Methods Article

The pragmatic classification of upper extremity motion in neurological patients: a primer

Avinash Parnandi[1,*], **Jasim Uddin**[2], **Dawn M. Nilsen**[2], **and Heidi Schambra**[1]

[1] NYU School of Medicine; (avinash.parnandi and heidi.schambra)@nyulangone.org

[2] Columbia University Medical Center; (ju2189 and dmn12)@cumc.columbia.edu

* Correspondence: Avinash Parnandi, avinash.parnandi@nyulangone.org

Received: date; Accepted: date; Published: date

Keywords: Machine learning algorithms; wearable sensors; inertial measurement unit; accelerometers, functional primitives; stroke rehabilitation

Abstract

Recent advances in wearable sensor technology and machine learning (ML) have allowed for the seamless and objective study of human motion in clinical applications, including Parkinson's disease and stroke. Using ML to identify salient patterns in sensor data has the potential for widespread application in neurological disorders, so understanding how to develop this approach for one's area of inquiry is vital. We previously proposed an approach that combined wearable inertial measurement units (IMUs) and ML to classify motions made by stroke patients. However, our approach had computational and practical limitations. We address these limitations here in the form of a primer, presenting how to optimize a sensor-ML approach for clinical implementation. First, we demonstrate how to identify the ML algorithm that maximizes classification performance and pragmatic implementation. Second, we demonstrate how to identify the motion capture approach that maximizes classification performance but reduces cost. We used previously collected motion data from chronic stroke patients wearing off-the-shelf IMUs during a rehabilitation-like activity. To identify the optimal ML algorithm, we compared the classification performance, computational complexity, and tuning requirements of four off-the-shelf algorithms. To identify the optimal motion capture approach, we compared the classification performance of various sensor configurations (number and location on the body) and sensor type (IMUs versus accelerometers). Of the algorithms tested, linear discriminant analysis had the highest classification performance, low computational complexity, and modest tuning requirements. Of the sensor configurations tested, seven sensors on the paretic arm and trunk led to the highest classification performance, and IMUs outperformed accelerometers. Overall, we present a refined sensor-ML approach that maximizes both classification performance and pragmatic implementation. In addition, with this primer, we showcase important considerations for appraising off-the-shelf algorithms and sensors for quantitative motion assessment.

---

1 Introduction

Wearable sensors, such as inertial measurement units (IMUs) and accelerometers, provide an opportunity for the objective and seamless capture of human motion. Machine learning (ML) enables computers to learn without being explicitly programmed, and provides an opportunity to rapidly identify patterns in data. Given recent technological and computational advances, combining wearable sensor data with ML algorithms has the potential for rapid, automated, and accurate classification of motion.

Researchers have begun using this combined sensor-ML approach in a number of applications. These include human activity recognition [1-3], gesture analysis [4], assessment of bradykinesia in Parkinson's disease [5], motor function assessment in multiple sclerosis [6], and differentiating between functional and non-functional arm usage in stroke patients [7, 8]. While many of these studies showcase the application of sensors and ML in clinical populations, no previous work has detailed the various hardware and software considerations for using the sensor-ML approach. Furthermore, no guide currently exists to advise investigators in building and troubleshooting this approach, which sits at the intersection of human movement science, data science, and neurology. With the potential for the sensor-ML approach to have widespread applicability to neurological disorders, understanding how to develop this approach for one's own area of inquiry is paramount.

One possible application of the combined sensor-ML approach is the monitoring of rehabilitation dose in stroke patients. Quantifying the dose of rehabilitation entails classifying units of measurement, which are subsequently tallied. In our previous proof-of-principle study, we used IMUs worn by stroke subjects performing a structured tabletop activity to capture motion data. Our units of measurement were functional primitives, elemental motions that cannot be further decomposed by a human observer. We applied an ML algorithm (hidden Markov model with logistic regression) to the IMU motion data to recognize functional primitives embedded in this activity, achieving an overall classification performance of 79% [9]. While promising, this sensor-ML approach had variable classification performance among the primitives (62-87% accuracy). It also did not address research implementation challenges such as the computational complexity and computational costs of the ML approach, or clinical implementation challenges such as the expense [10] and electromagnetic intolerance of the IMUs.

In the present study, we address these limitations in the form of a primer, outlining deliberations that researchers developing their own sensor-ML approach would need to consider. We describe our rationale and steps for identifying (1) an algorithm that is highly accurate but computationally tractable, and (2) the type and array of sensors that minimize cost but maximize accuracy. We use functional primitives as the motion type to be classified, and describe our approach for both capturing and identifying these motions. We also use off-the-shelf algorithms and sensors, providing an accessible framework for investigators seeking to address new research and clinical questions with the sensor-ML approach.

2 Methods

To demonstrate the steps in identifying the optimal ML algorithm and sensor array, we use data collected from previous work [9]. Briefly, six mild-to-moderately impaired stroke patients (Table 1) moved a toilet paper roll and aluminum can over a horizontal array of targets (Fig. 1).

Subjects performed 5 trials moving the object between a center target and eight radially arrayed targets (20 cm away). The task generates the following functional primitives: *reach* (to move into contact with a target object); *transport* (to convey a target object); *reposition* (to move proximate to a target object); and *idle* (to stand at the ready near target object). Functional primitives are discrete, object-oriented motions with a single goal. Functional primitives are non-divisible and are largely invariant across individuals [11], may be represented cortically [12-14], and provide a finer-grained capture of performance in stroke patients who may be unable to accomplish a full activity. Akin to words, functional primitives are combined to make a functional movement [15] (analogous to a sentence), which in turn are combined to make an activity (analogous to a paragraph) [16]. For example, a series of *reach-transport-reposition* primitives could constitute a functional movement for zipping up a jacket, within the activity of dressing.

Motion data were recorded with 11 IMUs (XSens Technology) worn on the head, sternum, pelvis, and bilateral hands, forearms, arms, and scapulae. 3D linear accelerations, 3D angular velocities, and quaternions were generated at 240 Hz. To segment and label the motion data as constituent primitives, we synchronously recorded motion (30 Hz) with a single video camera. Trained coders used the video recording to label the beginning and end of each functional primitive, which also labeled the corresponding IMU data. These labels served as the ground truth. This step enabled us to train ML algorithms on motion data and test their classification performance against the ground-truth labels. IMU data were z-score normalized and statistical features were extracted. The statistical features were the following: mean, standard deviation, minimum, maximum, entropy, skewness, energy, and root mean square. These statistical features have been shown to capture human motion efficiently, reducing the computational burden [17-19]. Following prior work, we selected a window size of 0.25 s sliding by 0.1 s [9], from which to derive the statistical features. The statistical feature data were fed to the ML algorithms.

The dataset consisted of 2881 primitives, consisting of 810 *reaches*, 708 *transports*, 781 *repositions*, and 582 *idles*. It is important to note that this is the sample size of interest (not the number of subjects). Accounting for repeated measures within-subject and at each target, and using this dataset of 2881 primitives with $\alpha = 0.05$, we have 81% power to detect a classification performance of at least 79% (positive predictive value, section 3.2.1 below). We used 79% accuracy as the benchmark for sufficient classification performance as achieved in our previous study [9].

3 Computational details

3.1 ML methods for classification

In the present study, we sought to identify an ML algorithm that performs well for identifying primitives, i.e. has a high classification performance, but that also is practical, i.e. has low computational overhead and minimal tuning requirements. Supervised ML algorithms work in two phases: training and testing. During training, ML algorithms learn the relationship between a pattern of data characteristics (here, the statistical features) and its class (here, its primitive label). During testing, the trained ML algorithm uses the pattern of data characteristics to identify a new data sample as one of the primitives. This identification is checked against the ground-truth human label, thus reading out classification performance.

We considered both generative and discriminative algorithms. Generative algorithms model the underlying distribution of data for each class, seeking to identify data characteristics that enable matching of new data samples to a given class. In contrast, discriminative algorithms model the boundaries between classes and not the data themselves. They seek to identify the plane separating the classes so that, based on location relative to the plane, a new data sample is assigned to the appropriate class.

We selected four algorithms that have been found to provide high classification performance in human activity recognition: linear discriminant analysis (LDA) [18], Naïve Bayes classifier (NBC) [17], support vector machine (SVM) [20], and k-nearest neighbors (KNN) [19]. LDA and NBC are generative algorithms, whereas SVM and KNN are discriminative algorithms. We used off-the-shelf versions of these algorithms without any special permutations; in other words, the algorithms are widely available in most machine learning libraries such as scikit-learn [21, 22].

3.2 Algorithm performance metrics

3.2.1 Classification performance of algorithms

We first evaluated how well the algorithms could classify primitives, measuring classification performance by comparing algorithm-chosen labels against ground-truth human labels. Primitives were classified as true positive (*TP*, labels agreed) and false positive (*FP*, labels disagreed). We used 60% of the data to train the algorithm and 40% to test it, repeating the process 10 times. A validation dataset was not used because we were not optimizing algorithm architectures, and the test dataset provides an unbiased estimate of algorithm performance. Data were randomly selected for each primitive proportional to its prevalence in the complete dataset (i.e., stratified proportional sampling). This ensured that each dataset adequately represented the entire sample population. In addition, to examine the possibility that within-subject dependencies in the training and testing sets leads to an overestimation of classification performance, we also performed a leave-one-subject-out analysis i.e., training the algorithms using data from all but one subject and testing its performance on the data from the remaining subjects. This process was repeated 6 times, once for each subject, and classification performances were averaged.

The first metric for classification performance was positive predictive value (PPV; *TP/(TP+FP)*). PPV reflects how often a primitive was actually performed when the algorithm labeled it as such; in other words, PPV is how often a primitive was correctly classified. We generated primitive-level PPVs in a one-versus-all analysis (e.g., *reach* vs. *transport* + *reposition* + *idle* combined). We also generated an overall PPV by combining data for all primitives and tallying all true and false positives. We prefer PPV because it takes into account the prevalence of the primitive in the dataset [23].

The second metric for classification performance was the receiver operating characteristic (ROC) curve. ROC curves depict the relative tradeoff between true positive rate (sensitivity; y-axis) and false positive rate (1-specificity; x-axis) and identify the optimal operating point of an algorithm [24]. Perfect classification would lead to a ROC curve that passes through the upper left corner, with an area under the ROC curve (AUC) equal to 1 and an operating point at 100% sensitivity and 100% specificity [24].

3.2.2 Practical performance of algorithms

We next considered the computational complexity of the algorithms in terms of their training and testing times and their tuning requirements. Having a high computational complexity means that specialized computing hardware and advanced expertise would be needed, potentially hindering widespread implementation in research.

3.2.2.1 Training and testing times of the algorithms

The time required to train and test the algorithms was measured for datasets of different sizes. If training time is fast, rapid appraisal and optimization of the algorithm are possible, favoring rapid development and deployment. If the testing time is fast, real-time classification and online feedback are possible, favoring clinical implementation.

We first used 20-100% of the dataset (n=2881 primitives) in randomly selected 10% increments. At each increment, we measured (1) the time required to train the algorithm (training time), and

(2) the time required for a trained algorithm to classify a primitive (testing time). At each 10% increment, the algorithms were trained *de novo* to avoid overfitting and to provide unbiased estimates.

Given the modest size of our dataset, we next used a simulated dataset that could be expected from a typical sample size of 50 subjects performing a variety of activities. The simulated dataset had 300,000 primitives with same proportion, mean, and variance as our original dataset. We used 25-100% of the dataset in randomly selected 25% increments. At each increment, we measured the training and testing times, training the algorithm *de novo* as above. Of note, the simulated dataset was used only to generate training and testing times, and was not used for classification performance assessments.

3.2.2.2 Tuning requirements of the algorithms

We also assessed the algorithm's need for tuning, the adjustment of algorithm parameters to maximize classification performance. A high tuning requirement requires the extensive analysis of the algorithm to identify its optimal parameters, potentially limiting implementation in settings that lack domain expertise. Of note, tuning requirements were only used to index complexity, but we did not tune the algorithms themselves in the assessment of classification performance.

We operationalized the algorithms' tuning requirements as the number of parameters that can be adjusted. We also qualitatively classified the level of domain knowledge required to implement and tune the algorithms. Based on typical US educational programs, "low" domain expertise indicates a basic knowledge of statistics, "medium" indicates undergraduate-level knowledge of machine learning, and "high" indicates graduate-level knowledge of machine learning.

3.2.3 Optimal sensor characteristics

We then focused on the hardware side, seeking the best balance between ease of motion capture and high classification performance. We first considered the use of IMUs compared to accelerometers alone. IMUs are a combination of sensors, including accelerometers, gyroscopes, and magnetometers. Many IMU hardware-software systems generate 3D linear accelerations, 3D angular velocities, 3D magnetic heading, and 4D quaternions, resulting in 10 data dimensions per sensor. We used accelerations, angular velocities, and quaternions for derivation of statistical features (section 2), as these data types have been used previously for human activity recognition [19, 25, 26]. In contrast, 3D accelerometers generate only 3D linear accelerations, resulting in 3 data dimensions per sensor.

While IMUs are data-rich, they are challenged by electromagnetic drift. Magnetic environments lead to potentially inaccurate gyroscopic measurements and therefore necessitate frequent recalibration. While accelerometers are data-sparse, they are largely unaffected by a magnetic environment.

Another practical consideration for sensor choice is system expense. IMU systems can cost on the order of thousands of dollars [10] whereas accelerometry systems cost in the hundreds [27]. It is possible that cost and set-up time could be optimized by reducing the number of sensors or by using accelerometers alone. Although simplified and less expensive motion capture would favor clinical implementation, it may come at the cost of reduced classification performance.

In this analysis, we subsampled data from the IMUs to extract accelerometry data, ensuring that comparisons were based on identical sensor locations and primitive motions. LDA was trained and tested on the separate datasets to read out effects on classification performance.

3.2.3.1 Optimal sensor number and configuration for classification

We first evaluated how the number of sensors and their location on the body affects classification performance. We used exhaustive search to systematically test all possible sensor configurations [28]. This approach provides an unbiased appraisal of all sensor combinations for each incremental reduction in sensor number.

3.2.3.2 Optimal sensor type for classification

We also evaluated how sensor type affected classification performance. We compared classification accuracies using IMU data versus accelerometry-only data. This allowed us to determine whether accelerometers, with their reduced dimensionality, could enable sufficient accuracy to warrant their use in lieu of IMUs.

4 Results

4.1 Classification performance of algorithms

We first determined the classification performance of multiple ML algorithms using PPVs (Table 2). LDA and SVM had high classification performance for all primitives (overall PPV 92.5% and 92%, respectively). KNN had intermediate performance (PPV 87.5%) and NBC had the lowest performance (PPV 80.2%), particularly for *reaches* (PPV 77%) and *transports* (PPV 71%). In the leave-one-subject-out analysis, which addressed the possibility of within-subject dependencies, similar overall classification performances were identified (PPVs of 89% for LDA, 90% for SVM, 83% for KNN, and 75% for NBC).

To further characterize classification performance, we generated ROC curves for each primitive (Fig. 2). All algorithms detected *idle* with high accuracy (AUC > 0.87). For the other primitives, LDA and SVM had AUCs 0.95-0.99, indicating very high classification performance. KNN also had high classification performance for *reach* (AUC 0.94) and *transport* (AUC 0.90) and intermediate classification performance for *reposition* (AUC 0.87). In contrast, NBC had the lowest classification performance on the remaining primitives (AUC 0.80-0.85). We also identified the optimal operating point, indicating the best tradeoff between sensitivity and specificity, for each algorithm (Fig. 2). At their respective optimal operating points, LDA and SVM achieved high sensitivities (0.83-0.95) and specificities (0.83-0.95) for all primitives. KNN achieved a high sensitivity (0.91) and specificity (0.86) for *transport*, but had moderate sensitivities (0.80-0.88) and specificities (0.79-0.86) for other primitives. NBC had the lowest sensitivities (0.74-0.81) and specificities (0.74-0.79) for all primitives. In sum, these findings indicate that LDA and SVM have the highest classification performance of the algorithms tested.

4.2 Training and testing times of the algorithms

We next evaluated the pragmatic aspects of implementing the algorithm to gauge real-world applicability. We first calculated the time required to train and test the algorithm on increasing quantities of data (Fig. 3) from our dataset of 2880 primitives. In terms of training times, NBC and LDA were on the order of seconds (12 s and 26 s, respectively), with training times growing linearly with increasing data quantity. SVM was on the order of minutes (5.6 min), with training times growing quadratically with increasing data quantity. KNN required no time to train as an inherent property of the model. In terms of testing, LDA, NBC, and SVM required sub-millisecond times (approximately 0.03 ms), whereas KNN required the longest time (1.5 ms) with testing times growing linearly with increasing dataset size.

To investigate the real-world ramifications of training and testing requirements, we generated a dataset with 300,000 primitives (Fig. 4). Training times became prohibitively long for SVM (up to 23 h) but were manageable for the other algorithms (up to 13 min). Testing time was relatively high for KNN (up to 2.3 min), whereas LDA, NBC, and SVM required nominal testing times (<0.03 ms). Given their consistently low training and testing times, LDA and NBC have the best practical performance of the algorithms tested.

4.3 Tuning requirements of the algorithms

To gauge the difficulty of algorithm implementation, we characterized their tuning requirements (Table 3). NBC has the lowest number of parameters (1) and requires a low amount of domain knowledge in machine learning to optimize it. KNN has a moderate number of parameters (5), but their optimization is reasonably intuitive and requires a low level of domain knowledge. LDA has fewer parameters (3), but they require a medium level of domain knowledge. SVM has many parameters (9) and requires a high level of domain knowledge to build an accurate and efficient model. In sum, these findings indicate that NBC and KNN are the easiest to implement, and LDA implementation requires a modestly higher skillset.

4.4 Optimal sensor characteristics

4.4.1 Optimal sensor number and configuration

To evaluate the effect of the sensor number and configuration on classification performance, we used an exhaustive search process, which evaluated all combinations of sensor number and location. We note that exhaustive search arrived at the same optimal configurations for IMUs as for accelerometers. Seven sensors on the head, sternum, pelvis, and UE of the active side resulted in the highest classification performance (IMU PPV 92.5%; accelerometer PPV 84%). In comparison, when exhaustive search progressively added sensors to the non-active forearm, then hand, then upper arm, then scapula, classification performance worsened (IMU PPV 88%; accelerometer PPV 80%) (Fig. 5). When exhaustive search progressively removed sensors on the trunk and then head, performance also worsened. Subsequent removal of sensors from the scapula, then arm, and then hand further worsened performance, arriving at PPVs of 71% and 62% for IMUs and accelerometers, respectively, for the remaining forearm sensor.

4.4.2 Optimal sensor type for classification

To finish, we evaluated classification performance using IMU versus accelerometry data only. Classification performance using accelerometry data was consistently lower than for IMU data for all sensor configurations (Fig. 5; Table 4). Classification performance with accelerometers was lower especially for *reaches* (PPV 77% vs. 93%; Table 4), which include different arm configurations to grasp the objects (e.g. supinating to side-grasp the aluminum can versus pronating to overhand grasp the toilet paper roll). These findings indicate that IMU data enable a superior level of classification, particularly with more variable motions involving forearm rotations.

5 Discussion

The combination of wearable sensors and machine learning offers exciting opportunities in numerous applications, including human activity recognition [1-3] and assessment of impaired motion [5, 7, 8]. We recently proposed an approach that uses wearable sensors and ML algorithms to classify functional primitives, which could be summed to quantify rehabilitation dose. In this study, we aimed to address limitations in this previous work, including a modest computational performance, high computational complexity, and hardware drawbacks. We present our analyses as a primer for considering software and hardware variables in the capture and classification of motion data. We sought to identify—from both performance and practical standpoints—the best machine learning algorithm, sensor configuration, and sensor type to classify functional primitives in stroke patients.

Among the ML algorithms, LDA represented the best balance of classification performance and pragmatic implementation. Among sensor configurations, seven sensors on the paretic arm and trunk enabled better classification performance than more or fewer sensors on the body. Among sensor types, IMU data enabled better classification performance than accelerometers. To our knowledge, this is the first study to systematically outline the steps of identifying optimal ML algorithms, sensor configurations, and sensor types to automatically classify motion patterns of neurological patients.

Optimal performer in classification. Evaluating the ability of the ML algorithms to classifying functional primitives, we found that LDA and SVM had the highest classification performance. LDA performs well because it aims to reduce dimensionality while preserving as much discriminatory information as possible. This approach leads to tight clusters and high separation between the classes [29]. SVM performs well because it projects training data to a high-dimensional space. This approach leads to maximal separation between classes that may not be possible in the original feature space [30]. Overall, LDA aims to find commonalities within and differences between data classes, whereas SVM aims to find a classification boundary that is furthest from the data classes. Importantly, these algorithms maximize rigor in the training phase by being less susceptible to noisy or outlier data [31, 32]. LDA accomplishes this by using the clusters' centers and ignoring outlier samples to classify [31], while SVM uses the most closely spaced data (i.e., the most difficult to discriminate) to define class boundaries [32]. It is worth noting that LDA assumes that the underlying classes are normally distributed (unimodal Gaussians) with the same covariance matrix [29]. If real-world motion data are significantly non-Gaussian, LDA may not capture the complex data structures required for accurate classification. In this case, classification performance can be tuned by allowing the covariance matrices among classes to vary, resulting in a regularized discriminant analysis [33].

By comparison, KNN showed a marginally lower classification performance, likely due to its susceptibility to noise [34]. KNN relies on the assumption that samples from the same class exist in close proximity. Given a new sample, KNN assigns it to the class with the majority of closest neighbors [35]. In our current setup of KNN, all nearest sample points are given the same weighting. Therefore when assigning a class label, a noisy sample will be weighted the same as other statistically important samples. KNN classification performance can be tuned by choosing an appropriate weighting metric (e.g., inverse squared weighing) [36], which ensures that samples closer to the test sample contribute more to classifying it. Performance may also be tuned by using mutual nearest neighbors, where noisy samples are detected using pseudo-neighbors (neighbors of neighbors) and are assigned lower weights [37].

Finally, NBC had the lowest performance compared to other algorithms. NBC uses Bayes' rule and prior information to classify a new sample, using the posterior probability of it belonging to a class [38]. Its lower performance may be attributed to its underlying assumption of conditional independence between data features [39]. This assumption is violated for data streams that are correlated, such as data from adjacent sensors on the body, like the hand and wrist. The performance of NBC could be improved by applying principal components analysis to the dataset as a pre-processing step, and then training the NBC [40].

Comparing these results with our prior work [9], we found that the four algorithms outperformed the hidden Markov model-logistic regression (HMM-LR) classifier for identifying the functional primitives in stroke patients. The improved performance may be due in part to differences in the training datasets. Our previous study trained the algorithm on healthy control data and tested on stroke patient data to examine the generalizability of the model. It is conceivable that if the HMM-LR classifier been trained and tested in stroke patients only, its performance would have been higher.

Optimal performer in practicality. We also determined the most pragmatic algorithms with respect to their training and testing times and their tuning requirements. In terms of training times, KNN did not have any computational overhead. This is expected, since KNN requires no training and shifts its computations to the testing phase. Training times for LDA and NBC grew gradually with dataset size, but took at most minutes with a real world-sized dataset. LDA had lower training times than NBC on a smaller dataset, but required more training time as the dataset increased. This is explained by the scatter matrix computations and optimization of LDA, which become computationally expensive as the dataset size increases [18]. By contrast, SVM training time increased quadratically with dataset size, because finding an optimal hyperplane between classes entails solving a quadratic programming problem [20]. Complex algorithms such as SVM thus require more processing time for large datasets, which limits real-world application. For example, for a modestly sized study, training times for SVM may be on the order of days. This lag would be prohibitive for rapid tuning, significantly delaying algorithm optimizations. Conversely, performance of LDA and NBC could be rapidly appraised after training, alerting an investigator to further tune the algorithm or to move on from it.

In terms of testing times, SVM, LDA, and NBC required sub-milliseconds to classify primitives, whereas KNN took seconds-minutes and testing times grew linearly with dataset size. This can be explained by the exhaustive and computationally expensive search performed by KNN [41]. During testing, the KNN algorithm searches for the k nearest neighbors that have similar data characteristics as the test sample. With increasing samples and dimensionality of the data, the search broadens and takes more time. If an investigator wishes to classify primitives offline, KNN testing times may be acceptable. For applications requiring near- or real-time classification (e.g. for online feedback), the other algorithms should be considered instead. Alternatively, the computational complexity of KNN can be reduced by selecting an efficient search algorithm (e.g., KD tree) [42], which limits the search space during testing.

In terms of ease of tuning to increase classification performance and reduce training/testing time, we determined that NBC had the lowest parameter complexity and requirement for domain knowledge in machine learning. KNN has a moderate number of tuning parameters, but they are relatively straightforward to understand and address. LDA has fewer tuning parameters than KNN, but moderate domain knowledge is required to select the amount of regularization allowing the covariance among classes to vary [33]. SVM requires the highest amount of parameter tuning, and necessitates a deep understanding of statistics, optimization, probability theory, and machine learning [43]. This level of domain knowledge is prohibitive for SVM use in an unsupported research setting.

Weighing classification performance and pragmatic implementation, we judged LDA to be the best choice for our application. Investigators will similarly need to weigh their performance goals, time resources, and available level of expertise for ML implementation in their own motion classification questions.

Optimal IMU configuration. On the hardware side, we determined the optimal sensor location and configuration to facilitate data capture while maintaining high classification performance. Seven sensors (not more or fewer) enabled optimal classification performance, and the best sensor configuration was placed on the active limb and trunk. This result is expected, given that the participants performed a unimanual task. Interestingly, accuracy worsened with more sensors, likely because of the increased dimensionality of the dataset. This may cause the ML algorithm to overfit the training data, resulting in lower classification performance during the testing phase [44]. Finally, we found that if only one sensor was available, the forearm location was the most informative, although classification performance was modest. This location is nonetheless appealing, given recent advances in smartwatches capable of capturing motion.

Optimal data characteristics. Finally, we determined the sensor type that led to the highest classification performance. Accelerometry data consistently generated lower accuracies than IMU data, likely due to its fewer dimensions. Although IMUs enable higher classification performance than accelerometers, they also have some drawbacks: a higher risk of electromagnetic drift leading to inaccurate data estimates and the need for more frequent recalibrations, a higher consumption of energy [45], and a higher cost [10]. Thus there is a tradeoff between robust motion capture and practical motion capture. We believe that the benefits of richer data and better classification of IMUs outweigh their practical limitations. However, there currently exist no benchmarks for the level of classification accuracy needed to justify clinical implementation. If these accuracy benchmarks are lower than those achieved by IMUs, and if investigators are constrained by financial resources or the magnetic noisiness of an environment, accelerometers could be appropriate.

5.1 Limitations and future work

Our study has some limitations to be considered. We importantly do not suggest that we have found the definitive approach for classifying primitives in rehabilitating stroke patients, for two reasons. First, our analysis was performed on a dataset of mild-to-moderately impaired stroke patients, limiting generalization to stroke patients with severe impairment. To achieve high classification performance across the range of stroke impairment, separately trained ML models may be needed for different impairment levels. Second, the activity used in this study was highly structured. The motion characteristics of the resulting primitives were thus more consistent and limited than what would be found in a real-world rehabilitation setting. The training and testing of algorithms on functional primitives with an array of kinematic characteristics is still required, and is ongoing in our laboratory. We use this circumscribed dataset here to showcase the practical deliberations required in the development of a sensor-ML approach for motion classification.

6 Conclusion

In summary, we present a primer that details how one can optimize both the software and hardware facets of motion capture. This work outlines computational and practical considerations for implementing a sensor-ML approach in quantitative research. Specific to our application, we demonstrate how to refine a strategy that builds towards the precise and pragmatic classification of functional primitives in stroke patients. We found that LDA had the best combination of classification performance and pragmatic performance. We also found that seven sensors on the paretic UE and trunk maximized classification performance, and that IMUs enabled superior classification compared to accelerometers.

7 Declarations

Ethics approval and consent to participate: Institutional Review Board-approved testing occurred at Columbia University. Subjects gave written informed consent to participate in this study, in accordance with the Declaration of Helsinki.

Consent for publication: Not applicable

Availability of data and material: The dataset analyzed for the current study are available from the corresponding author on reasonable request.

Pre-print: This manuscript has been released as an online preprint on arXiv (https://arxiv.org/abs/1902.08697)

Competing interests: All authors report no financial or non-financial competing interests.

Funding: The work was supported by K02NS104207 (HS) and K23NS078052 (HS).

Authors' contributions: AP analyzed and interpreted the data and wrote the paper. JU collected and coded the data. DN created the activities battery and interpreted the data. HS collected, labeled, and interpreted the data and wrote the paper.

Acknowledgements: We would like to acknowledge Dr. Sunil Agarwal for provision of the sensor system and Jorge Guerra for his early machine learning analysis. We thank Shatif Hughes and Dr. Hyoungseop Kim for their assistance with data labeling.

8 Figure legends

Figure 1. Tabletop activity set-up. Healthy individual wearing the sensors and transporting the object from center to a target.

Figure 2. Performance characteristics of machine learning algorithms for (A) Reach, (B) Transport, (C) Reposition, and (D) Idle. Receiver operating characteristic (ROC) curves show the trade-off between true positive rate (or sensitivity) and false positive rate (1-specificity). Curves closer to the top-left corner indicate a better classification performance. The optimal operating point for each algorithm (solid circles), reflect the best tradeoff between sensitivity and specificity for an algorithm. The area under the curve (AUC), a measure of classification performance, is shown in parenthesis for each algorithm. AUC=1 represents perfect classification. LDA had the highest AUCs followed closely by SVM, indicating high classification performances. NBC had consistently the lowest AUCs, indicating the weakest classification performance.

Figure 3. Algorithm (A) training times and (B) testing times on sample dataset. The dataset is comprised of 2880 primitives. We computed times to train and test each algorithm on 20-100% of the dataset in increments of 10%. To avoid overfitting and compute an unbiased estimate of training and testing times, ML algorithms were trained and tested de novo with each incremental increase. For training with the complete sample dataset, SVM required the most time (336 s) while the other algorithms finished training rapidly (<30 s). For testing, KNN required the most time (1.5 ms), while the other algorithms finished testing rapidly (<0.03 ms). Please note break in the y-axis to highlight the difference in the algorithm testing times.

Figure 4. Algorithm (A) training times and (B) testing times on real world-sized dataset. The dataset is comprised of 300,000 simulated primitives. We evaluated training and testing times required by each algorithm for quartile increases in dataset size. Please note the break in the y-axes to highlight differences in training and testing times. To avoid overfitting and compute unbiased estimates, the algorithms were trained and tested de novo at each quartile. For training with the entire dataset, SVM required the most time (1380 min) while the other algorithms required less time (LDA: 13 min; NBC: 2.5 min; KNN: 0 min, as per model property). For testing, KNN required the most time (2.3 min). The remainder of algorithms (LDA, NBC, and SVM) needed a testing time of <0.09 ms, which grew marginally with increasing sample sizes.

Figure 5. Classification performance for full and reduced sensor counts. Performance was computed using LDA and data from with progressively reduced sensor counts. Seven sensors (pelvis, sternum, head, and the active shoulder, upper arm, forearm, and hand) gave the best classification performance, with a performance drop-off at more or fewer sensors. IMU data consistently supported higher classification than accelerometery data, achieving PPV 92.5% vs. 82% at the seven sensors.

9 References

[1] R. J. Lemmens, Y. J. Janssen-Potten, A. A. Timmermans, R. J. Smeets, and H. A. Seelen, "Recognizing complex upper extremity activities using body worn sensors," *PLoS One,* vol. 10, no. 3, p. e0118642, 2015.

[2] D. Biswas *et al.*, "Recognition of elementary arm movements using orientation of a tri-axial accelerometer located near the wrist," *Physiol Meas,* vol. 35, no. 9, pp. 1751-68, Sep 2014.

[3] N. A. Capela, E. D. Lemaire, and N. Baddour, "Feature selection for wearable smartphone-based human activity recognition with able bodied, elderly, and stroke patients," *PLoS One,* vol. 10, no. 4, p. e0124414, 2015.

[4] H. Junker, O. Amft, P. Lukowicz, and G. J. P. R. Tröster, "Gesture spotting with body-worn inertial sensors to detect user activities," vol. 41, no. 6, pp. 2010-2024, 2008.

[5] B. M. Eskofier *et al.*, "Recent machine learning advancements in sensor-based mobility analysis: Deep learning for Parkinson's disease assessment," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2016, pp. 655-658: IEEE.

[6] I. Carpinella, D. Cattaneo, and M. Ferrarin, "Quantitative assessment of upper limb motor function in Multiple Sclerosis using an instrumented Action Research Arm Test," *Journal of neuroengineering rehabilitation,* vol. 11, no. 1, p. 67, 2014.

[7] E. M. Bochniewicz, G. Emmer, A. McLeod, J. Barth, A. W. Dromerick, and P. Lum, "Measuring Functional Arm Movement after Stroke Using a Single Wrist-Worn Sensor and Machine Learning," *Journal of Stroke and Cerebrovascular Diseases,* vol. 26, no. 12, pp. 2880-2887, 2017.

[8] K. Leuenberger *et al.*, "A method to qualitatively assess arm use in stroke survivors in the home environment," *Medical & Biological Engineering & Computing,* vol. 55, no. 1, pp. 141-150, 2017.

[9] J. Guerra *et al.*, "Capture, learning, and classification of upper extremity movement primitives in healthy controls and stroke patients," (in eng), *IEEE Int Conf Rehabil Robot,* vol. 2017, pp. 547-554, Jul 2017.

[10] *Analog Devices Inertial Measurement Units*. Available: https://www.mouser.com/ProductDetail/Analog-Devices/ADIS16497-3BMLZ?qs=sGAEpiMZZMt8xBNXJng%252BZAmSdp7Aftz5BZDzjPI1t2Q%2FH2F5lOOUxA%3D%3D
[11] G. Sumbre, G. Fiorito, T. Flash, and B. Hochner, "Neurobiology: motor control of flexible octopus arms," *Nature,* vol. 433, no. 7026, pp. 595-6, Feb 10 2005.
[12] M. Desmurget *et al.*, "Neural representations of ethologically relevant hand/mouth synergies in the human precentral gyrus," *Proc Natl Acad Sci U S A,* vol. 111, no. 15, pp. 5718-22, Apr 15 2014.
[13] M. S. A. Graziano, "Ethological Action Maps: A Paradigm Shift for the Motor Cortex," *Trends Cogn Sci,* vol. 20, no. 2, pp. 121-132, Feb 2016.
[14] F. A. Mussa-Ivaldi and S. A. Solla, "Neural primitives for motion control," *IEEE Journal of Oceanic Engineering,* vol. 29, no. 3, pp. 640-650, 2004.
[15] C. E. Lang, D. F. Edwards, R. L. Birkenmeier, and A. W. Dromerick, "Estimating minimal clinically important differences of upper-extremity measures early after stroke," *Arch Phys Med Rehabil,* vol. 89, no. 9, pp. 1693-700, Sep 2008.
[16] C. Bregler, "Learning and recognizing human dynamics in video sequences," in *IEEE Conference on Computer Vision and Pattern Recognition*, 1997, pp. 568-574.
[17] L. Bao and S. Intille, "Activity recognition from user-annotated acceleration data," *Pervasive computing,* pp. 1-17, 2004.
[18] A. M. Khan, Y.-K. Lee, S. Y. Lee, and T.-S. Kim, "A triaxial accelerometer-based physical-activity recognition via augmented-signal features and a hierarchical recognizer," *IEEE transactions on information technology in biomedicine,* vol. 14, no. 5, pp. 1166-1172, 2010.
[19] N. Ravi, N. Dandekar, P. Mysore, and M. L. Littman, "Activity recognition from accelerometer data," in *Aaai*, 2005, vol. 5, no. 2005, pp. 1541-1546.
[20] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, 2004, vol. 3, pp. 32-36: IEEE.
[21] Mathworks. (2018). *Analyze and model data using statistics and machine learning*. Available: https://www.mathworks.com/products/statistics.html
[22] (2018). *scikit-learn Machine Learning in Python*. Available: http://scikit-learn.org/stable/
[23] R. Parikh, A. Mathai, S. Parikh, G. C. Sekhar, and R. J. I. j. o. o. Thomas, "Understanding and using sensitivity, specificity and predictive values," vol. 56, no. 1, p. 45, 2008.
[24] B. JC, "Statistical Analysis and Presentation of Data," in *Evidence-Based Laboratory Medicine; Principles, Practice and Outcomes*, C. Price and R. Christenson, Eds. 2 ed. Washington DC, USA: AACC Press, 2007, pp. 113-40.
[25] S. Ha and S. Choi, "Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors," in *2016 International Joint Conference on Neural Networks (IJCNN)*, 2016, pp. 381-388: IEEE.
[26] Y. Shin, W. Choi, and T. Shin, "Physical activity recognition based on rotated acceleration data using quaternion in sedentary behavior: A preliminary study," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2014, pp. 4976-4978: IEEE.
[27] *Analog Devices Accelerometers*. Available: https://www.mouser.com/ProductDetail/Analog-Devices/ADXL357BEZ?qs=sGAEpiMZZMvwE4h8i4g3cg1FWtTpVweBl4PgoJiZDJ%2FRVLPerCg45w%3D%3D
[28] E. W. Weisstein. *Exhaustive Search. From MathWorld--A Wolfram Web Resource.* Available: http://mathworld.wolfram.com/ExhaustiveSearch.html
[29] A. J. Izenman, "Linear discriminant analysis," in *Modern multivariate statistical techniques*: Springer, 2013, pp. 237-280.
[30] C. Burges, "A tutorial on support vector machines for pattern recognition," *Journal Data mining knowledge discovery,* vol. 2, no. 2, pp. 121-167, 1998.
[31] W. Zhao, R. Chellappa, and N. Nandhakumar, "Empirical performance analysis of linear discriminant classifiers," in *Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No. 98CB36231)*, 1998, pp. 164-169: IEEE.
[32] T. B. Trafalis and S. A. Alwazzi, "Support vector machine classification with noisy data: a second order cone programming approach," *J International Journal of General Systems,* vol. 39, no. 7, pp. 757-781, 2010.
[33] J. H. Friedman, "Regularized discriminant analysis," *Journal of the American statistical association,* vol. 84, no. 405, pp. 165-175, 1989.

[34] C. M. Van der Walt and E. Barnard, "Data characteristics that determine classifier performance," 2006.
[35] L. E. Peterson, "K-nearest neighbor," *Scholarpedia,* vol. 4, no. 2, p. 1883, 2009.
[36] D. Bridge, "Classification: k Nearest Neighbours," ed, 2007.
[37] H. Liu and S. Zhang, "Noisy data elimination using mutual k-nearest neighbor for classification mining," *Journal of Systems Software,* vol. 85, no. 5, pp. 1067-1074, 2012.
[38] K. P. Murphy, "Naive bayes classifiers," *University of British Columbia,* vol. 18, 2006.
[39] M. Tom, "Generative and discriminative classifiers: Naive bayes and logistic regression," ed, 2005.
[40] L. Fan and K. L. Poh, "A comparative study of PCA, ICA and class-conditional ICA for naïve bayes classifier," in *International Work-Conference on Artificial Neural Networks*, 2007, pp. 16-22: Springer.
[41] P. Cunningham and S. J. J. M. C. S. Delany, "k-Nearest neighbour classifiers," vol. 34, no. 8, pp. 1-17, 2007.
[42] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Communications of the ACM,* vol. 18, no. 9, pp. 509-517, 1975.
[43] B. Scholkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.
[44] O. M. Giggins, K. T. Sweeney, and B. Caulfield, "Rehabilitation exercise assessment using inertial sensors: a cross-sectional analytical study," *J Neuroeng Rehabil,* vol. 11, p. 158, Nov 27 2014.
[45] Q. Liu *et al.*, "Gazelle: Energy-efficient wearable analysis for running," no. 9, pp. 2531-2544, 2017.

10 Table legends

Table 1. Demographic and clinical characteristics of patients. Shown are number of participants, mean age (range), gender, race, hand dominance, paretic side, mean Fugl-Meyer assessment score at first assessment (range; maximum 66), and time since stroke (range). Inclusion criteria were age ≥18 years; premorbid right-hand dominance; unilateral motor stroke; contralateral arm weakness with Medical Research Council score <5/5 in a major muscle group. Exclusion criteria were traumatic brain injury; musculoskeletal, medical, or non-stroke neurological condition interfering with assessment of motor function; contracture at shoulder, elbow, or wrist; moderate dysmetria or truncal ataxia; visuospatial neglect; apraxia; global inattention; blindness.

Table 2. Classification performance of machine learning algorithms for functional primitives. Positive predictive values (PPV) with associated 95% confidence intervals are shown. PPV reflects how often a primitive was actually made when the algorithm identified it as such, was calculated for the primitives of reach, transport, reposition, and idle. Primitive-level PPVs were computed in one-versus-all analysis (e.g., reach vs. transport + reposition + idle combined). The overall PPV was assessed by combining data for all primitives and tallying all true and false positives. Overall classification performance was highest for linear discriminant analysis (LDA) and support vector machine (SVM), moderately high for k-nearest neighbors (KNN), and lowest for Naïve Bayes classifier (NBC).

Table 3. Complexity of algorithm implementation. Algorithm parameter tuning is necessary to achieve optimal classification performance. Shown are algorithm tuning characteristics, as indicated by number and specifics of the tuning parameters. Also shown is a graded estimate of the level of domain knowledge required to tune these parameters. NBC is considered the simplest to tune while SVM is the most difficult. LDA has a handful of parameters that require medium domain knowledge to negotiate. KNN has a moderate number of parameters that are intuitive to tune and require little domain knowledge. Level of domain knowledge: low, basic knowledge of statistics; medium, undergraduate-level knowledge of ML; high, graduate-level knowledge of ML.

Table 4. Primitive-level classification using IMU or accelerometer data. Classification performance is shown using the 7-sensor configuration (pelvis, sternum, head, and the active

shoulder, upper arm, forearm, and hand). Accelerometers provided systematically poorer classification performance compared to IMUs across all primitives. Classification performance using accelerometry data was particularly low for reach (PPV 77%) and relatively higher for idle (PPV 88%).

11 Tables

N	**6**
Age (years)	61.7 (46.5 -71.0)
Gender (Female/Male)	2F/4M
Dominant arm (Right/Left)	5R/1L
Paretic side (Right/Left)	6R
Impairment (Fugl-Meyer score)	52.8 (45 - 62)
Time since stroke (years)	12.0 (2.0 - 31.1)

Table 1. Demographic and clinical characteristics of patients. Shown are number of participants, mean age (range), gender, race, hand dominance, paretic side, mean Fugl-Meyer assessment score at first assessment (range; maximum 66), and time since stroke (range). Inclusion criteria were age ≥18 years; premorbid right-hand dominance; unilateral motor stroke; contralateral arm weakness with Medical Research Council score <5/5 in a major muscle group. Exclusion criteria were traumatic brain injury; musculoskeletal, medical, or non-stroke neurological condition interfering with assessment of motor function; contracture at shoulder, elbow, or wrist; moderate dysmetria or truncal ataxia; visuospatial neglect; apraxia; global inattention; blindness.

Algorithm	PPVs for functional primitives				Overall PPV
	Reach	Transport	Reposition	Idle	
LDA	93 ± 1.47%	91 ± 1.65%	93 ± 1.47%	92 ± 1.56%	92.5 ± 1.52%
NBC	77 ± 2.42%	71 ± 2.61%	83 ± 2.16%	85 ± 2.06%	80.2 ± 2.30%
SVM	92 ± 1.56%	90 ± 1.73%	92 ± 1.56%	93 ± 1.47%	92 ± 1.56%
KNN	86 ± 2.00%	87 ± 1.94%	85 ± 2.06%	89 ± 1.80%	87.5 ± 1.90%

Table 2. Classification performance of machine learning algorithms for functional primitives. Positive predictive values (PPV) with associated 95% confidence intervals are shown. PPV reflects how often a primitive was actually made when the algorithm identified it as such, was calculated for the primitives of *reach*, *transport*, *reposition*, and *idle*. Primitive-level PPVs were computed in one-versus-all analysis (e.g., *reach* vs. *transport* + *reposition* + *idle* combined). The overall PPV was assessed by combining data for all primitives and tallying all true and false positives. Overall classification performance was highest for linear discriminant analysis (LDA) and support vector machine (SVM), moderately high for k-nearest neighbors (KNN), and lowest for Naïve Bayes classifier (NBC).

Algorithm	# tuning parameters	Tuning parameters	Level of domain knowledge
LDA	3	Prior probability, regularization term, optimizer	Medium
NBC	1	selection of prior distribution	Low
SVM	9	Kernel function, kernel parameters (scale, offset), regularization term, # of iterations, Nu, prior probability, convergence parameter, optimizer	High
KNN	5	# of neighbors (K), distance metric, search algorithm, tie breaker, weighing criterion	Low

Table 3. Complexity of algorithm implementation. Algorithm parameter tuning is necessary to achieve optimal classification performance. Shown are algorithm tuning characteristics, as indicated by number and specifics of the tuning parameters. Also shown is a graded estimate of the level of domain knowledge required to tune these parameters. NBC is considered the simplest to tune while SVM is the most difficult. LDA has a handful of parameters that require medium domain knowledge to negotiate. KNN has a moderate number of parameters that are intuitive to tune and require little domain knowledge. Level of domain knowledge: low, basic knowledge of statistics; medium, undergraduate-level knowledge of ML; high, graduate-level knowledge of ML.

Primitives	Classification performance (PPV)	
	IMU	Accelerometer
Reach	93%	77%
Transport	91%	80%
Reposition	93%	82%
Idle	92%	88%
Average	92.5%	82%

Table 4. Primitive-level classification using IMU or accelerometer data. Classification performance is shown using the 7-sensor configuration (pelvis, sternum, head, and the active shoulder, upper arm, forearm, and hand). Accelerometers provided systematically poorer classification performance compared to IMUs across all primitives. Classification performance using accelerometry data was particularly low for *reach* (PPV 77%) and relatively higher for *idle* (PPV 88%).