The Use of Machine Learning and Deep Learning Techniques to Assess Proprioceptive Impairments of the Upper Limb after Stroke

Delowar Hossain
University of Calgary
Stephen Scott
Queen's University
Tyler Cluff
University of Calgary
Sean Dukelow (spdukelo@ucalgary.ca)
University of Calgary

Research Article

Keywords: Stroke, Proprioception, Robotics, Position Sense, Machine Learning, Deep Learning

Posted Date: April 12th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1514146/v1

Additional Declarations: Competing interest reported. SHS is the co-founder and Chief Scientific Officer of Kinarm (formally known as BKIN Technologies), the company that commercializes the Kinarm robotic device used in this study. All other authors confirm no conflict of interest.

Version of Record: A version of this preprint was published at Journal of NeuroEngineering and Rehabilitation on January 27th, 2023. See the published version at https://doi.org/10.1186/s12984-023-01140-9.

Abstract

Background

Proprioception is commonly impaired after stroke. Robotic tools precisely measure multiple attributes of position sense and create large datasets. Previously, we quantified individual performance based on single measured robotic parameters and an overall task score in an arm position matching (APM) task. In the present manuscript, we used machine learning and deep learning techniques to classify whether individuals had a stroke or not based on their robotic APM task performance.

Methods

Participants performed an APM task in the Kinarm exoskeleton robot that produced 12 parameters to quantify multiple attributes of position sense. We first quantified impairment in individual parameters and overall task score by determining if participants with stroke fell outside of the 95% cut-off score of control (normative) values. Then, we applied five machine learning algorithms: Logistic Regression, Decision Tree, Random Forest, Random Forest with Hyperparameters Tuning, and Support Vector Machine; and a deep learning algorithm: Deep Neural Network, to classify individual participants as to whether or not they had a stroke based only on robotic assessment results using a 10-fold cross-validation approach.

Results

We recruited 429 participants with neuroimaging confirmed stroke (< 35 days post-stroke) and 465 healthy controls. Depending on the APM parameter, we observed that 10.9–48.4% of stroke participants were impaired. Using the overall task score, 44% were impaired. The mean performance metrics of machine learning and deep learning models were: accuracy 82.4%, precision 85.6%, recall 76.5%, and F1 score 80.6%. All machine learning and deep learning models displayed similar classification accuracy; however, the Random Forest model had the highest numerical accuracy (83%). Our models showed high sensitivity and specificity (AUC = 0.89) in classifying individual participants based on their performance in the APM task. We also found that variability was the most important feature of classifying performance in the APM task.

Conclusion

Our machine learning and deep learning models displayed similar classification accuracy. Each model classified more participants correctly as stroke or control than classification of impairment based on individual parameters or overall task score using a cut-off score. Machine learning and deep learning techniques may provide opportunities to better understand proprioceptive impairments after stroke.

Introduction

Proprioception is the sense of body position, motion, and force based on information from muscle spindles, Golgi tendon organs, cutaneous receptors, joint receptors, and efference copy of motor commands [1–4]. Proprioceptive impairments are very common after stroke [5–7] and occur in as many as 64% of stroke survivors [8]. These impairments are associated with deficits in learning sequences of button presses [9], as well as decreased independence, quality of life, and poor functional recovery [10].

Clinical assessments of proprioception have traditionally relied on coarse observer-based examinations. Most often, patients are asked to close their eyes while an examiner moves the distal aspect of the patient's finger up or down. The patient is asked to report the position of their fingertip. Alternatively, some clinicians administer the Thumb Localization Test [11]. This is a simple test in which the clinician passively moves the patient's hand to a random position overhead while the patient's eyes are closed, and the patient must then reach to grasp their passively moved thumb with the opposite hand. Unfortunately, these clinical tests have issues with reliability, lack resolution, and display ceiling effects [18–21]. Some research groups have designed standardized clinician-administered tests such as the Nottingham Sensory Assessment [12], Wrist Position Sense Test (WPST) [13–14], and Rivermead Assessment of Somatosensory Performance (RASP) [15] in attempts to deal with the issues outlined above.

However, much of the field studying proprioception has moved to the use of automated measurement tools [16–17]. Robotic and instrumented assessments are commonly used in research studies of upper extremity proprioception [13, 22–24]. Some authors have used passive movement threshold detection paradigms [25–26], whereas others have used single limb position-matching [27–28, 103–104] or mirror-matching tasks [29–30]. Our group has significant experience using a robotic arm position matching task in individuals after stroke [31–34]. The arm position matching task, which we used in this study, relies on mirror matching and can measure various aspects of an individual's position sense, including variability in matching positions, systematic shifts in a perceived workspace, and perceived contraction or expansion of the workspace. This task can be administered quickly (~ 3 minutes) and has several advantages over typical clinical measures, including generating reliable, continuous measures of position sense, lack of floor or ceiling effects, and the fact that the interpretation of human examiners is not required.

Robotic assessments of proprioception can generate a large volume of data that may eventually be useful in predicting outcomes and planning for treatment after stroke. Machine Learning may be helpful in this regard. Several past studies have attempted to predict clinical outcomes following stroke (e.g., discharge from a rehabilitation unit to home, risk of medical complications, risk of readmission to hospital) using standardized observer-based clinical scales [35–41]. Many of these studies relied on Logistic Regression, although a few used of Machine Learning (ML) techniques [51–54]. ML techniques are highly effective algorithms that are driven by large volumes of data and can aid in prognosis. They are a set of powerful algorithms capable of modeling hidden and complex relationships between clinical variables and treatment outcomes without necessarily relying on any formal statistical assumptions [42].

Recently, Deep Learning-based approaches such as Deep Neural Networks, a branch of ML techniques, have achieved impressive results across a variety of Artificial Intelligence (AI) fields [43–47]. Deep Learning approaches are inspired by the human brain's ability to abstract high-level representations from low-level sensory stimuli [48]. These multi-leveled approaches can be mathematically represented as multi-layered neural networks and recently are able to be trained in layer-wise backpropagation to obtain tractable optimization [49]. These techniques are currently state-of-the-art in applications of speech recognition, image processing, computer vision, and natural language processing [50].

In the present study, we examine the performance of individuals with stroke and healthy controls on a robotic Arm Position Matching (APM) task. The goals of this study were: (1) to compare the different ML and DL techniques with a more traditional model that relied on the 95% cut-off score of normative data for different attributes of position sense, determining which technique flagged the highest number of stroke participants as abnormal, (2) to compare different Machine Learning (ML) and Deep Learning (DL) techniques, and their ability to retrospectively predict whether someone has had a stroke or not, and (3) to examine the relative importance of different parameters measured in the Arm Position Matching (APM) task and their usefulness in retrospectively predicting whether or not someone has had a stroke.

Methods

Participants

Participants with stroke were recruited from the inpatient acute stroke or stroke rehabilitation units at the Foothills Medical Centre, the Dr. Vernon Fanning Care Centre in Calgary, Alberta, Canada, and Providence Care, St Mary's of the Lake Hospital, Kingston, Ontario, Canada. Inclusion criteria for participants with stroke were: recent onset (< 35 days) of first clinical stroke and age > = 18 years. Exclusion criteria for participants with stroke were: other underlying neurological conditions (e.g. Parkinson's, Multiple Sclerosis), upper limb orthopedic impairments, inability to understand task instructions or evidence of apraxia [55]. Neurologically intact control participants who also met the inclusion and exclusion criteria above, but had no history of stroke, were recruited from the communities of Calgary, Alberta, and Kingston, Ontario, Canada. This study was reviewed and approved by the University of Calgary Conjoint Health Research Ethics Board and the Queen's University Research Ethics Board. All participants gave written informed consent before performing the assessment.

Robotic Assessment

Robotic Device. The robotic assessment of position sense was performed using a Kinarm Exoskeleton robotic device (Fig. 1A; Kinarm, Kingston, Ontario, Canada), which permits movements of the arm in the horizontal plane involving horizontal abduction/adduction of the shoulder and flexion/extension of the elbow. Participants were seated in a height-adjustable wheelchair base with their arms supported against gravity. The device was fit to each participant's arm by research staff who were trained to conduct the robotic assessment. The robot was wheeled to a 2D virtual/augmented reality display. The visual display is capable of projecting virtual targets into the plane of the participant's arm during calibration and task

performance. Given the focus on proprioceptive function, visual stimuli were not displayed on the screen during the experiment. Direct vision of the upper extremities was occluded by a shutter and a bib. The set-up and calibration procedures took between 6–8 minutes for each participant.

Arm Position Matching Task. The Arm Position Matching (APM) task was used to assess the individual's position sense of their arm and has been described previously [31–34][56]. Participants were instructed to relax one arm (passive hand) and let the robot passively move the hand to one of four/nine spatial locations separated by 20/10 cm (Fig. 1B, 9-target task). The 4-target protocol is spaced on a 2x2 grid with targets spaced at 20 cm intervals in the X- and Y-directions. The 9-target protocol is the same as the 4-target protocol but includes nine targets spaced on a 3x3 grid at 10 cm intervals. Target locations were pseudo-randomized within a block. Each block contained one trial at each target location and participants completed six blocks. The robot moved the passive hand using a bell-shaped speed profile (max speed < 1 m/s). After the robot completed the passive movement, participants were asked to move their other arm (active hand; Fig. 1B) to mirror-match the spatial position of the passive hand. Participants were granted as much time as necessary to match the active hand position with the passive hand. Participants notified the examiner when they had matched their hand position, and the examiner triggered the next trial. Each participant completed either the 4-target or 9-target task protocol [105]. For the stroke participants, the affected arm was always the passive hand. Healthy control participants completed the task twice, where each arm served as the passive hand once and we consider each arm data as a separate participant in the analysis [117].

Robotic Task Parameters

The following parameters were used to quantify task performance after completing all trials: (a) trial-to-trial *Variability* (Var) of the active hand, (b) *Spatial Contraction/Expansion* (Cont/Exp) of the area matched by the active hand, (c) *Systematic Spatial Shifts* (Shift) between the passive and active hands, and (d) *Absolute Error* (AE).

Variability. *Variability* in Arm Position Matching (APM) describes the trial-to-trial consistency of the active hand in matching the same target position (Fig. 1C). It was calculated as the standard deviation of the active hand's position for each target location. The mean of the standard deviations was then calculated across all target positions in the x-coordinate (Var_x), y-coordinate (Var_y), and resultant linear variability of both coordinates (Var_{xy}):

$$Var_{xy} = \sqrt{Var_x{}^2 + Var_y{}^2}$$

1

Spatial Contraction/Expansion Ratio. Spatial *Contraction/Expansion Ratio* describes whether a participant displayed contraction or expansion of their perceived workspace (Fig. 1D). It was calculated as the matched area/range of the workspace of the active hand relative to the passive hand. This parameter was calculated for the matched x-coordinates ($Cont/Exp_x$) by finding the difference between

the mean x-coordinate of the three left and three right targets for the active hand compared with the passive hand:

$$Cont/Exp_x = \frac{range_{x_{active}}}{range_{x_{passive}}}$$

2

A similar procedure was to calculate contraction/expansion in the y-coordinate (*Cont/Exp*$_y$) using the range of the top and bottom three targets. Spatial contraction/expansion in both the x- and y- coordinates (*Cont/Exp*$_{xy}$) was calculated by finding the area spanned by the active hand for the eight border targets and then normalized by the total spatial area spanned by these same targets using the passive hand.

Systematic Spatial Shifts. Systematic Spatial *Shifts* describe constant errors between the active and passive hands (Fig. 1E). These errors were calculated as the mean error between the passive and active hands for each target position. The mean was then calculated using the means for all target locations. Systematic shifts were calculated in the x-coordinate (*Shift*$_x$), y-coordinate (*Shift*$_y$), and combined across both coordinates to provide a measure of the resultant shift in matched positions (*Shift*$_{xy}$):

$$Shift_{xy} = \sqrt{{Shift_x}^2 + {Shift_y}^2}$$

3

Absolute Error. *Absolute Error* describes the mean absolute distance error between the position of the active and passive hands. The mean absolute distance errors were calculated in the x-coordinate (*AE*$_x$), y-coordinate (*AE*$_y$), and combined across both coordinates (*AE*$_{xy}$) of all trials between the active hand and the target position:

$$AE_{xy} = \sqrt{{AE_x}^2 + {AE_y}^2}$$

4

A total of 12 parameters were used to measure performance in the arm position matching task.

Z-score. For each of the parameters above, we relied on the Dexterit-E software [106] associated with the Kinarm to calculate a Z-score. The Z-score or standardized score, is the distance, measured in standard deviations, that a data point falls from the mean of the healthy cohort. Kinarm (Kinarm, Kingston, ON) [71] uses a consistent methodology for developing normal models to calculate the Z-scores of each parameter. Parameter scores from the distribution of the normative data set (developed from neurologically intact controls) are transformed using a Box-Cox power transformation [72]. The transformed data are fitted by accounting for age, sex, handedness, and robotic platform (exoskeleton, endpoint robot) using Multiple Linear Regression (MLR). After the first regression, the standard deviation

of the residuals is then modeled using a second MLR with the same factors such as age, sex, handedness, and robotic platform. Z-scores are calculated using the residuals of the first regression and standard deviation modeled by second regression for each parameter. Z-scores are the particular values from the mean, i.e., a Z-score of 1 means a value was 1 standard deviation above the mean, and a Z-score of -1 means a value was 1 standard deviation below the mean of the healthy control data.

To ensure the distribution is “close-to-normal”, the skew and kurtosis of the final distribution were calculated and compared to the following criteria (Equations 5 and 6). These criteria were selected from Pearson and Please [73] so that it is statistically valid to use parametric tests with the Z-scores.

$$skew : abs\left(\sqrt{\beta_1}\right) \leq 0.8, \sqrt{\beta_1} = \frac{\mu_3}{\sigma^3}$$

5

$$kurtosis : 2.4 \leq \beta_2 \leq 3.6, \beta_2 = \frac{\mu_4}{\sigma^4}$$

6

where σ is the standard deviation, and μ_3 and μ_4 are the third and fourth moments of the mean.

Task Score. A task score gives a global measure of a participant’s performance for a given task. It measures how far the participant’s performance is from the best performance. For calculating the task score, the first stage is converting the task parameter scores into standardized Z-scores (described above). The second stage is to identify whether the best performance for a given metric reflects large negative Z-scores, large positive Z-scores, or near zero Z-scores. The Z-scores are transformed into Zeta scores using Eq. 7 for those parameters in which best performance is one-sided (i.e., large negative or large positive Z-scores).

$$\varsigma = \sqrt{2} \bullet erfc^{-1}(\frac{1}{2} \bullet erfc\left(\frac{\pm z}{\sqrt{2}}\right))$$

7

Where ‘+’ is used when poor performance is positive and ‘-’ is used when poor performance is negative.

In the final stage, task scores are calculated based on the performance of healthy controls. The root-sum-square (RSS) distance of Z-scores and Zeta scores are calculated using Eq. 8 for healthy controls. RSS distance is also known as the Euclidean distance and is transformed into a Z-score using a Box-Cox transform. The Z-score of the RSS distance is then transformed using Eq. 7 to a one-sided statistic.

$$rssDistance = \sqrt{\sum_i z_i^2 + \sum_j \varsigma_j^2}$$

Where $\sum_i z_i^2$ includes all two-sided parameters and $\sum_j \varsigma_j^2$ includes all one-sided parameters.

Task scores are always positive. A score of 0 corresponds to the best performance, and increasing values represent poorer performance. If the task score is > 3.29 (that is normally 1 in 1000) for control participants, then that participant was classified as an outlier for the task and removed. Outliers were removed to improve the robustness of the modeling process of normative data sets.

Clinical Assessments

A broad range of clinical assessments was performed to characterize the impairment of stroke participants in this study. The assessments served to quantify sensation, movement, cognition, and functional abilities. The assessments were performed by a physician or physiotherapist who had expertise in stroke rehabilitation. They were blinded to the results of the robotic assessment.

Position sense was clinically assessed using the Thumb Localization Test (TLT) [11]. It was chosen because it has been used to quantify whole-limb position sense in many studies involving stroke [57–65]. In this test, the examiner moves the participant's stroke-affected arm to a position in front of the participant at or above eye level, lateral to the midline with the participant's eyes closed. The participant is then asked to pinch the thumb of that limb with the opposite thumb and forefinger (reaching limb). Participants were scored as 0 if their performance was considered normal (completed task perfectly) to 3, which is considered markedly abnormal (the participant was unable to find his or her thumb and did not climb up the affected arm to locate it).

Motor impairment was assessed using the Purdue Peg Board test (PPB) (Lafayette Instrument Co., Lafayette, IN, USA) [66] and the Chedoke-McMaster Stroke Assessment (CMSA) [67]. In the PPB assessment, participants placed as many small pegs as possible into holes in a board over 30 seconds using one hand. The participant is required to use the proximal upper extremity to keep the hand in the appropriate position to retrieve and insert each peg as a test of fine motor skills. The CMSA relies on the concept of stages of motor recovery, which was first introduced by Twitchell [68]. The CMSA classifies participants into subgroups based on the stage of motor recovery. The 7-point scale corresponds to seven stages of motor recovery, where the score of 1 is considered the most abnormal and a score of 7 is normal.

Functional abilities were assessed using the Functional Independence Measure (FIM) [69]. It was used as a metric for independence within activities of daily living. Within the 18-item scale, 13 items are considered as motor tasks, and 5 items are considered as cognitive tasks. In the current manuscript, we present the total FIM score (measured out of 126) and the FIM score for the motor component (measured out of 91).

Data Analysis

Data analysis was done using Machine Learning and Deep Learning techniques in the Python programming language (version 3.7.4) [70]. In the first step of our analysis, we determined when stroke participants were impaired on robotic parameters using the Z-scores described above. We determined the 95% cut-off score of control performance (Task score > 1.96 is considered as impaired and Task score < = 1.96 is considered as unimpaired) on each robotic parameter to find whether an individual participant failed on a given parameter. When a stroke/control participant's score fell outside of the control range, they were classified as impaired on that robotic task. Our primary analysis was concerned with comparing the impairment rate found on individual parameters and the overall task score (so called cut-off score technique) versus the ability of Machine Learning and Deep Learning techniques to determine impairment.

Machine Learning and Deep Learning

Flowchart of the Classification Models. The workflow blueprint of the data classification models is shown in Fig. 2. The K-fold cross-validation (K = 10, CV) training and testing data represent the outcome measures (features) derived from the Arm Position Matching (APM) task (12 parameters) of each control and stroke participant. K-fold CV training and testing data were classified and labeled into two different categories ("control" and "stroke"). This data was passed through feature extraction and scaling processes. It was then fitted to the supervised machine learning and deep learning models. After evaluation, we applied the mean and standard deviation of the K-fold CV of all model performance metrics. At the last step, we showed the receiver operating characteristic curves (ROC curves) for the mean of the K-fold cross-validated result of each model.

K-fold Cross-Validation (CV). The K-fold Cross-Validation procedure randomly divided the dataset into K-disjoint folds. One-fold was used for testing and remain K-l folds were used for training the model. This process was repeated K-times until the testing was performed on all K folds. All folds contained equal data points unless otherwise specified. We applied K-fold cross-validation (where K = 10) to estimate the performance and reliability of each classification algorithm and enable meaningful comparison between classification models. The performance of the classification models was evaluated by the mean and standard deviation across the K-fold datasets.

Features. A feature represents a measurable piece of data that can be used for analysis. It is also known as "attributes" or "variables". In our case, features were the Z-score data of the 12 task parameters (Variability X, Variability Y, Variability XY, etc.) such that all features were selected for our analysis. The features were then normalized using the min-max normalization (where the minimum value of that feature got transformed into 0, the maximum value got transformed into 1, and every other value got transformed between 0 and 1) so that the variance of the features was in the same range. Then, features were trained and tested using machine learning and deep learning models. After that, we tested classification models for prediction and evaluation in the testing phase.

Classification Methods

We applied five Machine Learning (ML) techniques: Logistic Regression (LR) [108], Decision Tree (DT) [109], Random Forest (RF) [110], Random Forest with Hyperparameters Tuning (RFT) [111], and Support Vector Machine (SVM) [112]. We also applied one Deep Learning technique: Deep Neural Network (DNN) [113] for the classification (or supervised learning) of stroke and control data.

Logistic Regression (LR). We used a Logistic Regression model to classify each participant as a stroke or control based on their performance in the arm position matching task. For that purpose, we implemented a logistic regression classifier that was fitted in the binary logistic regression regularization. This regularization added a penalty as model complexity increased to ensure the model generalized the data and prevented overfitting with an increase in parameters. LR model assumes a linear relationship between the input features and output. The binary logistic model had a dependent variable with two possible outcomes as healthy control and stroke. We used a tolerance of 0.0001 and the maximum number of iterations of 100 as criteria to stop network training.

Decision Tree (DT). We implemented a Decision Tree classifier as one of predictive modeling. It uses a tree-like model in which each internal node (non-leaf) is labeled with an input feature. The arcs coming from a node (branch) labeled with an input feature are labeled with each of the possible values of the target feature or the arcs leads to a subordinate decision node on a different input feature. Each leaf node is labeled with a class either healthy control or stroke. This model splits the nodes of all available features/parameters and then selects the splits, which results in the most homogeneous sub-nodes.

Our decision tree classifier implementation consisted of the following parameters: Gini impurity as a criterion to measure the quality of split, best as a splitter to choose the best split, the maximum depth of the tree as 4, and the minimum number of samples at the leaf node as 1.

Random Forest (RF). We implemented an ensemble learning model (i.e., a Random Forest classifier). It is a classification algorithm consisting of many decision trees, which uses bagging and feature randomness when building each individual tree. It tries to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree. The output of the random forest model was the class selected by most trees.

The parameters included in our implementation were: the number of estimators (the number of trees in the forest) was 100, Gini impurity as the criterion for the information gain, the minimum number of samples required to split an internal node was 2, and the minimum number of samples required to be a leaf node was 1.

Random Forest with Hyperparameters Tuning (RFT). We tuned the hyperparameters (a hyperparameter is a parameter whose value is used to control the learning process) of the Random Forest model to determine the best hyperparameters. It relies more on experimental results than theory, and thus the best model to determine the optimal settings was by trying many different combinations to evaluate each model's performance.

The tuned hyperparameters of the random forest model were: the number of trees in the forest, the maximum number of levels in each decision tree, the maximum number of features considered for spotting a node, the minimum number of data points placed in a node before the node is split, and the minimum number of data points allowed in a leaf node.

Support Vector Machine (SVM). We implemented a Support Vector Machine (SVM) classifier. It constructed a set of hyperplanes (hyperplanes are decision boundaries that help to classify the data points) in high-dimensional space to perform the classification task. The model transformed the data to find an optimal boundary between outputs (control or stroke). A good separation is achieved by the hyperplane that had the largest distance, or functional margin, to the nearest training data point of any class.

Our implementation included the following parameters: the regularization parameter that must strictly positive, the Radial Basis Function (RBF) type kernel, the size of the kernel cache as 200 MB, the pseudo-random number generator was used for shuffling the data for probability estimators, and tolerance of 0.001 was applied as the network stopping criterion.

Deep Neural Network (DNN). We also implemented a Deep Learning technique, namely, Deep Neural Network (DNN). It is a part of a broader family of machine learning techniques based on artificial neural networks.

Our DNN classifier implementation consisted of three hidden layers between input and output layers. The first hidden layer had 12 units with the Rectified Linear Unit (ReLU) as the activation function, the second hidden layer had 8 units with the ReLU as the activation function, and the third hidden layer had 1 unit with the sigmoid function as the activation function. We also used: binary cross-entropy as loss function, the Root Means Square propagation optimizer (RMSprop), the batch size of 10, and the number epoch of 100. An epoch refers to the number of passes of the entire training dataset the deep learning technique has completed. The input layer had 12 units for 12 features, and the output had 1 unit to predict a 0 or 1 that maps back to the “healthy control” or “stroke” class. Each layer of nodes trained a distinct set of features based on the output of the previous layer. The feature hierarchical process of our DNN model made it capable of handling very large, and high-dimensional datasets with billions of parameters passed through nonlinear functions.

Feature Importance

Feature importance [114] in machine learning refers to techniques that assign a score to each feature based on their usefulness in the classification task. The score is expressed in a percentage. We applied different feature importance techniques/calculations for the different machine learning techniques. For LR and SVM models, feature importance was based on an information-theoretic criterion, measuring the entropy in the changes of predictions, and perturbation of a given feature [115]. For the DT, RF, and RFT models, feature importance was computed as the mean and standard deviation of the impurity decrease (the total decrease in node impurity (weighted by the probability of reaching node) averaged over all trees

of the ensemble) within each tree [116]. In general, a higher score of feature importance means the specific feature has a large effect (importance) on the model that is being used to classify participants as “stroke” and “control”, and a lower score means the specific feature has less impact on the classification model.

Results

Participant Demographics. Data were collected from 429 stroke participants and 465 healthy control participants. Demographics and clinical features of all groups are summarized in Table 1. Ninety-three percent of control and 92% of stroke participants were right-hand dominant. Two control and three stroke participants were scored as mixed handedness on the Edinburgh Handedness Inventory [74]. Forty-eight percent of stroke participants were observed to have proprioceptive impairment based on the Thumb Localization Test (TLT) test. Seventy-six percent of stroke participants demonstrated motor impairments on their affected arm of the Chedoke-McMaster Stroke Assessment (CMSA) test.

Table 1
Demographic and clinical information for the sample of 894 participants of healthy control and stroke. Data are presented as the mean (range) unless otherwise noted. Square brackets for TLT, CMSA scores indicate the actual number of individuals who obtained a given score on the test, e.g., 210 individuals scored 0 on the TLT.

	Control (n = 465)	Stroke (n = 429)
Age	51 (20–88)	63 (18–92)
Sex	244 M, 221 F	280 M, 149 F
Dominant Hand	434 R, 29 L, 2 A	393 R, 33 L, 3 A
Days since Stroke	. . .	17 (1–34)
Types of Stroke	. . .	370 I, 59 H
TLT [0, 1, 2, 3]		
Affected Side	. . .	[210, 104, 73, 30]*
CMSA [1, 2, 3, 4, 5, 6, 7]		
Affected Arm	. . .	[35, 53, 60, 31, 82, 63, 105]
Unaffected Arm	. . .	[0, 0, 0, 0, 15, 94, 320]
CMSA [1, 2, 3, 4, 5, 6, 7]		
Affected Hand	. . .	[45, 33, 39, 35, 106, 95, 76]
Unaffected Hand	. . .	[0, 0, 0, 0, 7, 125, 297]
PPB		
Affected Side	. . .	6.9 (0-17.5)
Unaffected Side	. . .	10.5 (2.5–19)
FIM (Total Score)	. . .	93.7 (37–126)
FIM (Motor)	. . .	65 (13–91)
M – Male; F – Female; R – Right; L – Light; A - Ambidextrous; H Hemorrhagic; I – Ischemic; TLT – Thumb Localizing Test; CSMA – Chedoke-McMaster Stroke Assessment; PPB – Purdue Peg Board; and FIM – Functional Independence Measure.		
*12 scores were missing.		

Data Visualization. To visualize the distribution of scores on the robotic task parameters, we plotted histograms of each parameter. *Variability Y* for stroke and control participants is presented in Fig. 3. This exemplar figure demonstrated that the distribution of values of the *Variability Y* parameter, not surprisingly, overlapped between stroke and control participants. We chose to present *Variability Y* because this parameter had the most influence on the classification tasks (see Fig. 6). Similar findings were seen when examining the distributions of the other parameters (not shown). The overlap of stroke and control data highlighted the challenge of differentiating normal from abnormal behavior based on a single parameter.

Cut-off Score Technique, Machine Learning, and Deep Learning Classifier Models. We first examined the impairment rates for individual robotic parameters and the overall task score using the 95% cut-off score technique. The mean and standard deviation of 10-fold cross-validation of individual parameters (Variability, Contraction/Expansion ratio, Shift, and Absolute Error) based classifiers performance metrics, and the overall task score to find the number of impaired participants is shown in Fig. 4A. The result indicates that the highest number of participants were impaired on the parameter *Variability XY* (48.4%) followed by the other two variabilities: *Variability Y* (47.1%) and *Variability X* (45.2%) parameters using the cut-off technique based on individual parameter. The least number of participants were impaired on parameter *Shift X* (10.9%), followed by *Shift Y* (11.5%) and *Shift XY* (20.3%). The overall task score impairment rate by the 95% cut-off score technique was 44%. *Contraction/Expansion ratio X* had the highest standard deviation of 7.8%, followed by *Contraction/Expansion ratio Y* (7.4%) and *Contraction/Expansion ratio XY* (7.1%). *Absolute Error X* had the least standard deviation of 3.6%, followed by *Shift XY* (4.7%) and *Variability Y* (4.8%).

We then implemented five Machine Learning classifier models, namely Logistic Regression (LR), Decision Tree (DT), Random Forest (RT), Random Forest with Hyperparameters Tuning (RFT), Support Vector Machine (SVM), and one Deep Learning classifier model, namely Deep Neural Network (DNN) to classify data into two categories: “control” and “stroke”. The mean and standard deviation of 10-fold cross-validated performance metrics i.e., accuracy, precision, recall, and F1 score [107] of Machine Learning and Deep Learning models are shown in Fig. 4B.

The results of this analysis indicated that RF and RFT had the highest and nearly similar accuracy (83% and 82.8%). In the case of precision, LR had a higher value of 86.6% than any other classifiers. Again, for recall and F1 score metrics, RF had a higher value of 78% and 81.4% than any other classifiers. In terms of standard deviation, LR had the highest spread out over a range of 4.1% for accuracy, 7.8% for recall, and 4.9% for F1 score, whereas DT had the highest spread out over a range of 5.6% for the precision compared with other classifiers. We examined with the different number of maximum depths of the tree for the DT model and found that the maximum depth of four gave us the best classification accuracy because the lower number made our model faster but not as accurate; higher number gave more accuracy but slow and risk of overfitting. Also, we examined the different number of epochs (i.e., 50, 100, 200, 500, 1000, 2000, 3000, 5000, 10000, 20000, 50000) for the DNN model and found that 100 epochs gave us the best performance metrics.

Receiver Operating Characteristic (ROC) Curve. The mean ROC curve and Area Under the Curve (AUC) value for the classification performance of LR, DT, RF, RFT, SVM, and DNN models is shown in Fig. 5. The best possible classifier model would yield a point in the upper-left coordinate (0,1) of the ROC space, representing 100% sensitivity and 100% specificity, which is called a perfect classification. The LR model had the highest AUC (AUC = 0.900) value for the classification task, suggesting that LR had the best separation capability between control and stroke, followed by the DNN model (AUC = 0.898). RF (AUC = 0.892) performed slightly better at classification than RFT (AUC = 0.887). SVM model performed similarly (AUC = 0.892) like RF. DT performed the worst at classification task (AUC = 0.848) among all models. In summary, LR had the highest level of sensitivity and specificity, whereas DT had the lowest level of specificity and sensitivity among models.

Feature Importance. The mean and standard deviation of 10-fold cross-validation of feature importance based on individual parameters (Variability, Contraction/Expansion ratio, Shift, Absolute Error) obtained using LR, DT, RF, RFT, and SVM models for the classification task is shown in Fig. 6. We were not able to plot feature importance using the DNN model because of its complex structure. We can see that different models had different feature importance scores in percentage. Across the models, some features tended to have higher feature importance scores, whereas others tended to have lower feature importance scores. For instance, we observed that *Variability Y* was the most important feature of all models for the classification task, followed by *Variability XY* and *Variability X*. The least important feature tended to be *Shift XY*, followed by *Shift X* and *Shift Y*. Although, many features contributed similarly for the classification task using LR, RF, RFT, and SVM models, the relative importance of the features appeared to be different in the DT model (see Fig. 6).

Comparison between Cut-off Score Technique and Machine Learning/Deep Learning Models. The impairment rate based on overall task score was 44% using the 95% cut-off score technique. In comparison, the average accuracy of Machine Learning and Deep Learning models was 82.4% and their average standard deviation was 3.56%. Also, all machine learning and deep learning models showed high sensitivities and specificities (AUC > 0.84, i.e., more than 84% sensitivity and specificity) in the classification task.

Discussion

Proprioceptive impairment is a common consequence of stroke. Traditional clinical approaches to assess proprioception have known issues with reliability [75] and tend to rely on simplistic observer-based ordinal scales. This has led to the development of different instrumented assessment tools [31] [76–77]. Assessments such as robotics, which can provide detailed kinematic measures, have the potential to offer new insights into the nature and severity of the proprioceptive impairments that occur after stroke. Employing machine learning techniques to better understand the complex datasets that are generated by robotic assessments may prove valuable in this regard.

The current manuscript represents a first foray into this venue. We attempted to compare the prevalence of proprioceptive deficits as identified by a previously used [31] standardized technique that relies on cut-off scores based on the 95% distribution of healthy control performance to the information derived from several different machine learning techniques. Using the cut-off score technique, we observed that the percentage of individuals classified as impaired on any given task parameter was between 10.9% and 48.4%. The percentage of individuals impaired on the overall task score was 44%. The machine learning techniques, on the other hand, demonstrated accuracies that ranged from 82.0–83.0% (depending on the given technique) when trying to classify whether participants had or had not suffered a stroke based on their performance in the robot. While one might infer that the machine learning techniques are obviously better, it is critical to remember that they are attempting to determine two slightly different things. The cut-off score technique tries to determine who does or does not have a proprioceptive impairment based on robotic performance (i.e., prevalence of proprioceptive deficits), whereas the machine learning techniques use the same information to determine whether or not someone has had a stroke.

The cut-off score technique allows examination of individual kinematic variables in comparison to healthy controls (e.g., $Variability_{xy}$), whereas the machine learning techniques, as employed, do not. In itself, individual parameters may be important for appreciating the nature of a patient's deficits at the granular level required to precisely design an intervention. However, working with clinicians and researchers over the years, our group has been repeatedly asked to develop an overall task score that can be quickly and easily interpreted and potentially used as a primary outcome for clinical trials. The overall task score that was developed relies on summing the individual components and assuming equal weighting for each parameter and determining a single score [78]. Despite the mathematical complexity in generating normative scores, this method is simplistic in its implementation as all parameters are equally weighted which may or may not be the most appropriate method to generate an overall score, particularly as some parameters may be highly correlated.

Machine learning techniques, as we employed them, did not calculate the prevalence of proprioceptive deficits like our cut-off score technique. Rather, we calculate and compare the accuracy, precision, recall and F1 score of the different ML techniques in attempting to retrospectively predict whether someone had a stroke or not based on participant performance in the behavioural task. In general, all of the machine learning techniques we employed performed reasonably well. We interpret this to mean that our overall task score is likely underestimating the number of individuals with impairments after stroke. Perhaps this is not surprising as the machine learning techniques developed weighting values for the individual parameters (see Fig. 6), unlike the equal weighting assigned when generating overall task scores. When exposed to new data, machine learning models learn, grow, modify, change, and develop by themselves. Simply put, machine learning and deep learning techniques do this process by leveraging algorithms that learn from data in an iterative process, which is not possible using traditional data analysis methods.

Determining whether someone has had a stroke or not based on their performance in a robotic task, on the surface, may seem a bit pedantic as the majority of cases of stroke already have a diagnosis that has been accurately made based on clinical observations and confirmed with some form of neuroimaging

(either computed tomography or magnetic resonance imaging). In the current manuscript, however, knowledge of the diagnosis provided a ground truth for us to test the performance of machine learning and make comparisons to the way which we have historically analyzed robotic data. While we do not see using machine learning to make the diagnosis of stroke using kinematics to be practical, our study demonstrates the potential usefulness of machine learning tools.

In the present manuscript, the machine learning techniques we employed, for the most part, performed similarly. While these similarities were perhaps not so surprising based on other published studies that have shown relatively low variability in the results produced by different machine learning techniques [79–83], there are fairly sizable differences between how some of these techniques were mathematically operationalized (e.g., Logistic Regression vs. Support Vector Machines). While overfitting can be a concern when using machine learning, we used a rather low number of features and relatively large dataset with a cross-validation approach to minimize the risk of this. In the end, the similar performance of the different models may simply stem from the fact that the underlying dataset used was the same. If pushed to recommend a given model going forward for the type of data we examined, we would recommend using the “Random Forest” model. A random forest is an ensemble learning technique that fits several decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy, precision, recall, F1 score, and control overfitting. This technique also provides an indicator of feature importance in the dataset. Given the similarities in performance of the models we tested, however, any of the techniques used seem to work reasonably well.

Machine learning is not without its limitations. The techniques we employed required hundreds of datasets from individual participants which required thousands of hours to collect. In general, these techniques do not respond well to missing or unknown data. Like most analysis techniques used, there is always some chance of error inherent in the predictions that are made. If there is bias in the data that is used to train the machine learning techniques, this can be carried over to when a model is deployed for testing. In the present paper, we employed 10-fold cross-validated datasets as this has been recommended to achieve stable results compared to traditional statistical modelling [84]. We then tested the model on a dataset that was not used in training to decrease the risk of bias.

As neurorehabilitation begins to incorporate more technology in both clinical practice and research settings, the opportunities for employing techniques like machine learning will continue to grow and evolve. Thoughtful application of these techniques may provide new insights into longstanding clinical problems. We could potentially see these types of techniques being used to identify prognostic factors for recovery following stroke or other disease states. There are already examples of this elsewhere in healthcare: clinical variables to predict post-stroke outcomes [83, 85–89], medical imaging diagnosis [90–94], drug discovery and manufacturing [95–97], identifying diseases and diagnosis [98–100], outbreak prediction [101–102], etc. Further, artificial intelligence may be helpful in guiding rehabilitation for individual patients based on their characteristics using information gleaned from thousands of patients’ journeys. Machine learning provides us with a newer set of analysis tools to enhance our

understanding of data. Careful implementation could lead to significant changes in the way we carry out stroke rehabilitation in the future.

Conclusions

In this work, we applied five machine learning techniques (i.e., LR, DT, RF, RFT, and SVM) and one deep learning technique (i.e., DNN) to classify stroke patients from control participants and assessed the proprioception impairment of the upper limb after stroke. The classification performances were compared with the cut-off score technique based on individual parameters, and our models outperformed the cut-off score technique. The resulting AUC of the ROC curve can range up to 90% depending on the classifiers used. Also, we were able to find the important features, which contributed significantly to the classification task. The machine learning and deep learning models we demonstrate here can be readily applied to other clinical and medical research.

Abbreviations

AE: Absolute Error; APM: Arm Position Matching; AUC: Area Under the Curve; AI: Artificial Intelligence; CMSA: Chedoke-McMaster Stroke Assessment; Cont/Exp: Contraction/Expansion; CV: Cross-Validation; DL: Deep Learning; DNN: Deep Neural Network; DT: Decision Tree; FIM: Functional Independence Measure; LR: Logistic Regression; ML: Machine Learning; MLR: Multiple Linear Regression; PPB: Purdue Peg Board; RASP: Rivermead Assessment of Somatosensory Performance; ReLU: Rectified Linear Unit; RMSprop: Root Means Square propagation optimizer; RF: Random Forest; RFT: Random Forest with Hyperparameters Tuning; ROC: Receiver Operating Characteristics; RSS: Root-Sum-Square; SD: Standard Deviation; SVM: Support Vector Machine; TLT: Thumb Localization Test; Var: Variability; WSPT: Wrist Position Sense Test.

Declarations

Acknowledgments

We would like to acknowledge the efforts and support from M Piitz and H Bretzke in data collection and technical assistance.

Authors contributions

DH analyzed the collected data, carried out the machine learning and deep learning analysis, drafted the manuscript. SHS participated in the task design, data analysis, and drafting manuscript. TC participated in data analysis and drafting of the manuscript. SPD was involved in the task design and patient recruitment, participated in data analysis, and drafted the manuscript. All authors read the final manuscript and approved it.

Corresponding author

Sean P. Dukelow: spdukelo@ucalgary.ca

Funding

This research was supported by CIHR in Stroke Rehabilitation Research (MOP 106662), and a Heart and Stroke Foundation of Canada Grant-in-Aid (G-13-0003029) to SPD and an Ontario Research Foundation – Research Excellence grant (ORE-RE 04-47) to SHS.

Availability of data and materials

The dataset generated for the current study are not publicly available. Data may be available from the corresponding author on reasonable request.

Ethics approval and consent to participate

The protocol was approved by the University of Calgary Conjoint Health Research Ethics Board and the Queen's University Research Ethics Board. All participants gave written informed consent before performing the assessment.

Competing of interests

SHS is the co-founder and Chief Scientific Officer of Kinarm (formally known as BKIN Technologies), the company that commercializes the Kinarm robotic device used in this study. All other authors confirm no conflict of interest.

Consent for publication

Consent to publish was obtained from all participants; and also, consent was obtained for Fig. 1A.

References

1. Campbell, W.W. and DeJong, R.N., 2005. *DeJong's the neurologic examination* (No. 2005). Lippincott Williams & Wilkins.
2. Lephart, S.M., 2000. Introduction to the sensorimotor system. *Proprioception and neuromuscular control in joint stability*, pp.16–26.
3. Riemann, B.L. and Lephart, S.M., 2002. The sensorimotor system, part I: the physiologic basis of functional joint stability. *Journal of athletic training*, *37*(1), p.71.
4. Riemann, B.L. and Lephart, S.M., 2002. The sensorimotor system, part II: the role of proprioception in motor control and functional joint stability. *Journal of athletic training*, *37*(1), p.80.
5. Kusoffsky, A., Wadell, I. and Nilsson, B.Y., 1982. The relationship between sensory impairment and motor recovery in patients with hemiplegia. Scandinavian journal of rehabilitation medicine, *14*(1), pp.27–32.

6. Campfens, S.F., Zandvliet, S.B., Meskers, C.G., Schouten, A.C., van Putten, M.J. and van der Kooij, H., 2015. Poor motor function is associated with reduced sensory processing after stroke. Experimental brain research, *233*(4), pp.1339–1349.
7. Carey, L.M., 1995. Somatosensory loss after stroke. Critical Reviews™ in Physical and Rehabilitation Medicine, *7*(1).
8. Connell, L.A., Lincoln, N.B. and Radford, K.A., 2008. Somatosensory impairment after stroke: frequency of different deficits and their recovery. Clinical rehabilitation, *22*(8), pp.758–767.
9. Vidoni, E.D. and Boyd, L.A., 2009. Preserved motor learning after stroke is related to the degree of proprioceptive deficit. Behavioral and Brain Functions, *5*(1), pp.1–10.
10. Carey, L.M. and Matyas, T.A., 2011. Frequency of discriminative sensory loss in the hand after stroke in a rehabilitation setting. Journal of rehabilitation medicine, *43*(3), pp.257–263.
11. Hirayama, K., Fukutake, T. and Kawamura, M., 1999. 'Thumb localizing test'for detecting a lesion in the posterior column–medial lemniscal system. Journal of the neurological sciences, *167*(1), pp.45–49.
12. Lincoln, N.B., Jackson, J.M. and Adams, S.A., 1998. Reliability and revision of the Nottingham Sensory Assessment for stroke patients. Physiotherapy, *84*(8), pp.358–365.
13. Carey, L.M., Oke, L.E. and Matyas, T.A., 1996. Impaired limb position sense after stroke: a quantitative test for clinical use. Archives of physical medicine and rehabilitation, *77*(12), pp.1271–1278.
14. Carey, L.M., Matyas, T.A. and Oke, L.E., 2002. Evaluation of impaired fingertip texture discrimination and wrist position sense in patients affected by stroke: comparison of clinical and new quantitative measures. Journal of Hand Therapy, *15*(1), pp.71–82.
15. Winward, C.E., Halligan, P.W. and Wade, D.T., 2002. The Rivermead Assessment of Somatosensory Performance (RASP): standardization and reliability data. Clinical rehabilitation, *16*(5), pp.523–533.
16. Squeri, V., Zenzeri, J., Morasso, P. and Basteris, A., 2011, June. Integrating proprioceptive assessment with proprioceptive training of stroke patients. In *2011 IEEE International Conference on Rehabilitation Robotics* (pp. 1–6). IEEE.
17. Niessen, M.H., Veeger, D.H., Meskers, C.G., Koppe, P.A., Konijnenbelt, M.H. and Janssen, T.W., 2009. Relationship among shoulder proprioception, kinematics, and pain after stroke. Archives of physical medicine and rehabilitation, *90*(9), pp.1557–1564.
18. Lincoln, N.B., Crow, J.L., Jackson, J.M., Waters, G.R., Adams, S.A. and Hodgson, P., 1991. The unreliability of sensory assessments. Clinical rehabilitation, *5*(4), pp.273–282.
19. Dellon, A.L., Mackinnon, S.E. and Crosby, P.M., 1987. Reliability of two-point discrimination measurements. The Journal of hand surgery, *12*(5), pp.693–696.
20. Lincoln, N.B., Jackson, J.M. and Adams, S.A., 1998. Reliability and revision of the Nottingham Sensory Assessment for stroke patients. Physiotherapy, *84*(8), pp.358–365.
21. Rinderknecht, M.D., Lambercy, O., Raible, V., Büsching, I., Sehle, A., Liepert, J. and Gassert, R., 2018. Reliability, validity, and clinical feasibility of a rapid and objective assessment of post-stroke deficits

in hand proprioception. *Journal of neuroengineering and rehabilitation*, *15*(1), pp.1–15.

22. Goble, D.J. and Brown, S.H., 2009. Dynamic proprioceptive target matching behavior in the upper limb: effects of speed, task difficulty and arm/hemisphere asymmetries. Behavioural brain research, *200*(1), pp.7–14.
23. Scheidt, R.A., Lillis, K.P. and Emerson, S.J., 2010. Visual, motor and attentional influences on proprioceptive contributions to perception of hand path rectilinearity during reaching. Experimental brain research, *204*(2), pp.239–254.
24. Goble, D.J., Coxon, J.P., Van Impe, A., Geurts, M., Van Hecke, W., Sunaert, S., Wenderoth, N. and Swinnen, S.P., 2012. The neural basis of central proprioceptive processing in older versus younger adults: an important sensory role for right putamen. Human brain mapping, *33*(4), pp.895–908.
25. Bengtson, M.C., Mrotek, L.A., Stoeckmann, T., Ghez, C. and Scheidt, R.A., 2014, August. The arm motion detection (AMD) test. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 5349–5352). IEEE.
26. Simo, L.S., Ghez, C., Botzer, L. and Scheidt, R.A., 2011, August. A quantitative and standardized robotic method for the evaluation of arm proprioception after stroke. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 8227–8230). IEEE.
27. Deblock-Bellamy, A., Batcho, C.S., Mercier, C. and Blanchette, A.K., 2018. Quantification of upper limb position sense using an exoskeleton and a virtual reality display. Journal of neuroengineering and rehabilitation, *15*(1), pp.1–11.
28. Lowrey, C.R., Blazevski, B., Marnet, J.L., Bretzke, H., Dukelow, S.P. and Scott, S.H., 2020. Robotic tests for position sense and movement discrimination in the upper limb reveal that they each are highly reproducible but not correlated in healthy individuals. Journal of NeuroEngineering and Rehabilitation, *17*(1), pp.1–13.
29. Goble, D.J., 2010. Proprioceptive acuity assessment via joint position matching: from basic science to general practice. Physical therapy, *90*(8), pp.1176–1184.
30. Smorenburg, A.R., Ledebt, A., Deconinck, F.J. and Savelsbergh, G.J., 2013. Practicing a matching movement with a mirror in individuals with spastic hemiplegia. Research in developmental disabilities, *34*(9), pp.2507–2513.
31. Dukelow, S.P., Herter, T.M., Moore, K.D., Demers, M.J., Glasgow, J.I., Bagg, S.D., Norman, K.E. and Scott, S.H., 2010. Quantitative assessment of limb position sense following stroke. Neurorehabilitation and neural repair, *24*(2), pp.178–187.
32. Dukelow, S.P., Herter, T.M., Bagg, S.D. and Scott, S.H., 2012. The independence of deficits in position sense and visually guided reaching following stroke. Journal of neuroengineering and rehabilitation, *9*(1), pp.1–13.
33. Kenzie, J.M., Semrau, J.A., Hill, M.D., Scott, S.H. and Dukelow, S.P., 2017. A composite robotic-based measure of upper limb proprioception. Journal of neuroengineering and rehabilitation, *14*(1), pp.1–12.

34. Findlater, S.E., Hawe, R.L., Semrau, J.A., Kenzie, J.M., Amy, Y.Y., Scott, S.H. and Dukelow, S.P., 2018. Lesion locations associated with persistent proprioceptive impairment in the upper limbs after stroke. NeuroImage: Clinical, *20*, pp.955–971.
35. Roth, E.J., Lovell, L., Harvey, R.L., Heinemann, A.W., Semik, P. and Diaz, S., 2001. Incidence of and risk factors for medical complications during stroke rehabilitation. Stroke, *32*(2), pp.523–529.
36. Ottenbacher, K.J., Smith, P.M., Illig, S.B., Fiedler, R.C., Gonzales, V. and Granger, C.V., 2001. Characteristics of persons rehospitalized after stroke rehabilitation. Archives of Physical Medicine and Rehabilitation, *82*(10), pp.1367–1374.
37. Wilson, D.B., Houle, D.M. and Keith, R.A., 1991. Stroke rehabilitation: a model predicting return home. *Western Journal of Medicine*, *154*(5), p.587.
38. Paolucci, S., Bragoni, M., Coiro, P., De Angelis, D., Fusco, F.R., Morelli, D., Venturiero, V. and Pratesi, L., 2006. Is sex a prognostic factor in stroke rehabilitation? A matched comparison. Stroke, *37*(12), pp.2989–2994.
39. Berlowitz, D.R., Hoenig, H., Cowper, D.C., Duncan, P.W. and Vogel, W.B., 2008. Impact of comorbidities on stroke rehabilitation outcomes: does the method matter?. Archives of physical medicine and rehabilitation, *89*(10), pp.1903–1906.
40. Kristensen, H.K., Tistad, M., Koch, L.V. and Ytterberg, C., 2016. The importance of patient involvement in stroke rehabilitation. PloS one, *11*(6), p.e0157149.
41. Picena, M.C., Recupero, E., Finocchiaro, F., Santagati, A., Greco, S., Longo, P., Manca, M., Cosentino, E., Mayer, F., Biondi, T. and Mugelli, C., 2008. Outcome predictors of rehabilitation for first stroke in the elderly. European journal of physical and rehabilitation medicine, *44*, pp.3–11.
42. Bishop, C.M., 1995. *Neural networks for pattern recognition*. Oxford university press.
43. Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, *25*, pp.1097–1105.
44. Ciregan, D., Meier, U. and Schmidhuber, J., 2012, June. Multi-column deep neural networks for image classification. In *2012 IEEE conference on computer vision and pattern recognition* (pp. 3642–3649). IEEE.
45. Shen, D., Wu, G. and Suk, H.I., 2017. Deep learning in medical image analysis. Annual review of biomedical engineering, *19*, pp.221–248.
46. Dong, D., Wu, H., He, W., Yu, D. and Wang, H., 2015, July. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1723–1732).
47. Radford, A., Metz, L. and Chintala, S., 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.
48. Zador, A.M., 2019. A critique of pure learning and what artificial neural networks can learn from animal brains. Nature communications, *10*(1), pp.1–7.

49. LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. nature, *521*(7553), pp.436–444.
50. Du, H., Ghassemi, M.M. and Feng, M., 2016, August. The effects of deep network topology on mortality prediction. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 2602–2605). IEEE.
51. Yang, G., Deng, J., Pang, G., Zhang, H., Li, J., Deng, B., Pang, Z., Xu, J., Jiang, M., Liljeberg, P. and Xie, H., 2018. An IoT-enabled stroke rehabilitation system based on smart wearable armband and machine learning. IEEE journal of translational engineering in health and medicine, *6*, pp.1–10.
52. Mohanty, R., Sinha, A.M., Remsik, A.B., Dodd, K.C., Young, B.M., Jacobson, T., McMillan, M., Thoma, J., Advani, H., Nair, V.A. and Kang, T.J., 2018. Machine learning classification to identify the stage of brain-computer interface therapy for stroke rehabilitation using functional connectivity. *Frontiers in neuroscience*, *12*, p.353.
53. Lee, M.H., Siewiorek, D.P., Smailagic, A. and Bernardino, A., 2020. Opportunities of a machine learning-based decision support system for stroke rehabilitation assessment. *arXiv preprint arXiv:2002.12261*.
54. Miao, S., Shen, C., Feng, X., Zhu, Q., Shorfuzzaman, M. and Lv, Z., 2021. Upper limb rehabilitation system for stroke survivors based on multi-modal sensors and machine learning. IEEE Access, *9*, pp.30283–30291.
55. Vanbellingen, T., Kersten, B., Van de Winckel, A., Bellion, M., Baronti, F., Müri, R. and Bohlhalter, S., 2011. A new bedside test of gestures in stroke: the apraxia screen of TULIA (AST). Journal of Neurology, Neurosurgery & Psychiatry, *82*(4), pp.389–392.
56. Semrau, J.A., Herter, T.M., Scott, S.H. and Dukelow, S.P., 2013. Robotic identification of kinesthetic deficits after stroke. Stroke, *44*(12), pp.3414–3421.
57. Head, H. and Holmes, G., 1911. Sensory disturbances from cerebral lesions. Brain, *34*(2–3), pp.102–254.
58. Fukutake, T., Hirayama, K. and Komatsu, T., 1993. Transient unilateral catalepsy and right parietal damage. Psychiatry and Clinical Neurosciences, *47*(3), pp.647–650.
59. Hiraga, A., Sakakibara, R., Mizobuchi, K., Asahina, M., Kuwabara, S., Hayashi, Y. and Hattori, T., 2005. Putaminal hemorrhage disrupts thalamocortical projection to secondary somatosensory cortex: case report. Journal of the neurological sciences, *231*(1–2), pp.81–83.
60. Ihori, N., Kawamura, M., Araki, S. and Kawachi, J., 2002. Kinesthetic alexia due to left parietal lobe lesions. European neurology, *48*(2), pp.87–96.
61. Kwakkel, G., Kollen, B.J., van der Grond, J. and Prevo, A.J., 2003. Probability of regaining dexterity in the flaccid upper limb: impact of severity of paresis and time since onset in acute stroke. Stroke, *34*(9), pp.2181–2186.
62. Rand, D., Weiss, P.L. and Gottlieb, D., 1999. Does proprioceptive loss influence recovery of the upper extremity after stroke?. Neurorehabilitation and neural Repair, *13*(1), pp.15–21.
63. Rand, D., Gottlieb, D. and Weiss, P.L., 2001. Recovery of patients with a combined motor and proprioception deficit during the first six weeks of post stroke rehabilitation. Physical & Occupational

Therapy in Geriatrics, *18*(3), pp.69–87.

64. Yoshida, H., Kondo, T. and Nakasato, N., 2008. Neuromagnetic investigation of somatosensory cortical reorganization in hemiplegic patients after thalamic hemorrhage. Journal of Physical Therapy Science, *20*(2), pp.123–127.
65. Sakakibara, R., Kishi, M., Ogawa, E. and Shirai, K., 2009. Isolated facio-lingual hypoalgesia and weakness after a hemorrhagic infarct localized at the contralateral operculum. Journal of the neurological sciences, *276*(1–2), pp.193–195.
66. Tiffin, J. and Asher, E.J., 1948. The Purdue Pegboard: norms and studies of reliability and validity. *Journal of applied psychology*, *32*(3), p.234.
67. Gowland, C., Stratford, P., Ward, M., Moreland, J., Torresin, W., Van Hullenaar, S., Sanford, J., Barreca, S., Vanspall, B. and Plews, N., 1993. Measuring physical impairment and disability with the Chedoke-McMaster Stroke Assessment. Stroke, *24*(1), pp.58–63.
68. Twitchell, T.E., 1951. The restoration of motor function following hemiplegia in man. Brain, *74*(4), pp.443–480.
69. Keith, R.A., Granger, C.V., Hamilton, B.B., Sherwin, F.S., 1987. The functional independence measure: a new tool for rehabilitation. Advances in Clinical Rehabilitation, 1:6–18.
70. Van Rossum, G. & Drake, F.L., 2009. Python 3 Reference Manual, Scotts Valley, CA: CreateSpace.
71. Kinarm Lab: https://kinarm.com/kinarm-products/kinarm-exoskeleton-lab/. Last visited on December 16, 2021.
72. Box, G.E. and Cox, D.R., 1964. An analysis of transformations. Journal of the Royal Statistical Society: Series B (Methodological), *26*(2), pp.211–243.
73. Pearson, E.S. and Please, N.W., 1975. Relation between the shape of population distribution and the robustness of four simple test statistics. Biometrika, 62(2), pp.223–241.
74. Veale, J.F., 2014. Edinburgh handedness inventory–short form: a revised version based on confirmatory factor analysis. Laterality: Asymmetries of Body, Brain and Cognition, *19*(2), pp.164–177.
75. Lincoln, N.B., Crow, J.L., Jackson, J.M., Waters, G.R., Adams, S.A. and Hodgson, P., 1991. The unreliability of sensory assessments. Clinical rehabilitation, *5*(4), pp.273–282.
76. Carey, L.M., Oke, L.E. and Matyas, T.A., 1996. Impaired limb position sense after stroke: a quantitative test for clinical use. Archives of physical medicine and rehabilitation, *77*(12), pp.1271–1278.
77. Goble, D.J., Lewis, C.A. and Brown, S.H., 2006. Upper limb asymmetries in the utilization of proprioceptive feedback. Experimental brain research, *168*(1–2), pp.307–311.
78. Dexterit-E Explorer User Guide 3.9: https://kinarm.com/download/dexterit-e-explorer-3-9-user-guide/. Last visited on December 16, 2021.
79. Kim, S.J., Cho, K.J. and Oh, S., 2017. Development of machine learning models for diagnosis of glaucoma. PloS one, *12*(5), p.e0177726.

80. Gromiha, M.M. and Suresh, M.X., 2008. Discrimination of mesophilic and thermophilic proteins using machine learning algorithms. Proteins: Structure, Function, and Bioinformatics, *70*(4), pp.1274–1279.
81. Govindarajan, P., Soundarapandian, R.K., Gandomi, A.H., Patan, R., Jayaraman, P. and Manikandan, R., 2020. Classification of stroke disease using machine learning algorithms. Neural Computing and Applications, *32*(3), pp.817–828.
82. Hung, C.Y., Chen, W.C., Lai, P.T., Lin, C.H. and Lee, C.C., 2017, July. Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 3110–3113). IEEE.
83. Chang, S.C., Chu, C.L., Chen, C.K., Chang, H.N., Wong, A.M., Chen, Y.P. and Pei, Y.C., 2021. The Comparison and Interpretation of Machine-Learning Models in Post-Stroke Functional Outcome Prediction. *Diagnostics*, *11*(10), p.1784.
84. van der Ploeg, T., Austin, P.C. and Steyerberg, E.W., 2014. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC medical research methodology, 14(1), pp.1–13.
85. Lin, W.Y., Chen, C.H., Tseng, Y.J., Tsai, Y.T., Chang, C.Y., Wang, H.Y. and Chen, C.K., 2018. Predicting post-stroke activities of daily living through a machine learning-based approach on initiating rehabilitation. International journal of medical informatics, *111*, pp.159–164.
86. Ge, Y., Wang, Q., Wang, L., Wu, H., Peng, C., Wang, J., Xu, Y., Xiong, G., Zhang, Y. and Yi, Y., 2019. Predicting post-stroke pneumonia using deep neural network approaches. *International journal of medical informatics*, *132*, p.103986.
87. Li, X., Wu, M., Sun, C., Zhao, Z., Wang, F., Zheng, X., Ge, W., Zhou, J. and Zou, J., 2020. Using machine learning to predict stroke-associated pneumonia in Chinese acute ischaemic stroke patients. European journal of neurology, *27*(8), pp.1656–1663.
88. Rondina, J.M., Filippone, M., Girolami, M. and Ward, N.S., 2016. Decoding post-stroke motor function from structural brain imaging. NeuroImage: Clinical, *12*, pp.372–380.
89. Wang, W., Kiik, M., Peek, N., Curcin, V., Marshall, I.J., Rudd, A.G., Wang, Y., Douiri, A., Wolfe, C.D. and Bray, B., 2020. A systematic review of machine learning models for predicting outcomes of stroke with structured data. PloS one, *15*(6), p.e0234722.
90. Erickson, B.J., Korfiatis, P., Akkus, Z. and Kline, T.L., 2017. Machine learning for medical imaging. Radiographics, *37*(2), pp.505–515.
91. Magoulas, G.D. and Prentza, A., 1999, July. Machine learning in medical applications. In *Advanced course on artificial intelligence* (pp. 300–307). Springer, Berlin, Heidelberg.
92. De Bruijne, M., 2016. Machine learning approaches in medical image analysis: From detection to diagnosis.
93. Magoulas, G.D. and Prentza, A., 1999, July. Machine learning in medical applications. In *Advanced course on artificial intelligence* (pp. 300–307). Springer, Berlin, Heidelberg.

94. Lundervold, A.S. and Lundervold, A., 2019. An overview of deep learning in medical imaging focusing on MRI. Zeitschrift für Medizinische Physik, *29*(2), pp.102–127.
95. Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M. and Zhao, S., 2019. Applications of machine learning in drug discovery and development. Nature Reviews Drug Discovery, *18*(6), pp.463–477.
96. Lavecchia, A., 2015. Machine-learning approaches in drug discovery: methods and applications. Drug discovery today, *20*(3), pp.318–331.
97. Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K. and Tekade, R.K., 2021. Artificial intelligence in drug discovery and development. *Drug Discovery Today*, *26*(1), p.80.
98. Myszczynska, M.A., Ojamies, P.N., Lacoste, A.M., Neil, D., Saffari, A., Mead, R., Hautbergue, G.M., Holbrook, J.D. and Ferraiuolo, L., 2020. Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. Nature Reviews Neurology, *16*(8), pp.440–456.
99. Battineni, G., Sagaro, G.G., Chinatalapudi, N. and Amenta, F., 2020. Applications of machine learning predictive models in the chronic disease diagnosis. *Journal of personalized medicine*, *10*(2), p.21.
100. Myszczynska, M.A., Ojamies, P.N., Lacoste, A.M., Neil, D., Saffari, A., Mead, R., Hautbergue, G.M., Holbrook, J.D. and Ferraiuolo, L., 2020. Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. Nature Reviews Neurology, *16*(8), pp.440–456.
101. Ardabili, S.F., Mosavi, A., Ghamisi, P., Ferdinand, F., Varkonyi-Koczy, A.R., Reuter, U., Rabczuk, T. and Atkinson, P.M., 2020. Covid-19 outbreak prediction with machine learning. *Algorithms*, *13*(10), p.249.
102. Kim, J. and Ahn, I., 2021. Infectious disease outbreak prediction using media articles with machine learning models. Scientific Reports, *11*(1), pp.1–13.
103. Gurari, N., Drogos, J.M. and Dewald, J.P., 2017. Individuals with chronic hemiparetic stroke can correctly match forearm positions within a single arm. Clinical Neurophysiology, *128*(1), pp.18–30.
104. Goble, D.J., 2010. Proprioceptive acuity assessment via joint position matching: from basic science to general practice. Physical therapy, *90*(8), pp.1176–1184.
105. Mochizuki, G., Centen, A., Resnick, M., Lowrey, C., Dukelow, S.P. and Scott, S.H., 2019. Movement kinematics and proprioception in post-stroke spasticity: assessment using the Kinarm robotic exoskeleton. Journal of neuroengineering and rehabilitation, *16*(1), pp.1–13.
106. Dexterit-E Software: https://kinarm.com/kinarm-products/dexterit-e/. Last visited on February 07, 2022.
107. Confusion Matrix, Accuracy, Precision, Recall, F1 Score: https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd. Last Visited on February 08, 2022.
108. Kleinbaum, D.G., Dietz, K., Gail, M., Klein, M. and Klein, M., 2002. *Logistic regression* (p. 536). New York: Springer-Verlag.
109. Song, Y.Y. and Ying, L.U., 2015. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, *27*(2), p.130.

110. Breiman, L., 2001. Random forests. Machine learning, *45*(1), pp.5–32.
111. Probst, P., Wright, M.N. and Boulesteix, A.L., 2019. Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: data mining and knowledge discovery, *9*(3), p.e1301.
112. Noble, W.S., 2006. What is a support vector machine?. Nature biotechnology, *24*(12), pp.1565–1567.
113. Bengio, Y., 2009. *Learning deep architectures for AI*. Now Publishers Inc.
114. König, G., Molnar, C., Bischl, B. and Grosse-Wentrup, M., 2021, January. Relative feature importance. In *2020 25th International Conference on Pattern Recognition (ICPR)* (pp. 9318–9325). IEEE.
115. P. Choudhary, A. Kramer, and c. datascience.com team, "Skater: Model interpretation library," Mar. 2018. Available: https://oracle.github.io/Skater/reference/interpretation.html? Last Visited on 4 March 2022.
116. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. the Journal of machine Learning research, *12*, pp.2825–2830.
117. Wood, M.D., Simmatis, L.E., Jacobson, J.A., Dukelow, S.P., Boyd, J.G. and Scott, S.H., 2021. Principal Components Analysis Using Data Collected From Healthy Individuals on Two Robotic Assessment Platforms Yields Similar Behavioral Patterns. *Frontiers in human neuroscience*, *15*.

Figures

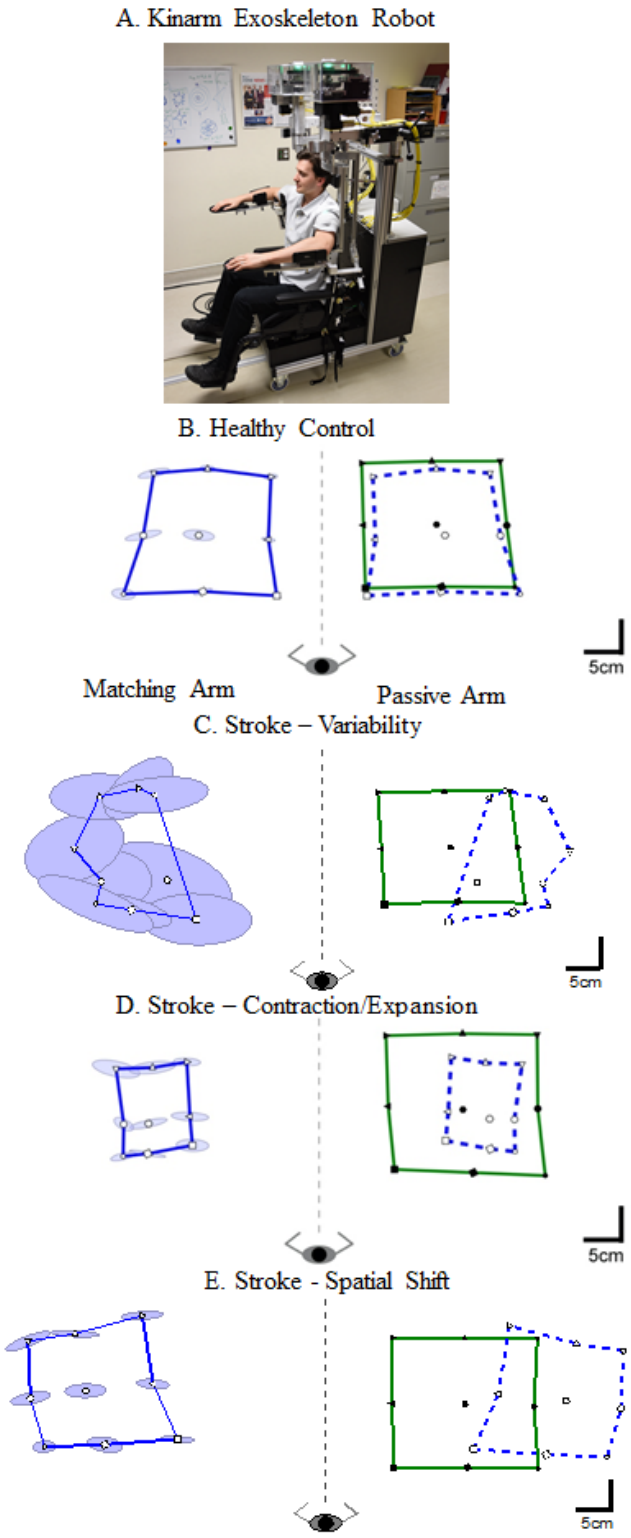

Figure 1

Arm Position Matching (APM) task.

A) The Kinarm exoskeleton robot.

B) Typical healthy control participant data. The robot moved the participant's passive right hand to one of 9 spatial locations (filled symbols). The participant then attempted to mirror match with the left active hand (open symbols). The solid blue line connects the average final positions of the outer eight target locations of the matching hand (active hand). Solid green line connects the outer eight targets for the robot moved passive hand. The dashed blue line is the mirror reflection of solid blue line, which allows a visual comparison of the average final outer 8 positions of the active and passive (robot-moved) hands. Ellipses represent one standard deviation of the matched positions. The ellipses represent trial-to-trial variability, where a larger ellipse means the participant was less consistent (i.e., more variable) in matching the position of their passive hand with the active hand.

C) An exemplar stroke participant who demonstrated high variability in position matching.

D) An exemplar stroke participant who demonstrated a contracted sense of their workspace.

E) An exemplar stroke participant who demonstrated a spatial shift of their workspace.

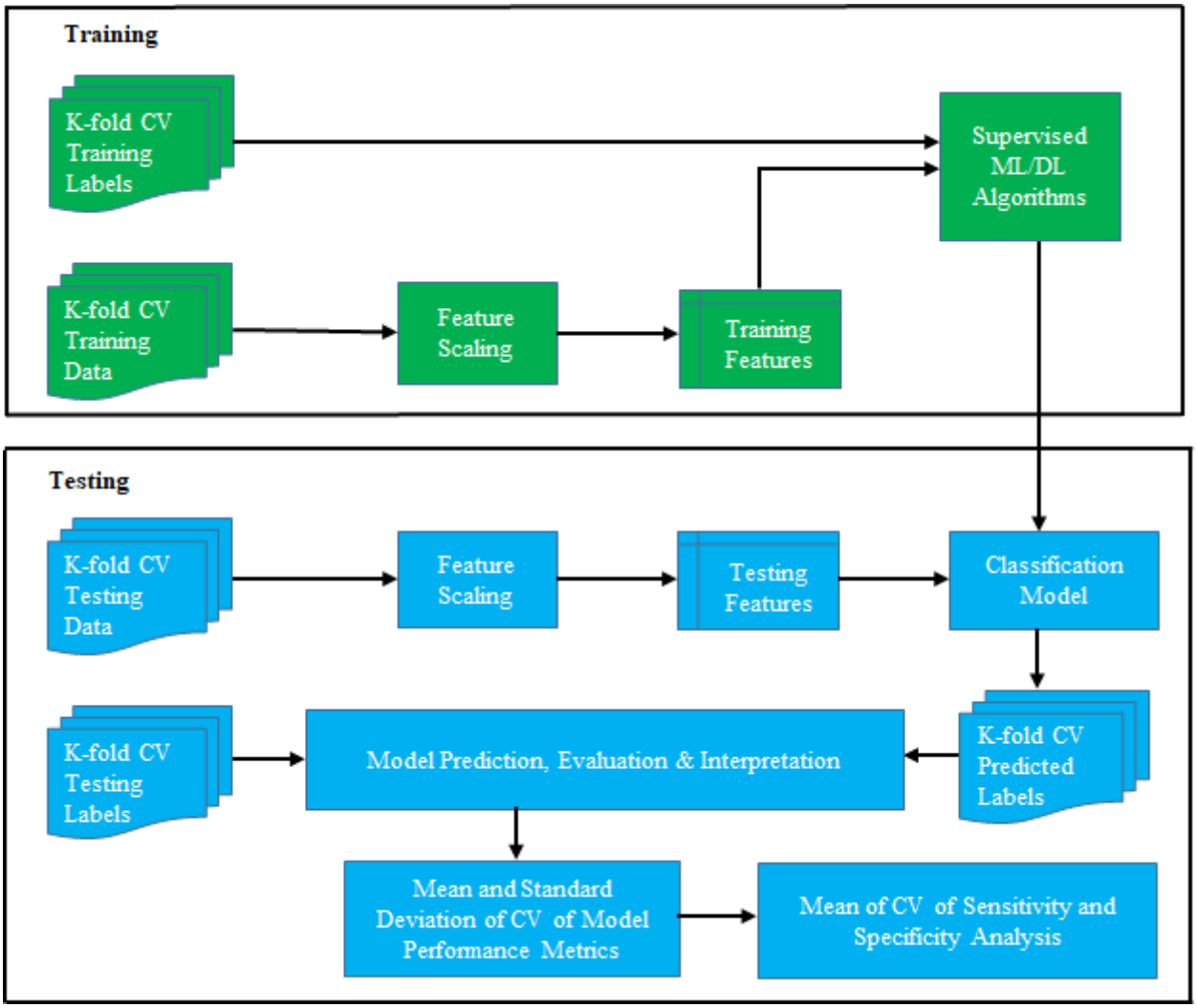

Figure 2

The workflow of K-fold (K=10) cross-validation (CV) of the machine learning and deep learning models. The training and testing data refer to outcome measures derived from the position matching task in each stroke and control participant. The model generates a label that classified each individual participant as a control or participant with stroke.

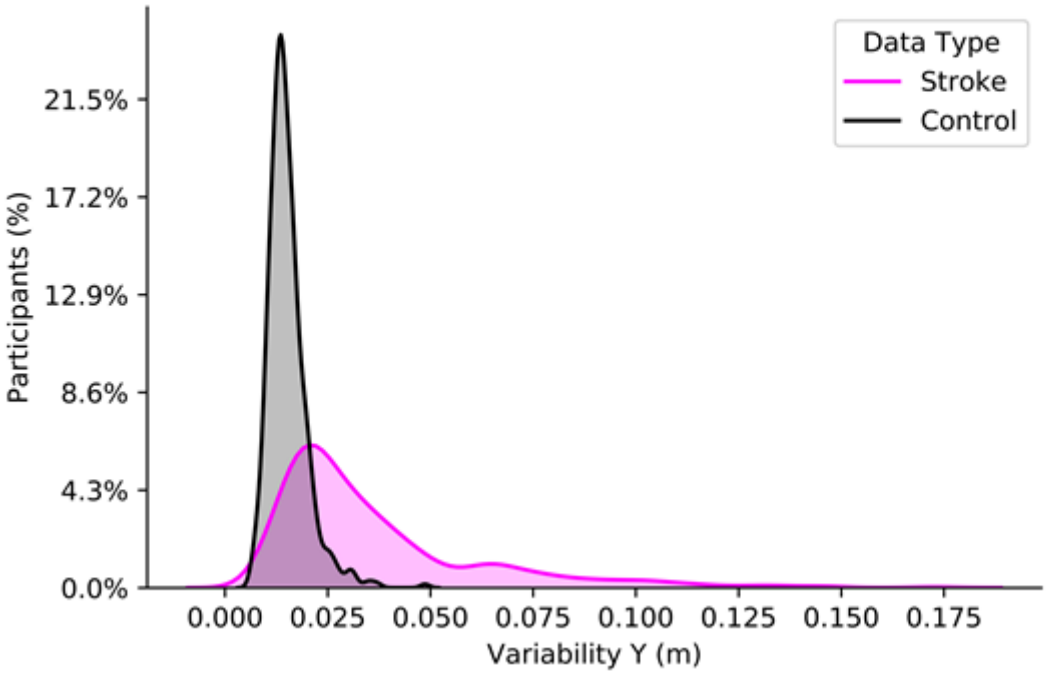

Figure 3

Histogram showing the distribution of *Variability Y* values in healthy controls and participants with stroke. The percentage on the y-axis is the participant count in each bin normalized to the number of participants with stroke (n=429) and control (n=465).

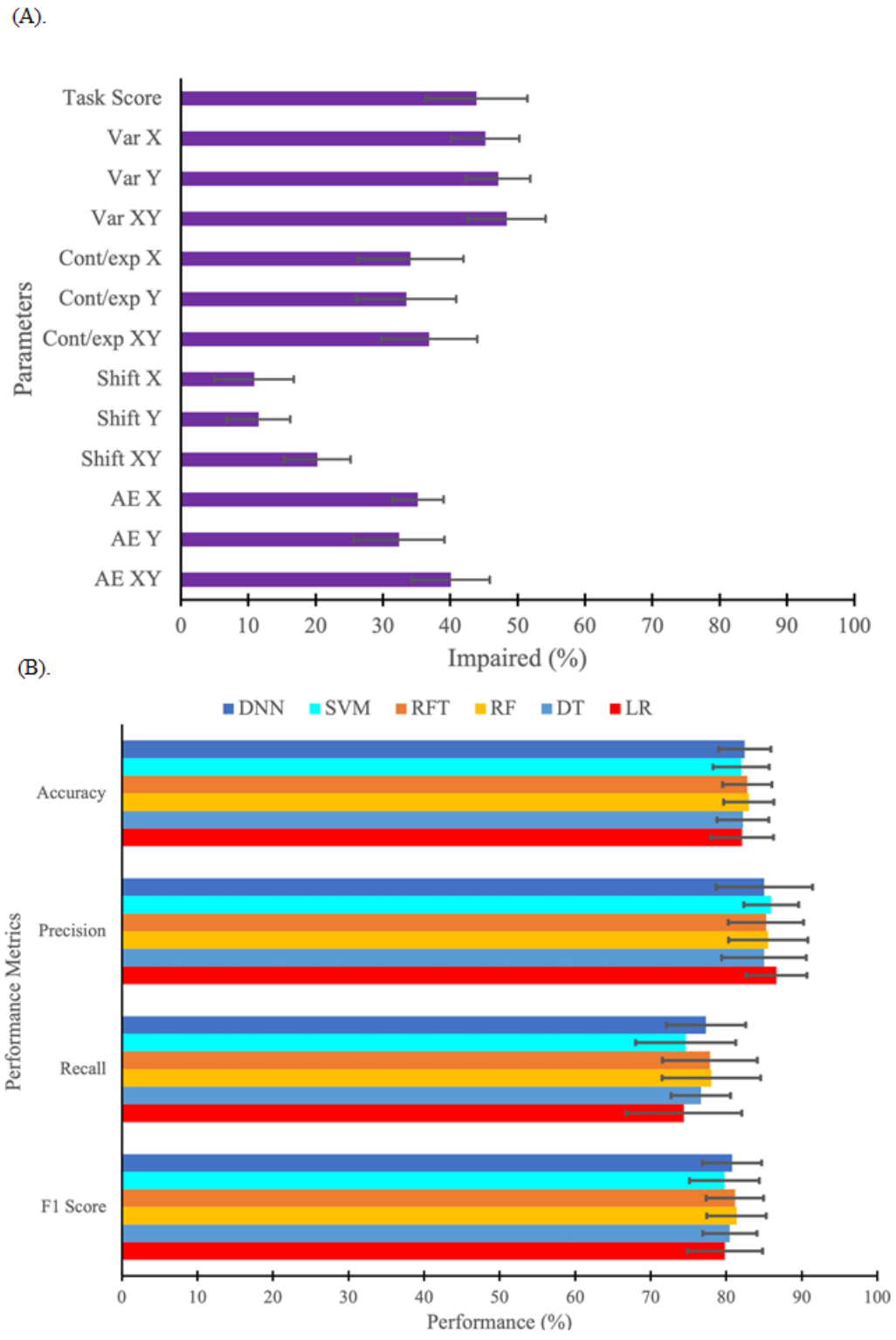

Figure 4

Comparison of the 95% cut-off score technique based on individual parameters and the mean and standard deviation of 10-fold cross-validation (CV) of performance metrics of machine learning and deep learning models.

A) Performance metrics when classified based on individual parameters (Var: variability, Cont/Exp: contraction/expansion, Shift, AE: absolute error) of arm position matching task, as well as overall task score to find the number of impaired participants.

B) Performance metrics (accuracy, precision, recall, and F1 score) are shown for the machine learning and deep learning models (LR: Logistic Regression, DT: Decision Tree, RF: Random Forest, RFT: Random Forest with Hyperparameters Tuning, SVM: Support Vector Machine, DNN: Deep Neural Network).

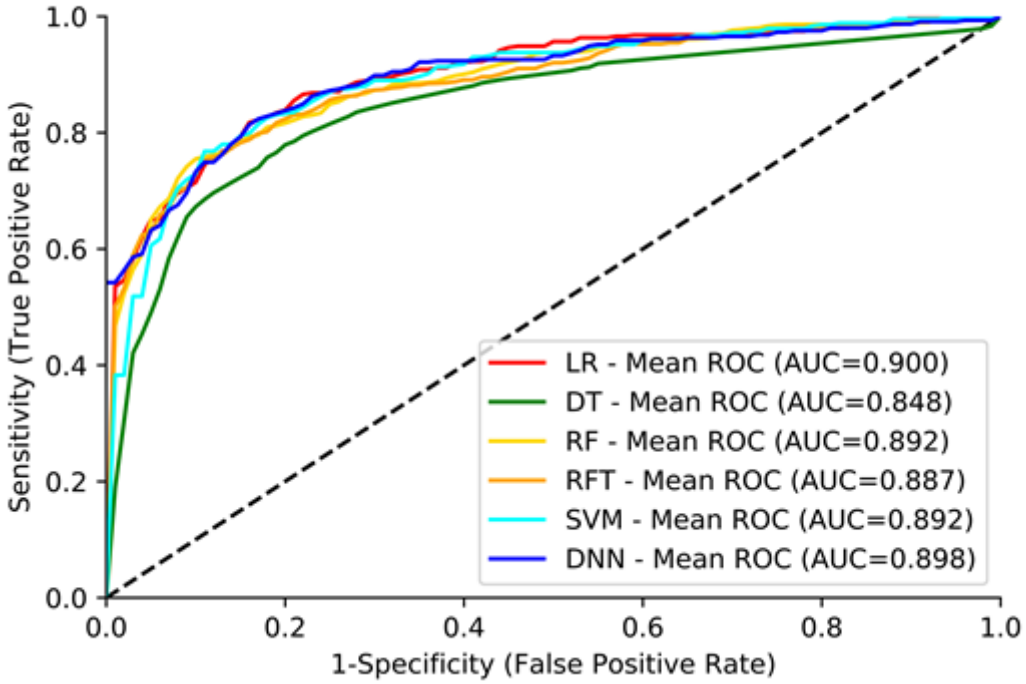

Figure 5

Mean of 10-fold cross-validation of the receiver operating characteristics (ROC) curve and area under the curve (AUC) for the classification performance of LR, DT, RF, RFT, SVM, and DNN models. The dashed line corresponds to classification due to random chance (AUC=0.5, i.e., 50% sensitivity and 50% specificity).

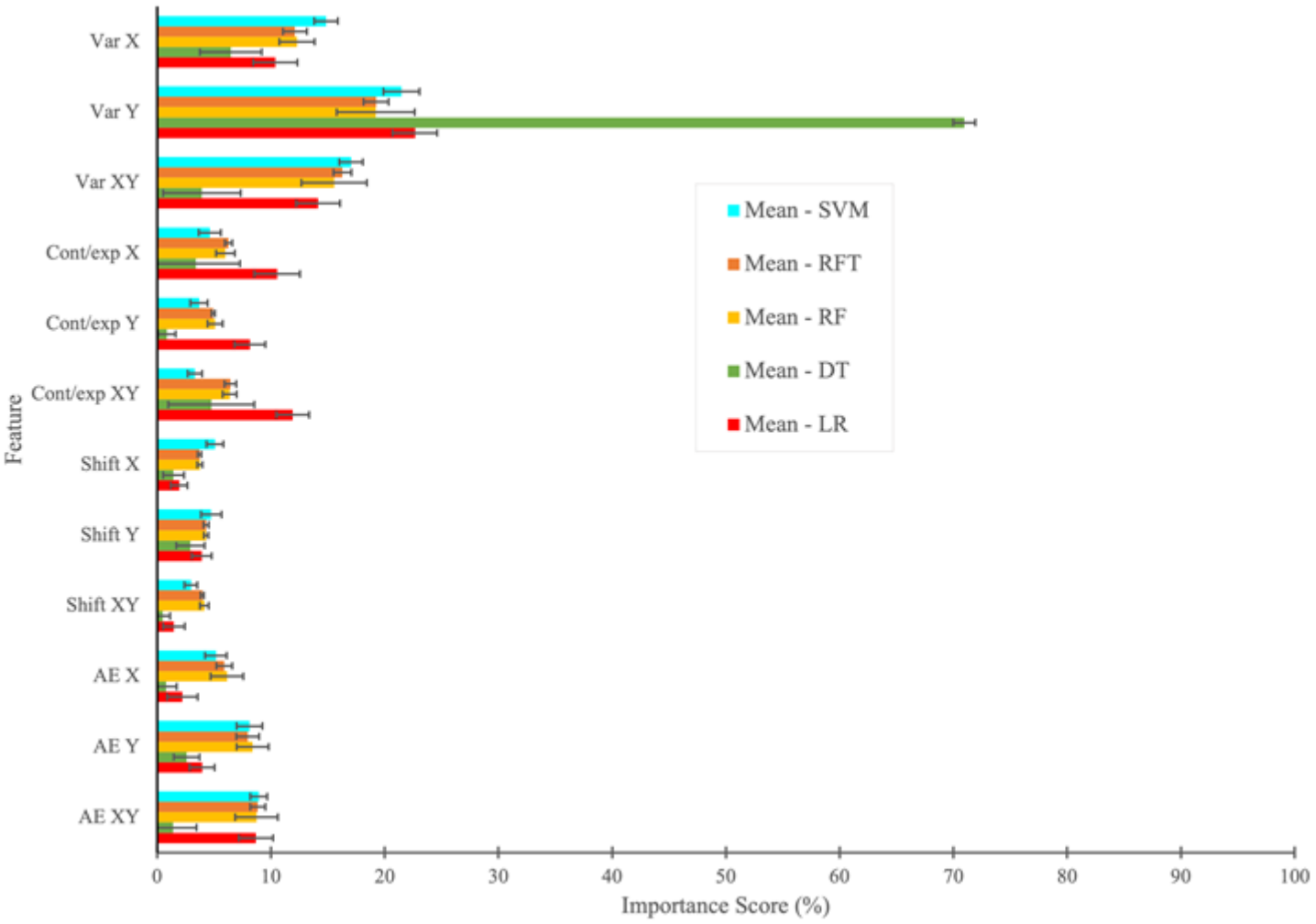

Figure 6

Mean and standard deviation of K-fold (K=10) cross-validation of feature importance based on individual parameters (Var, Cont/exp, Shift, and AE) obtained from LR, DT, RF, RFT, and SVM models. Due to the complex structure of Deep Neural Network (DNN) model, we could not plot the feature importance using DNN model.