Ergonomics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/terg20

Assessing workload in using electromyography (EMG)-based prostheses

Junho Park, Joseph Berman, Albert Dodson, Yunmei Liu, Matthew Armstrong, He Huang, David Kaber, Jaime Ruiz & Maryam Zahabi

To cite this article: Junho Park, Joseph Berman, Albert Dodson, Yunmei Liu, Matthew Armstrong, He Huang, David Kaber, Jaime Ruiz & Maryam Zahabi (2023): Assessing workload in using electromyography (EMG)-based prostheses, Ergonomics, DOI: 10.1080/00140139.2023.2221413

To link to this article: https://doi.org/10.1080/00140139.2023.2221413

Published online: 12 Jun 2023.

Submit your article to this journal

Article views: 52

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=terg20

ERGONOMICS
https://doi.org/10.1080/00140139.2023.2221413

Taylor & Francis
Taylor & Francis Group

Check for updates

Assessing workload in using electromyography (EMG)-based prostheses

Junho Park[a], Joseph Berman[b], Albert Dodson[c,d], Yunmei Liu[e], Matthew Armstrong[f], He Huang[c,d], David Kaber[e], Jaime Ruiz[g] and Maryam Zahabi[a]

[a]Wm Michael Barnes '64 Department of Industrial & Systems Engineering, Texas A&M University, College Station, TX, USA; [b]Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC, USA; [c]Joint Department of Biomedical Engineering, North Carolina State University, Raleigh, NC, USA; [d]Joint Department of Biomedical Engineering, University of North Carolina at Chapel Hill, NC, USA; [e]Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL, USA; [f]Intercollegiate School of Engineering Medicine, Texas A&M University, Houston, TX, USA; [g]Department of Computer & Information Science & Engineering, University of Florida, Gainesville, FL, USA

ABSTRACT

Using prosthetic devices requires a substantial cognitive workload. This study investigated classification models for assessing cognitive workload in electromyography (EMG)-based prosthetic devices with various types of input features including eye-tracking measures, task performance, and cognitive performance model (CPM) outcomes. Features selection algorithm, hyperparameter tuning with grid search, and k-fold cross-validation were applied to select the most important features and find the optimal models. Classification accuracy, the area under the receiver operation characteristic curve (AUC), precision, recall, and F1 scores were calculated to compare the models' performance. The findings suggested that task performance measures, pupillometry data, and CPM outcomes, combined with the naïve bayes (NB) and random forest (RF) algorithms, are most promising for classifying cognitive workload. The proposed algorithms can help manufacturers/clinicians predict the cognitive workload of future EMG-based prosthetic devices in early design phases.

Practitioner summary: This study investigated the use of machine learning algorithms for classifying the cognitive workload of prosthetic devices. The findings suggested that the models could predict workload with high accuracy and low computational cost and could be used in assessing the usability of prosthetic devices in the early phases of the design process.

Abbreviations: 3d: 3 dimensional; ADL: Activities for daily living; ANN: Artificial neural network; AUC: Area under the receiver operation characteristic curve; CC: Continuous control; CPM: Cognitive performance model; CPM-GOMS: Cognitive-Perceptual-Motor GOMS; CRT: Clothespin relocation test; CV: Cross validation; CW: Cognitive workload; DC: Direct control; DOF: Degrees of freedom; ECRL: Extensor carpi radialis longus; ED: Extensor digitorum; EEG: Electroencephalogram; EMG: Electromyography; FCR: Flexor carpi radialis; FD: Flexor digitorum; GOMS: Goals, Operations, Methods, and Selection Rules; LDA: Linear discriminant analysis; MAV: Mean absolute value; MCP: Metacarpophalangeal; ML: Machine learning; NASA-TLX: NASA task load index; NB: Naïve Bayes; PCPS: Percent change in pupil size; PPT: Purdue Pegboard Test; PR: Pattern recognition; PROS-TLX: Prosthesis task load index; RF: Random forest; RFE: Recursive feature selection; SHAP: Southampton hand assessment protocol; SFS: Sequential feature selection; SVC: Support vector classifier

ARTICLE HISTORY
Received 21 October 2022
Accepted 31 May 2023

KEYWORDS
Mental workload; prosthesis; classification; machine learning

1. Introduction

Amputee patients experience severe functional disability in activities of daily living (ADLs) due to the lack of prosthetic device usability (Bowker 2004; Montagnani, Controzzi, and Cipriani 2015). More than two million amputees live in the U.S., and about 185,000 amputations occur each year. This number is expected to be doubled by 2050 due to the increasing rate of contributing diseases (Amputee Coalition 2021). Amputees use prosthetic devices regularly to perform ADLs. These activities may not be possible without prosthetic devices or require additional effort and time (Gaskins et al. 2018; Lusardi, Jorge, and Nielsen 2013). However, the devices are often reported to be

CONTACT Maryam Zahabi mzahabi@tamu.edu Wm Michael Barnes '64 Department of Industrial & Systems Engineering, Texas A&M University, Emerging Technologies Building, College Station, TX 77843-3131, USA

© 2023 Informa UK Limited, trading as Taylor & Francis Group

challenging to use, which can lead to reduced utilisation and device rejection (Engdahl et al. 2015). About 53% of passive hand users, 50% of body-powered hook users, and 39% of myoelectric hand users reject prosthetic arms (Kannenberg and Zacharias 2014). The main reasons for rejection were poor dexterity, glove durability, and lack of sensory feedback (Biddiss, Beaton, and Chau 2007; Bowker 2004; Montagnani, Controzzi, and Cipriani 2015).

Using prostheses requires substantial cognitive resources (Geurts and Mulder 1994; Geurts et al. 1991; Heller, Datta, and Howitt 2000; Hofstad et al. 2009; Williams et al. 2006). Cognitive resources are used to compensate for the loss of motor control and mitigate the damage of somatosensory feedback from the amputated limb (Childress 1980; Heller, Datta, and Howitt 2000; Herberts and Körner 1979; Krewer et al. 2007; Williams et al. 2006; Witteveen et al. 2012). Therefore, using prostheses can cause a lack of cognitive capacity to perform other mental activities (Heller, Datta, and Howitt 2000; Williams et al. 2006). High mental workload can also reduce the primary task performance (Duysens et al. 2012). In case of upper limb amputation, most of the current control strategies use limited information (i.e. shoulder movements or recorded electromyography (EMG) signals) for activating several degrees of freedom (DOF) of the prosthetic devices, which is non-intuitive and unnatural, and can result in high cognitive workload (CW) (Cordella et al. 2016). Therefore, it is essential to assess CW of prosthetic devices early in the design and development process to improve device usability.

1.1. Cognitive workload classification

Using machine learning (ML) algorithms for the classification of CW has several advantages as compared to inferential statistics. First, ML algorithms can deal with the ambiguity and uncertainty associated with non-linear factors, which is not possible with inferential statistics (Moustafa, Luz, and Longo 2017). For example, multivariate analysis of variance (MANOVA) is still limited to model multifaced nature of CW (Matthews et al. 2015; Wickens 2017). ML algorithms can also be used to find relationships in high dimensional spaces compared to statistical modelling (Hillege et al. 2020). Finally, ML approaches can be used to predict CW of using prosthetic devices in real-time (Braarud et al. 2021). With the recent development in experimental devices and ML techniques, it is possible to design an interface to adapt in real-time based on the classified CW level (Zahabi, Wang, and Shahrampour 2021).

A recent review of literature found four types of CW measures in prosthetic device studies including physiological measures [e.g. electroencephalogram (EEG), heart rate variability, pupil diameter change], subjective measures [e.g. NASA-Task Load (NASA-TLX) Index], performance measures, and cognitive performance model (CPM) outcomes (e.g. number of cognitive operators) (Park and Zahabi 2022). Although physiological, subjective, and performance measures were used more frequently in previous studies as compared to CPM measures, Zahabi et al. (2019) found that cognitive models, such as the Goals, Operations, Methods, and Selection Rules (GOMS) method can be used to assess cognitive workload and usability of using upper-limb prosthetic devices.

Several ML algorithms have been used to classify CW in different domains. The most frequently used methods were support vector classifier (SVC) (Meyer 2017), random forest (RF) (Liaw and Wiener 2002), and Naïve Bayes (NB) (Majka 2018). Furthermore, the SVC (Ghaderyan and Abbasi 2016; Pettersson et al. 2020) and RF algorithms were found to have the highest classification accuracy in prior studies (Badarna et al. 2018; Skaramagkas et al. 2021). Neural network (Sharma et al. 2021; Walambe et al. 2021) and NB (Nourbakhsh, Wang, and Chen 2013b; Raufi 2019b) algorithms have been used in prior studies with large datasets. While a majority of studies used physiological measurements as input features (e.g. heart rate, pupil diameter, respiration rate, brain signal or skin conductance) (Meteier et al. 2021; Momeni et al. 2019; Wang et al. 2013), some investigations used task performance outcomes (e.g. response time in mental arithmetic task) as input features for classifying CW (Appel et al. 2019; Li et al. 2020). Although several optimisation methods exist including the ensemble method (Skaramagkas et al. 2021), recursive feature elimination (RFE) (Raufi 2019b), k-fold cross validation (CV) (Momeni et al. 2019), grid search (Mock et al. 2016) and random search (Skaramagkas et al. 2021), previous studies did not clearly state their feature selection methods (Hillege et al. 2020; Li et al. 2020) or hyperparameter tuning (Momeni et al. 2019; Wang et al. 2013), which are essential for optimising the ML algorithms.

1.2. Research gaps and objective

There has not been any investigation on the classification of CW for prosthetic devices, although high CW is one of the major challenges with existing prosthetic devices. In addition, although several measures, such as physiological responses and subjective measures have been used as input features in CW classification algorithms, no study used CPM outcomes as input features to classify CW. CPM models can be generated by

observation of different tasks and using knowledge elicitation approaches with small sample sizes, do not require extensive human-subject experiments, and therefore can be used in the early stages of the design cycle. Lastly, a majority of previous studies skipped optimisation processes in developing their models. Therefore, the objective of this study was to classify the level of CW in using EMG-based prosthetic devices based on various features, metrics, and tasks.

In our preliminary study (Park et al. 2022), we explored a subset of features using an example of ADL (i.e. clothespin relocation test; CRT). We classified cognitive workload into two classes (low or high) and compared the accuracy of models to investigate the feasibility of CW estimation in this domain. In this study, we explored a more comprehensive list of features, evaluated the algorithms using two tasks (CRT and Southampton Hand Assessment Protocol—SHAP) with a larger dataset, explored more detailed classes of workload, and used five metrices [accuracy, area under the receiver operation characteristic curve (AUC), Precision, Recall, and F1 score] to evaluate the model outcomes.

2. Methods

2.1. Participants

Thirty able-bodied participants (18 males and 12 females) were recruited for this study (Age: $M = 22.9$ years; $SD = 2.8$ years). All participants had 20/20 or corrected vision with no prior experience using a prosthetic arm or a myoelectric exoskeleton for upper limbs. Participants' dexterity level was assessed using the Purdue Pegboard Test (PPT) (Tiffin and Asher 1948). The experiment protocol was approved by the Institutional Review Board at the University of North Carolina at Chapel Hill. The number of participants was determined based on pilot tests. Pilot tests were conducted with six participants (randomly assigned two participants to each of the device configuration). Effect sizes (d) were calculated for each of the dependent variables. Among all the calculated effect sizes, the minimum effect size (0.6) (associated with the percent change of pupil size (PCPS) responses) was selected to have the most conservative approach, which is between the medium and large effect size of Cohen's d (Cohen 1988). With this effect size and including other parameters, such as alpha ($\alpha = .05$) and power ($1-\beta = .8$), the sample size for each device configuration was calculated as 9.43 using the 'pwr' package in R 4.2.2 (Champely et al. 2018; Sakai and Sakai 2018). Finally, we round up this value to 10 and recruited in a total of 30 participants (i.e. 10 participants per each device configuration). This sample size was larger than the average number of participants used in prior studies assessing the cognitive workload of prosthetic devices with able-bodied subjects (i.e. $M = 13.46$, $SD = 6.49$) (Park and Zahabi 2022).

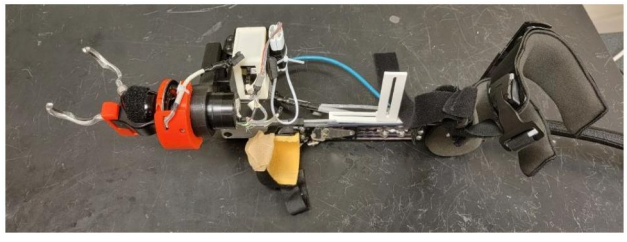

Figure 1. The prosthetic device was used for the experiment.

2.2. Experiment setup

A commercial prosthetic device (Motion Control ETD, Filauer) with 2-DOF of actuation in hand open/close and wrist pronation/supination was used to test three control schemes: direct control (DC), pattern recognition (PR), or continuous control (CC). A custom prosthetic hand adapter was designed and fabricated as a bypass device, as shown in Figure 1. The weight of the device was 4.54 lb.

For the DC control scheme, muscle activation levels were estimated with the EMG signals from two channels [hand close/wrist pronation for the flexor carpi radialis (FCR) channel; hand open/wrist supination for the extensor carpi radialis longus (ECRL) channel; Figure 2] based on the mean absolute value (MAV)

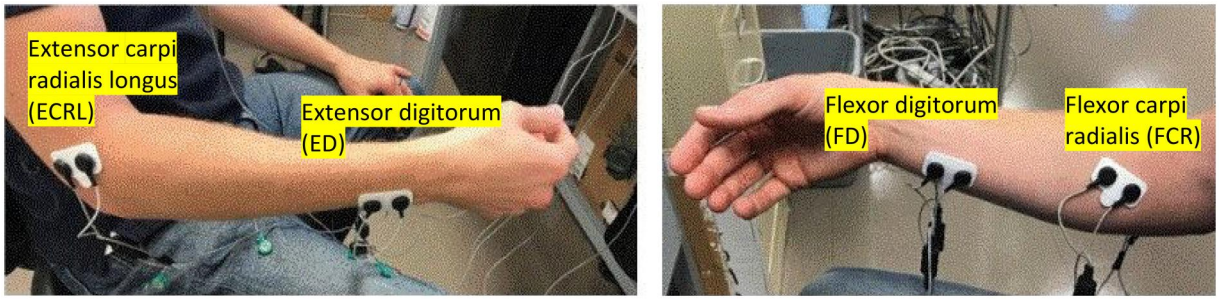

Figure 2. EMG sensor placement.

from each channel (Resnik et al. 2018; White et al. 2017). Participants could control only one DOF or mode [either rotation mode (wrist pronation/supination) or open/close mode (hand flexion/extension)] at a time. If one of the MAV from two channels exceeded its preconfigured threshold, the motor for the currently active DOF would move in the corresponding mode (hand close/wrist pronation for FCR channel; hand open/wrist supination for ECRL channel) at speed proportional to the magnitude of the EMG signal. If the thresholds of both channels were exceeded *via* co-contraction of forearm muscles by the participants (i.e. power grip), the active mode was switched. The experimenter manually adjusted thresholds and proportional control gains for each channel based on feedback from the participants.

For training of the PR configuration, participants performed five hand gestures (labelled as hand close, hand open, wrist pronation, wrist supination, and no movement). Each movement class was held for 4 s and was followed by a 5 s rest period. All movement classes were performed twice. EMG data were simultaneously collected and labelled with the current movement class. Four commonly used time domain features (MAV, number of zero crossings, waveform length, and number of slope sign changes) were extracted from EMG signals following the methods used in prior studies (Resnik et al. 2018; White et al. 2017). The features and labels were used to train a Linear discriminant analysis (LDA)-based classifier to predict one of the five-movement classes from the input features. The speed was set proportional to the sum of the magnitudes of the four EMG signals. During the calibration, the experimenter manually adjusted control gains based on the classification performance and feedback from the participants.

In the CC control scheme, EMG data were recorded simultaneously with kinematic data from a Leap Motion Controller (Leap Motion, Inc., USA). The device uses a camera to accurately estimate the positions of segments in the hand and forearm (Butt et al. 2018; Dyshel et al. 2015). Estimates of the positions of the phalangeal, palm, and forearm segments were recorded at 120 Hz and used to calculate wrist pronation/supination and metacarpophalangeal (MCP) flexion/extension joint angles. Muscle activations were estimated from the recorded EMG signals by calculating the MAV using a sliding window incremented in 10 ms steps resulting in 100 Hz input EMG data. The kinematics data were downsampled to 100 Hz to match the EMG data. Training data were collected from participants while they performed three motion types: MCP flexion/extension only, wrist pronation/supination only, and simultaneous wrist and MCP. All motions were performed in a pattern in which participants moved their wrist/MCP between one of the five positions (fully flexed/pronated, relaxed, fully extended/supinated) to a metronome set at a 1 Hz frequency. Three 10s trials were recorded for each motion type to be used for training. An artificial neural network (ANN) algorithm was developed for each participant for both the wrist and MCP using the Deep Learning Toolbox in MATLAB 2018b (Mathworks Inc., USA). The ANNs were trained to map processed EMG signals to joint positions. Velocity was estimated by differentiating the estimated positions.

A Pupil-core eye tracking system (Pupil Labs, Germany) was used to collect pupil data. The system hardware included one world camera and two eye cameras. The eye cameras detected and tracked the pupil with 3-dimensional models. Gaze parameters were gathered in normalised 3D gaze positions and binocular vergence. Eye movements were recorded with .6 degrees accuracy [i.e. the average angular offset (distance) (in degrees of visual angle) between fixations locations and the corresponding locations of the fixation targets], .02 precision [i.e. the root mean square of the angular distance (in degrees of visual angle) between successive samples], and frequency of 200 Hz. The eye-tracking system was calibrated using Apriltag markers. Dismissing rate during the calibration was consistently controlled to be $<$20% based on the criteria defined by the manufacturer (Pupil Labs). The pupil size was calculated by measuring the relative size of eye camera pixels in millimetre unit in the 3D eye model.

2.3. Task

For assessing CW of prosthetic devices, two ADLs were used in this study including the CRT and the SHAP (Figure 3). CRT is a commonly applied ADL for assessing the usability of upper limb prostheses (Stubblefield et al. 2005; Zahabi et al. 2019). SHAP—door handle task was selected as another appropriate testbed based on our previous study as it includes a combination of upper limb movements including shoulder elevation/depression, arm abduction/adduction, arm flexion/extension, arm medial/lateral rotation, forearm flexion/extension, and wrist supination/pronation (Park et al. 2020). The CRT requires participants to move as many pins as possible from the horizontal rod to the vertical rod in 2 min. The SHAP task requires participants to rotate the door handle using a

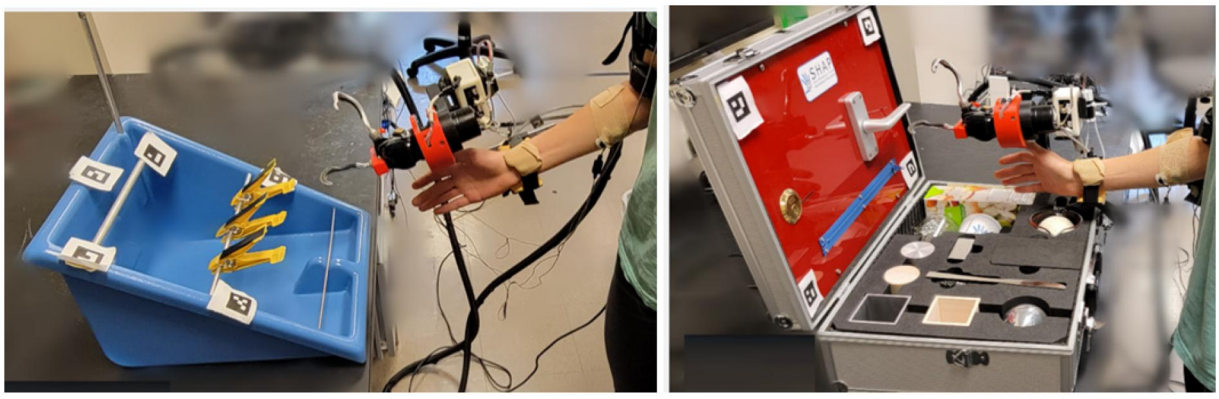

Figure 3. Clothespin relocation task and Southampton Hand Assessment Procedure Task.

power grip until it is fully open, then release the handle as quickly as possible. The SHAP form-board was placed in front of the participant with the blue side facing upward, ~8 cm from the front edge of the table. The door handle task was demonstrated to the participant using slow, precise movements, ensuring that the participant was aware of the proper grip for completing the task. The demonstration was carried out using the corresponding hand under assessment to avoid any confusion for the participant.

2.4. Experiment design and variables

The experiment followed a between-subject design in which each participant was randomly assigned to one of the three prosthetic configurations (i.e. DC, PR, or CC). This approach was selected to reduce learning effects that might occur for participants as a result of working with different prostheses across multiple test trials. Upon being assigned to a specific type of prosthesis, all participants experienced two tasks (i.e. CRT and SHAP door handle tasks), including three trials for each task.

CW of participants was measured while performing the tasks using task performance measures, pupillary measures, CPM outcomes, and perceived workload ratings. Task performance was captured by watching recorded videos of participants performing the tasks and using measures including the numbers of pins moved (for the CRT) and time to rotate door handle five times (for the SHAP task). In addition, the time to complete one cycle of each task was measured in milliseconds [i.e. the best (fastest) task completion time to move one pin from one bar to another bar in CRT or the best (fastest) task completion time to rotate the door handle once in SHAP]. Pupillary measures included the percent change in pupil size (PCPS) and blink rate. PCPS has been used in previous studies to assess CW of prosthetic devices (Zhang et al. 2016). Blink rate has also been frequently used as an indicator of CW (Cardona and Quevedo 2014; Fogarty and Stern 1989; Martins and Carvalho 2015). Blink rate is defined as the number of eye closures in a given period (White et al. 2017). Eye blinks and blink duration decreases as visual workload increases (De Waard and Brookhuis 1996).

For developing CPMs, task analysis was initially conducted for each type of configuration and task. Based on the findings of the task analysis, six Cognitive-Perceptual-Motor GOMS (CPM-GOMS) (John 1990) models were developed in Cogulator (Estes 2017). The models generated outcomes including task completion time for one cycle, number of cognitive, perceptual, and motor operators, and the number of memory chunks. A list of features used in this study is shown in Table 1.

NASA-TLX score was used as a ground truth or target variable to compare with the findings of CW classification algorithms, as this measure has been used extensively in prior studies using prosthetic devices (Connan et al. 2016; Deeny et al. 2014; Markovic et al. 2018). Participants were asked to rate their perceived workload using the NASA-TLX questionnaire after each trial.

2.5. Procedure

Before the experiment, participants signed the informed consent form, an informed consent form addendum for research during the COVID-19 pandemic, and a demographic questionnaire. After the participants signed all documents, they were asked to complete the Edinburgh Handedness Test (Oldfield 1971) and the Purdue Pegboard Test (PPT) (Tiffin and Asher 1948; White et al. 2017). The PPT was conducted three times to determine if they fell within the range of 'normal' manipulative dexterity. Participants were recruited for the experiment if they received a right-hand dominance score of 0.7 or greater based on the Edinburgh Handedness Test and

Table 1. List of input features and their description.

Category	Features	Data type	Description
Device configuration	Device control scheme (i.e. DC, PR, CC)	Categorical	Type of device configuration
Task performance measures	Task performance	Continuous	Number of pins moved in 2 min
	Number of training trials	Discrete	Number of training sessions needed to pass the training criteria
	Tasks completion time of one cycle	Continuous	Best (fastest) task completion time to move one pin from one bar to another bar
Pupillary measures	Percent change in pupil size	Continuous	Percent change in pupil size
	Blink rate	Continuous	Blinks per minute
CPM outcomes	Number of cognitive operators	Discrete	Number of cognitive operators in CPM
	Number of perceptual operators	Discrete	Number of perceptual operators in CPM
	Number of motor operators	Discrete	Number of motor operators in CPM
	Tasks completion time estimate from model	Continuous	Best (Fastest) task completion time to move one pin from one bar to another bar from CPM
	Memory chunks	Continuous	Number of memory chunks

their PPT score was no more than one standard deviation below the normal mean dexterity for their age and gender group (Tiffin and Asher 1948).

Once participants completed the PPT test, they donned the prosthetic adapter, and EMG electrodes were placed on their skin based on the assigned control mode. A verbal description of the prosthesis DOF and control strategy was provided. For participants assigned to the DC group, the prosthesis was activated during the EMG threshold configuration procedure. Participants were allowed to practice controlling the device until they reported comfort with the DC control. Participants then advanced to the formal training period. Participants assigned to the PR group were instructed to perform specific arm motions and to observe the feedback (the classified gestures including hand open, hand closed, wrist pronated, wrist supinated, and relaxed hand and wrist) from the experimenter's laptop screen. Five seconds of rest were allowed between each posture. Participants assigned to the CC group were asked to perform 10s trials three times for each movement type—isolated hand open/close, isolated wrist pronation/supination, and simultaneous movements—at a 0.25 Hz tempo set on a metronome, resulting in 9 total trials. Angles of the metacarpophalangeal joints and the wrist's rotation angle were recorded using a Leap Motion Controller placed ~4″ below the participant's hand at 120 Hz simultaneously with EMG data. The MAV of the EMG was calculated with a 200 ms sliding window adjusted in 10 ms increments, and the joint angle data were down-sampled to 100 Hz to match the EMG data. The processed EMG and motion data were used to train two neural networks for the 2 DOF. Gains for the controller's output and thresholds to reduce small unintentional movements from the user were adjusted using feedback from them. After the classifier was trained, participants were allowed to practice controlling the device until they reported comfort with the control.

Once the participants received training for their assigned control mode, they were trained on the task-specific training, which assessed mastery of device handling and the respective control mode. The training session required participants to use the prosthesis to move three clothespins from a horizontal bar at the base of the workstation to a vertical bar extending upward on the clothespin apparatus. They began with the movement of the rightmost clothespin and, as quickly as possible, completed all pins. An experimenter recorded the time to move the three consecutive clothespins. If participants dropped a clothespin, they were required to restart the trial. A training criterion was established based on pilot test data generated from learning curve analysis, including when participants achieved asymptotic performance with the device and at what level (task time). If the average task completion time of three sequential trials was within 15–25s for the PR, 20–35s for the DC, and 16–23s for the CC mode, the participant passed the training and proceeded to the actual experimental trials. Upon completion of the training trials, the eye-tracking system was calibrated for the participants, and they could begin the actual experiment trials after having 5 min of rest.

Participants were provided instructions on how to complete the two tasks. The order of tasks was randomised to avoid any learning effect from one task to another. For CRT trials, the instruction included moving as many clothespins as possible from the horizontal rod to the vertical rod and back within 2 min. The number of successfully relocated clothespins was recorded at the end of each trial. For SHAP—Door Handle task, participants were instructed to rotate the handle five times as fast as possible. The participant's eyes were tracked throughout each trial. All

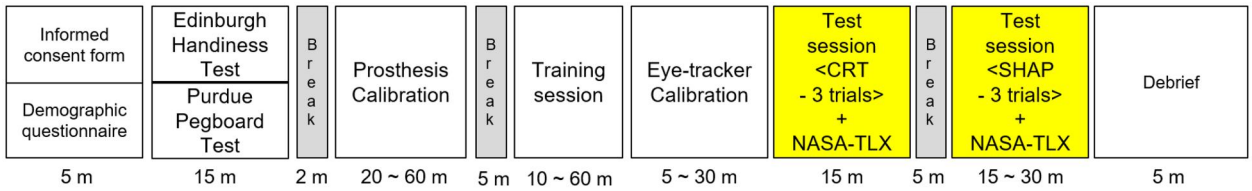

Figure 4. Experiment procedure.

Table 2. Distribution of data points in each class.

Task	Target (number of classes)	Class	Number of data points
CRT	Two	High	54
		Low	36
SHAP	Two	High	50
		Low	40
	Three	High	24
		Moderate	45
		Low	21

participants completed a total of three trials for each task and were provided with a 5-min rest period after each trial. After each actual trial, participants filled out the NASA-TLX questionnaire. Figure 4 illustrates a summary of the experimental procedure.

2.6. Cognitive workload classification

2.6.1. Data labeling

Participants' NASA-TLX scores and weights for each dimension were collected based on the procedure described in Hart and Steveland (1988). We collected the weights before the first trial of the experiment by asking the participants to complete the pairwise comparison rating form. After each trial, participants completed the workload ratings for each dimension based on what they experienced during that trial. Using these weights, the weighted average was calculated for each trial to have a single and overall score of NASA-TLX and then the overall scores were clustered into different classes. Since this target variable [i.e. the overall NASA-TLX score (0–100%)] was a continuous variable, there was a need to group the data into different categories before classification.

A clustering analysis was conducted on all participants' NASA-TLX scores to find the optimal number of classes of CW using the NbClust package in R. There are several clustering analysis approaches, and each algorithm generates different results based on specific indices or methods (e.g. kmeans). We tested all the combinations of clustering methods and indices and found that the most frequent optimal number of classes determined from different methods were two, four, and three clusters, respectively. Although we could simply select the most frequent optimal number of classes (which was having two classes of workload), we decided to include the top three selected classes as having more detailed classification (e.g. low, medium, high workload) would provide more precise estimate of workload. However, due to the lack of sufficient number of data points in some of these classes, only two or three classes of CW were used in our analysis. Table 2 illustrates the distribution of data points for each class.

2.6.2. Algorithm selection

Three algorithms of Random Forest (RF), Support Vector Classifier (SVC), and Naïve Bayes (NB) were selected to classify CW since (1) they were used extensively in recent studies (Braarud et al. 2021; Kaczorowska, Plechawska-Wójcik, and Tokovarov 2021; Meteier et al. 2021; Shao et al. 2021; Sharma et al. 2021; Walambe et al. 2021), (2) included physiological data (e.g. pupillometry) and task performance (e.g. response time on secondary task) measures as their input features, and (3) exhibited high prediction accuracy (>80%) in small datasets (Kaczorowska, Plechawska-Wójcik, and Tokovarov 2021).

2.6.3. Optimisation and validation

Given the small dataset (i.e. 90 datapoints for each task = 10 participants per control scheme × 3 control schemes × 3 trials), overfitting was the major concern for establishing the ML structure. Therefore, we first split our dataset into training (70% of the data) and testing (30% of the data) groups. We randomly partitioned the data from 30 participants into the training and testing datasets (i.e. the data points of one participant only appeared either in training or testing dataset). Then, 10-fold CV was employed to optimise the hyperparameters (Götze, Gürtler, and Witowski 2020b). A hyperparameter grid search method was conducted using the *sklearn* Python library (Pedregosa et al. 2011) and a *Pipeline* function to streamline testing across three different model types (i.e. RF, SVC, and NB). RF has a wide range of applications and is noted to perform well for classification tasks, even with default hyperparameter input (Donges 2021). Of the many configurable inputs to the random forest model, the three most notable and influential variables are the number of trees in the model forest, the maximum tolerable depth of each tree, and

the number of features necessary at each branching point (Probst 2019). Limiting the number and depth of trees reduces overfitting of the data; otherwise, though a model may be ideal for the training data if allowed to infinitely grow, out-of-sample performance would be extremely poor. Considering the number of data points at each branching point in the tree is another means of limiting the shunting of model performance towards narrow-minded behaviour. In preliminary testing, however, the number of features necessary at each breaking point continuously output its default value of 2, and thus it was not considered in the final grid search.

SVC employs a spatial approach to delineating class margins and has a reputation for being computationally expedient in rudimentary modelling. Many studies with similar dataset challenges have employed SVC to classify data efficiently (Braarud et al. 2021; Raihan-Al-Masud and Mondal 2020). In these situations, a linear kernel type was used, specifying which subtype of SVC to employ (Meteier et al. 2021). In doing so, the chief remaining hyperparameter was the regularisation variable ('c' in Table 3). This parameter calculates the amount of tolerable error the algorithm considers before passing a model as output. Like the tree count for random forest, a regularisation constant that is too small could massively overfit the data.

For NB, given our small and unbalanced dataset, a complement NB model was implemented as this method is more appropriate for imbalanced dataset (Rennie et al. 2003). Hyperparameter grid searching was performed only for the 'alpha' parameter with the values contained in Table 3 as it determines the portion of the largest variance of all features that are added to variances for calculating stability (Jain 2021; Rennie et al. 2003). Controlling the degree of smoothness permitted by the model in delineating different classes allowed a balance to be obtained between cross-validated performances in the grid search k-folding.

2.6.4. Feature selection

To make modelling more efficient, feature selection methods were used to eliminate less-contributory features from the training data set. Each of the selection methods attempted to increase testing performance. Therefore, the K-Best method of selection was employed as the representative method of the univariate filter class of selectors (Aggarwal 2018). For more multivariate methods, the recursive feature selection (RFE) and forward feature selection methods were employed (Ferreira and Figueiredo 2012; Raihan-Al-Masud and Mondal 2020). RFE considers multivariate feature contribution as a whole and iteratively eliminates the least contributory features until the desired count is obtained (Guyon et al. 2002). Sequential forward selection (SFS) adds features by order of significance until the number of features is obtained. RFE and SFS have demonstrated a decent performance in improving model accuracy and efficiency in prior studies (Ferreira and Figueiredo 2012). Each of the three algorithms was employed for each model type and was executed and tested for specified feature counts 1–13 (i.e. the total number of features in the dataset).

2.6.5. Model evaluation

The validation and test datasets were used for model evaluation. Cross-validation score, classification accuracy, area under the receiver operating characteristic curve (AUC), precision, recall, and F1-score were calculated as measures of model performance (Ding et al. 2020; Skaramagkas et al. 2021). Cross-validation is a technique for evaluating ML models with split datasets (Hastie et al. 2009; Kuhn 2008). Accuracy is the ratio of correctly classified samples. F1-score is the harmonic

Table 3. Classifiers and hyperparameters.

Classifier	Hyperparameter	Definition	Range	References
RF	n_estimators	Number of trees in the forest	[start: 100, end: 1000, step size: 100]	Götze, Gürtler, and Witowski 2020a, 2020b
	max_depth	Maximum number of layers of decisions tolerated	[1, 13, 1]	Mullainathan and Spiess 2017; Nadi and Moradi 2019
	min_samples_split	Number of samples necessary to be present in the creation of a branching point in the tree (default: 2)	Fixed as default value 2	Götze, Gürtler, and Witowski 2020a, 2020b
SVC	c	Regularisation parameter—i.e. how much error tolerable in producing model	[0.5, 0.6, 0.7, 0.8, 0.9, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10]	Raihan-Al-Masud and Mondal 2020
	Kernel	Specifies which kernel to use in the program	Fixed as linear	Braarud et al. 2021; Meteier et al. 2021
Naive-Bayes	Alpha	Additive (Laplace/Lidstone) smoothing parameter	20 points from [1, 10] spaced evenly in log-space	Jain 2021; Rennie et al. 2003

mean of recall (i.e. probability of detecting each class) and precision (i.e. reliability of results in each class). The F1-score was obtained by calculating recall and precision separately for each class and averaging them, weighted by the number of samples in each class. We used F1, recall, and precision because they are useful metrics for both balanced and imbalanced datasets, while accuracy is usually a good metric for a balanced dataset (Jeni, Cohn, and De La Torre 2013). In addition, the computation time for grid search was calculated (Intel® Core i7-8700 @ 3.20 GHz). We calculated grid search time because grid search was the most demanding and the dataset was extremely small. To improve the reliability and generalisability of ML results, we ran each of the models with 15 random seeds per suggestion from Colas, Sigaud, and Oudeyer (2019) and calculated the average prediction performance. In every run, 21 participants (70% of the total dataset) were randomly partitioned into training/validation sets.

3. Results

The average performance of each algorithm over 15 runs with random seeds is presented in Table 4. Considering AUC, for the CRT, the best model was NB with two classes, and it resulted in 0.78 of AUC (model No. 9 in Table 4). The model exhibited decent performance across other metrics including test accuracy, cross-validation score, precision, recall, and F1 score (Grandini, Bagli, and Visani 2020; Sokolova and Lapalme 2009). For the SHAP, the RF model with two classes (model No. 10) exhibited the best performance (AUC: 0.74). Although the AUC of model 12 is the highest among all SHAP results, the model selected only one feature (task completion time for one cycle), which is the indication of underfitting. The RF and NB algorithms exhibited accuracies above random guessing (i.e. 0.5) in all 15 runs for both CRT and SHAP tasks when the target variable (i.e. CW) was classified in two clusters. Meanwhile, the SVC algorithm performed better (i.e. higher accuracy) than random guessing in 60% of all runs. Regarding the AUC, the RF and NB algorithms had better performance than random guessing in 80 and 73% of all runs, respectively, while the SVC algorithm had higher AUC than random guessing only in 40% of times.

If we consider AUC as the criterion to determine the best model, the important features to classify CW were PCPS, task performance, number of cognitive operators, and device configuration (i.e. model No. 9 in Table 4). For the SHAP, task performance, task completion time for one cycle, number of training trials, and number of cognitive operators were selected in the best model (i.e. model 10). Considering the test accuracy, the important (or selected) features in the best CRT model (i.e. model No. 7) included: blink rate, TCT of one cycle, number of cognitive operators,

Table 4. Summary of average classification performance by taking different classes as targets.

Task	No.	Target	Classifier	Feature selector	CV score	Test accuracy	AUC	Precision	Recall	F1-Score
CRT	1	Two classes	RF	K-Best	0.94	0.72	0.65	0.70	0.69	0.68
	2			RFE	0.89	0.71	0.51	0.58	0.61	0.59
	3			SFS	0.92	0.57	0.61	0.62	0.60	0.57
	4		SVC	K-Best	0.57	0.65	0.49	0.44	0.56	0.48
	5			RFE	0.71	0.45	0.45	0.49	0.50	0.41
	6			SFS	0.54	0.74	0.63	0.72	0.78	0.72
	7		NB	K-Best	0.64	0.83	0.73	0.83	0.81	0.82
	8			RFE	0.57	0.72	0.61	0.58	0.63	0.58
	9			SFS (best)	0.64	0.80	0.78	0.79	0.81	0.78
SHAP	10	Two classes	RF	K-Best (best)	0.86	0.69	0.74	0.72	0.71	0.67
	11			RFE	0.82	0.72	0.72	0.66	0.69	0.67
	12			SFS	0.76	0.74	0.81	0.76	0.77	0.74
	13		SVC	K-Best	0.54	0.63	0.51	0.52	0.57	0.49
	14			RFE	0.51	0.57	0.45	0.54	0.60	0.52
	15			SFS	0.44	0.58	0.31	0.58	0.57	0.54
	16		NB	K-Best	0.63	0.67	0.53	0.79	0.70	0.65
	17			RFE	0.49	0.61	0.59	0.52	0.57	0.51
	18			SFS	0.54	0.63	0.68	0.65	0.67	0.62
	19	Three classes	RF	K-Best	0.72	0.52	0.50	0.47	0.51	0.44
	20			RFE	0.72	0.56	0.60	0.39	0.38	0.38
	21			SFS	0.45	0.53	0.38	0.18	0.33	0.23
	22		SVC	K-Best	0.55	0.53	0.46	0.24	0.40	0.29
	23			RFE	0.45	0.56	0.52	0.38	0.41	0.35
	24			SFS	0.49	0.67	0.38	0.40	0.54	0.46
	25		NB	K-Best	0.46	0.43	0.45	0.50	0.39	0.41
	26			RFE	0.42	0.48	0.58	0.41	0.44	0.39
	27			SFS	0.44	0.54	0.49	0.44	0.46	0.43

CV: cross-validation; AUC: area under curve; CRT: Clothespin Relocation Test; SHAP: Southampton Hand Assessment Procedure; RF: random forest; SVC: support vector classifier; NB: Naïve Bayes; RFE: recursive feature elimination; SFS: sequential forward selection.

Table 5. Hyperparameter values for random forest models.

Task	No.	Target	Feature selector	n_estimators	max_depth
CRT	1	Two classes	K-Best	389	8
	2		RFE	175	6
	3		SFS	150	5
SHAP	10	Two classes	K-Best	278	7
	11		RFE	460	7
	12		SFS	200	5
	19	Three classes	K-Best	700	10
	20		RFE	100	8
	21		SFS	125	5

CV: cross-validation; AUC: area under curve; CRT: Clothespin Relocation Test; SHAP: Southampton Hand Assessment Procedure; RFE: recursive feature elimination; SFS: sequential forward selection.

Table 6. Hyperparameter values of models with support vector classifier.

Task	No.	Target	Feature selector	Regularisation parameter (c)
CRT	4	Two classes	K-Best	0.88
	5		RFE	0.83
	6		SFS	0.50
SHAP	13	Two classes	K-Best	1.89
	14		RFE	0.50
	15		SFS	1.00
	22	Three classes	K-Best	4.25
	23		RFE	0.71
	24		SFS	0.50

CV: cross-validation; AUC: area under curve; CRT: Clothespin Relocation Test; SHAP: Southampton Hand Assessment Procedure; RFE: recursive feature elimination; SFS: sequential forward selection.

Table 7. Hyperparameter values of Naïve Bayes models.

Task	No.	Target	Feature selector	Alpha
CRT	7	Two classes	K-Best	11.29
	8		RFE	10.48
	9		SFS	1.00
SHAP	16	Two classes	K-Best	43.50
	17		RFE	11.89
	18		SFS	127.43
	25	Three classes	K-Best	127.43
	26		RFE	11.17
	27		SFS	5.12

CV: cross-validation; AUC: area under curve; CRT: Clothespin Relocation Test; SHAP: Southampton Hand Assessment Procedure; RFE: recursive feature elimination; SFS: sequential forward selection.

number of training trials, and the device configuration. The important features in the best SHAP model (i.e. model 10) included the task completion time for one cycle, number of training trials, and the device configuration. Hyperparameters values of all models including the best models are shown in Tables 5–7.

CV scores of the RF algorithm were higher as compared to other algorithms. This could be because RF is an ensemble learning method that combines multiple decision trees to make more accurate predictions (Hastie et al. 2009; Kuhn 2008). Furthermore, it can handle both numerical and categorical data, and can deal with the missing data and outliers more effectively than some other algorithms (Kuhn 2008). However, AUC of the RF algorithm was worse than that of other algorithms, which means the model is not able to distinguish between positive and negative samples with high accuracy, which may indicate that it is not good at discriminating between the two classes (Gareth et al. 2013).

Table 8. Grid search time (seconds).

		Feature selector		
Classifier	Target	RFE	K-Best	SFS
RF	Two	4092.6	1282.2	21,767.4
	Three	5107.2	2023.2	13,851.6
SVC	Two	25.2	22.8	747.6
	Three	33.6	35.4	428.4
NB	Two	23.4	26.4	582
	Three	20.4	19.8	303.6

Regarding the target variable, in general, classifying the NASA-TLX scores into smaller number of classes led to better algorithm performance than having larger number of classes under the clustering algorithms (i.e. algorithms in *NbClust* package).

The grid search time for every combination of classifiers, targets, and feature selectors suggested that the SVC and NB algorithms outperformed the RF in terms of computational cost (Table 8). Both SVC and NB performed within a few seconds. Among the three feature selectors, SFS exhibited significantly longer computational time as compared to other two selectors.

Therefore, considering all the metrics and computational costs, the NB algorithm with two classes was selected as the best model for CRT (model No. 9 in Table 4). For the SHAP task, the RF algorithm performed better than other algorithms although its computational time was much longer than other methods.

The best models (models 9 and 10) were released to Github (https://github.com/hsilab/pros_cw). Users can download the uploaded files and estimate CW based on the instructions in the readme file. However, it is important to note that the models were generated based on the performance of able-bodied participants and with two ADL testbeds which might limit the generalisability of the models to other applications.

4. Discussion

4.1. Classification performance

The findings suggested that CW of using prosthetic devices can be classified with reasonable accuracy and low computational cost. This study is the first investigation that included CPM outcomes as input features in ML algorithms. Some CPM outcomes (i.e. number of cognitive operators) and task performance features were included in the best models. This can suggest the

possibility of predicting CW of prosthetic devices without conducting human-subject experiments because task performance can also be modelled from the CPM outcomes. Some CPM outcomes, such as the number of perceptual operators were not selected in the best models. This might be because the perceptual operators only appeared in the DC control scheme. In PR and CC configurations, there were no perceptual operators in the outcome of cognitive models because all perceptual operators were in parallel with cognitive or motor operators. However, if the task is more complex or with other prosthetic device configurations, more CPM outcomes might be included as important features in the algorithm. There are several advantages of using CPM over human-subject experiments. For example, the analyst can conduct CPM in the early design process. It is a faster and safer approach than the experimental approach as it can minimises human participant's involvement. It can also quantify and predict human behaviour in natural tasks with simple tools, such as Cogulator (Estes 2017) or CogTool (John and Suzuki 2009) based on human information processing theory. Lastly, CPM can also generate task performance-related features without the need of conducting human-subject experiments and by using the results of task analysis and operator times from the literature (Estes 2017).

This study suggested that multiple metrics should be considered to evaluate the ML algorithms and find the best model(s). For example, although the accuracy of some models was above 70% (e.g. model No. 1 in Table 4), their AUC was relatively low (e.g. 0.65). Precision and Recall were also helpful to test the robustness of ML algorithms and to avoid 'accuracy paradox' (due to unbalanced classes) (Afonja 2017; Valverde-Albacete, Carrillo-de-Albornoz, and Peláez-Moreno 2013). For example, model 24 exhibited reasonable accuracy (0.67) among other algorithms for the SHAP task. However, its recall percentage was low (around 0.5), which implies that those models are not useful for classifying CW when the target variable is not well-balanced. Considering only precision or recall scores individually is also not sufficient for evaluating ML algorithms. For example, we can have a recall score of 100% even though the accuracy of the model is low. In this case, precision will be close to 0. Thus, F1-score should be used to reflect the imbalance between precision and recall because it is a harmonic average between these two measures.

The results also revealed that task performance measures were more promising in predicting CW as compared to other features that were collected from the experiment. This finding is in line with the results of prior studies that found primary task measures as a key indicator of CW for prosthetic devices. Wood and Parr (2022) recently developed a questionnaire for measuring CW of prostheses as an extension of NASA-TLX, which is called prosthesis task load index (PROS-TLX). While validating their questionnaire, the authors used task performance as an indicator of CW as there was a high correlation between the task performance and the evaluated scores on PROS-TLX. Deeny et al. (2014) also found high positive correlation under the complicated task condition between the task performance and the self-report workload score. Task performance measures have advantages in that they evaluate participants' performance on the task of interest directly. However, these measures often lack scientific rigour, making interpretation of the results difficult as unknown or uncontrolled factors may affect results rather than the intended manipulations in the study (Park and Zahabi 2022; Wilson and Schlegel 2004; Wood and Parr 2022). Therefore, some studies suggested using physiological measures of workload instead (Cain 2007). We found that pupillometry measures were selected as important features in the models. The results support the findings of previous studies that used eye-tracking data for measuring CW of prosthetic devices (White et al. 2017; Zahabi et al. 2019; Zhang et al. 2016). Eye-tracking measures have been widely applied to other domains to measure CW of operators, such as simulations for emergency responders (Appel et al. 2019), construction (Li et al. 2020), and foetal ultrasound examination (Sharma et al. 2021).

It was also found that the models with two classes performed better than models with three classes. This is intuitive from a general classification stance since two classes are simpler than several classes to be classified as it has only one threshold. This is in line with previous studies that found smaller number of labels led to high classification accuracy (Nourbakhsh, Wang, and Chen 2013b; Wang et al. 2013).

Although the sample size was small, the NB algorithm exhibited reasonable average performance across multiple runs, which is in line with prior studies that found NB was more accurate than the SVM algorithm in classifying CW (Nourbakhsh, Wang, and Chen 2013a; Raufi 2019a). There are several advantages of NB that resulted in classification accuracy above 70%. First, NB can compensate for class imbalance (Murphy 2006). Second, NB can perform well with small datasets (Huang and Li 2011) and it is a fast and

computationally effective approach (Jadhav and Channe 2016; McCallum and Nigam 1998).

The RF model did not perform as well as NB and some of the models had overfitting issues, which was mainly due to the detailed hyperparameter tuning on an extremely small dataset. Prior studies found that with small and imbalanced datasets, RF could generate either poor results due to a lack of diversity in the dataset or might cause overfitting (Tang, Garreau, and von Luxburg 2018). SVC also performs poorly when the dataset is imbalanced. This is mainly due to the weakness of the soft margin optimisation (Batuwita and Palade 2013) that allows SVC to make a certain number of mistakes and keep margin as wide as possible so that other points can still be classified correctly. This could result in the hyperplanes being skewed to the minority class when imbalanced data is used for training. The second reason is related to the issue of an imbalanced support vector ratio. That is, the ratio between the positive and negative support vectors becomes imbalanced and as a result, datapoints at the decision boundaries of the hyperplanes have a higher chance of being classified as negative. The major reason why RF generated longer computational time is that it included more hyperparameters, especially the number of trees in the forest and their levels, than the other two algorithms. Basically, training time complexity of RF is faster than SVC (Kumar 2019). However, RF took much longer time than SVC due to the burden of hyperparameter tuning. In addition, the main limitation of RF is that a large number of trees can make the algorithm too slow and ineffective for real-time predictions (Donges 2021). SFS demanded extensive computational time because it is a wrapper method that needs to train the classifier for each feature subset, and therefore the method can be impractical.

The findings suggested two ML algorithms (RF or NB) for the classification of CW for prostheses. Our intention was not to propose one specific algorithm or feature selector which should be used for all types of tasks mainly because depending on the characteristics of the dataset, several factors can affect the algorithm performance, including size and quality of the dataset, complexity of the models, and potential biases in the dataset (Dietterich 2000; Goodfellow, Bengio, and Courville 2016; Murphy 2012). We suggest researchers to use the findings of this study as a starting point in estimating CW of prosthetic devices and explore other models depending on the characteristics of their dataset.

4.2. Practical implications

There are several merits of having a model to estimate CW of upper limb prostheses for clinicians, device designers, and other researchers in the ergonomics field. Clinicians can test CW of prostheses before recommending them to amputees. For instance, using the model, clinicians can estimate CW of a specific prosthesis by measuring PCPS, training, and task performance, and adding specific features of the device configuration (e.g. control algorithm), as they are the most important features of the best models. The model can also be used by device developers or designers to estimate CW before developing the physical prototype. Whenever they design a novel control scheme, they can test it with this model to see any improvement in CW.

By estimating the CW of prostheses in advance, the model could contribute to ergonomically-designed upper limb prostheses. If the model can accurately predict CW, it could be used to identify situations where users are at risk of experiencing mental fatigue or injury due to excessive CW. This information could be used to modify prostheses or to provide users with training on how to manage their CW more effectively (Hudgins, Parker, and Scott 1993; Kaczmarek et al. 1991). By reducing cognitive workload and fatigue, users of upper limb prostheses may be able to work more efficiently and effectively and use these devices for ADLs, which could lead to increased productivity (Biddiss and Chau 2007; Oskoei and Hu 2008) and better quality of life and greater independence (Biddiss and Chau 2007). Lastly, developing an ML model to predict CW for upper limb prostheses can advance the research and innovation in this area, which has previously relied only on human subject experiments. We have released the codes for the best models on Github (https://github.com/hsilab/pros_cw) so that other researchers can use or update the model based on their application.

4.3. Limitations and future work

The first limitation of this study was the small dataset that was used for training the models. Future studies with larger datasets are necessary to validate the findings of this investigation. Second, the models were generated based on the performance of able-bodied participants. The decision to work with an able-bodied population was made due to the limited number of trans-radial amputees in the surrounding area. In addition, since most patients currently use devices with DC modes (commonly used in myoelectric control),

recruiting such patients could have produced a bias in their performance. Therefore, there is a need for further investigation with amputees, as an actual user population, to validate the models.

5. Conclusion

This study classified CW of prostheses considering different features, evaluation metrics, and tasks. The findings suggested that the NB and RF algorithms are most promising for classifying CW into two classes (high *vs.* low). It was found that including some of the CPM outcomes in the model could improve the algorithm performance. The proposed algorithms can help manufacturers/clinicians predict CW of future prosthetic devices in the early design phases.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This project was funded by the National Science Foundation (No. IIS-1856676/1856441/1900044). The opinions expressed in this report are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Afonja, T. 2017. "Accuracy Paradox." *Towards Data Science*.

Aggarwal, C.C. 2018. *Machine Learning for Text*. Vol. 848. Springer.

Amputee Coalition 2021. "Limb Loss & Limb Difference in the U.S."

Appel, T., N. Sevcenko, F. Wortha, K. Tsarava, K. Moeller, M. Ninaus, E. Kasneci, P. Gerjets, and M. Assoc Comp. 2019. "Predicting Cognitive Load in an Emergency Simulation Based on Behavioral and Physiological Measures." Paper presented at the 21st ACM international conference on multimodal interaction (ICMI), Suzhou, October 14–18. doi: 10.1145/3340555.3353735.

Badarna, M., I. Shimshoni, G. Luria, and S. Rosenblum. 2018. "The Importance of Pen Motion Pattern Groups for Semi-Automatic Classification of Handwriting into Mental Workload Classes." *Cognitive Computation* 10 (2): 215–227. doi:10.1007/s12559-017-9520-2.

Batuwita, R., and V. Palade. 2013. "Class Imbalance Learning Methods for Support Vector Machines." In *Imbalanced Learning: Foundations, Algorithms, and Applications*, 83–99. John Wiley & Sons, Inc.

Biddiss, E., D. Beaton, and T. Chau. 2007. "Consumer Design Priorities for Upper Limb Prosthetics." *Disability and Rehabilitation. Assistive Technology* 2 (6): 346–357. doi:10.1080/17483100701714733.

Biddiss, E.A., and T.T. Chau. 2007. "Upper Limb Prosthesis Use and Abandonment: A Survey of the Last 25 Years." *Prosthetics and Orthotics International* 31 (3): 236–257. doi: 10.1080/03093640600994581.

Bowker, J. 2004. "The Art of Prosthesis Prescription." In *American Academy of Orthopaedic Surgeons. Atlas of Amputations and Limb Deficienciens Surgical, Prosthetic and Rehabilitation Principles*. 3rd ed., edited by D.G. Smith, J.W. Michael, and J.H. Bowker, 739–744. Rosemont, IL: Bone and Joint Decade.

Braarud, P.O., T. Bodal, J.E. Hulsund, M.N. Louka, C. Nihlwing, E. Nystad, H. Svengren, and E. Wingstedt. 2021. "An Investigation of Speech Features, Plant System Alarms, and Operator-System Interaction for the Classification of Operator Cognitive Workload during Dynamic Work." *Human Factors* 63 (5): 736–756. doi:10.1177/0018720820961730.

Butt, A.H., E. Rovini, C. Dolciotti, G. De Petris, P. Bongioanni, M. Carboncini, and F. Cavallo. 2018. "Objective and Automatic Classification of Parkinson Disease with Leap Motion Controller." *BioMedical Engineering OnLine* 17 (1): 1–21. doi:10.1186/s12938-018-0600-7.

Cain, B. 2007. *A Review of the Mental Workload Literature*.

Cardona, G., and N. Quevedo. 2014. "Blinking and Driving: The Influence of Saccades and Cognitive Workload." *Current Eye Research* 39 (3): 239–244. doi:10.3109/02713683.2013.841256.

Champely, S., C. Ekstrom, P. Dalgaard, J. Gill, S. Weibelzahl, A. Anandkumar, C. Ford, R. Volcic, H. De Rosario, and M.H. De Rosario. 2018. "Package 'Pwr'." *R Package Version* 1 (2): 1–22.

Childress, D.S. 1980. "Closed-Loop Control in Prosthetic Systems: historical Perspective." *Annals of Biomedical Engineering* 8 (4–6): 293–303. doi:10.1007/BF02363433.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Erlbaum.

Colas, C., O. Sigaud, and P.-Y. Oudeyer. 2019. "A Hitchhiker's Guide to Statistical Comparisons of Reinforcement Learning Algorithms." *arXiv preprint* arXiv:1904.06979.

Connan, M., E. Ruiz Ramírez, B. Vodermayer, and C. Castellini. 2016. "Assessment of a Wearable Force-and Electromyography Device and Comparison of the Related Signals for Myocontrol." *Frontiers in Neurorobotics* 10: 17. doi:10.3389/fnbot.2016.00017.

Cordella, F., A.L. Ciancio, R. Sacchetti, A. Davalli, A.G. Cutti, E. Guglielmelli, and L. Zollo. 2016. "Literature Review on Needs of Upper Limb Prosthesis Users." *Frontiers in Neuroscience* 10: 209–209. doi:10.3389/fnins.2016.00209.

De Waard, D., and K. Brookhuis. 1996. "The Measurement of Drivers' Mental Workload."

Deeny, S., M. Barstead, C. Chicoine, L. Hargrove, T. Parrish, and A. Jayaraman. 2014. "EEG as an Outcome Measure for Cognitive Workload during Prosthetic Use."

Dietterich, T.G. 2000. "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization." *Machine Learning* 40 (2): 139–157. doi:10.1023/A:1007607513941.

Ding, Y., Y.Q. Cao, V.G. Duffy, Y. Wang, and X.F. Zhang. 2020. "Measurement and Identification of Mental Workload during Simulated Computer Tasks with Multimodal Methods and Machine Learning." *Ergonomics* 63 (7): 896–908. doi: 10.1080/00140139.2020.1759699.

Donges, N. 2021. "Random Forest Algorithm: A Complete Guide." https://builtin.com/data-science/random-forest-algorithm

Duysens, J., Z. Potocanac, J. Hegeman, S. Verschueren, and B.J. McFadyen. 2012. "Split-Second Decisions on a Split Belt: Does Simulated Limping Affect Obstacle Avoidance?" *Experimental Brain Research* 223 (1): 33–42. doi:10.1007/s00221-012-3238-x.

Dyshel, M., D. Arkadir, H. Bergman, and D. Weinshall. 2015. "Quantifying Levodopa-Induced Dyskinesia Using Depth Camera." Paper presented at the proceedings of the IEEE international conference on computer vision workshops.

Engdahl, S.M., B.P. Christie, B. Kelly, A. Davis, C.A. Chestek, and D.H. Gates. 2015. "Surveying the Interest of Individuals with Upper Limb Loss in Novel Prosthetic Control Techniques." *Journal of Neuroengineering and Rehabilitation* 12 (1): 53. doi:10.1186/s12984-015-0044-2.

Estes, S. 2017. *Cogulator*. The MITRE Corporation.

Ferreira, A.J., and M.A. Figueiredo. 2012. "Efficient Feature Selection Filters for High-Dimensional Data." *Pattern Recognition Letters* 33 (13): 1794–1804. doi:10.1016/j.patrec.2012.05.019.

Fogarty, C., and J.A. Stern. 1989. "Eye Movements and Blinks: Their Relationship to Higher Cognitive Processes." *International Journal of Psychophysiology* 8 (1): 35–42. doi:10.1016/0167-8760(89)90017-2.

Gareth, J., W. Daniela, H. Trevor, and T. Robert. 2013. *An Introduction to Statistical Learning: With Applications in R*. Springer.

Gaskins, C., K. Kontson, E.P. Shaw, I.M. Shuggi, M.J. Ayoub, J.C. Rietschel, M.W. Miller, and R. Gentili. 2018. "Mental Workload Assessment during Simulated Upper Extremity Prosthetic Performance." *Archives of Physical Medicine and Rehabilitation* 99 (10): e33. doi:10.1016/j.apmr.2018.07.115.

Geurts, A.C., and T.H. Mulder. 1994. "Attention Demands in Balance Recovery following Lower Limb Amputation." *Journal of Motor Behavior* 26 (2): 162–170. doi:10.1080/00222895.1994.9941670.

Geurts, A.C., T.W. Mulder, B. Nienhuis, and R.A. Rijken. 1991. "Dual-Task Assessment of Reorganization of Postural Control in Persons with Lower Limb Amputation." *Archives of Physical Medicine and Rehabilitation* 72 (13): 1059–1064.

Ghaderyan, P., and A. Abbasi. 2016. "An Efficient Automatic Workload Estimation Method Based on Electrodermal Activity Using Pattern Classifier Combinations." *International Journal of Psychophysiology* 110: 91–101. doi:10.1016/j.ijpsycho.2016.10.013.

Goodfellow, I., Y. Bengio, and A. Courville. 2016. *Deep Learning*. MIT Press.

Götze, T., M. Gürtler, and E. Witowski. 2020a. "How to Deal with Small Data Sets in Machine Learning: An Analysis on the CAT Bond Market." Available at SSRN 3528082.

Götze, T., M. Gürtler, and E. Witowski. 2020b. "Improving CAT Bond Pricing Models via Machine Learning." *Journal of Asset Management* 21 (5): 428–446. doi:10.1057/s41260-020-00167-0.

Grandini, M., E. Bagli, and G. Visani. 2020. "Metrics for Multi-Class Classification: An Overview." *arXiv preprint* arXiv: 2008.05756.

Guyon, I., J. Weston, S. Barnhill, and V. Vapnik. 2002. "Gene Selection for Cancer Classification Using Support Vector Machines." *Machine Learning* 46 (1/3): 389–422. doi:10.1023/A:1012487302797.

Hart, S.G., and L.E. Staveland. 1988. "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research." In *Advances in psychology*. Vol. 52, pp. 139–183. North-Holland.

Hastie, T., R. Tibshirani, J.H. Friedman, and J.H. Friedman. 2009. *The Elements of Statistical Learning: data Mining, Inference, and Prediction*. Vol. 2. Springer.

Heller, B.W., D. Datta, and J. Howitt. 2000. "A Pilot Study Comparing the Cognitive Demand of Walking for Transfemoral Amputees Using the Intelligent Prosthesis with That Using Conventionally Damped Knees." *Clinical Rehabilitation* 14 (5): 518–522. doi:10.1191/0269215500cr345oa.

Herberts, P., and L. Körner. 1979. "Ideas on Sensory Feedback in Hand Prostheses." *Prosthetics and Orthotics International* 3 (3): 157–162. doi:10.3109/03093647909103104.

Hillege, R.H.L., J.C. Lo, C.P. Janssen, and N. Romeijn. 2020. "The Mental Machine: Classifying Mental Workload State from Unobtrusive Heart Rate-Measures using Machine Learning." Paper presented at the 2nd international conference on adaptive instructional systems, AIS 2020, held as part of the 22nd international conference on human–computer interaction, HCII 2020, Copenhagen, July 19–24.

Hofstad, C.J., V. Weerdesteyn, H. van der Linde, B. Nienhuis, A.C. Geurts, and J. Duysens. 2009. "Evidence for Bilaterally Delayed and Decreased Obstacle Avoidance Responses While Walking with a Lower Limb Prosthesis." *Clinical Neurophysiology* 120 (5): 1009–1015. doi:10.1016/j.clinph.2009.03.003.

Huang, Y., and L. Li. 2011. "Naive Bayes Classification Algorithm Based on Small Sample Set." Paper presented at the 2011 IEEE international conference on cloud computing and intelligence systems.

Hudgins, B., P. Parker, and R.N. Scott. 1993. "A New Strategy for Multifunction Myoelectric Control." *IEEE Transactions on Bio-Medical Engineering* 40 (1): 82–94. doi:10.1109/10.204774.

Jadhav, S.D., and H. Channe. 2016. "Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques." *International Journal of Science and Research* 5 (1): 1842–1845.

Jain, K. 2021. "How to Improve Naive Bayes?" https://medium.com/analytics-vidhya/how-to-improve-naive-bayes-9fa698e14cba

Jeni, L.A., J.F. Cohn, and F. De La Torre. 2013. "Facing Imbalanced Data–Recommendations for the Use of Performance Metrics." Paper presented at the 2013 Humaine Association conference on affective computing and intelligent interaction.

John, B.E. 1990. "Extensions of GOMS Analyses to Expert Performance Requiring Perception of Dynamic Visual and Auditory." Paper presented at the empowering people: CHI'90 conference proceedings [on] human factors in computing systems, Seattle, WA, April 1–5.

John, B.E., and S. Suzuki. 2009. "Toward Cognitive Modeling for Predicting Usability." Paper presented at the 13th international conference on human–computer interaction, HCI International 2009, San Diego, CA, July 19–24.

Kaczmarek, K.A., J.G. Webster, P. Bach-y-Rita, and W.J. Tompkins. 1991. "Electrotactile and Vibrotactile Displays

for Sensory Substitution Systems." *IEEE Transactions on Bio-Medical Engineering* 38 (1): 1–16. doi:10.1109/10.68204.

Kaczorowska, Monika, Małgorzata Plechawska-Wójcik, and Mikhail Tokovarov. 2021. "Interpretable Machine Learning Models for Three-Way Classification of Cognitive Workload Levels for Eye-Tracking Features." *Brain Sciences* 11 (2): 210. doi:10.3390/brainsci11020210.

Kannenberg, A., and B. Zacharias. 2014. "Difficulty of Performing Activities of Daily Living with the Michelangelo Multigrip and Traditional Myoelectric Hands." Paper presented at the American Academy of Orthotists & Prosthetists 40th academy annual meeting & scientific symposium, FPTH14.

Krewer, C., F. Müller, B. Husemann, S. Heller, J. Quintern, and E. Koenig. 2007. "The Influence of Different Lokomat Walking Conditions on the Energy Expenditure of Hemiparetic Patients and Healthy Subjects." *Gait & Posture* 26 (3): 372–377. doi:10.1016/j.gaitpost.2006.10.003.

Kuhn, M. 2008. "Building Predictive Models in R Using the Caret Package." *Journal of Statistical Software* 28 (5): 1–26. doi:10.18637/jss.v028.i05.

Kumar, P. 2019. "Computational Complexity of ML Models." https://medium.com/analytics-vidhya/time-complexity-of-ml-models-4ec39fad2770

Li, J., H. Li, W. Umer, H.W. Wang, X.J. Xing, S.K. Zhao, and J. Hou. 2020. "Identification and Classification of Construction Equipment Operators' Mental Fatigue Using Wearable Eye-Tracking Technology." *Automation in Construction* 109: 103000. doi:10.1016/j.autcon.2019.103000.

Liaw, A., and M. Wiener. 2002. "Classification and Regression by Random Forest." *R News* 2 (3): 18–22. http://CRAN.R-project.org/doc/Rnews

Lusardi, M.M., M. Jorge, and C.C. Nielsen. 2013. *Orthotics and Prosthetics in Rehabilitation-E-Book*. Elsevier Health Sciences.

Majka, M. 2018. *Naive Bayes: High Performance Implementation of the Naive Bayes Algorithm*. R package version 0.9.2.

Markovic, M., M.A. Schweisfurth, L.F. Engels, T. Bentz, D. Wüstefeld, D. Farina, and S. Dosen. 2018. "The Clinical Relevance of Advanced Artificial Feedback in the Control of a Multi-Functional Myoelectric Prosthesis." *Journal of Neuroengineering and Rehabilitation* 15 (1): 28. doi:10.1186/s12984-018-0371-1.

Martins, R., and J. Carvalho. 2015. "Eye Blinking as an Indicator of Fatigue and Mental Load-a Systematic Review." *Occupational Safety and Hygiene III* 10: 231–235.

Matthews, G., L. Reinerman-Jones, R. Wohleber, J. Lin, J. Mercado, and J. Abich. 2015. "Workload Is Multidimensional, Not Unitary: What Now?" Paper presented at the international conference on augmented cognition.

McCallum, A., and K. Nigam. 1998. "A Comparison of Event Models for Naive Bayes Text Classification." Paper presented at the AAAI-98 workshop on learning for text categorization.

Meteier, Q., M. Capallera, S. Ruffieux, L. Angelini, O. Abou Khaled, E. Mugellini, M. Widmer, and A. Sonderegger. 2021. "Classification of Drivers' Workload Using Physiological Signals in Conditional Automation." *Frontiers in Psychology* 12: 18. doi:10.3389/fpsyg.2021.596038.

Meyer, D. 2017. *Support Vector Machines: The Interface to libsvm in Package e1071*. R package version 1.6-8. https://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf.

Mock, P., P. Gerjets, M. Tibus, U. Trautwein, K. Moller, and W. Rosenstiel. 2016. "Using Touchscreen Interaction Data to Predict Cognitive Workload." Paper presented at the 18th ACM international conference on multimodal interaction (ICMI), Tokyo November 12–16. doi:10.1145/2993148.2993202.

Momeni, N., Dell'Agnola, F. Arza, A. Atienza, and D. Ieee. 2019. "Real-Time Cognitive Workload Monitoring Based on Machine Learning Using Physiological Signals in Rescue Missions." Paper presented at the 41st annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, July 23–27.

Montagnani, F., M. Controzzi, and C. Cipriani. 2015. "Is It Finger or Wrist Dexterity That Is Missing in Current Hand Prostheses?" *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 23 (4): 600–609. doi:10.1109/tnsre.2015.2398112.

Moustafa, K., S. Luz, and L. Longo. 2017. "Assessment of Mental Workload: A Comparison of Machine Learning Methods and Subjective Assessment Techniques." Paper presented at the 1st international symposium on human mental workload: models and applications, H-WORKLOAD 2017, Dublin, June 28–30.

Mullainathan, S., and J. Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2): 87–106. doi:10.1257/jep.31.2.87.

Murphy, K.P. 2006. *Naive Bayes Classifiers*. Vol. 18, 1–8. University of British Columbia.

Murphy, K.P. 2012. *Machine Learning: A Probabilistic Perspective*. MIT Press.

Nadi, A., and H. Moradi. 2019. "Increasing the Views and Reducing the Depth in Random Forest." *Expert Systems with Applications* 138: 112801. doi:10.1016/j.eswa.2019.07.018.

Nourbakhsh, N., Y. Wang, and F. Chen. 2013a. "GSR and Blink Features for Cognitive Load Classification." Paper presented at the IFIP conference on human–computer interaction.

Nourbakhsh, N., Y. Wang, and F. Chen. 2013b. "GSR and Blink Features for Cognitive Load Classification." Paper presented at the 14th IFIP TC 13 INTERACT international conference on designing for diversity, Cape Town, September 2–6.

Oldfield, R.C. 1971. "The Assessment and Analysis of Handedness: The Edinburgh Inventory." *Neuropsychologia* 9 (1): 97–113. doi:10.1016/0028-3932(71)90067-4.

Oskoei, M.A., and H. Hu. 2008. "Support Vector Machine-Based Classification Scheme for Myoelectric Control Applied to Upper Limb." *IEEE Transactions on Bio-Medical Engineering* 55 (8): 1956–1965. doi:10.1109/TBME.2008.919734.

Park, J., J. Berman, A. Dodson, Y. Liu, A. Matthew, H. Huang, D. Kaber, J. Ruiz, and M. Zahabi. 2022. "Cognitive Workload Classification of Upper-Limb Prosthetic Devices." Paper presented at the 2022 IEEE 3rd international conference on human–machine systems (ICHMS). doi:10.1109/ICHMS56717.2022.9980676.

Park, J., and M. Zahabi. 2022. "Cognitive Workload Assessment of Prosthetic Devices: A Review of Literature and Meta-Analysis." *IEEE Transactions on Human-Machine Systems* 52 (2): 181–195. doi:10.1109/THMS.2022.3143998.

Park, J., M. Zahabi, D. Kaber, J. Ruiz, and H. Huang. 2020. "Evaluation of Activities of Daily Living Tesbeds for

Assessing Prosthetic Device Usability." Paper presented at the 2020 IEEE international conference on human–machine systems (ICHMS). doi:10.1109/ICHMS49158.2020.9209553.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg. 2011. "Scikit-Learn: Machine Learning in Python." *The Journal of Machine Learning Research* 12: 2825–2830.

Pettersson, K., Tervonen, J. Narvainen , J. Henttonen, P. Maattanen, and I. Mantyjarvi. 2020. "Selecting Feature Sets and Comparing Classification Methods for Cognitive State Estimation." Paper presented at the 20th IEEE international conference on bioinformatics and bioengineering (BIBE), Electr Network, October 26–28.

Probst, P. 2019. *Hyperparameters, Tuning and Meta-Learning for Random Forest and Other Machine Learning Algorithms*. LMU.

Raihan-Al-Masud, M., and M.R.H. Mondal. 2020. "Data-Driven Diagnosis of Spinal Abnormalities Using Feature Selection and Machine Learning Algorithms." *PLoS One* 15 (2): e0228422. doi:10.1371/journal.pone.0228422.

Raufi, B. 2019a. "Hybrid Models of Performance Using Mental Workload and Usability Features via Supervised Machine Learning." Paper presented at the international symposium on human mental workload: models and applications.

Raufi, B. 2019b. "Hybrid Models of Performance Using Mental Workload and Usability Features via Supervised Machine Learning." Paper presented at the 3rd international symposium on human mental workload: models and applications, H-WORKLOAD 2019, Rome, November 14–15.

Rennie, J.D., L. Shih, J. Teevan, and D.R. Karger. 2003. "Tackling the Poor Assumptions of Naive Bayes Text Classifiers." Paper presented at the proceedings of the 20th international conference on machine learning (ICML-03).

Resnik, L., H.H. Huang, A. Winslow, D.L. Crouch, F. Zhang, and N. Wolk. 2018. "Evaluation of EMG Pattern Recognition for Upper Limb Prosthesis Control: A Case Study in Comparison with Direct Myoelectric Control." *Journal of Neuroengineering and Rehabilitation* 15 (1): 23. doi:10.1186/s12984-018-0361-3.

Sakai, T., and T. Sakai. 2018. *Power Analysis Using R. Laboratory Experiments in Information Retrieval: Sample Sizes, Effect Sizes, and Statistical Power*, 133–145.

Shao, S., T. Wang, Y. Li, C. Song, Y. Jiang, and C. Yao. 2021. "Comparison Analysis of Different Time-Scale Heart Rate Variability Signals for Mental Workload Assessment in Human-Robot Interaction." *Wireless Communications and Mobile Computing* 2021: 1–12. doi:10.1155/2021/8371637.

Sharma, H., L. Drukker, A.T. Papageorghiou, and J.A. Noble. 2021. "Machine Learning-Based Analysis of Operator Pupillary Response to Assess Cognitive Workload in Clinical Ultrasound Imaging." *Computers in Biology and Medicine* 135: 104589. doi:10.1016/j.compbiomed.2021.104589.

Skaramagkas, V., E. Ktistakis, D. Manousos, N.S. Tachos, E. Kazantzaki, E.E. Tripoliti, D.I. Fotiadis, and M. Tsiknakis. 2021. "Cognitive Workload Level Estimation Based on Eye Tracking: A Machine Learning Approach." Paper presented at the 21st IEEE international conference on bioinformatics and bioengineering, BIBE 2021, Kragujevac, October 25–27.

Sokolova, M., and G. Lapalme. 2009. "A Systematic Analysis of Performance Measures for Classification Tasks." *Information Processing & Management* 45 (4): 427–437. doi:10.1016/j.ipm.2009.03.002.

Stubblefield, K., R. Lipschutz, M. Phillips, C. Heckathorne, and T. Kuiken. 2005. "Occupational Therapy Outcomes with Targeted Hyper-Reinnervation Nerve Transfer Surgery: Two Case Studies." Paper presented at the proceedings of the myoelectric controls/powered prosthetics symposium.

Tang, C., D. Garreau, and U. von Luxburg. 2018. "When Do Random Forests Fail?" In *32nd Conference on Neural Information Processing Systems, Montreal, Canada*. Vol. 31. https://proceedings.neurips.cc/paper/2018/file/204da255aea2cd4a75ace6018fad6b4d-Paper.pdf

Tiffin, J., and E.J. Asher. 1948. "The Purdue Pegboard: Norms and Studies of Reliability and Validity." *The Journal of Applied Psychology* 32 (3): 234–247. doi:10.1037/h0061266.

Valverde-Albacete, F.J., J. Carrillo-de-Albornoz, and C. Peláez-Moreno. 2013. "A Proposal for New Evaluation Metrics and Result Visualization Technique for Sentiment Analysis Tasks." Paper presented at the international conference of the cross-language evaluation forum for European languages.

Walambe, R., P. Nayak, A. Bhardwaj, and K. Kotecha. 2021. "Employing Multimodal Machine Learning for Stress Detection." *Journal of Healthcare Engineering* 2021: 1–12. doi:10.1155/2021/9356452.

Wang, W., Z. Li, Y. Wang, and F. Chen. 2013. "Indexing Cognitive Workload Based on Pupillary Response under Luminance and Emotional Changes." Paper presented at the 18th international conference on intelligent user interfaces, IUI 2013, Santa Monica, CA, March 19–22. doi:10.1145/2449396.2449428.

White, M.M., W. Zhang, A.T. Winslow, M. Zahabi, F. Zhang, H. Huang, and D.B. Kaber. 2017. "Usability Comparison of Conventional Direct Control versus Pattern Recognition Control of Transradial Prostheses." *IEEE Transactions on Human–Machine Systems* 47 (6): 1146–1157. doi:10.1109/THMS.2017.2759762.

Wickens, C.D. 2017. "Mental Workload: Assessment, Prediction and Consequences." Paper presented at the international symposium on human mental workload: models and applications.

Williams, R.M., A.P. Turner, M. Orendurff, A.D. Segal, G.K. Klute, J. Pecoraro, and J. Czerniecki. 2006. "Does Having a Computerized Prosthetic Knee Influence Cognitive Performance during Amputee Walking?" *Archives of Physical Medicine and Rehabilitation* 87 (7): 989–994. doi:10.1016/j.apmr.2006.03.006.

Wilson, G., and R. Schlegel. 2004. *Operator Functional State Assessment*. Paris: North Atlantic Treaty Organisation (NATO). Research and Technology Organisation (RTO) BP, 25.

Witteveen, H., L. de Rond, J.S. Rietman, and P.H. Veltink. 2012. "Hand-Opening Feedback for Myoelectric Forearm Prostheses: Performance in Virtual Grasping Tasks Influenced by Different Levels of Distraction." *Journal of Rehabilitation Research and Development* 49 (10): 1517–1526. doi:10.1682/jrrd.2011.12.0243.

Wood, G., and J. Parr. 2022. "A Tool for Measuring Mental Workload during Prosthesis Use: The Prosthesis Task Load Index (PROS-TLX)."

Zahabi, M., Y. Wang, and S. Shahrampour. 2021. "Classification of Officers' Driving Situations Based on Eye-Tracking and Driver Performance Measures." *IEEE Transactions on Human–Machine Systems* 51 (4): 394–402. doi:10.1109/THMS.2021.3090787.

Zahabi, M., M.M. White, W. Zhang, A.T. Winslow, F. Zhang, H. Huang, and D.B. Kaber. 2019. "Application of Cognitive Task Performance Modeling for Assessing Usability of Transradial Prostheses." *IEEE Transactions on Human–Machine Systems* 49 (4): 381–387. doi:10.1109/THMS.2019.2903188.

Zhang, W., M. White, M. Zahabi, A.T. Winslow, F. Zhang, H. Huang, and D. Kaber. 2016. "Cognitive Workload in Conventional Direct Control vs. pattern Recognition Control of an Upper-Limb Prosthesis." Paper presented at the 2016 IEEE international conference on systems, man, and cybernetics (SMC).