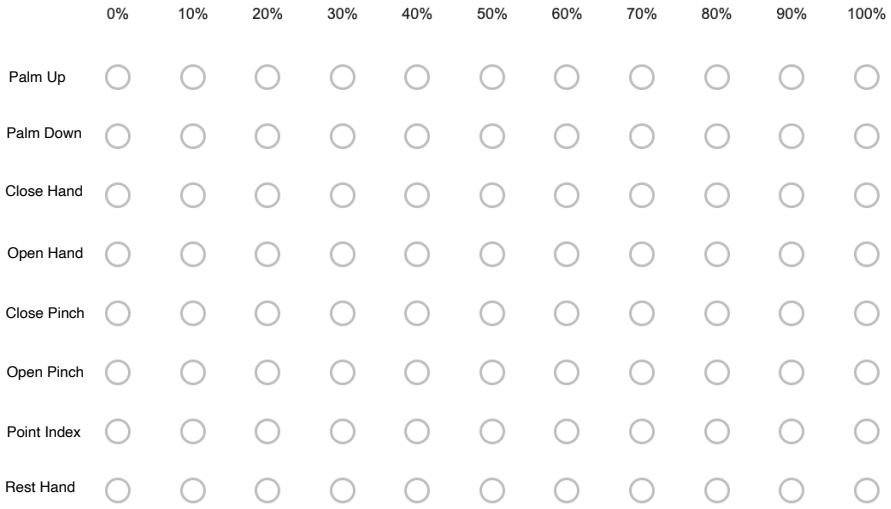


Fig. 9. Questionnaire