A Multi-Variate Approach to Predicting Myoelectric Control Usability

Jena L. Nawfel, Kevin B. Englehart, *Senior Member, IEEE*, and Erik J. Scheme, *Senior Member, IEEE*

***Abstract*— Pattern recognition techniques leveraging the use of electromyography signals have become a popular approach to provide intuitive control of myoelectric devices. Performance of these control interfaces is commonly quantified using offline classification accuracy, despite studies having shown that this metric is a poor indicator of usability. Researchers have identified alternative offline metrics that better correlate with online performance; however, the relationship has yet to be fully defined in the literature. This has necessitated the continued trial-and-error-style online testing of algorithms developed using offline approaches. To bridge this information divide, we conducted an exploratory study where thirty-two different metrics from the offline training data were extracted. A correlation analysis and an ordinary least squares regression were implemented to investigate the relationship between the offline metrics and six aspects online use. The results indicate that the current offline standard, classification accuracy, is a poor indicator of usability and that other metrics may hold predictive power. The metrics identified in this work also may constitute more representative evaluation criteria when designing and reporting new control schemes. Furthermore, linear combinations of offline training metrics generate substantially more accurate predictions than using individual metrics. We found that the offline metric feature efficiency generated the best predictions for the usability metric throughput. A combination of two offline metrics (mean semi-principal axes and mean absolute value) significantly outperformed feature efficiency alone, with a 166% increase in the predicted R^2 value (i.e., VEcv). These findings suggest that combinations of metrics could provide a more robust framework for predicting usability.**

***Index Terms*— electromyography, myoelectric control, offline training, online performance, pattern recognition**

I. Introduction

Mobility impairments are the leading cause of disability in the United States, affecting one in seven adults [1], and are the third highest cause of disability in Canada, affecting one in fourteen adults [2]. These impairments can be caused by disease, injury, or congenital defects and can often have significant implications on an individual's ability to perform activities of daily living (ADLs). An inability to perform ADLs can hinder a person's independence and potentially diminish their quality of life.

Consequently, assistive and rehabilitation technologies are commonly used to increase the physical capabilities of impaired individuals. An essential component of these technologies is the ability for the user to intuitively interact with and control the device. Both assistive and rehabilitation technologies, therefore, have leveraged pattern recognition approaches to decipher user intent. One such method is through the use of electromyography (EMG) signals from residual functioning muscles [3]–[5]. The patterns generated during muscular contractions can be decoded and used as input for a human computer interface (HCI), prosthesis, or orthosis, by mapping intent to control multiple degrees of freedom (DOFs).

For decades, the performance of pattern recognition-based myoelectric control has predominantly been assessed using offline classification accuracy. Increasingly, however, studies have found that this measure has little to no correlation with online usability [6]–[8]. A recent study claimed that global offline accuracy was highly correlated with the completion rate of an online usability test ($r = 0.90$, $p < 0.05$) [9]; but, most studies suggest a more complex relationship between offline classification accuracy and online usability. Hargrove *et al.* observed that including transient contractions in the training data set decreases offline classification accuracy but increases online usability during a virtual clothespin task [7]. Similar findings demonstrated that a multiple binary classifier with a statistically higher classification error rate than a linear discriminant analysis (LDA) classifier produced a more controllable system with faster clothespin placement times and a higher completion rate [10]. Although there is some evidence suggesting offline classification accuracy provides meaningful information with regard to online performance, the exact relationship between this metric and real-time control has yet to be fully defined in the literature.

Researchers also have investigated the use of other offline training metrics as indicators of usability, such as separability- and repeatability-based metrics. A correlation analysis between the usability metric completion time and the separability metrics modified separability index and Bhattacharyya distance yielded correlation coefficients of $r = 0.54$ and $r = 0.45$, respectively [11]. This same study identified no meaningful relationship between completion time and the repeatability index ($r = 0.018$) [11]. Another study yielded a correlation coefficient of $r = 0.53$ between the separability index and testing error [12]. Although these results suggest a moderate relationship between pattern separability and online performance, there remains little consensus in the literature; for example, a more recent study demonstrated no significant correlation between separability and online accuracy [13].

Because offline training metrics fail to provide the necessary information to evaluate online myoelectric control, the most accurate performance assessments remain those that incorporate the use of the end device. This is necessary because implementing the physical device's control system introduces many challenges associated with the stability of the EMG recordings, the interference from non-targeted muscle groups, the effects of tissue loading and arm dynamics, and the fit of a socket [14]. Prostheses and orthoses, however, can be quite expensive and often require a clinical population group to test on, making them impractical for use in some experiments.

In an attempt to bypass the need for a physical myoelectric device, researchers have proposed and implemented alternative usability assessments that leverage virtual testing environments [7], [14]–[17]. Recently, Hargrove *et al.* justified the continued use of virtual testing environments by demonstrating a significant correlation between virtual and physical outcome measures [18]. Virtual testing environments also allow researchers to evaluate their control scheme without the influence of all the physical factors that come with implementing a device. The following three virtual assessments incorporate the user in the control loop and are among the more commonly cited tests in the literature: the motion test [14], the Target Achievement Control (TAC) test [16], and the Fitts' Law usability test [17].

It is generally agreed upon in the literature that the motion test is an oversimplified version of real-time use. This is because misclassifications and unintended movements are not registered in the testing environment. The TAC test and the widely accepted Fitts' Law

Submitted on Sept. 28, 2020. This work was supported in part by NSERC Discovery Grants 2014-04920 and 217354-15.

J. Nawfel, K. Englehart, and E. Scheme are with the Institute of Biomedical Engineering, University of New Brunswick, Fredericton, NB, E3B 5A3, CA (e-mail: escheme@unb.ca, jena.nawfel@unb.ca).

test are more challenging virtual assessments compared to the motion test. Both assessments provide users with the ability to modulate muscle activity, contraction intensity, and the output of the test in real-time. Although the Fitts' Law and TAC tests have much in common, they cannot be considered interchangeable [19]. A study comparing the two methods suggests significantly higher user error and reported confusion for the TAC test, concluding that the Fitts' Law test may be a more reliable tool for performance evaluation [19].

While researchers have established the link between virtual and physical device usability [18], the literature still lacks an established link between offline training metrics and virtual outcome measures. Performing usability testing in a clinical environment, whether it be virtual or physical assessments, takes time and resources. It is often impractical to evaluate online performance for every individual fitted with a myoelectric device. Therefore, when patients use their devices at home, they may experience erroneous motions and limitations in the dexterity of control, which have been cited as common attributing factors for the abandonment of myoelectric devices [3], [20]–[23]. If users knew that the outcome of their training protocols might result in poor practical use, they could retrain the system immediately to avoid unnecessary frustration during activities of daily living. Furthermore, training protocols could be targeted toward improving training data characteristics known to be valuable predictors of online performance. This would help users better understand what is necessary for successful myoelectric control. Establishing more representative metrics may also help to improve the design for the actual use case rather than for classification accuracy, which does not translate well to real-time myoelectric control. More reliable use may in turn lead to higher user satisfaction and acceptance of these devices.

To the best of our knowledge, there has been no prior work investigating a multi-variate relationship between users' training data and their online usability. Past attempts to quantify this relationship only consider a small set of offline metrics and their individual correlations with online performance. Influenced by the feature analysis presented by Phinyomark *et al.* [24], this paper presents an exploratory and unconstrained analysis using 32 offline metrics and six online usability metrics to draw out and identify uni- and multi-variate relationships.

II. Methods

A. Participants

Twelve able-bodied subjects (9 male/3 female, age range: 22-63 yrs., mean and standard deviation of age: 33±15.2 yrs., median age: 25.5 yrs.) took part in this study. Ten participants reported right hand dominance and two reported left hand dominance. The procedures were approved by the University of New Brunswick's research ethics board (REB #2020-016), and subjects provided written informed consent prior to participating in the experiment.

B. Experimental Setup

EMG signals were recorded using a standard, non-invasive, wireless EMG collection system (Trigno^{TM} Wireless system, Delsys Inc., USA). The signals were sampled at 2000 Hz and filtered to remove power-line and digital interference with 2nd-order Butterworth band-stop filters at 60, 180, 250, and 300 Hz. A 3rd-order Butterworth high-pass filter with a cutoff frequency of 20 Hz was also implemented to remove motion artifact. Prior to positioning the electrodes, the skin was cleansed with an alcohol swab to remove excess skin oil and debris. Six electrodes were uniformly spaced around the proximal third of the dominant forearm. Participants sat in a chair with their dominant arm held unsupported, but comfortably, at a 90-degree angle by their side and with their forearm parallel to the floor.

C. Experimental Protocol

The experiment consisted of one 20-30 minute session involving a training phase and a testing phase.

1) Training: EMG signals for five motion classes were collected: no movement, wrist flexion, wrist extension, power grip, and hand open. Each cycle through these five movements constituted a trial. In total, eight trials were conducted over the user training period.

A screen guided training approach was implemented to guide users through the training process [25]. An image of a hand gesture prompted the user to perform a given movement, and a progress bar informed the user how long to hold their contraction. Subjects began movements at rest, transitioned into the desired movement, and then maintained the contraction for the duration of the repetition. Users were given minimal instruction with regards to their contraction intensity. They were told to perform contractions at an intensity for which they felt comfortable and would not fatigue over the course of the experiment. The system recorded four seconds of EMG data for each prompt followed by a two-second delay period during which no data were recorded. The delay allowed users to return to a resting position before the next prompted movement.

This experiment employed adaptive LDA classifier training based on the maximum likelihood output of the classifier [26]. An initial classifier was trained following the completion of the first trial. The data from this trial were segmented into 160 ms windows with a 64 ms overlap [27]. The four commonly-used time domain features described by Hudgins [28] were extracted at each of the six electrode channels for a total of 24 features. These features were then used to train the LDA classifier.

The data collected in the next trial were classified to determine the windows of data that would be used to adapt the classifier. Bayesian classification theory was used to provide a score based on the likelihood outputs of the classifier [29], [30]. A data window was concatenated to the existing classifier data set if the class with the maximum likelihood matched that of the training class. The classifier was retrained after each trial.

This process of appending data to the classifier data set continued for subsequent trials, with a forgetting factor of four trials to limit the amount of data being used to train the classifier. Following this approach, only data from the four most recent trials were included in the classifier training set, as shown in Figure 1. We adopted this adaptive procedure in place of a classical static data collection to reduce the potential impact of user learning. Adaptive algorithms have been shown to significantly reduce classification error, reinforce good decisions, account for slow drifts in the boundaries of the classifier, and ultimately reflect changes in user behavior [26].

2) Testing: The classifier generated during training was tested in a Fitts' Law environment to determine its usability during a virtual target acquisition task. Fitts' Law was introduced in 1954 by Paul Fitts and uses principles derived from Shannon's work in communication theory to demonstrate that any human motor task exhibits a trade-off between speed and accuracy [31]. Fitts' Law-style testing has become an international standard (ISO9341–9) for validating human-computer interfaces, including mice, joysticks, touchpads, and human motion. A Fitts' Law usability test maps specific motions to control the movement of a cursor in a virtual environment during a target acquisition task. The user must respond and correct for system misclassifications to successfully acquire the target. Over the last decade, researchers have verified the use of EMG as a control input using Fitts' Law, making this a popular approach in the literature for evaluating myoelectric control [17], [19], [32].

The four active motion classes collected during training were mapped to control the movement of the cursor on the computer screen. In one DOF, hand open moved the cursor up and power grip

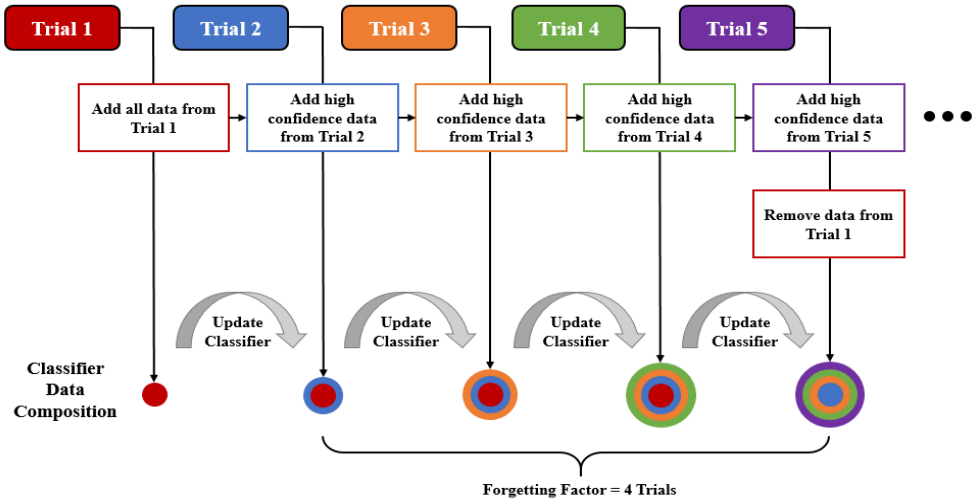


Fig. 1: Adaptive classifier procedure.

moved the cursor down. In the horizontal DOF, the direction of wrist extension and wrist flexion was mapped to match that of the subject's reported hand dominance. Proportional control was implemented using the procedures outlined in [33] and [34]. Class-specific gains that map the average class-specific amplitudes to 50% of full speed were calculated during training. This allowed the strength of the user's muscle contraction to regulate the speed of the cursor.

The Fitts' Law test positioned targets at varying target distances and in locations that required activation of one or both DOFs. The user successfully completed a trial by acquiring the target and maintaining the cursor within the target boundaries for one second. The allotted time to reach each target was ten seconds, after which the trial timed out, and the test automatically moved to the next target.

The testing phase included 32 single DOF targets (e.g. right or down) and 32 dual DOF targets (e.g., left AND up) for a total of 64 targets. The incoming test data were segmented into 160 ms windows from which features were extracted and classifications were made. This window length is within the optimal range found by Smith *et al.* for myoelectric control [35] and is the preferred setting for the software package used in this study [25]. To ensure sufficient time to process the data, a conservative update interval of 16 ms was selected [34], [36].

D. Offline Training Metrics

Several offline training metrics proposed in the literature were assembled into one expansive set and were designated as potential predictors of usability. Currently, there is no accepted means of determining how different types of offline metrics relate to online performance. A total of 32 offline training metrics were calculated using the 24-dimensional feature space (i.e., four features x six electrode channels) obtained from the training data outlined in Section II-C.1. Collectively, these metrics provide a comprehensive view of the feature space populated during training. Of the 32 metrics, seven were *variability* measures, eleven were *separability* measures, nine were *complexity* measures, three were *classification* measures, and two were *neighborhood* measures. The analysis of these measures leveraged the full training repetitions rather than only the portions applied to the adaptive classifier to fully evaluate the user's behavior throughout training.

- The *variability metrics* quantified intra-class characteristics. A full list of these metrics and formulations can be found in Section VI.
- The *separability metrics* assessed inter-class attributes. A full list of these metrics and formulations can be found in Section VII.
- The *complexity metrics* leveraged feature space partitioning algorithms to examine regional class discriminability. A full list of these metrics and formulations can be found in Section VIII.
- The *classification metrics* were based on classification performance using an LDA classifier. A full list of these metrics and formulations can be found in Section IX.
- The *neighborhood metrics* considered nearest neighbor relationships. A full list of these metrics and formulations can be found in Section X.

E. Online Usability Metrics

Fitts' Law has been widely adopted to describe the information bandwidth of a control scheme, such as the movement of a pointer or cursor in a virtual environment.

The following usability metrics (i.e., virtual outcomes) were extracted in the present study to evaluate online myoelectric control:

- **Throughput (TP, bits/sec)** is the Fitts' Law summary metric and is considered to be the rate of information transfer [17], [31], [37]. It is characterized by the target's index of difficulty (ID) and the movement time (MT) of the cursor averaged across N trials.

$$TP = \frac{1}{N}\sum_{i=1}^{N}\frac{ID_i}{MT_i} \tag{1}$$

The index of difficulty, defined as a function of the target's width (W) and distance (D), was calculated using Shannon's formulation [38]. The distance (D) was measured as the distance between the starting point of the cursor and the center of the target.

$$ID = \log_2\left(\frac{D}{W}+1\right) \tag{2}$$

- **Effective Throughput (eTP, bits/sec)** is a modified version of throughput where the distance to the target during the calculation of ID is adjusted based on the actual distance the cursor travels (D_e). If a user consistently stops the cursor on the inner edge of the target, the effective distance to the target becomes smaller. Likewise, if a user stops the cursor on the outer edge of the target, the effective distance increases.

$$ID_e = \log_2\left(\frac{D_e}{W}+1\right) \tag{3}$$

- **Path Efficiency (PE, %)** describes the system's quality of control and is calculated as the ratio between the shortest path to the target and the actual path traveled [17], [37].
- **Overshoot (OS)** measures a user's ability to stop on a target by counting the number of times per task the user acquired then lost the target before the dwell time was reached [17], [37].
- **Average Speed (AS, pixels/sec)** highlights a user's gross ability to control the cursor and is computed as the average non-zero speed of the cursor for each task [17], [37].
- **Stopping Distance (SD, pixels)** evaluates a user's ability to maintain no motion to stop within a target. It is calculated as the total distance traveled by the cursor during the dwell time [17].

F. Linear Regression

Linear regression models were generated using sets of one, two, and three predictor variables on a given response variable. An ordinary least squares regression was used rather than a higher order polynomial regression to reduce the potential of overtuning and to encourage generalizability. Overfitting concerns were further addressed by restricting the number of predictors to three or less. The relationship between offline and online myoelectric control was evaluated using the offline training metrics as predictors and the online usability metrics as response variables. Both predictors and responses were normalized to be between zero and one prior to training the models. Normalization allows for the interpretation of the coefficient weights in the regression models. These weights can be used to determine the relative contributions of the predictors on the prediction. The relative coefficient weightings in this study were calculated by dividing the absolute value of each individual predictor weight by the sum of the absolute values of all the assigned predictor weights. Multicollinearity was assessed using the variance inflation factor (VIF). While there is no universal agreement for the VIF cut-off value that should be used to detect multicollinearity, a VIF greater than 5 is often considered to be problematic [39].

1) Predictor Selection: We implemented a predictor selection approach similar to the consensus nested cross-validation technique recently proposed by Parvandeh *et al.* for feature selection [40]. This approach finds consistent and stable features, with the goal of providing a more generalizable model [40]. The technique also has been shown to be effective for small sample sizes [40].

Following this selection procedure, we performed a 12-choose-11 subject calculation to establish twelve subsets with eleven subjects each. Predictor selection was then performed using a leave-one-subject-out (LOSO) cross-validation technique within each subset. The minimization criterion was the average mean squared error (MSE) between the model's outputs and the true validation values.

The number of predictor sets that were evaluated varied depending on the number of predictors used in the model. The one-predictor models assessed 32 predictor sets, equalling the total number of offline metrics. The two-predictor models considered 496 predictor combinations, which encompassed all possible combinations of 32-choose-2 predictors. Finally, the three-predictor models assessed 4960 predictor combinations, based on 32-choose-3 predictors.

For the two- and three-predictor models, we identified the top 50 combinations of offline metrics with the lowest average MSE across the LOSO cross validations for each subset. A consensus in the top predictor combinations was then required across at least six of the twelve subsets. The predictor combinations that met the consensus requirements were selected for further evaluation.

2) Performance Evaluation: The predictive performance of each selected offline metric combination was determined using a LOSO cross validation across all twelve subjects. This was repeated to assess predictor sensitivity for the one-, two-, and three-predictor models against each of the response variables. The predictor set demonstrating the lowest predictive MSE was selected as having the "best" performance. Therefore, each response variable had a corresponding model with one, two, and three predictors. The goodness of fit and predictive accuracy of these final models were evaluated using the following metrics.

a) Measures of Goodness of Fit: Goodness of fit, also known as the training error, refers to a model's ability to predict the samples used during parameter estimation. The list below describes each of the measures of goodness of fit used in assessing the performance of the trained prediction models.

- **Mean Absolute Error (MAE)**: The MAE is an interpretable metric that provides information about the average magnitude of error between the true and predicted values [41]. All errors are equally weighted, and the units match that of the response variable.
- **Mean Squared Error (MSE)**: The MSE measures the average squared error between the true and predicted values [41]. MSE assigns higher weights to larger errors, and consequently, is more sensitive to outliers. MSE has units equalling the square of the response variable, making it arguably less interpretable than MAE.
- **Root Mean Squared Error (RMSE)**: The RMSE provides an estimate for the standard deviation of the associated error distribution. It is the square root of the MSE and has the same units as the response variable [41].
- **Adjusted Coefficient of Determination** (R^2_{adj}): The R^2_{adj} is a recommended measure of goodness of fit when multiple predictors are used in model development. It accounts for the number of predictors by increasing in value only when the addition of a predictor significantly improves the fit of the model [42]. The R^2_{adj} decreases when the model improvements are not greater than what would be expected by chance.
- **Corrected Akaike Information Criterion (AICc)**: AICc measures the relative quality of a model by balancing the tradeoff between goodness of fit and number of predictors [43]. AICc is recommended for small sample sizes and incorporates a correction or penalty to address overfitting [43].

Smaller values of MAE, MSE, RMSE, and AICc imply that the generated model more closely resembles the "true model". Higher adjusted coefficient of determination values suggest that the parameters of the model better fit the observations.

b) Measures of Predictive Accuracy: Predictive accuracy is concerned with the model's ability to predict new instances, previously unseen by the model. The following list provides descriptions of each measure of predictive accuracy used to evaluate how well the developed models predicted usability.

- **Normalized (n) MAE, MSE, and RMSE**: Unit- and scale-independent versions of MAE, MSE, and RMSE were obtained by dividing the metrics by their corresponding range of true response values observed in the trained model. Normalization allows comparisons to be made across datasets.
- **Mean Absolute Percent Error (MAPE)**: MAPE is an interpretable and scale-independent metric commonly used to measure forecasting accuracy [44]. It is calculated as the average of the absolute percentage errors. It is important to note that MAPE becomes undefined as the true values approach zero.
- **Variance Explained (VEcv)**: The VEcv metric, sometimes referred to as the predicted R^2, is based on cross-validation and allows direct comparisons between accuracies of predictive models for data with different units, scale, and variation [41], [45]. Negative VEcv values indicate that the predictions generated by the model are less accurate than using the mean of the validation

data as predictions [45]. Positive VEcv values demonstrate that the predictions generated by the corresponding model are more accurate than using the validation mean [45]. The maximum obtainable VEcv value is 100% and occurs when the model predictions are equal to their corresponding validation values. According to Li, the performance of predictive models based on VEcv measures can be divided into the following five categories: *very poor:* VEcv $\leq 10\%$, *poor:* $10\% <$ VEcv $\leq 30\%$, *average:* $30 <$ VEcv $\leq 50\%$, *good:* $50 <$ VEcv $\leq 80\%$, *excellent:* VEcv $> 80\%$.

Increases in VEcv and reductions in nMAE, nMSE, nRMSE, and MAPE indicate better predictive accuracy.

III. Results

A. Offline Classification Performance

Table I displays the average leave-one-trial-out cross-validation classification accuracy (CA) and active classification accuracy (ACA) across the eight training trials for each subject. Because ramp contractions were collected during training, the movement data contained trace amounts of the no motion class. The ACA metric removes misclassifications due to no motion from the accuracy calculation. Subject 4 obtained the highest average CA at $96.4\% \pm 2.8\%$ and subject 12 demonstrated the lowest average CA at $78.5\% \pm 12.4\%$. Subject 4 also exhibited the highest average ACA at $100.0\% \pm 0.1\%$ while subject 2 displayed the lowest average ACA at $98.3\% \pm 3.5\%$.

TABLE I: Subject-wise breakdown of classification accuracy (CA) and active classification accuracy (ACA) results. A ceiling effect is observed for the active classification accuracy metric.

Metric	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**
CA	90.6	89.8	99.0	96.4	91.2	86.7
ACA	99.1	98.3	99.8	100.0	99.8	99.9
Metric	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**
CA	92.2	91.4	90.5	92.0	79.8	78.5
ACA	100.0	99.8	99.4	99.7	99.7	99.6

B. Online Performance

The range of the average usability metrics across subjects is shown in Table II. Subject 7 achieved the highest average throughput, effective throughput, path efficiency, and average speed with values of $2.0 \pm 0.7, 4.1 \pm 1.1, 95.4\% \pm 8.5\%, 65.2 \pm 15.9$, respectively. Subject 12 obtained the lowest average throughput, effective throughput, average speed, and stopping distance with values of $0.9 \pm 0.3, 1.8 \pm 0.8, 24.0 \pm 9.2, 3.9 \pm 2.2$, respectively. The lowest average path efficiency $(79.3\% \pm 23.6\%)$ and the highest average stopping distance (7.3 ± 3.4) were achieved by subject 2. Subject 6 had the highest average overshoot (0.7 ± 2.0) and subject 1 had the lowest average overshoot (0.05 ± 0.2).

TABLE II: The minimum and maximum average values across subjects for each usability metric.

Metric	**Min**	**Max**
Throughput (TP)	0.9 ± 0.3	2.0 ± 0.7
Effective Throughput (eTP)	1.8 ± 0.8	4.1 ± 1.1
Path Efficiency (PE)	79.3 ± 23.6	95.4 ± 8.5
Overshoot (OS)	0.05 ± 0.2	0.7 ± 2.0
Average Speed (AS)	24.0 ± 9.2	65.2 ± 15.9
Stopping Distance (SD)	3.9 ± 2.2	7.3 ± 3.4

C. Correlation Analysis

Shapiro-Wilk tests were performed on each offline and online metric to evaluate whether they obeyed a normal distribution. When the assumption of normality was not violated, the Pearson correlation coefficient was calculated between the corresponding offline and online metrics. When the assumption of normality was rejected, Kendall's coefficient of rank correlation was implemented [46]. The resulting significant correlations are summarized in Table III.

TABLE III: Significant correlations between the offline training metrics and the online usability metrics. Bold text indicates significant correlations at the 95% confidence level. Metrics followed by an asterisk violated the assumption of normality.

Metric Type	Offline Metrics	Sig. Correlation
Separability	Bhattacharyya Distance (BD)	**PE** (-0.60)
	Squared Hellinger Distance (HD)*	**PE** (-0.55)
	Fisher's Discriminant Ratio (FDR)	**AS** (0.61), **SD** (0.71)
	Feature Efficiency (FE)*	**TP** (0.52), **eTP** (0.49)
Complexity	Rescaled Purity (rPU)	**AS** (0.62), **SD** (0.62)
Neighborhood	Intra-Inter Fraction (IIF)*	**TP** (-0.46)

Of the 32 offline training metrics investigated in this work, at most two correlations for each online metric were significant. Feature efficiency (FE) and intra-inter fraction (IIF) both demonstrated moderate correlations with throughput, (r = 0.52, p = 0.02) and (r = -0.46, p = 0.04), respectively. FE also exhibited a moderate association with effective throughput, (r = 0.49, p = 0.03). Bhattacharyya distance (BD), (r = -0.60, p = 0.04), and Hellinger Distance (HD), (r = -0.55, p = 0.01), were significantly correlated with path efficiency. Fisher's discriminant ratio (FDR) and the rescaled purity metric (rPU) had significant correlations with average speed, (r = 0.61, p = 0.03) and (r = 0.62, p = 0.03), respectively, and stopping distance, (r = 0.71, p = 0.01) and (r = 0.62, p = 0.03), respectively. While there is no consensus as to how the strength of the correlation coefficient should be interpreted, Akoglu presents three commonly used scales in [47]. The highest correlation coefficient observed in Table III is 0.71 which may be considered very strong, strong, or moderate depending on the scale. The lowest correlation coefficient observed in Table III is 0.46 which may be considered strong, moderate, or fair.

D. Predictive Modeling

Table IV shows the results of the measures of goodness of fit and the measures of predictive accuracy for the six response variables. The models highlighted in Table IV are the baseline model using classification accuracy (CA) alone and the previously selected one-, two-, and three-predictor models for each response variable. Table IV also displays the relative coefficient weightings for each predictor in the selected models, indicating how important each metric was in generating predictions. The measures of goodness of fit in the table generally support the idea that the three-predictor models provide the best fit. The general consensus across the measures of predictive accuracy is also that the selected three-predictor models generate the best predictive performance. A graphical representation of the VEcv metric for the CA-predictor model along with the selected one-, two-, and three-predictor models is illustrated in Figure 2.

a) Throughput: The model using CA as a single predictor for throughput demonstrated lower predictive accuracy across all measures than the selected one-, two-, and three-predictor models. The best performing individual predictor was feature efficiency (FE). This metric rendered a significant correlation with throughput (r = 0.52, p = 0.02). Although the positive VEcv value of 17.4% indicates that this metric generates predictions with lower errors than predicting using the mean of the data, the predictive performance is

TABLE IV: Measures of goodness of fit and predictive accuracy for each response variable using the selected models. The relative coefficient weightings for each predictor are shown in parenthesis. Positive and negative signs indicate a direct and indirect relationship, respectively, with the response variable. Predictors with a variance inflation factor above 5 are indicated with an asterisk (*).

Model Specifications		Goodness of Fit					Predictive Accuracy				
Response	Predictors	MAE	MSE	RMSE	R^2_{adj}	AICc	nMAE	nMSE	nRMSE	MAPE	VEcv
Throughput	CA (+100%)	0.29	0.11	0.32	-0.04	11.86	0.35	0.19	0.40	34.67	-50.13
	FE (+100%)	0.25	0.08	0.27	0.27	7.89	0.27	0.10	0.30	26.96	17.41
	MSA* (-58%), MAV* (+42%)	0.16	0.03	0.18	0.62	3.27	0.21	0.07	0.24	21.12	46.31
	MSA* (-51%), MAV* (+36%), C* (+13%)	0.12	0.02	0.15	0.72	3.60	0.17	0.04	0.19	16.64	64.67
Effective Throughput	CA (+100%)	0.29	0.11	0.32	-0.03	11.89	0.36	0.38	0.40	34.28	-49.52
	FE (+100%)	0.25	0.08	0.28	0.26	8.16	0.28	0.21	0.30	27.15	15.54
	MSA* (-59%), MAV* (+41%)	0.16	0.03	0.18	0.63	3.28	0.21	0.13	0.24	20.35	47.42
	MSA* (-52%), MAV* (+35%), C* (+13%)	0.12	0.02	0.15	0.72	3.86	0.16	0.09	0.19	15.12	65.63
Path Efficiency	CA (-100%)	0.24	0.08	0.29	0.00	9.45	0.27	1.60	0.32	4.89	-14.95
	BD (-100%)	0.19	0.06	0.24	0.29	5.58	0.21	1.19	0.27	3.80	15.02
	mwRI (+50%), IIF (+50%)	0.16	0.04	0.19	0.51	4.00	0.19	0.82	0.23	3.49	41.16
	mwRI* (+37%), CD* (-23%), CDM (+40%)	0.12	0.02	0.14	0.70	2.18	0.15	0.50	0.18	2.67	64.39
Overshoot	CA (+100%)	0.29	0.11	0.33	-0.09	12.58	0.35	0.10	0.39	172.9	-33.35
	BD (+100%)	0.23	0.09	0.30	0.11	10.32	0.26	0.08	0.34	134.2	-3.04
	FE (-49%), IIF (-51%)	0.23	0.09	0.30	0.02	14.09	0.28	0.08	0.34	121.0	-3.25
	BD* (+16%), NS* (+39%), rCE* (-45%)	0.12	0.03	0.18	0.58	8.25	0.17	0.04	0.24	76.50	50.55
Average Speed	CA (+100%)	0.28	0.09	0.31	0.05	10.70	0.34	5.70	0.37	36.85	-29.68
	rPU (+100%)	0.22	0.07	0.26	0.32	6.92	0.25	3.50	0.29	25.17	20.29
	FDR (+53%), RI (+47%)	0.14	0.03	0.17	0.66	2.03	0.20	2.47	0.25	20.05	43.72
	FDR (+40%), FE* (+25%), RI* (+35%)	0.10	0.01	0.12	0.81	-0.82	0.18	1.77	0.21	16.95	59.77
Stopping Distance	CA (+100%)	0.17	0.05	0.23	0.13	4.57	0.19	0.22	0.26	12.32	-3.53
	rPU (+100%)	0.15	0.04	0.20	0.32	1.74	0.17	0.17	0.23	10.52	18.66
	mwRI (-47%), CDM (-53%)	0.12	0.02	0.15	0.60	-1.52	0.16	0.13	0.20	10.04	38.96
	mwRI* (-14%), PU* (+50%), rCE* (-36%)	0.09	0.01	0.12	0.72	-1.40	0.14	0.10	0.17	8.21	52.59

Predictors Abbrv. BD: Bhattacharyya Distance, C: Compactness, CA: Classification Accuracy, CD: Centroid Drift, CDM: Class Discriminability Measure, FDR: Fisher's Discriminant Ratio, FE: Feature Efficiency, IIF: Intra-Inter Fraction, MAV: Mean Absolute Value, MSA: Mean Semi-principal Axes, mwRI: mean within-repetition Repeatability Index, NS: Neighborhood Separability, PU: Purity, rCE: rescaled Collective Entropy , RI: Repeatability Index, rPU: rescaled Purity

still considered to be poor according to Li [45]. The two-predictor model selected mean semi-principal axes (MSA) and mean absolute value (MAV) as predictors. Both metrics obtained non-significant individual correlation coefficients with throughput, ($r_{MSA} = -0.51$, p = 0.09; $r_{MAV} = -0.03$, p = 0.92). Using these two predictors separately in single-predictor models yielded very poor VEcv values of 3.7% and -56.3% for MSA and MAV, respectively. However, pairing these predictors in a multiple regression generated a model with a VEcv value of 46.3%. This is a 166% increase in VEcv compared to the FE-predictor model and suggests average predictive accuracy according to Li [45]. Similarly, even though the compactness measure (C) had a correlation of (r = 0.20, p = 0.54) with throughput and an individual predictive VEcv of -67.1%, adding it to the two-predictor model increased the VEcv from 46.3% to 64.7% — a 39.6% increase.

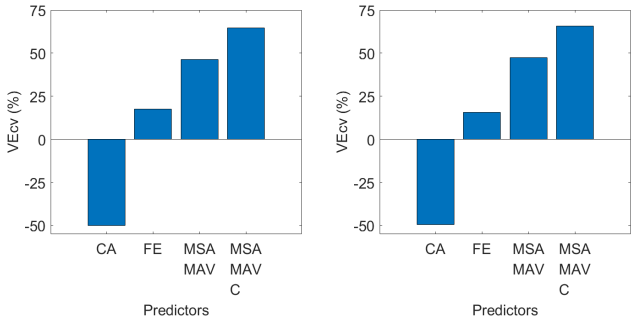
b) Effective Throughput: The models generated for effective throughput exhibited similar predictive behavior and goodness of fit as those for throughput. Furthermore, the same offline metrics were chosen for the selected models with near equal coefficient weightings. The three-predictor VEcv value for effective throughput was 65.6%, which is slightly higher than that for throughput (VEcv = 64.7%).

c) Path Efficiency: The selected one-, two-, and three-predictor models for path efficiency all exhibited higher predictive accuracy than when using CA as a single predictor. The Bhattacharyya distance (BD) was the preferred offline metric for the individual-predictor model. Although it rendered a significant correlation with path efficiency (r = -0.60, p = 0.04), the predictive accuracy of the model based on the measures in Table IV was categorized as poor (VEcv = 15.0%). The two-predictor model specified the mean within-repetition repeatability index (mwRI) and intra-inter fraction (IIF) as predictors, both of which demonstrated low and insignificant correlations with path efficiency, (r = 0.31, p = 0.33) and (r = 0.12, p = 0.64), respectively. The individual predictive performance of the corresponding single-predictor models generated very poor VEcv values of -46.1% for mwRI and -19.3% for IIF. Similar to the effects seen for throughput and effective throughput, the combination of two predictors led to a substantial 175% increase in VEcv compared to the BD-predictor model. This same trend was also observed for the offline metrics in the three-predictor model, which exhibited the highest predictive performance and goodness of fit.

d) Overshoot: The CA-predictor model, single-predictor model, and two-predictor model were each unable to reliably predict overshoot. However, a combination of three predictors generated a model with good predictive capacity (VEcv = 50.6%). This model selected Bhattacharrya distance (BD), neighborhood separability (NS), and rescaled collective entropy (rCE) as its predictors. Their individual correlations with overshoot were (r = 0.44, p = 0.15), (r = 0.22, p = 0.49), and (r = 0.14, p = 0.67) for BD, NS, and rCE, respectively. Even though the offline metrics had limited predictability on their own, combining these metrics led to a functional predictive model.

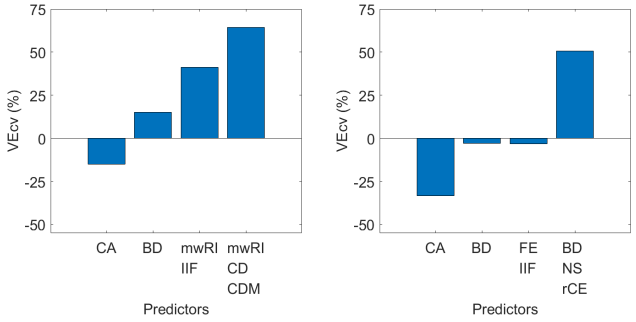
e) Average Speed: The selected one-, two-, and three-predictor models exhibited higher predictive accuracy compared to the CA-predictor model. The best performing single predictor was rescaled purity (rPU), leading to a VEcv of 20.3%. A 115% increase in VEcv was observed when Fisher's discriminant ratio (FDR) and repeatability index (RI) were combined in a two-predictor model. The three-predictor model generated the highest VEcv with a score of 59.8% — a 195% increase compared to the rPU-predictor model.

f) Stopping Distance: The stopping distance predictions also improved as the number of selected predictors increased from one to three. The CA-predictor model exhibited a VEcv score of -3.5%, indicating worse predictability than simply using the mean of the validation data. The rescaled purity (rPU) metric led to a performance increase with a VEcv of 18.7%. The two-predictor model, which used



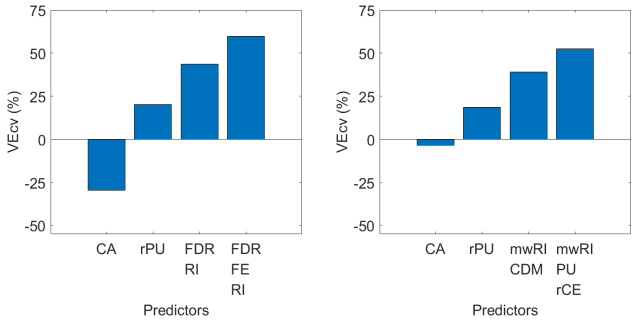
(a) Throughput

(b) Effective Throughput



(c) Path Efficiency

(d) Overshoot



(e) Average Speed

(f) Stopping Distance

Fig. 2: Model predictive performance based on VEcv (%).

the mean within-repetition repeatability index (mwRI) and the class discriminability metric (CDM), further generated a 109% increase in VEcv from the single-predictor model. Both metrics had low individual predictive capacity, as was observed from their low and insignificant correlations with stopping distance as well as from their insufficient VEcv scores (mwRI: r = -0.28, p = 0.38, VEcv = -45.2; CDM: r = -0.30, p = 0.23, VEcv = 3.7%) Similar results were obtained for the three-predictor model using mwRI, purity (PU), and rescaled collective entropy (rCE). Although the individual predictability of these offline metrics was limited, the interaction of the predictors produced a good VEcv score of 52.6%.

E. Predictor Sensitivity

The best predictor set identified during the selection procedure for each response variable (as shown in Table IV) was further investigated to determine the degree to which it outperformed other offline metric combinations that satisfied the consensus requirement as described in Section II-F.1. A representative analysis for the response variable throughput is shown in Figure 3, where the top predictor combinations are plotted against their corresponding predictive nMSE values. A dotted line indicating the nMSE of the CA-predictor model is superimposed onto the graphs to highlight the relative improvement in the model's predictive capacity when appropriate offline metrics are chosen for evaluation. It is important to note that many of the top two-predictor combinations in Figure 3b contain MSA while many of the top three-predictor combinations in Figure 3c contain MSA and MAV. This suggests that, although different combinations of two and three predictors may yield similar predictive error, certain metrics appear to be more important than others in predicting the response variable. Figure 3 also displays a narrowing tendency of the 95% confidence interval as the number of predictors increases from one to three and as the predictor combinations become more indicative of the response variable. The additional five usability metrics follow similar trends as those seen in Figure 3.

IV. Discussion

This study investigated the ability to use offline classification accuracy and alternative training metrics to predict online usability. The relationship between usability and user satisfaction has been established [48], and so these predicative measures may directly inform user experience in the real-world. Customized protocols targeting improvements of informative metrics could be implemented to ensure efficient and effective training sessions. A "training score" could be assigned to a training session based on the offline training metrics, providing an indication of future online performance.

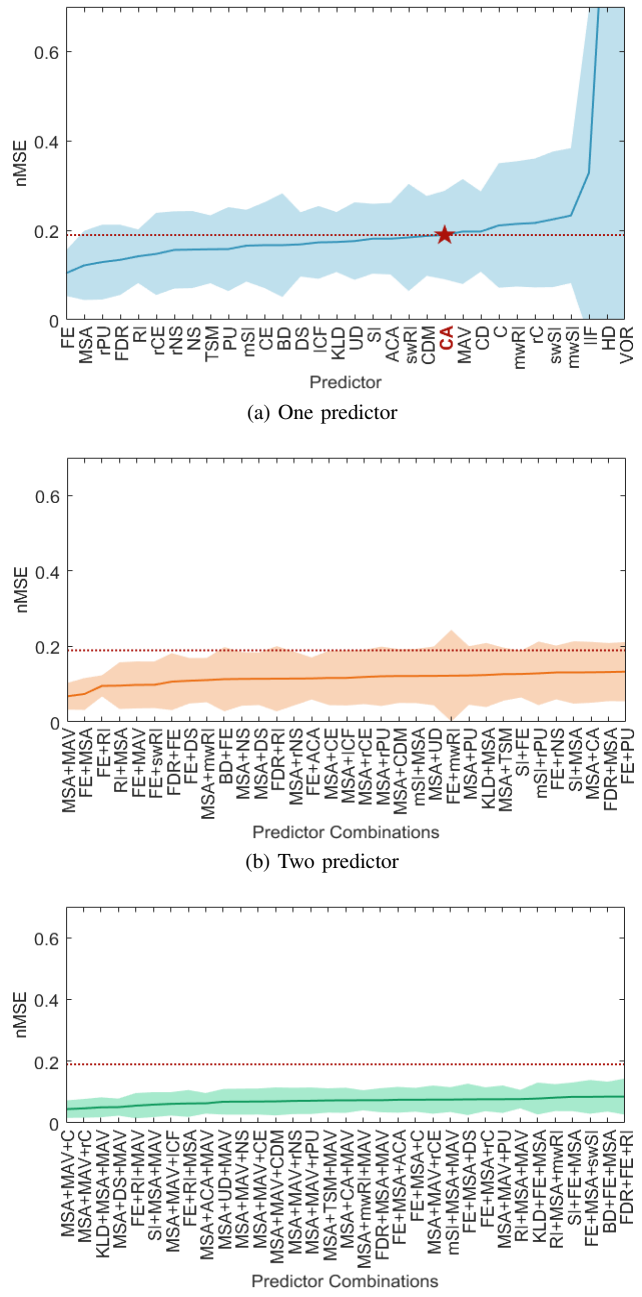
In addition, the findings of this research could be used to improve the use of offline data in the design of algorithms. For example, rather than using offline classification accuracy as an objective function for feature selection, other metrics that were shown here to be more valuable indicators of usability could be used instead. This may allow for more representative EMG features to be selected, a more predictable and usable myoelectric control system, and a reduction in the gap between research and clinical results.

A. Offline accuracy as a predictor of online usability

Under the conditions of this study, offline classification accuracy fails to accurately predict online usability. For each response variable, the CA-predictor model produced greater error than simply predicting the mean of the validation data for each subject. This is illustrated in Figure 2 by the negative VEcv values. The same can be said for a highly related metric, active classification accuracy (ACA), which removes misclassifications due to no movement.

The poor predictive performance given by these two metrics could be the result of an observed ceiling effect. This effect is especially prominent for ACA in Table I. All twelve subjects produced minimal error with average active accuracies ranging from 98.3% to 100.0%. Since subjects obtained accuracies close to 100%, the accuracy values may not offer a complete domain representation. Although these results are representative of real-world use, the ceiling effect likely limited the influence of accuracy on the usability metrics. Because several studies have shown that both long-term and short-term user practice results in increased classification performance [4], [12], [49], ceiling effects may be common when testing experienced users. This supports the need to identify alternative metrics that provide the necessary information to predict usability.

A recent study by Lv *et al.* investigating the correlation between offline classification accuracy and online usability found that offline accuracy had a strong and significant correlation with completion rate [9]. This same study failed to find a significant correlation between classification accuracy and completion time and between classification accuracy and path efficiency [9]. We avoided the usability metric



(a) One predictor

(b) Two predictor

(c) Three predictors

Fig. 3: Predictive normalized mean squared error (nMSE) for the top feature combinations for the throughput response variable. Shaded regions indicate a 95% confidence interval.

completion rate in our study because almost all users were able to acquire every target; therefore, minimal information about online use could be extracted from this metric. Additionally, completion rate is highly dependent on the nature of the task and on the adopted completion rules (such as the chosen timeout), making completion time and path efficiency arguably more informative usability metrics. Completion time was not directly used in our study; however we implemented the throughput metric which is a function of both completion time and the difficulty rating of the task. Our results support the findings of Lv *et al.* by demonstrating no significant correlations between offline classification accuracy and the usability metrics throughput and path efficiency.

B. Alternative metrics as predictors of online usability

The correlation analysis outlined in Table III differs slightly compared to other evaluations presented in the literature [11], [12]. For example, the separability index and modified separability index have yielded significant or near significant correlations in previous works [11], [12]; however, the results of this study do not support these observations. Similar to the study conducted by Kristofferson *et al.*, we found no significant correlations between interpretations of the separability index and usability [13]. In addition, online performance in other studies has been evaluated by counting correctly classified movements while ignoring the effects of incorrect decisions [11]–[13]. In the present study, however, online testing was evaluated in a Fitts' Law environment where users were actively involved and required to correct for misclassifications made by the control system. Similar environments also have been shown to correlate with functional prosthesis use [18]. Consequently, the user behavior is reflected in the online usability metrics and may be a cause for the observed differences in some correlation results.

The offline metrics showing significant correlations with online usability in Table III are characterized as either separability, neighborhood, or complexity measures. These categories of offline metrics provide similar information by examining the relationship between clusters in feature space. The separability measures in this work present a global analysis of feature space while the complexity measures present a local analysis through the use of feature space partitioning algorithms. Neighborhood measures differ slightly by considering samples located along the class boundaries.

Bhattacharyya distance (BD) and Hellinger distance (HD) both demonstrated significant correlations with the online metric path efficiency. BD and HD were also significantly correlated with each other ($r = 0.70$, $p < 0.001$). Similarly, Fisher's discriminant ratio (FDR) and rescaled purity (rPU), both of which were significantly correlated with average speed and stopping distance, demonstrated a significant correlation with each other ($r = 0.67$, $p = 0.02$). The two final offline metrics in Table III, feature efficiency (FE) and intra-inter fraction (IIF), both of which were significantly correlated with throughput, also rendered significant correlations with each other ($r = -0.7$, $p<0.001$). These results indicate that the pairs of offline metrics described above are not independent.

Although only six offline metrics were significantly correlated with online usability, other metrics exhibited moderate correlations, including the mean semi-principal axes (MSA), collective entropy (CE) and its rescaled version (rCE), neighborhood separability (NS) and its rescaled version (rNS), and inter-class fraction (ICF). Sample size plays a key role in determining whether a result is significant [50]. As sample size increases, both random error and variability decrease, resulting in more precise measurements [50]. Therefore, studies with larger sample sizes are more likely than those with smaller sample sizes to find a significant relationship given one exists [50]. Because of the limited sample size in the current experiment ($n = 12$), additional significant results may have gone undetected.

C. Combinations of metrics as predictors of online usability

In this study, the individual offline metrics did not possess enough information to effectively represent the online use case. A more robust outlook on usability was established when combinations of offline metrics were used as predictors. Importantly, the metrics chosen as part of these predictor sets did not necessarily show significant individual relationships with the response variable. Measures with poor individual correlations were often combined in ways to provide meaningful predictive information, suggesting a more complex relationship between offline performance and online usability. Additionally, the

predictor chosen in a single-predictor model was not necessarily favored in corresponding two- or three-predictor models. Likewise, the predictors in the two-predictor models may not have been selected for the three-predictor models. This indicates that the interaction between the predictors may be just as important as the predictors themselves. One way to determine the relative importance of the predictors is to examine the coefficient weightings assigned to each offline metric. The coefficients for the two-predictor models in Table IV convey relatively equal weightings, implying that both predictors have comparable importance in the linear model. The coefficients for the three-predictor models display a more varied weighting profile; however, any one predictor contributes at least 10%. It should also be noted that the variance inflation factor (VIF) is greater than 5 for many of the predictors in Table IV,indicating correlation between predictors. Models with correlated predictors should be interpreted by looking at how well the combination of predictors predicts the outcome variable, not by looking at any individual predictor and its contribution to the model [39].

The relationship between the offline training metrics and the response variable for a given model differed depending on the total number of predictors. In the single-predictor models, the subjects that demonstrated the highest and lowest online performance generally produced the best and worst scores, respectively, for the corresponding offline predictor; for example, subject 7 achieved the highest throughput and effective throughput while also yielding the highest feature efficiency (FE). Subject 12 exhibited the lowest throughput and effective throughput while generating the lowest FE. These results support the direct relationship between the offline metric in the single-predictor model and the online usability metric.

The relationship between the predictor variables and the response variable became more involved when multiple offline metrics were used to predict usability; for example, subject 7, who yielded the best throughput and effective throughput, produced high MAV values but average MSA compared to the other subjects. Subject 12, who demonstrated the lowest throughput and effective throughput, produced high MAV and high MSA. These findings potentially indicate an interaction between the two predictor variables. Given the evidence of multicollinearity between the MSA and MAV predictors, it is difficult to interpret the individual effects of the predictors on the response variable. However, even though the individual effects cannot necessarily be determined, the fact that similar models were generated for throughput and effective throughput (which are highly related online metrics) support the idea that the chosen predictors are representative of online performance.

To the best of our knowledge, no other studies have assessed the relationship between offline training metrics and online performance using multiple linear regression. Based on the results of this experiment, grouping offline metrics together may be a more instructive approach than trying to discover a singular metric that encompasses all of the variability of online use. This could be due, in part, to the required coordination of pattern generation, proportional control, and target acquisition during online use. The best performing single predictors outlined in Table IV generated models that were classified as having poor predictive accuracy as measured by VEcv [45]. It is important to note, however, that although the predictive behavior was classified as poor, the selected single predictors generally produced better predictions than simply predicting using the mean of the data. Furthermore, our selected predictor resulted in a substantial increase in performance compared to the current offline standard, classification accuracy. The two-predictor models generally displayed average predictive performance while the selected three-predictor models exhibited good predictive behavior [45]. This improving trend was also present for the measures of goodness of fit, indicating that both the model parameters and the predictive capacity improve as the number of predictors increased from one to three.

The models generated for the response variable overshoot were the only models that deviated from this improving trend. It may be that overshoot is more difficult to predict compared to the other online metrics and requires information about the separability between motion classes (from BD), knowledge of the class boundaries (from NS), and details about the uncertainty and disorder of the dataset (from rCE) for accurate predictions. Furthermore, overshoot reflects aspects of the user's task planning and reaction time, which may not be sufficiently reflected within the SGT training approach.

The results of this study did not identify a common set of predictors across all usability metrics. This suggests that the online metrics are providing unique information regarding the usability of the system. As we have presented our results, users would likely have to prioritize one aspect of usability and target the offline metrics associated with the corresponding model.

Although models outlined in Table IV produced the lowest error, they were not the only acceptable predictor combinations Figures 3b and 3c show a gradual increase in nMSE as additional predictors were evaluated. The gradual increase is evidence that other sets of offline metrics can produce predictions comparable to those rendered by the top selected combination. Furthermore, the results in Figures 3b and 3c, suggest that the top performing predictor set was not assembled by chance. Many of the predictor combinations along the x-axis contain similar metrics (i.e, MSA for the two-predictor sets and MSA and MAV for the three-predictor sets). It is also important to note that the predictor combinations plotted in Figures 3b and 3c are the best performing sets out of 496 combinations for two predictors and 4960 combinations for three predictors. Not all predictor combinations led to acceptable performance. The worst combination of two metrics for predicting throughput was Hellinger distance (HD) and volume of overlap region (VOR), leading to a nMSE of 5757.

D. Limitations and Future Work

As with any experiment with a limited sample size, overfitting poses a major concern. Training of the different regression model folds was performed with eleven subjects and only one subject was included in the test set. Although studies have demonstrated that accurate regression models can be formed with as little as two [51] and five [52] samples per predictor, it is more accepted to have at least ten samples per predictor [53]. The basis for limiting the number of predictors in the linear models stemmed from concerns about overfitting. When evaluating models with four predictors during pilot testing, we observed a general drop in the AICc generalization performance as compared to the two- and three- predictor models. Consequently, because these models were more likely to be overfit, we limited the input space to three.

Our results demonstrated that the two- and three-predictor models generated average and good prediction accuracy, respectively. However, cross-validation was based on the prediction of only one subject. Therefore, the reported absolute predictive performance may be an inflated view of the true model behavior. For the best generalizability, an additional study with more subjects should be conducted with a proper training, validation, and test set. Additionally, because our main focus was on predicting usability, we did not attempt to solve multicollinearity among the predictors in the regression models. Multicollinearity makes it difficult to accurately investigate associations among the predictor variables, but it does not impact the fit of the model or the model's predictions [39]. For these reasons, we refrain from making model specific recommendations, but rather suggest that researchers move toward predicting usability using a variety of offline training metrics.

This study also did not take into account additional variables that could potentially have an influence on predicting usability, such as the user's experience level, age, and gender. Further expansion of the types of feedback and number of offline metrics may also be beneficial to the research community to provide a more comprehensive evaluation of predictive performance. Other areas of work could focus on extending the subject population group to those with neurological disorders or physical disabilities.

V. Conclusions

This work provides a foundation for using offline training metrics as predictors of online usability. Currently, classification accuracy is the most reported offline metric for describing myoelectric control performance [3], [5], [54]. The results of this work support many previous studies by showing that offline classification accuracy is a poor indicator of usability [6]–[8]. Unfortunately, there is little consensus in the literature about the use of alternative offline metrics to indicate online performance. This work identified metrics that, under the conditions of this study, were shown to be more powerful predictors than what has previously been used in the literature. To the best of our knowledge, no work has investigated a combination of offline metrics that embody the ability to predict usability. Our findings suggest that a combination of two and three offline metrics may provide a more robust framework for predicting online performance.

References

[1] C. A. Okoro, N. D. Hollis, A. C. Cyrus, and S. Griffin-Blake, "Prevalence of Disabilities and Health Care Access by Disability Status and Type Among Adults — United States, 2016," Department of Health and Human Services, Tech. Rep. 32, Aug. 2018.

[2] C. Bizier, G. Fawcett, and S. Gilbert, *Mobility disabilities among Canadians aged 15 years and older, 2012*. Statistics Canada, 2012.

[3] E. Scheme and K. Englehart, "Electromyogram pattern recognition for control of powered upper-limb prostheses: State of the art and challenges for clinical use," *J. Rehabil. Res. Dev.*, vol. 48, no. 6, pp. 643–660, 2011.

[4] M. A. Powell, R. R. Kaliki, and N. V. Thakor, "User training for pattern recognition-based myoelectric prostheses: Improving phantom limb movement consistency and distinguishability," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 22, no. 3, pp. 522–532, 2014.

[5] N. Seth, R. C. Freitas, M. Chaulk, C. O'Connell, K. Englehart, and E. Scheme, "EMG pattern recognition for persons with cervical spinal cord injury," in *IEEE 16th Int. Conf. Rehabil. Robot.* IEEE Computer Society, Jun. 2019, pp. 1055–1060.

[6] B. A. Lock, K. Englehart, and B. Hudgins, "Real-time myoelectric control in a virtual environment to relate usability vs. accuracy," in *MyoElectric Control. Prosthetics Symp.*, 2005.

[7] L. Hargrove, Y. Losier, B. Lock, K. Englehart, and B. Hudgins, "A real-time pattern recognition based myoelectric control usability study implemented in a virtual environment," in *Annu. Int. Conf. IEEE Eng. Med. Biol. - Proc.*, 2007, pp. 4842–4845.

[8] A. Krasoulis, S. Vijayakumar, and K. Nazarpour, "Effect of User Practice on Prosthetic Finger Control With an Intuitive Myoelectric Decoder," *Front. Neurosci.*, vol. 13, no. Sept., pp. 1–16, 2019.

[9] B. Lv, X. Sheng, D. Hao, and X. Zhu, "Relationship between Offline and Online Metrics in Myoelectric Pattern Recognition Control Based on Target Achievement Control Test," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, pp. 6595–6598, 2019.

[10] L. J. Hargrove, E. J. Scheme, K. B. Englehart, and B. S. Hudgins, "Multiple binary classifications via linear discriminant analysis for improved controllability of a powered prosthesis," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 18, no. 1, pp. 49–57, Feb. 2010.

[11] N. Nilsson, B. Håkansson, and M. Ortiz-Catalan, "Classification complexity in myoelectric pattern recognition," *J. Neuroeng. Rehabil.*, vol. 14, no. 1, Jul. 2017.

[12] N. E. Bunderson, T. A. Kuiken, and S. Member, "Quanti fi cation of Feature Space Changes With Experience During Electromyogram Pattern Recognition Control," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 20, no. 3, pp. 239–246, 2012.

[13] M. B. Kristoffersen, A. W. Franzke, C. K. V. D. Sluis, A. Murgia, and R. M. Bongers, "The effect of feedback during training sessions on learning pattern-recognition based prosthesis control," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 10, pp. 2087–2096.

[14] T. A. Kuiken, G. Li, B. A. Lock, R. D. Lipschutz, L. A. Miller, K. A. Stubblefield, and K. B. Englehart, "Targeted muscle reinnervation for real-time myoelectric control of multifunction artificial arms," *JAMA - J. Am. Med. Assoc.*, vol. 301, no. 6, pp. 619–628, Feb. 2009.

[15] R. J. Smith, D. Huberdeau, F. Tenore, and N. V. Thakor, "Real-time myoelectric decoding of individual finger movements for a virtual target task," in *Proc. 31st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBC 2009*. IEEE Computer Society, 2009, pp. 2376–2379.

[16] A. M. Simon, L. J. Hargrove, B. A. Lock, and T. A. Kuiken, "Target achievement control test: Evaluating real-time myoelectric pattern-recognition control of multifunctional upper-limb prostheses," *J. Rehabil. Res. Dev.*, vol. 48, no. 6, pp. 619–628, 2011.

[17] E. J. Scheme and K. B. Englehart, "Validation of a selective ensemble-based classification scheme for myoelectric control using a three-dimensional fitts' law test," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 21, no. 4, pp. 616–623, 2013.

[18] L. Hargrove, L. Miller, K. Turner, and T. Kuiken, "Control within a virtual environment is correlated to functional outcomes when using a physical prosthesis," *J. Neuroeng. Rehabil.*, vol. 15, no. S1, p. 60, 2018.

[19] J. Gusman, E. Mastinu, and M. Ortiz-Catalán, "Evaluation of computer-based target achievement tests for myoelectric control," *IEEE J. Transl. Eng. Heal. Med.*, vol. 5, Nov. 2017.

[20] D. Atkins, D. Heard, and W. Donovan, "Epidemiologic Overview of Individuals with Upper-Limb Loss and Their Reported Research Priorities," *J. Prosthetics Orthot.*, vol. 8, no. 1, 1996.

[21] A. Chadwell, L. Kenney, S. Thies, A. Galpin, and J. Head, "The reality of myoelectric prostheses: Understanding what makes these devices difficult for some users to control," *Front. Neurorobot.*, vol. 10, no. Aug., 2016.

[22] S. Millstein, H. Heger, and G. Hunter, "A Review of the Failures In Use of the Below Elbow Myoelectric Prosthesis," *Orthot. Prosthetics*, vol. 36, no. 2, pp. 29–34, 1982.

[23] E. Biddiss and T. Chau, "Upper limb prosthesis use and abandonment: A survey of the last 25 years," *Prosthet. Orthot. Int.*, vol. 31, no. 3, pp. 236–257, Sept. 2007.

[24] A. Phinyomark, P. Phukpattaranont, and C. Limsakul, "Feature reduction and selection for EMG signal classification," *Expert Syst. Appl.*, vol. 39, no. 8, pp. 7420–7431, jun 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417412001200

[25] E. Scheme and K. Englehart, "A Flexible User Interface for Rapid Prototyping of Advanced Real-time Myoelectric Control Schemes," in *MyoElectric Control. Prosthetics Symp.*, 2008.

[26] J. W. Sensinger, B. A. Lock, and T. A. Kuiken, "Adaptive pattern recognition of myoelectric signals: Exploration of conceptual framework and practical algorithms," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 17, no. 3, pp. 270–278, June 2009.

[27] R. Menon, G. Di Caterina, H. Lakany, L. Petropoulakis, B. A. Conway, and J. J. Soraghan, "Study on Interaction between Temporal and Spatial Information in Classification of EMG Signals for Myoelectric Prostheses," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 10, pp. 1832–1842, 2017.

[28] B. Hudgins, P. Parker, and R. N. Scott, "A New Strategy for Multifunction Myoelectric Control," *IEEE Trans. Biomed. Eng.*, vol. 40, no. 1, pp. 82–94, 1993.

[29] E. J. Scheme, B. S. Hudgins, and K. B. Englehart, "Confidence-based rejection for improved pattern recognition myoelectric control," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 6, pp. 1563–1570, 2013.

[30] E. Scheme and K. Englehart, "A comparison of classification based confidence metrics for use in the design of myoelectric control systems," in *Proc. 2015 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*. Milan, Italy: IEEE, 2015, pp. 7278–7283.

[31] Paul M. Fitts, "The Information Capacity of the Human Motor System in Controlling the Amplitude of Movement," *J. Exp. Psychol.*, vol. 47, no. 6, pp. 381–391, 1954.

[32] J. Park, W. Bae, H. Kim, and S. Park, "EMG - Force correlation considering Fitts' law," in *IEEE Int. Conf. Multisens. Fusion Integr. Intell. Syst.*, 2008, pp. 644–649.

[33] E. Scheme, B. Lock, L. Hargrove, W. Hill, U. Kuruganti, and K. Englehart, "Motion normalized proportional control for improved pattern recognition-based myoelectric control," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 22, no. 1, pp. 149–157, 2014.

[34] E. J. Scheme, B. S. Hudgins, and K. B. Englehart, "Confidence-based rejection for improved pattern recognition myoelectric control," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 6, pp. 1563–1570, 2013.
[35] L. H. Smith, "Classification Error and Controller Delay," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 19, no. 2, pp. 186–192, 2011.
[36] E. N. Kamavuako, E. J. Scheme, and K. B. Englehart, "On the usability of intramuscular EMG for prosthetic control: A Fitts' Law approach," *J. Electromyogr. Kinesiol.*, vol. 24, no. 5, pp. 770–777, 2014.
[37] M. R. Williams and R. F. Kirsch, "Evaluation of Head Orientation and Neck Muscle EMG Signals as Command Inputs to a Human-Computer Interface for Individuals with High Tetraplegia," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 16, no. 5, pp. 485–496, 2008.
[38] C. E. Shannon and W. Weaver, "The Mathematical Theory of Communication," *Univ. Illinois Press. Urbana*, 1949.
[39] K. Vatcheva, M. Lee, J. McCormick, and M. Rahbar, "Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies," *Epidemiol.*, vol. 6, no. 2, 2016.
[40] S. Parvandeh, H. W. Yeh, M. P. Paulus, and B. A. McKinney, "Consensus features nested cross-validation," *Bioinformatics*, vol. 36, no. 10, pp. 3093–3098, 2020.
[41] J. Li, "Assessing the accuracy of predictive models for numerical data: Not r nor r2, why not? Then what?" *PLoS One*, vol. 12, no. 8, Aug. 2017.
[42] J. M. Wooldridge, "A note on computing r-squared and adjusted r-squared for trending and seasonal data," *Econ. Lett.*, vol. 36, no. 1, pp. 49–54, 1991.
[43] M. J. Brewer, A. Butler, and S. L. Cooksley, "The relative performance of AIC, AICc and BIC in the presence of unobserved heterogeneity," *Methods Ecol. Evol.*, vol. 7, no. 6, pp. 679–692, June 2016.
[44] S. Kim and H. Kim, "A new metric of absolute percentage error for intermittent demand forecasts," *Int. J. Forecast.*, vol. 32, no. 3, pp. 669–679, Jul. 2016.
[45] J. Li, "Assessing spatial predictive models in the environmental sciences: Accuracy measures, data variation and variance explained," *Environ. Model. Softw.*, vol. 80, pp. 1–8, June 2016.
[46] H. Khamis, "Measures of association: How to choose?" *J. Diagnostic Med. Sonogr.*, vol. 24, no. 3, pp. 155–162, 2008.
[47] H. Akoglu, "User's guide to correlation coefficients," *Turkish J. Emerg. Med.*, vol. 18, no. 3, pp. 91–93, 2018.
[48] L. Resnik, F. Acluche, M. Borgia, G. Latlief, and S. Phillips, "EMG pattern recognition control of the DEKA Arm: Impact on user ratings of satisfaction and usability," *IEEE J. Transl. Eng. Heal. Med.*, vol. 7, 2019.
[49] J. He, D. Zhang, N. Jiang, and X. Sheng, "User adaptation in long-term , open-loop myoelectric training : Implications for EMG pattern recognition in prosthesis control," *J. Neural Eng.*, vol. 12, 2015.
[50] M. S. Thiese, B. Ronna, and U. Ott, "P value interpretations and considerations," *J. Thorac. Dis.*, vol. 8, no. 9, pp. E928–E931, 2016.
[51] P. C. Austin and E. W. Steyerberg, "The number of subjects per variable required in linear regression analyses," *J. Clin. Epidemiol.*, vol. 68, no. 6, pp. 627–636, June 2015.
[52] B. G. Tabachnick and L. S. Fidell, *Multivariate Statistics*, 2nd ed. New York: Harper & Row, 1989.
[53] F. E. Harrell, *Regression modeling strategies.* Springer, 2001.
[54] Y. Fang, D. Zhou, S. Member, K. Li, and S. Member, "Classifier-Feedback-Based User Training," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 11, pp. 2575–2583, 2017.
[55] N. E. Krausz, B. H. Hu, and L. J. Hargrove, "Subject-and environment-based sensor variability for wearable lower-limb assistive devices," *Sensors (Switzerland)*, vol. 19, no. 22, nov 2019.
[56] A. Bhattacharyya, "On a measure of divergence between two multinomial populations," *Indian J. Stat.*, vol. 7, no. 5, pp. 401–406, 1946.
[57] T. K. Ho and M. Basu, "Complexity measures of supervised classification problems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 289–300, Mar. 2002.
[58] A. F. Kohn, L. G. Nakano, and M. O. E. Silva, "A class discriminability measure based on feature space partitioning," *Pattern Recognition*, 1996.
[59] S. Singh, "PRISM-A novel framework for pattern recognition," *Pattern Anal Applic*, vol. 6, pp. 134–149, 2003.
[60] E. Campbell, A. Phinyomark, and E. Scheme, "Current trends and confounding factors in myoelectric control: Limb position and contraction intensity," *Sensors (Switzerland)*, vol. 20, no. 6, p. 1613, mar 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/6/1613/htm https://www.mdpi.com/1424-8220/20/6/1613

VI. Appendix: Variability Metrics

Repeatability Index (RI): A measure of the reproducibility of EMG patterns between repetitions [12].

$$RI = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{R}\sum_{r=1}^{R}\frac{1}{2}\sqrt{(\mu_{TR_j}-\mu_{r,j})'S_{TR_j}^{-1}(\mu_{TR_j}-\mu_{r,j})}$$

where μ_{TR_j} is the centroid of the class j training ellipsoid, $\mu_{r,j}$ is the centroid of a testing ellipsoid of the same class j from repetition r, S_{TR_j} is the covariance matrix of the class j training ellipsoid, R is the total number of repetitions, and N is the total number of active motion classes.

mean within-repetition Repeatability Index (mwRI): An interpretation of Bunderson and Kuiken's repeatability index [12]

$$mwRI = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{P}\sum_{p=1}^{P}\frac{1}{2}\times\sqrt{(\mu_{TR_{r,j}}-x_{TR_{p,r,j}})'S_{TR_j}^{-1}(\mu_{TR_{r,j}}-x_{TR_{p,r,j}})}$$

where $\mu_{TR_{r,j}}$ is the centroid of class j from a given repetition r, $x_{TR_{p,r,j}}$ is a data point in r, S_{TR_j} is the covariance matrix of the class j training ellipsoid, R is the total number of receptions, P is the total number of data points in R, and N is the total number of active motion classes.

standard deviation within-repetition Repeatability Index (swRI): A measure of the variation of the within-repetition repeatability across repetitions.

$$swRI = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}\left(\left(\frac{1}{N}\sum_{j=1}^{N}\frac{1}{P}\sum_{p=1}^{P}\frac{1}{2}\times\sqrt{(\mu_{TR_{r,j}}-x_{TR_{p,r,j}})'S_{TR_j}^{-1}(\mu_{TR_{r,j}}-x_{TR_{p,r,j}})}\right)-mwRI\right)^2}$$

where $\mu_{TR_{r,j}}$ is the centroid of class j from a given repetition r, $x_{TR_{p,r,j}}$ is a data point within that repetition, S_{TR_j} is the covariance matrix of the class j training ellipsoid, mwRI is the offline metric described above, R is the total number of receptions, P is the total number of data points in R, and N is the total number of active motion classes.

standard deviation within-trial Separability Index (swSI): A measure of the variability of the distinguishability of EMG patterns across trials.

$$swSI = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{j=1}^{N}\min_{i=1,..,j-1,j+1,..,N}\frac{1}{2}\times\sqrt{(\mu_{TR_{j,t}}-\mu_{TR_{i,t}})'S_{TR_{j,t}}^{-1}(\mu_{TR_{j,t}}-\mu_{TR_{i,t}})}\right)-mwSI)^2}$$

where $\mu_{TR_{j,t}}$ is the centroid of class j from trial t, $\mu_{TR_{i,t}}$ is the centroid of the nearest training ellipsoid of a different class i from trial t, $S_{TR_j,t}$ is the covariance matrix of the class j training ellipsoid from trial t, mwSI is the offline metric defined in Appendix VII, T is the total number of trials, and N is the total number of active motion classes.

Mean Semi-principal Axes (MSA): A measure that quantifies the size of a training ellipsoid. [12].

$$MSA = \frac{1}{N}\sum_{j=1}^{N}\left(\left(\prod_{k=1}^{D}a_{j,k}\right)^{1/D}\right)$$

where a_k is the geometric mean of each semi-principal axis (calculated using Principal Component Analysis (PCA)) in dimension k [55], D is the total dimensionality of the feature space, and N is the total number of active motion classes.

Centroid Drift (CD): A measure that quantifies the variation in centroid location of a training ellipsoid across subsequent repetitions.

$$CD = \frac{1}{N}\sum_{j=1}^{N}\left(\sum_{r=1}^{R-1}||\mu_{r,j}-\mu_{r+1,j}||\right)$$

where $\mu_{r,j}$ is the centroid of a training ellipsoid of class j in repetition r, $(\mu_{r+1,j})$ is the centroid from the next repetition of class j, R is the total number of repetitions, and N is the total number of active motion classes.

Mean Absolute Value (MAV): A measure that specifies the average amplitude of the EMG signal.

$$MAV = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{1}{E}\sum_{ch=1}^{E}\left(\frac{1}{n}\sum_{i=1}^{n}|x_{i,ch,j}|\right)\right)$$

where x is the raw EMG signal in the ith data frame, n is the total number of data frames, E is the total number of electrode channels, and N is the total number of active motion classes.

VII. Appendix: Separability Metrics

Separability Index (SI): A measure of interclass distance [12].

$$SI = \frac{1}{N}\sum_{j=1}^{N} \min_{i=1,..,j-1,j+1,..,N} \frac{1}{2} \times \sqrt{(\mu_{TR_j} - \mu_{TR_i})' S_{TR_j}^{-1} (\mu_{TR_j} - \mu_{TR_i})}$$

where μ_{TR_j} is the centroid of the class j training ellipsoid (includes all repetitions), μ_{TR_i} is the centroid of the nearest training ellipsoid of a different class i, S_{TR_j} is the covariance matrix of the class j training ellipsoid, and N is the total number of active motion classes.

modified Separability Index (SI): A measure similar to the separability index, except that it accounts for the covariance matrix of both distributions being compared [11].

$$mSI = \frac{1}{N}\sum_{j=1}^{N} \min_{i=1,..,j-1,j+1,..N} \frac{1}{2} \times \sqrt{(\mu_{TR_j} - \mu_{TR_i})' S^{-1} (\mu_{TR_j} - \mu_{TR_i})}$$

where μ_{TR_j} is the centroid of the class j training ellipsoid (includes all repetitions), μ_{TR_i} is the centroid of the nearest training ellipsoid of a different class i, S is the average covariance matrix of the class j covariance S_{TR_j} and the class i covariance S_{TR_i}, and N is the total number of active motion classes.

mean within-trial Separability Index (mwSI): A measure of the distinguishability of EMG patterns within a trial.

$$mwSI = \frac{1}{T}\sum_{t=1}^{T} \frac{1}{N}\sum_{j=1}^{N} \min_{i=1,..,j-1,j+1,..,N} \frac{1}{2} \times \sqrt{(\mu_{TR_{j,t}} - \mu_{TR_{i,t}})' S_{TR_{j,t}}^{-1} (\mu_{TR_{j,t}} - \mu_{TR_{i,t}})}$$

where $\mu_{TR_{j,t}}$ is the centroid of the class j training ellipsoid from trial t, $\mu_{TR_{i,t}}$ is the centroid of the nearest training ellipsoid of a different class i from trial t, $S_{TR_j,t}$ is the covariance matrix of the class j training ellipsoid from trial t, T is the total number of trials, and N is the total number of active motion classes.

Bhattacharyya Distance (BD): A measure of the statistical similarity between two distributions [56].

$$BD = \frac{1}{N}\sum_{j=1}^{N} \min_{i=1,...,j-1,j+1,...N} \frac{1}{8}(\mu_{TR_j} - \mu_{TR_i})' S^{-1} (\mu_{TR_j} - \mu_{TR_i}) - \frac{1}{2} \ln \left(\frac{|S|}{\sqrt{|S_{TR_j}||S_{TR_i}|}} \right)$$

where μ_{TR_j} is the centroid of the class j training ellipsoid, μ_{TR_i} is the centroid of the nearest training ellipsoid of a different class i, S is the average covariance matrix of the class j covariance S_{TR_j} and the class i covariance S_{TR_i}, and N is the total number of active classes.

Kullback-Leibler Divergence (KLD): A measure of how well a distribution can be approximated by a reference distribution [11].

$$KLD = \frac{1}{N}\sum_{j=1}^{N} \min_{i=1,...,j-1,j+1,...N} \frac{1}{2} \left(\mathrm{Tr}\left(S_{TR_j}^{-1}, S_{TR_i} \right) + (\mu_{TR_j} - \mu_{TR_i})' S_{TR_j}^{-1} (\mu_{TR_j} - \mu_{TR_i}) - D + \ln \left(\frac{|S_{TR_j}|}{|S_{TR_i}|} \right) \right)$$

where μ_{TR_j} is the centroid of the class j training ellipsoid, μ_{TR_i} is the centroid of the nearest training ellipsoid of a different class i, S_{TR_j} is the covariance matrix of class j, S_{TR_i} is the covariance matrix of class i, D is the dimensionality of feature space, and N is the total number of active classes.

Hellinger Distance (HD): A measure that quantifies the similarity between two probability distributions. The square of the Hellinger distance avoids the presence of complex numbers when the assumption of normality fails.

$$HD = \frac{1}{N}\sum_{j=1}^{N} \min_{i=1,...,j-1,j+1,...N} 1 - \frac{(|S_{TR_i}|)^{1/4}(|S_{TR_j}|)^{1/4}}{(|S|)^{1/2}} \times \exp\left\{ -\frac{1}{8}(\mu_{TR_j} - \mu_{TR_i})' S^{-1} (\mu_{TR_j} - \mu_{TR_i}) \right\}$$

where μ_{TR_j} is the centroid of the class j training ellipsoid, μ_{TR_i} is the centroid of the nearest training ellipsoid of a different class i, S is the average covariance matrix of the class j covariance S_{TR_j} and the class i covariance S_{TR_i}, and N is the total number of active classes.

Volume of Overlap Region (VOR): A measure of the degree of overlap between the tails of two class conditional distributions [57].

$$VOR = \frac{1}{N}\sum_{j=1}^{N} \max_{i=1,...,j-1,j+1,...N} \prod_{k} \frac{\min\left(\max(f_k|c_j), \max(f_k|c_i)\right) - \max\left(\min(f_k|c_j), \min(f_k|c_i)\right)}{\max\left(\max(f_k|c_j), \max(f_k|c_i)\right) - \min\left(\min(f_k|c_j), \min(f_k|c_i)\right)}$$

where $\max(f_k|c_j)$ is the maximum value of feature f in dimension k for class label j, $\max(f_k|c_i)$ is the maximum value of feature f in dimension k for class label i, $\min(f_k|c_j)$ is the minimum value of feature f in dimension k for class label j, $\min(f_k|c_i)$ is the minimum value of feature f in dimension k for class label i, and N is the total number of active motion classes.

Feature Efficiency (FE): A measure of the fraction of points separable by a particular feature [57].

$$FE = \frac{1}{N}\sum_{j=1}^{N} \max_{i=1,\ldots,j-1,j+1,\ldots N}\left(\max_{k=1,\ldots,D}\frac{n(C_i)+n(C_j)-n(S_k)}{n(C_i)+n(C_j)}\right)$$

$$S_k = p|p \in C_i \cup C_j, \min\left(\max(f_k|c_j), \max(f_k|c_i)\right) \geq p \geq \max\left(\min(f_k|c_j), \min(f_k|c_i)\right)$$

where S_k is the set of points not separable along feature dimension k, p is a D dimensional data point in class i or class j, $\max(f_k|c_j)$ is the maximum value of feature f in dimension k for class label j, $\max(f_k|c_i)$ is the maximum value of feature f in dimension k for class label i, $\min(f_k|c_j)$ is the minimum value of feature f in dimension k for class label j, $\min(f_k|c_i)$ is the minimum value of feature f in dimension k for class label i,$n(S_k)$ is the cardinality of the overlap set S_k, $n(C_i)$ is the cardinality of the set of points in class i, C_i, and $n(C_j)$ is the cardinality of the set of points in class j, C_j, and N is the total number of active classes.

Trace of the within-class and between-class Scatter Matrices (TSM): A measure of class discriminability [58].

$$TSM = \mathrm{Tr}\left(S_w^{-1}S_b\right)$$

$$S_w = \frac{1}{N}\sum_{j=1}^{N}\left(\sum_{i=1}^{n}(x_i-\mu_j)(x_i-\mu_j)'\right)$$

$$S_b = \frac{1}{C}\sum_{j=1}^{C}\left(n_j(\mu_j-\mu)(\mu_j-\mu)'\right)$$

where (S_w) is the within-class scatter matrix, (S_b) is the between class scatter matrix, N is the total number of motion classes, n is the total number of data frames, x is a data point in class j, (μ_j) is the centroid of the class j training ellipsoid, (μ) is the mean of the entire data set, and n_j is the number of data frames in class j.

Desirability Score (DS): A function of the separability index, the mean semi-principal axes, and the repeatability index [55].

$$DS = \frac{(SI)}{(RI)(MSA)}$$

where SI is the separability index defined above, RI is the repeatability index defined in Appendix VI, and MSA is the mean semi-principal axes defined in Appendix VI.

VIII. Appendix: Complexity Metrics

Class Discriminability Measure (CDM): A measure derived from the adaptive partitioning algorithm in [58] that provides information about the relationship between clusters in feature space.

$$CDM = \frac{1}{n}\sum_{i=1}^{M} h(i) - \max_{j} h(j|i)$$

where M is the total number of nonhomogeneous and not linearly separable cells, h(i) is the number of samples in the ith analysis cell, $h(j|i)$ is the number of samples from class j in the ith analysis cell, and n is the total number of samples in feature space.

Purity (PU): A measure derived from the PRISM framework in [59] that assess the homogeneity of the training data. Detailed formulations and implementation procedures are found in [59].

Neighborhood Separability (NS): A measure derived from the PRISM framework in [59] that focuses on the class decision boundaries by quantifying the relationship between nearest neighbors. Detailed formulations and implementation procedures are found in [59].

Collective Entropy (CE): A measure derived from the PRISM framework in [59] that represents the accumulated uncertainty in the data across different resolutions. Detailed formulations and implementation procedures are found in [59].

Compactness (C): A measure derived from the PRISM framework in [59] that provides an estimate of the spread of the data. Detailed formulations and implementation procedures are found in [59].

Weighted/rescaled versions of PU, NS, CE, and C were calculated by dividing by the maximum possible area under the weighted metric vs. normalized resolution curve, as described in [59].

IX. Appendix: Classification Metrics

Classification Accuracy (CA): A measure describing the fraction of predictions the classifier labelled correctly calculated using a leave-one-trial-out cross validation technique [60].

$$CA = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{y}_{i,t} == y_{i,t}\right)$$

Active Classification Accuracy (ACA): A measure describing the fraction of predictions the classifier labelled correctly excluding misclassifications due to no motion [60].

$$ACA = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{n}\sum_{i=1}^{n}\left((\hat{y}_{i,t} == y_{i,t}) \text{ OR } (\hat{y}_{i,t} == y_{NM})\right)\right)$$

where where "==" generates a Boolean value (0 or 1), n is the total number of data frames, T is the total number of trials, $\hat{y}_{i,t}$ is the predicted class label for data point i, y_{NM} is the no movement class label, and $y_{i,t}$ is the true class label for data point i. where "==" generates a Boolean value (0 or 1), n is the total number of data frames, T is the total number of trials, $\hat{y}_{i,t}$ is the predicted class label for data point i, and $y_{i,t}$ is the true class label for data point i.

Usable Data (UD): A measure describing the percentage of correctly classified decisions over the entire user training period using the adaptive classifier procedure outlined in Figure 1.

$$UD = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_{i,t} == y_{i,t}\right)$$

where "==" generates a Boolean value (0 or 1), n is the total number of data frames, $\hat{y}_{i,t}$ is the predicted class label for data point i, and $y_{i,t}$ is the true class label for data point i.

X. Appendix: Neighborhood Metrics

Inter-Class Fraction (ICF): A measure describing the ratio of the number of inter-class nearest neighbors to the total number of samples in the data set [57].

$$ICF = \frac{1}{n}\sum_{t=1}^{n}\left(y_t \neq y_e\right)$$

where "$\neq$" generates a Boolean value (0 or 1), x_e is the nearest neighbor calculated using the Euclidean distance to data point x_t, y_e is the class label associated with data point x_e, y_t is the class label associated with data point x_t, and n is the total number of data samples.

Intra-Inter Fraction (ICF): A measure describing the ratio of the average euclidean distance of intra-class nearest neighbors to the average euclidean distance of inter-class nearest neighbors [57].

$$IIF = \frac{\sum_{t=1}^{n}\left(d_{(x_t,x_e)} \times (y_t == y_e)\right)}{\sum_{t=1}^{n}\left(d_{(x_t,x_e)} \times (y_t \neq y_e)\right)}$$

where "==" and "$\neq$" both generate Boolean values (0 or 1), $d_{(x_t,x_e)}$ is the euclidean distance between a data point x_t with class label y_t and its nearest neighbor x_e with class label y_e.