HHS Public Access

Author manuscript

J Biomech. Author manuscript; available in PMC 2019 November 26.

Published in final edited form as:

J Biomech. 2018 November 16; 81: 1–11. doi:10.1016/j.jbiomech.2018.09.009.

Machine learning in human movement biomechanics: Best practices, common pitfalls, and new opportunities

Eni Halilaj[a,*], **Apoorva Rajagopal**[b], **Madalina Fiterau**[c], **Jennifer L. Hicks**[d], **Trevor J. Hastie**[e,f], **Scott L. Delp**[b,d,g]

[a]Department of Mechanical Engineering, Carnegie Mellon University, United States

[b]Department of Mechanical Engineering, Stanford University, United States

[c]Department of Computer Science, Stanford University, United States

[d]Department of Bioengineering, Stanford University, United States

[e]Department of Statistics, Stanford University, United States

[f]Department of Health Research and Policy, Stanford University, United States

[g]Department of Orthopaedic Surgery, Stanford University, United States

Abstract

Traditional laboratory experiments, rehabilitation clinics, and wearable sensors offer biomechanists a wealth of data on healthy and pathological movement. To harness the power of these data and make research more efficient, modern machine learning techniques are starting to complement traditional statistical tools. This survey summarizes the current usage of machine learning methods in human movement biomechanics and highlights best practices that will enable critical evaluation of the literature. We carried out a PubMed/Medline database search for original research articles that used machine learning to study movement biomechanics in patients with musculoskeletal and neuromuscular diseases. Most studies that met our inclusion criteria focused on classifying pathological movement, predicting risk of developing a disease, estimating the effect of an intervention, or automatically recognizing activities to facilitate out-of-clinic patient monitoring. We found that research studies build and evaluate models inconsistently, which motivated our discussion of best practices. We provide recommendations for training and evaluating machine learning models and discuss the potential of several underutilized approaches, such as deep learning, to generate new knowledge about human movement. We believe that cross-training biomechanists in data science and a cultural shift toward sharing of data and tools are essential to maximize the impact of biomechanics research.

Keywords

Machine learning; Data science; Musculoskeletal; Neuromuscular

[*]Corresponding author at: 5000 Forbes Avenue, Pittsburgh, PA 15213, United States. ehalilaj@andrew.cmu.edu (E. Halilaj).

1. Introduction

For most of the 20th century, inference in biomedical research was predominantly based on hypothesis testing using parametric tests, such as the Student's *t* test. The current surge of data, however, presents new challenges and opportunities that are shifting the data analytics landscape in many biomedical disciplines, including human movement biomechanics. Data characterizing human movement are high-dimensional, heterogeneous, and growing in volume with wearable sensing; often, they do not satisfy assumptions associated with parametric tests. Advanced analytical techniques to extract informative features from these data and model underlying relationships that cannot be modeled with traditional statistical tools could transform biomechanics research, as they have autonomous driving, speech recognition, and automated cancer detection.

Efforts to modernize biomechanical data analysis are exemplified by the use of feature extraction algorithms such as principal component analysis (PCA). The literature reflects an evolving awareness about the drawbacks of using only summary metrics (e.g., mean acceleration) or salient features (e.g., the peak knee adduction moment) to describe gait data, as summary metrics are not always the most informative with respect to outcomes of interest (e.g., disease status). PCA, which preserves the variability of multivariate datasets while reducing dimensionality to make analyses more tractable, has been used as an alternative (Deluzio and Astephen, 2007; Donoghue et al., 2008; Duhamel et al., 2006; Ryan et al., 2006). Although most biomechanics studies that employ these methods for dimensionality reduction continue to analyze the reduced data with traditional statistical tools, biomechanists are now also considering new problem formulations in which features extracted using PCA are used as inputs in machine learning models.

Two machine learning approaches, predictive modeling and data mining, serve different purposes than traditional inferential statistics. Predictive modeling is concerned with finding a function that optimally maps input data (e.g., kinematic waveforms) to a given output (e.g., disease status) with the goal of making accurate predictions in the future. One example of predictive modeling in biomechanics is myoelectric control of prostheses, where models are trained to recognize an individual's intention based on myoelectric signals and the predicted intention is used to control the prosthesis (Oskoei and Hu, 2008). More recent efforts have centered around diagnostic and prognostic predictive models for neuromuscular and musculoskeletal pathologies (e.g., Schwartz et al., 2013), fall prediction (e.g., Wei et al., 2017), activity recognition to facilitate out-of-clinic patient monitoring (e.g., Biswas et al., 2015), and event detection to guide interventions such as deep brain stimulation (e.g., Pérez-López et al., 2016). The goal of data mining, on the other hand, is to discover new patterns in the data. Using clustering methods to identify subpopulations that exhibit different types of pathological gaits is an example of data mining (e.g., Rozumalski and Schwartz, 2009).

While applications of machine learning methods are expanding in movement biomechanics, critical evaluation of studies that apply them remains difficult. Machine learning approaches differ from the traditional statistical tools that biomechanists are trained to apply and interpret based on established reporting standards (e.g., *p* value for statistical significance). As the field becomes more data-intense and the use of machine learning continues to

increase, good practices for conducting and reporting research at the intersection of biomechanics and machine learning are needed to ensure that conclusions are valid and reproducible. A discussion of this topic will also enable researchers to develop an intuition for the types of problems that machine learning can address more successfully than traditional statistics. Accordingly, the goal of this survey is to make machine learning efforts more visible and propose standards to increase the quality and impact of future research in this exciting area. To achieve this goal, we first review applications of machine learning that focus on neuromuscular and musculoskeletal diseases. We outline best practices for reporting the results of these analyses and common pitfalls we encountered in the literature. Finally, we offer suggestions for overcoming some of the challenges facing biomechanical data analytics and highlight opportunities where emerging techniques are likely to have great impact in upcoming years. Key terms are defined in Appendix A and our most important recommendations are summarized in the Conclusions section.

2. Methods

2.1. Literature search approach

We carried out a search for original research articles published up to December 31, 2017 using the PubMed/Medline database (1946-). Our search identified articles that used machine learning methods to study human movement biomechanics and was limited to studies of common musculoskeletal and neuromuscular diseases affecting mobility. We used search terms from three different categories to identify relevant studies: (1) movement biomechanics terms, such as gait, kinematics, and kinetics; (2) machine learning terms, such as support vector machines, neural network, and principal component analysis; (3) terms describing musculoskeletal and neuromuscular conditions, such as osteoarthritis, cerebral palsy, and stroke (Table 1). At least one term from each of these three categories had to appear in the title or abstract for the article to be considered.

2.2. Exclusion criteria

One of the authors reviewed all the titles and abstracts of articles retrieved from the database search. We excluded the following: dissertations, conference proceedings, conference abstracts, non-English articles, studies not involving human subjects, studies whose primary outcome was not one of musculoskeletal or neuromuscular function, and studies that utilized PCA for dimensionality reduction followed by traditional inferential statistics. We also excluded studies whose primary outcome or assessment of motor function was based on questionnaires, studies focusing on athletic performance that did not include injured patients, and studies focusing on myoelectric control of prostheses, because the use of machine learning in this area had been previously reviewed (Oskoei and Hu, 2008). A second author read the abstracts of the included articles to ensure that they met the inclusion criteria.

2.3. Assessed outcomes

The included articles were divided amongst four of the authors and the following information was extracted from each article: clinical condition studied, data sources (e.g., optical motion capture, electromyography, wearable devices), number of subjects, machine learning task (i.e., clustering, regression, classification), algorithms used, data shape (sample

size × number of features), data purity (noise, bias, missing data, single or multiple source, inconsistencies in data collection protocol across samples), incorporation of covariates (e.g., age, sex), feature engineering approach (e.g., domain knowledge, PCA), feature selection or additional dimensionality reduction methods (e.g., forward feature selection), use of validation data for model selection and hyper-parameter tuning, evaluation of performance (use of test data and performance metrics reported), efforts to interpret the resulting model (e.g., interpretation of clusters or predictive features), and overall strengths and weaknesses. In addition to reporting these data in the Results section, we used them along with other observations made during the review to guide the discussion and recommendations of best practices.

3. Results

Our search yielded 3193 research articles, out of which 129, dating from 1996 to 2017, satisfied the inclusion criteria (Fig. 1A; Supplementary Table 1). The majority of studies focused on predictive tasks—classification (80.6%) and regression (11.6%)—while a few focused on data mining, in particular clustering tasks (7.8%). The most used algorithms were support vector machines, artificial neural networks, and generalized linear models (linear or logistic regression) for predictive modeling and k-means clustering for data mining (Fig. 1B). The number of subjects used in these studies varied from 4 to 2956, with the median being 40 (Fig. 1C). Movement data from wearable sensors were the most common, followed by data from traditional motion capture systems (Fig. 1D). Three studies used data collected in natural environments (Punt et al., 2016a, 2016b; Raknim and Lan, 2016; Fisher et al., 2016), while eight others used data collected outside of the laboratory, but in a controlled environment and for a short period of time (10 min–2.5 h) (Bochniewicz et al., 2017; Keijsers et al., 2003a, 2003b; Laudanski et al., 2015; O'Brien et al., 2017; Rodriguez-Martin et al., 2017; Samà et al., 2017; Yu et al., 2016). The most common subject populations were patients with Parkinson's disease, cerebral palsy, spinal cord injury, osteoarthritis, running injuries, and stroke (Fig. 1E).

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jbiomech.2018.09.009.

3.1. Areas of application

The most common area of application was classification of movement patterns, with many studies focusing on distinguishing pathological kinematics from normal kinematics. A lack of quantitative methods to assess motor symptoms associated with Parkinson's disease has motivated the development of predictive models for automated detection of Parkinsonian gait, bradikynesia, tremor, freezing of gait, and medication-induced dyskinesia using data from accelerometers (Palmerini et al., 2011), inertial measurement units (IMUs) (Biswas et al., 2015; Buchman et al., 2014), pressure sensors (Alam et al., 2017), and digital handwriting tools (Drotár et al., 2015). Other diagnostic classifiers focused on detecting cerebral palsy in infants using data from videos (Rahmati et al., 2016) and in children using data from traditional motion capture systems (Kamruzzaman and Begg, 2006), determining cerebral palsy severity using motion capture data (Rozumalski and Schwartz, 2009; Sagawa

et al., 2013; Trost et al., 2016; Zwick et al., 2004), characterizing movement patterns in post-stroke patients to inform targeted therapy (Huang and Patton, 2016; Kaczmarczyk et al., 2012, 2009; Krebs et al., 2014; Yu et al., 2016), and diagnosing osteoarthritis (Laroche et al., 2014; Levinger et al., 2009; Moustakidis et al., 2010). A few identified gait features that are discriminative of osteoarthritis status (Astephen et al., 2008; Magalhães et al., 2013; Meyer et al., 2016, 2015; Nüesch et al., 2012) or differences in muscle function between impaired and able-bodied individuals (Kuntze et al., 2015; Liu et al., 2014; Nair et al., 2010). A different set of studies focused on automated recognition of daily activities to facilitate out-of-clinic patient monitoring during rehabilitation (Biswas et al., 2015; Capela et al., 2016; Fulk et al., 2012; Fulk and Sazonov, 2011; Laudanski et al., 2015; Masse et al., 2016; O'Brien et al., 2017; Redfield et al., 2013; Roy et al., 2009; Vakanski et al., 2016).

Another area of application was predictive modeling for prognosis; this included models to assess the risk of developing a disease or estimate the effects of an intervention. For example, a few studies built models to identify factors that increase fall risk in post-stroke, Parkinson's, and elderly patients, with the purpose of suggesting gait retraining and rehabilitation targets to mitigate risk (Kikkert et al., 2017; Punt et al., 2016a, 2016b; Wei et al., 2017), while another study built models to predict susceptibility to stroke based on features extracted from movement kinematics during writing tasks (O'Reilly et al., 2014). Finally, many studies focused on quantifying the effect of an intervention on gait and predicting which patients would benefit from interventions such as physical therapy (Duhamel et al., 2006; Jiang et al., 2017), gait modifications (Ardestani et al., 2014), orthoses prescriptions (Ries et al., 2014), and orthopedic surgery (Hicks et al., 2011; Schwartz et al., 2014, 2013).

3.2. Data characteristics

Most studies used time series data from a relatively small number of participants (median = 40 subjects), resulting in problems where the feature space was larger than the number of observations. Given that most studies were performed in controlled settings, the data did not have missing values. Noise was generally reduced using established filtering methods. Bias resulting from differences in subject characteristics was not addressed in many studies where it could have been. For example, several studies included case and control groups with unequal numbers of men and women or significantly different age groups. Generally, models were built using a single data type, but several studies integrated data from different sources (e.g., electromyography and optical motion capture) (Kamavuako et al., 2012; Moustakidis et al., 2010; Punt et al., 2017). Out of nine studies that fused data from different sources, four reported scaling or normalization of the features (Rahmati et al., 2016; Rissanen et al., 2008; Roy et al., 2013; Roy et al., 2009). For reasons pointed out in the Discussion, feature normalization is a practice we recommend when the algorithm is not scale invariant.

3.3. Feature engineering

Lower-dimensional features were most commonly engineered from time series using domain knowledge or PCA. Domain knowledge was used to extract key gait features, such as peak angles, moments, ranges of motion, muscle activity duration, cadence, and stride length. PCA was also performed to reduce dimensionality, while preserving most of the variation in

the data. The number of principal components used in models was based on the percentage of variability explained, with most studies aiming to preserve at least 90% of the variability. Fewer studies derived features using Information Theory (Drotár et al., 2015), fast Fourier transforms (Ahlrichs et al., 2016), hidden Markov models (Mannini et al., 2016), wavelet decomposition (Moustakidis et al., 2010; Nielsen et al., 2011), and dynamic time warping (Zhang et al., 2016). Using the full time series in models was less common. Studies that took this approach trained models using feature spaces that were much larger than the number of observations, which is not recommended.

3.4. Model learning and performance evaluation

Model selection procedures (e.g., use of cross-validation for feature selection) and performance evaluation metrics (e.g., sensitivity and specificity, in addition to accuracy) were reported inconsistently across studies. Many studies included feature selection and hyper-parameter tuning in the model learning process. Feature selection approaches included stepwise (forward and backward) feature selection, hill climbing, and informal selection of feature combinations that led to the best model performance. Another approach was to pre-screen features and include only the ones that were significantly different between classes, using the entire dataset. Selection of hyper-parameters, such as the soft-margin constant for support vector machines or the learning rate for neural networks, was carried out using a grid-search approach in some studies, while in others selection of these parameters was not discussed. Only 8 of over 50 studies selected features or hyper-parameters using validation data, which as we point out in Section 4.3 can lead to models that do not generalize well to new data.

Model performance was assessed using held-out data, with the most common technique being k-fold cross-validation. The majority of studies reported accuracy as the primary performance evaluation metric, while fewer studies included sensitivity, specificity, F-1 score, and area under the receiver operating characteristic curve (AUC) measurements. Class imbalance was generally not discussed, even in studies that included different numbers of cases and controls, which can artificially inflate the performance of a model when only overall accuracy is assessed.

4. Discussion

The use of machine learning methods in movement biomechanics research is on the rise (Fig. 1A). From passively monitoring post-stroke patients with wearable devices to predicting outcomes of interventions in children with cerebral palsy, the range of applications where advanced analytics can improve rehabilitation research will continue to expand, particularly as wearable sensing generates vast amounts of data. The aim of this review is to bring to light machine learning efforts in movement biomechanics research and initiate a discussion of best practices and reporting standards that will maximize the impact of future work in this area.

4.1. Initial considerations

An important consideration at the beginning of a study is choosing the most appropriate statistical tool to address the research question of interest. Many research questions can be answered using traditional inferential statistics. In the literature, predictive models were often built for studies whose primary aim was to explain pathology by determining if certain gait features are discriminative with respect to disease status. Predictive modeling is not the most appropriate approach for this goal because often features are correlated, which may affect whether they are selected by feature selection approaches. The primary aim of predictive modeling is to learn a mapping function that can be used to make predictions on new data, with feature interpretation being a secondary aim.

Another consideration is the size of the feature space (p) in comparison to the number of available subjects or observations (n) and its effect on the model selection process. When the number of features is larger than the number of observations ($n < p$), there is a high risk of overfitting a model (i.e., building a model that performs poorly on new data). The inability to learn an adequate model due to insufficient observations and a large feature space is often referred to as *the curse of dimensionality*. In addition to dimensionality reduction through feature engineering and feature selection, model complexity might also be considered in the context of n and p. Simple linear models, such as logistic regression, can often outperform non-linear models. Simple models are characterized by high bias and low variance, whereas complex models are characterized by low bias and high variance (i.e., complex models tend to be more accurate on training data, but less generalizable on new data; Fig. 2).

4.2. Feature engineering

Good feature engineering (Fig. 3A) is a key step in building models from high-dimensional data, as it can lead to high model performance even with simple algorithms, such as naïve Bayes or logistic regression. Reducing data dimensionality while retaining sufficient relevant information, however, is not trivial. Discriminative information with respect to a given outcome may not always be evident and captured by features engineered using domain knowledge (e.g., stride length). Automatic feature extraction techniques, such as PCA, may capture some of the saliency in the data, but may produce features that are difficult to interpret or that are not informative with respect to a given outcome. Generally, we recommend testing a combination of automatically extracted features and ones engineered by humans. When fusing features that represent different quantities (e.g., knee angles and moments), for scale-variant algorithms it is also important to rescale features to ensure that each one contributes equally to the objective function and the optimization algorithm converges faster. Z-score normalization and min-max scaling are two common techniques used to rescale data.

4.3. Model assessment and learning

The success of a model is measured by its ability to generalize to new data. It is thus critical for model performance to be evaluated on held-out observations not used in the model selection and training process (Fig. 3B and C). If the dataset is large and representative of the underlying population, then splitting the data into a training set used to learn the model parameters and a test set used to evaluate the model's performance is appropriate. Most

commonly, however, the number of observations is too small to generate sufficient training and testing datasets, and techniques such as k-fold cross-validation, leave-one-out cross-validation (when the sample is very small), and bootstrapping are used to increase the effective size of training and testing samples. When splitting the dataset, it is important to include all the data from a given subject in only one of the sets. For example, if 10 motion capture trials are recorded for each subject, then all the trials should be part of the same set (e.g., training or test, but not both), unless the task is to train a subject-specific, rather than a generalizable model.

We recommend assessing and reporting a comprehensive set of performance evaluation metrics and ensuring that any biases in the dataset are appropriately addressed. For classification models, accuracy is the most commonly used metric, but it often provides an insufficient assessment of the model's performance. For example, in a dataset with an unequal distribution of observations from each class (e.g., 70% versus 30%), the worst classifier—one that always predicts the majority class—would have an accuracy of 70%. Additional metrics, including the AUC, sensitivity and specificity, or the confusion matrix may be reported to enable evaluation and comparison with other models. For regression models, the mean squared error or r-squared value evaluated on test data are appropriate measures of a model's generalizability. However, similar to classification models, these measures can be biased assessments of generalizability if the data used to train the model come from a skewed sample. In both classification and regression models, bias can be reduced by adjusting the cost function to weigh errors from different observations disproportionally. Along similar lines, including important confounders as covariates in the model and analyzing their contribution, especially when subjects from different demographic groups are represented disproportionally in different classes (e.g., cases and controls), is another step that can lead to more robust models.

Model learning—the process during which an algorithm learns from training data—is a critical step whose thorough reporting is needed to allow assessment of the model's reliability. While for simple models this step focuses on learning model parameters, such as regression coefficients, in most cases it can include one or two additional steps: (1) feature selection and (2) hyper-parameter tuning (Fig. 3B). When these additional steps are taken, the data may be divided into three sets: a training set used to learn model parameters, a validation set used to select features or hyper-parameters, and a test set used to assess model performance, as done in a few of the reviewed studies (Palmerini et al., 2013; Rahmati et al., 2016). A common incorrect approach for feature selection is performing Student's t tests to determine which features are significantly different between cases and controls in the entire dataset and retaining only those features for model learning. Another incorrect approach is performing step-wise regression and using the test data both for feature selection and model performance evaluation. Using the entire dataset for feature selection is plausible only if the feature selection process is unsupervised. For example, performing PCA on gait waveforms from the entire dataset and retaining features that explain 90% of the variability is acceptable because the label to be predicted by the model (e.g., disease status) is not used in the feature selection process. Model hyper-parameters should also be selected using validation data. An example is the set of kernel parameters used by support vector machines to obtain a non-linear classification boundary (e.g., width of a Gaussian kernel or degree of a polynomial

kernel). Hyper-parameters are learned or tuned by sweeping through a pre-defined range, fitting different models to the training set, and tracking how the models perform on the validation set. We also recommend reporting the range of hyper-parameter sweeps, as is done by Kamruzzaman and Begg (2006). After the best model is selected, its performance is evaluated on the test set.

In unsupervised learning tasks, the model learning step may also involve selecting hyper-parameters. For example, in the case of k-means clustering, model learning involves choosing the number of clusters in addition to identifying cluster centers. The most common pitfall we encountered in the literature was the lack of rigorous approaches for selecting the number of clusters. Most studies performed clustering based on a pre-defined number. While it is appropriate to hypothesize that a dataset contains a given number of clusters, tests such as the Bayesian Information Criterion (Schwarz, 1978), Akaike Information Criterion (Akaike, 1974), Silhouette (Rousseeuw, 1987) may be used to confirm that such a model best represents the underlying structure of the data. One study, for example, used the Dunn test (Dunn, 1973) to determine that children with crouch gait exhibit distinct gait characteristics that fall into one of five subcategories (Rozumalski and Schwartz, 2009). These five categories were then related to distinct features of the underlying clinical pathology, which can guide the development of targeted treatments.

4.4. Sharing of data and tools

One of the powers of machine learning methods lies in their ability to draw insight from large, heterogeneous, noisy, and biased datasets. We found that models were nearly always trained and tested using data from a single study or laboratory, but many of the datasets could be pooled to enable the construction of more robust models. For example, pooling data from six studies that used optical motion capture data from patients with knee osteoarthritis (Astephen et al., 2008; Deluzio and Astephen, 2007; Favre et al., 2012; Levinger et al., 2009; Magalhães et al., 2013; Mezghani et al., 2017) would enable the construction of a database with 621 subjects, where protocol differences across sites could be accounted for in the model. Even if data were not pooled in the model learning phase, testing a model on data from different studies and allowing the community to improve it would increase its translational value. In recent years, a multitude of articles have called attention to a “reproducibility crisis” in biomedical research. Although a broader cultural shift toward open science is underway, sharing of data and tools has not been traditionally incentivized in biomechanics. Data curation and code documentation for public use require additional efforts and resources, and the return on investment for the researcher is not immediately apparent. Funding agencies have been promoting sharing of data and tools, but there is currently no evaluation system in place to reward researchers who contribute to open science. Despite the limitations of citation metrics such as the h-index, its increasing weight in promotion decisions makes it one of the few means through which researchers can be rewarded for open science. To that end, an internal analysis of publications from our group indicates that publications accompanied by public releases of biomechanical models (e.g., Rajagopal et al., 2016) and data (e.g., Hamner and Delp, 2013) garner higher citation counts. From 2007 to 2013, our group published 19 articles using open-source musculoskeletal modeling software OpenSim (Delp et al., 2007). Publications that included shared resources

(n = 9) had an average of 30 more citations than publications without shared resources, when controlling for years since publication and journal impact factor ($p < 0.05$). To promote and support open science, we developed a website, SimTK.org, which hosts hundreds of projects by biomedical scientists who wish to share data, musculoskeletal models, and links to source code repositories (e.g., GitHub.com). We encourage researchers working in biomechanics and data science to contribute.

4.5. Knowing when to trust a model

A majority of the studies we reviewed focused on predictive models that could serve as valuable diagnostic or prognostic tools, but currently there is no consensus on when a model is good enough for clinical translation. Historically, conclusions drawn on $p < 0.05$ have driven important health decisions, but recent discussions have called into question the appropriateness of an arbitrary significance threshold that can be easily—although unintentionally—hacked by scientists who have only basic training in statistics (Leek et al., 2017; Nuzzo, 2014). Past evidence thus cautions us against the use of arbitrary thresholds to evaluate the performance of predictive models. Instead, we encourage the adoption of standards that focus on rigor and transparency. The best practices for model learning, evaluation, and reporting presented here provide a baseline for critical evaluation of the literature. Ultimately, the decision on when a model is accurate enough should be application specific, made by a team of researchers and clinicians who have enough information and training to fully understand the implications of its performance. It is also important to note that most models are intended to enhance, rather than replace, human clinical decision-making. For example, a model that automates the Fugl-Meyer assessment of sensorimotor function in post-stroke patients using data from wearable sensors would enable remote monitoring of patients to supplement sparse in-clinic assessments. Even though model estimates of the Fugl-Meyer score would be noisy and less accurate than an expert clinical assessment, the high-frequency, low-cost monitoring may be valuable supplementary information for patients and clinicians.

4.6. Interpretability

As models become more complex, they also often become more difficult to interpret. Artificial neural networks, for example, can learn highly complex nonlinear relationships from large data and outperform humans at many tasks, yet their opaqueness inspires little confidence in biomedical scientists. Lack of interpretability is particularly challenging in biomedicine, where predictive modeling is often motivated by the need to improve prevention and rehabilitation efforts. If a “black box” model predicts with high confidence that a patient will develop osteoarthritis based on his or her gait pattern, but offers no insight into the specific features of gait that are driving osteoarthritis progression, it is unclear how this knowledge could be used to improve the patient’s health. While event detection and activity recognition tasks that prioritize predictive accuracy over interpretability can be tackled with complex models, diagnostic and prognostic tools are currently better served by transparent models. Interpretability is an active area of research in machine learning, and the trade-off between accuracy and interpretability may be soon mitigated by ongoing efforts to decipher what single artificial neurons have learned and how they communicate with each other to make a prediction (Doshi-Velez and Kim, 2017).

4.7. Training in data science

Successful application of machine learning methods requires understanding of the methods that are available and how to properly apply them. An increasing interest in data science across scientific disciplines has prompted the design of massive open online courses (MOOCs) on different aspects of data analytics (e.g., Statistical Learning (Hastie, 2016) and Machine Learning (Ng, 2018)), which are practical introductory resources. As a community, investing in cross-training efforts through webinars and workshops at scientific conferences will also be critical in educating the next generation of data-savvy biomechanists.

4.8. New opportunities

Reinforcement learning is a machine learning paradigm where an agent (e.g., a musculoskeletal model) learns to perform a task (e.g., walk) by interacting with the environment and selecting actions (e.g., muscle excitations) through trial and error to maximize reward (e.g., metabolic cost) (Sutton and Barto, 1998). This approach has been used to train a program to play the game Go well enough that it can outperform any human without any input from or gameplay against other humans (Silver et al., 2017). Efforts to model neural control of movement have been ongoing in the computer science and biomechanics communities, and they are poised to be even more successful as these communities start to work together to bridge advances in reinforcement learning with physiologically accurate biomechanical models (Kidziński et al., 2018). Future research in this area promises to advance our understanding of neural control of movement and how that adapts to new perturbations, ultimately informing the development of natural human-machine interfaces for rehabilitation.

As researchers mine data from wearable sensors, where device non-wear is a common problem that results in missing data, the use of algorithms that can learn well from incomplete curves will also become more relevant. *Functional data analysis*, which considers the dynamic nature of curves, has already been used in studies of human movement biomechanics (Donoghue et al., 2008; Duhamel et al., 2006; Ryan et al., 2006). Building on that foundation, researchers have proposed algorithms that can learn from incomplete (sparsely sampled) curves (Bachrach et al., 1999; James et al., 2000; James and Sugar, 2003). For example, these approaches have been used to track human movement in video sequences where visual occlusions result in motion trajectories with missing data (Ormoneit et al., 2005, 2000b, 2000a).

Shrinkage methods (Copas, 1983) are a class of algorithms that can be used to reduce model complexity and improve performance. Shrinkage methods shrink the coefficients associated with each feature by adding a penalty term to the cost function that encourages coefficients to be small. Intuitively, this approach improves model generalizability because it reduces extremely high coefficients, which are likely an artifact of the available data. LASSO is an example of such a method that can achieve both shrinkage and variable selection because some of the coefficients are forced to be zero (Tibshirani, 1996). Advanced versions of LASSO can also be used for specialized variable selection. For example, group LASSO (Meier et al., 2008) can be used to select groups of features from multiple channels or sensors at once, while fused LASSO can be used to induce similar penalties to features that

are spatially or temporally closer (e.g., select contiguous segments of a time series) (Tibshirani et al., 2005).

Artificial neural networks are a family of machine learning algorithms that aim to emulate the structure and connectivity of biological neural networks. They consist of an input, output, and one or more hidden layers of interconnected artificial neurons, or nodes. The number of hidden layers introduced in the model defines the depth of the network. How information flows between artificial neurons and through layers of the neural network is determined by model parameters that are learned using training data. *Deep networks*, defined as neural networks with multiple hidden layers, are successful at learning complex, nonlinear relationships and have already found wide use in speech recognition and image processing, where the data are high-dimensional. They also show promise in modeling movement time series data (Quisel et al., 2017), especially for applications where accuracy is more important than interpretability (e.g., event detection or activity classification). The number of model parameters to be learned grows rapidly with the depth of the network, however, requiring larger amounts of training data than are necessary for simpler models. Currently, there are no formal rules for determining the appropriate amount of data required to train a model with a desired depth. One approach to estimate if the available data are sufficient is to consider the learning curve. Typically, model performance has a logarithmic shape when plotted against sample size and eventually plateaus once the model reaches its limiting accuracy. In the latter stages of the curve, additional training data do not result in increased performance. Once pre-trained on large datasets, these networks can then be fine-tuned for tasks involving similar smaller datasets—a paradigm known as *transfer learning*. For example, networks such as VGG-16 have been pre-trained using millions of labeled images from the ImageNet database to recognize a multitude of different objects and are now being adapted with little effort for classification of specific types of medical images (Shin et al., 2016). Deep learning reduces the need for feature engineering, but it may only be appropriate when either a large dataset or a pre-trained network is available.

Supervised machine learning models are dependent on good feature engineering and labeled data. Most movement data being generated through mobile devices, however, are unlabeled. We have little information about what activities an individual is performing, and manual labeling of large time series is labor intensive. *Semi-supervised learning* is a paradigm that allows users to build models even if labels are available for only a small subset of the data (Chapelle et al., 2010). A related technique for leveraging unlabeled data is *weak supervision*, where a small gold-standard labeled set is used in conjunction with a large dataset with noisy labels obtained with comparatively little effort (e.g., through crowdsourcing or heuristic rules). For example, to train a supervised fall detection algorithm from several days of accelerometer data, the whole dataset must include labels on where falls occurred in the time series. Alternatively, when labels are not available, a user may take a weak supervision approach, writing a labeling function with a statement noting that if the vertical acceleration is greater than a given threshold, then the patient is likely to have experienced a fall. Since this threshold is not uniform across patients and falls, knowledge from multiple heuristics can be combined with a small set of gold-standard labels, which allows the model to learn when these heuristics agree and disagree. While the performance of weakly supervised models is partly dependent on the quality of noisy labels, these

approaches have already achieved state-of-the-art performance in natural language processing applications (Ratner et al., 2017) and may find wide use in wearable sensor data analysis.

4.9. Conclusions

Research at the intersection of machine learning and biomechanics offers great promise for advancing human movement research, improving clinical decision-making, and accelerating rehabilitation programs for neuromuscular and musculoskeletal diseases. To enable appropriate use of advanced analytical techniques and stay abreast of new developments in machine learning that are galvanizing research across other biomedical disciplines (e.g., imaging, genomics, neuroscience), open data, tools, and discussions must be actively encouraged within the movement biomechanics community. We offer the following summary for biomechanists to consider as they adopt these methods and review the literature.

- When the number of observations in a dataset is smaller than the number of features, reduce the feature space through feature engineering and feature selection approaches and use a simple model for better generalizability. A deep neural network built using gait data from 10 subjects will likely not achieve good predictive performance on new data.
- As an alternative to stepwise feature selection and Student's *t* tests, consider using regularization methods (e.g., LASSO) to reduce the number of features and improve model performance, especially when the feature space is very large and there is multicollinearity among features.
- Rescale heterogeneous data when using scale-variant algorithms, such as support vector machines, to ensure that each feature contributes equally to the objective function. Combining angular velocity and acceleration data from IMUs, for example, requires feature scaling.
- When developing a supervised model, use a training set for parameter selection, a validation set for hyper-parameter tuning and feature selection, and a test set for performance evaluation to ensure that the model generalizes well and the reported performance is not overestimated.
- When developing a model, include all the data from one subject (e.g., different trials) in only one set (training, validation, or test dataset) to ensure that the model generalizes well to new data.
- Include confounders as covariates in the model and analyze their contribution (e.g., regress them out). The model may be predicting a confounder rather than the outcome, if data from different classes are unbalanced in terms of that confounder. For example, if the diseased group contains a greater number of women than men compared to the normal group, a high-performing model may be predicting sex rather than disease status and will perform poorly if used to predict disease in the future.

- When classifying data into clusters, select the number of clusters using standard tests (e.g., Silhouette) to ensure that the model best represents the underlying structure of the data. Even if this number is based on a reasonable hypothesis or domain knowledge, a formal test will ensure that the model is plausible.
- Thoroughly report the feature-selection and model-learning steps, and include a comprehensive set of evaluation metrics (instead of accuracy alone). This will enable others to critically evaluate your work and place it in context with past and future studies.
- Make your data and code publicly available to allow others to reproduce and extend your models. Facilitating adoption will increase the impact of your work.

We hope that the discussion of practical guidelines, common pitfalls, and new opportunities highlighted here provide biomechanists with a scaffold upon which to build modern statistical models that enhance their research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was funded by the National Institutes of Health (NIH) Grant U54EB020405. The authors would like to thank Jessica Selinger, Rachel Jackson, Łukasz Kidziński, Wolf Thomsen, and Jennifer Yong for their insightful feedback.

Appendix A

The following terms are used in the literature review and discussion of best practices.

Label (also *response*, *dependent variable*)

The output or outcome that a model learns to predict. Disease status in a diagnostic model, outcome of surgery in prognostic model, and the occurrence of tremor in an automated event detection model are examples of labels.

Feature (also *predictor*, *attribute*, *independent* variable)

A variable used as input to a model. For example, the peak knee flexion angle during the swing phase is a feature taken from the trajectory of knee flexion angles.

Supervised learning

Techniques used to learn the relationship between a set of independent variables (features) and one or more dependent variables (labels). Labels are used as supervisors or teachers that enable the model to learn its parameters (e.g., regression coefficients) well enough that it can make accurate predictions from new data.

Unsupervised learning

Techniques to detect patterns in input data, without using any labels to supervise learning. Clustering algorithms used to find subgroupings and principal component analysis are two of the most widely used unsupervised learning methods.

Regression

A supervised learning approach where the label is a continuous variable.

Classification

A supervised learning approach where the label is a categorical variable. Binary and multi-class classification problems are subtypes where the label encodes either two or more classes, respectively.

Class imbalance

A disproportional distribution of observations that belong to different classes (e.g., many more subjects without than with impairment), which can bias the learning process.

Clustering

An unsupervised learning approach used to discover inherent groupings in the data.

Feature engineering

The process of creating new features from existing ones to reduce data dimensionality and improve model performance. Common approaches for feature engineering from time series data include summary metrics (e.g., min, max, average of the signal), extraction of important events based on domain knowledge (e.g., peak angle), PCA, fast Fourier transforms, etc. This includes all the data, but it is blind to data labels.

Feature scaling or normalization

Practices used to standardize the range of values for each feature to facilitate convergence and prevent some features, especially those spanning different ranges, from contributing unequally to the learning process in models where the objective is not scale invariant.

Feature selection

Techniques used to select the most informative features with respect to a given label from the whole set of features. This step may not include the test data.

Model parameters

Configuration variables intrinsic to the model, which are estimated based on training data. Examples include coefficients in a linear regression, weights in artificial neural networks, the support vectors in support vector machines.

Model hyper-parameters

Additional parameters that can be used to control the complexity and learning rate of a model. Examples include the penalty coefficient in penalized linear regression, learning rate in artificial neural networks, and kernel type (e.g., Gaussian, polynomial) in support vector machines. Hyper-parameters are not learned simultaneously with the primary model parameters, but must be either selected a priori and justified or selected through a process called “hyper-parameter tuning.”

Cost function (also *loss function*)

A function used to estimate how well a model’s predictions fit the true labels. Model parameters are selected to minimize the cost function. A typical cost function for linear regression is the mean squared error.

Accuracy

The rate of correct predictions made by a classifier.

Sensitivity (also *recall*, *true positive rate*) = $\frac{\text{number of true positives}}{\text{number of true positives+false negatives}}$; the proportion of actual positives that are correctly identified as such (e.g., the percentage of freezing episodes that are correctly detected by the model).

Specificity (also *true negative rate*) = $\frac{\text{number of true negatives}}{\text{number of true negatives+false positives}}$; the proportion of actual negatives that are correctly identified as such (e.g., the percentage of normal individuals predicted as not having OA).

Precision (also *positive predictive value*) = $\frac{\text{number of true positives}}{\text{number of true positives+false positives}}$; the proportion of identified positives that are indeed positive (e.g., the percentage of identified falls that are actually falls).

Confusion matrix

A $k \times k$ table, where k is the number of classes, often used to describe the performance of a classifier. The rows and columns represent predicted versus true classes.

F-1 score

The harmonic mean of precision and recall. Although not as intuitive as accuracy, it can be a more informative metric of model performance when class imbalance is present.

Receiver Operating Characteristic (ROC)

A commonly used plot that visualizes the tradeoff between sensitivity and specificity as the discrimination threshold is varied for a binary classifier.

Area Under the ROC Curve (AUC)

A commonly used quantitative summary of the ROC. A random classifier has an AUC of 0.5, while a perfect classifier has an AUC of 1.

Sparsity

Generally describes an entity (e.g., a dataset or model) with few non-zero values. A sparsely sampled dataset contains many missing values. A sparse model uses only a reduced subset of the initial features, typically by imposing constraints on the objective function.

Bias

In supervised learning tasks, bias is the error that results from incorrect assumptions (e.g., fitting a linear model when the underlying relationship is not linear), causing algorithms to miss important relationships between features and labels. Even when the model assumptions are correct, the fitting method (e.g., LASO, ridge regression) can sometimes lead to bias in the estimated parameters.

Variance

In supervised learning tasks, variance is the error that results from sensitivity to small variations, causing algorithms to model noise in the training data.

Bias-variance tradeoff

The process of minimizing bias and variance in order to train models that can generalize well to nontraining data. Simple models may be characterized by high bias and low variance (underfit to the training data), while complex models may be characterized by low bias and high variance (overfit to the training data).

References

Ahlrichs C, Samà A, Lawo M, Cabestany J, Rodríguez-Martín D, Pérez-López C, Sweeney D, Quinlan LR, Laighin GÒ, Counihan T, Browne P, Hadas L, Vainstein G, Costa A, Annicchiarico R, Alcaine S, Mestre B, Quispe P, Bayes À, Rodríguez-Molinero A, 2016 Detecting freezing of gait with a tri-axial accelerometer in Parkinson's disease patients. Med. Biol. Eng. Comput 54, 223–233. 10.1007/s11517-015-1395-3. [PubMed: 26429349]

Akaike H, 1974 A new look at the statistical model identification. IEEE Trans. Autom. Control 19, 716–723. 10.1109/TAC.1974.1100705.

Alam MN, Garg A, Munia TTK, Fazel-Rezai R, Tavakolian K, 2017 Vertical ground reaction force marker for Parkinson's disease. PLoS ONE 12 10.1371/journal.pone.0175951.

Ardestani MM, Chen Z, Wang L, Lian Q, Liu Y, He J, Li D, Jin Z, 2014 A neural network approach for determining gait modifications to reduce the contact force in knee joint implant. Med. Eng. Phys 36, 1253–1265. 10.1016/j.medengphy.2014.06.016. [PubMed: 25066584]

Astephen JL, Deluzio KJ, Caldwell GE, Dunbar MJ, Hubley-Kozey CL, 2008 Gait and neuromuscular pattern changes are associated with differences in knee osteoarthritis severity levels. J. Biomech 41, 868–876. 10.1016/j.jbiomech.2007.10.016. [PubMed: 18078943]

Bachrach LK, Hastie T, Wang M-C, Narasimhan B, Marcus R, 1999 Bone mineral acquisition in healthy Asian, Hispanic, black, and Caucasian youth: a longitudinal study. J. Clin. Endocrinol. Metab 84, 4702–4712. 10.1210/jcem.84.12.6182. [PubMed: 10599739]

Biswas D, Cranny A, Gupta N, Maharatna K, Achner J, Klemke J, Jobges M, Ortmann S, 2015 Recognizing upper limb movements with wrist worn inertial sensors using k-means clustering classification. Hum. Mov. Sci 40, 59–76. 10.1016/j.humov.2014.11.013. [PubMed: 25528632]

Bochniewicz EM, Emmer G, McLeod A, Barth J, Dromerick AW, Lum P, 2017 Measuring functional arm movement after stroke using a single wrist-worn sensor and machine learning. J. Stroke Cerebrovasc. Dis. Off. J. Natl. Stroke Assoc 26, 2880–2887. 10.1016/j.jstrokecerebrovasdis.2017.07.004.

Buchman AS, Leurgans SE, Weiss A, Vanderhorst V, Mirelman A, Dawe R, Barnes LL, Wilson RS, Hausdorff JM, Bennett DA, 2014 Associations between quantitative mobility measures derived from components of conventional mobility testing and Parkinsonian gait in older adults. PloS One 9, e86262 10.1371/journal.pone.0086262. [PubMed: 24465997]

Capela NA, Lemaire ED, Baddour N, Rudolf M, Goljar N, Burger H, 2016 Evaluation of a smartphone human activity recognition application with ablebodied and stroke participants. J. NeuroEng. Rehabil. JNER 13, 1–10. 10.1186/s12984-016-0114-0. [PubMed: 26728632]

Chapelle O, Schlkopf B, Zien A, 2010 Semi-Supervised Learning. The MIT Press.

Copas JB, 1983 Regression, prediction and shrinkage. J. R. Stat. Soc. Ser. B Methodol 45, 311–354.

Delp SL, Anderson FC, Arnold AS, Loan P, Habib A, John CT, Guendelman E, Thelen DG, 2007 OpenSim: open-source software to create and analyze dynamic simulations of movement. IEEE Trans. Biomed. Eng 54, 1940–1950. 10.1109/TBME.2007.901024. [PubMed: 18018689]

Deluzio KJ, Astephen JL, 2007 Biomechanical features of gait waveform data associated with knee osteoarthritis: an application of principal component analysis. Gait Post. 25, 86–93. 10.1016/j.gaitpost.2006.01.007.

Donoghue OA, Harrison AJ, Coffey N, Hayes K, 2008 Functional data analysis of running kinematics in chronic Achilles tendon injury. Med. Sci. Sports Exerc 40, 1323–1335. 10.1249/MSS.0b013e31816c4807. [PubMed: 18580414]

Doshi-Velez F, Kim B, 2017 Towards a rigorous science of interpretable machine learning. ArXiv Prepr. ArXiv170208608.

Drotár P, Mekyska J, Rektorová I, Masarová L, Smékal Z, Faundez-Zanuy M, 2015 Decision support framework for Parkinson's disease based on novel handwriting markers. IEEE Trans. Neural Syst. Rehabil. Eng. Publ. IEEE Eng. Med. Biol. Soc 23, 508–516. 10.1109/TNSRE.2014.2359997.

Duhamel A, Devos P, Bourriez JL, Preda C, Defebvre L, Beuscart R, 2006 Functional data analysis for gait curves study in Parkinson's disease. Stud. Health Technol. Inform 124, 569–574. [PubMed: 17108578]

Dunn JC, 1973 A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J. Cybern 3, 32–57. 10.1080/01969727308546046.

Favre J, Hayoz M, Erhart-Hledik JC, Andriacchi TP, 2012 A neural network model to predict knee adduction moment during walking based on ground reaction force and anthropometric measurements. J. Biomech 45, 692–698. 10.1016/j.jbiomech.2011.11.057. [PubMed: 22257888]

Fisher JM, Hammerla NY, Ploetz T, Andras P, Rochester L, Walker RW, 2016 Unsupervised home monitoring of Parkinson's disease motor symptoms using body-worn accelerometers. Parkinson. Relat. Disord 33, 44–50. 10.1016/j.parkreldis.2016.09.009.

Fulk GD, Edgar SR, Bierwirth R, Hart P, Lopez-Meyer P, Sazonov E, 2012 Identifying activity levels and steps of people with stroke using a novel shoebased sensor. J. Neurol. Phys. Ther 36, 100–107. 10.1097/NPT.0b013e318256370c. [PubMed: 22592067]

Fulk GD, Sazonov E, 2011 Using sensors to measure activity in people with stroke. Top. Stroke Rehabil 18, 746–757. 10.1310/tsr1806-746. [PubMed: 22436312]

Hamner SR, Delp SL, 2013 Muscle contributions to fore-aft and vertical body mass center accelerations over a range of running speeds. J. Biomech 46, 780–787. 10.1016/j.jbiomech.2012.11.024. [PubMed: 23246045]

Hastie TJ 2016 StatLearning - SELF PACED [WWW Document]. Lagunita URL <https://lagunita.stanford.edu/courses/HumanitiesSciences/StatLearning/Winter2016/> (accessed 5.26.18).

Hicks JL, Delp SL, Schwartz MH, 2011 Can biomechanical variables predict improvement in crouch gait? Gait Post. 34, 197–201. 10.1016/j.gaitpost.2011.04.009.

Huang FC, Patton JL, 2016 Movement distributions of stroke survivors exhibit distinct patterns that evolve with training. J. Neuroeng. Rehabil 13, 23 10.1186/s12984-016-0132-y. [PubMed: 26961682]

James GM, Hastie TJ, Sugar CA, 2000 Principal component models for sparse functional data. Biometrika 87, 587–602. 10.1093/biomet/87.3.587.

James GM, Sugar CA, 2003 Clustering for sparsely sampled functional data. J. Am. Stat. Assoc 98, 397–408.

Jiang N, Luk KD-K, Hu Y, 2017 A machine learning based surface electromyography topography evaluation for prognostic prediction of functional restoration rehabilitation in chronic low back pain. Spine. 10.1097/BRS.0000000000002159.

Kaczmarczyk K, Wit A, Krawczyk M, Zaborski J, 2009 Gait classification in post-stroke patients using artificial neural networks. Gait Post. 30, 207–210. 10.1016/j.gaitpost.2009.04.010.

Kaczmarczyk K, Wit A, Krawczyk M, Zaborski J, Gajewski J, 2012 Associations between gait patterns, brain lesion factors and functional recovery in stroke patients. Gait Post. 35, 214–217. 10.1016/j.gaitpost.2011.09.009.

Kamavuako EN, Farina D, Yoshida K, Jensen W, 2012 Estimation of grasping force from features of intramuscular EMG signals with mirrored bilateral training. Ann. Biomed. Eng 40, 648–656. 10.1007/s10439-011-0438-7. [PubMed: 22006428]

Kamruzzaman J, Begg RK, 2006 Support vector machines and other pattern recognition approaches to the diagnosis of cerebral palsy gait. IEEE Trans. Biomed. Eng 53, 2479–2490. 10.1109/TBME.2006.883697. [PubMed: 17153205]

Keijsers NLW, Horstink MWIM, Gielen SCAM, 2003a Automatic assessment of levodopa-induced dyskinesias in daily life by neural networks. Mov. Disord. Off. J. Mov. Disord. Soc 18, 70–80. 10.1002/mds.10310.

Keijsers NLW, Horstink MWIM, Gielen SCM, 2003b Movement parameters that distinguish between voluntary movements and levodopa-induced dyskinesia in Parkinson's disease. Hum. Mov. Sci 22, 67–89. 10.1016/S0167-9457(02)00179-3. [PubMed: 12623181]

Kidziński Ł, Mohanty SP, Ong C, Hicks JL, Carroll SF, Levine S, Salathé M, Delp SL, 2018 Learning to Run challenge: Synthesizing physiologically accurate motion using deep reinforcement learning. ArXiv180400198 Cs.

Kikkert LHJ, De Groot MH, Van Campen JP, Beijnen JH, Hortobágyi T, Vuillerme N, Lamoth CCJ, 2017 Gait dynamics to optimize fall risk assessment in geriatric patients admitted to an outpatient diagnostic clinic. PLoS ONE 12 10.1371/journal.pone.0178615.

Krebs HI, Krams M, Agrafiotis DK, Di Bernardo A, Chavez JC, Littman GS, Yang E, Byttebier G, Dipietro L, Rykman A, McArthur K, Hajjar K, Lees KR, Volpe BT, 2014 Robotic measurement of arm movements after stroke establishes biomarkers of motor recovery. Stroke 45, 200–204. 10.1161/Strokeaha.113.002296/-/DC1. [PubMed: 24335224]

Kuntze G, von Tscharner V, Hutchison C, Ronsky JL, 2015 Multi-muscle activation strategies during walking in female post-operative total joint replacement patients. J. Electromyogr. Kinesiol. Off. J. Int. Soc. Electrophysiol. Kinesiol 25, 715–721. 10.1016/j.jelekin.2015.04.001.

Laroche D, Tolambiya A, Morisset C, Maillefert JF, French RM, Ornetti P, Thomas E, 2014 A classification study of kinematic gait trajectories in hip osteoarthritis. Comput. Biol. Med 55, 42–48. 10.1016/j.compbiomed.2014.09.012. [PubMed: 25450217]

Laudanski A, Brouwer B, Li Q, 2015 Activity classification in persons with stroke based on frequency features. Med. Eng. Phys 37, 180–186. 10.1016/j.medengphy.2014.11.008. [PubMed: 25559935]

Leek J, McShane BB, Gelman A, Colquhoun D, Nuijten MB, Goodman SN, 2017 Five ways to fix statistics. Nature 551, 557–559. 10.1038/d41586-017-07522-z. [PubMed: 29189798]

Levinger P, Lai DTH, Begg RK, Webster KE, Feller JA, 2009 The application of support vector machines for detecting recovery from knee replacement surgery using spatio-temporal gait parameters. Gait Post. 29, 91–96. 10.1016/j.gaitpost.2008.07.004.

Liu J, Li X, Li G, Zhou P, 2014 EMG feature assessment for myoelectric pattern recognition and channel selection: a study with incomplete spinal cord injury. Med. Eng. Phys 36, 975–980. 10.1016/j.medengphy.2014.04.003. [PubMed: 24844608]

Magalhães CMB, Resende RA, Kirkwood RN, 2013 Increased hip internal abduction moment and reduced speed are the gait strategies used by women with knee osteoarthritis. J. Electromyogr. Kinesiol. Off. J. Int. Soc. Electrophysiol. Kinesiol 23, 1243–1249. 10.1016/j.jelekin.2013.05.013.

Mannini A, Trojaniello D, Cereatti A, Sabatini AM, 2016 A machine learning framework for gait classification using inertial sensors: application to elderly, post-stroke and Huntington's disease patients. Sensors 16 10.3390/s16010134.

Masse F, Gonzenbach R, Paraschiv-Ionescu A, Luft A, Aminian K, 2016 Wearable barometric pressure sensor to improve postural transition recognition of mobility-impaired stroke patients. IEEE Trans. Neural Syst. Rehabil. Eng. Publ. IEEE Eng. Med. Biol. Soc 10.1109/TNSRE.2016.2532844.

Meier L, van de Geer S, Bühlmann P, 2008 The group lasso for logistic regression. J. R. Stat. Soc. Ser. B Stat. Methodol 70, 53–71.

Meyer CAG, Corten K, Fieuws S, Deschamps K, Monari D, Wesseling M, Simon J-P, Desloovere K, 2016 Evaluation of stair motion contributes to new insights into hip osteoarthritis-related motion pathomechanics. J. Orthop. Res. Off. Publ. Orthop. Res. Soc 34, 187–196. 10.1002/jor.22990.

Meyer CAG, Corten K, Fieuws S, Deschamps K, Monari D, Wesseling M, Simon J-P, Desloovere K, 2015 Biomechanical gait features associated with hip osteoarthritis: Towards a better definition of clinical hallmarks. J. Orthop. Res. Off. Publ. Orthop. Res. Soc 33, 1498–1507. 10.1002/jor.22924.

Mezghani N, Ouakrim Y, Fuentes A, Mitiche A, Hagemeister N, Vendittoli P-A, de Guise JA, 2017 Mechanical biomarkers of medial compartment knee osteoarthritis diagnosis and severity grading: discovery phase. J. Biomech 52, 106–112. 10.1016/j.jbiomech.2016.12.022. [PubMed: 28088304]

Moustakidis SP, Theocharis JB, Giakas G, 2010 A fuzzy decision tree-based SVM classifier for assessing osteoarthritis severity using ground reaction force measurements. Med. Eng. Phys 32, 1145–1160. 10.1016/j.medengphy.2010.08.006. [PubMed: 20875766]

Nair SS, French RM, Laroche D, Thomas E, 2010 The application of machine learning algorithms to the analysis of electromyographic patterns from arthritic patients. IEEE Trans. Neural Syst. Rehabil. Eng. Publ. IEEE Eng. Med. Biol. Soc 18, 174–184. 10.1109/TNSRE.2009.2032638.

Ng A 2018 Machine Learning [WWW Document]. Coursera URL <https://www.coursera.org/learn/machine-learning> (accessed 5.26.18).

Nielsen JLG, Holmgaard S, Jiang N, Englehart KB, Farina D, Parker PA, 2011 Simultaneous and proportional force estimation for multifunction myoelectric prostheses using mirrored bilateral training. IEEE Trans. Biomed. Eng 58, 681–688. 10.1109/TBME.2010.2068298. [PubMed: 20729161]

Nüesch C, Valderrabano V, Huber C, von Tscharner V, Pagenstert G, 2012 Gait patterns of asymmetric ankle osteoarthritis patients. Clin. Biomech. Bristol Avon 27, 613–618. 10.1016/j.clinbiomech.2011.12.016.

Nuzzo R, 2014 Scientific method: statistical errors. Nat. News 506, 150 10.1038/506150a.

O'Brien MK, Shawen N, Mummidisetty CK, Kaur S, Bo X, Poellabauer C, Kording K, Jayaraman A, 2017 Activity recognition for persons with stroke using mobile phone technology: toward improved performance in a home setting. J. Med. Int. Res 19, e184 10.2196/jmir.7385.

O'Reilly C, Plamondon R, Lebrun L-H, 2014 Linking brain stroke risk factors to human movement features for the development of preventive tools. Front. Aging Neurosci 6, 150 10.3389/fnagi.2014.00150. [PubMed: 25071559]

Ormoneit D, Black MJ, Hastie T, Kjellström H, 2005 Representing cyclic human motion using functional analysis. Image Vis. Comput 23, 1264–1276. 10.1016/j.imavis.2005.09.004.

Ormoneit D, Sidenbladh H, Black MJ, Hastie TJ, 2000a Learning and tracking cyclic human motion. In: Proceedings of NIPS 2000 Presented at the Neural Information Processing Systems, pp. 894–900.

Ormoneit D, Sidenbladh Hedvig, Sidenbladh H, Black MJ, Hastie T, Fleet DJ, 2000b Learning and tracking human motion using functional analysis. In: IEEE Workshop on Human Modeling, Analysis and Synthesis, pp. 2–9.

Oskoei MA, Hu H, 2008. Support vector machine-based classification scheme for myoelectric control applied to upper limb. IEEE Trans. Biomed. Eng 55, 1956–1965. 10.1109/TBME.2008.919734. [PubMed: 18632358]

Palmerini L, Mellone S, Avanzolini G, Valzania F, Chiari L, 2013 Quantification of motor impairment in Parkinson's disease using an instrumented timed up and go test. IEEE Trans. Neural Syst. Rehabil. Eng. Publ. IEEE Eng. Med. Biol. Soc 21, 664–673. 10.1109/TNSRE.2012.2236577.

Palmerini L, Rocchi L, Mellone S, Valzania F, Chiari L, 2011 Feature selection for accelerometer-based posture analysis in Parkinson's disease. IEEE Trans. Inf. Technol. Biomed. Publ. IEEE Eng. Med. Biol. Soc 15, 481–490. 10.1109/TITB.2011.2107916.

Pérez-López C, Samà A, Rodríguez-Martín D, Moreno-Aróstegui JM, Cabestany J, Bayes A, Mestre B, Alcaine S, Quispe P, Laighin GÓ, Sweeney D, Quinlan LR, Counihan TJ, Browne P, Annicchiarico R, Costa A, Lewy H, RodríguezMolinero A, 2016 Dopaminergic-induced dyskinesia assessment based on a single belt-worn accelerometer. Artif. Intell. Med 67, 47–56. 10.1016/j.artmed.2016.01.001. [PubMed: 26831150]

Punt M, Bruijn SM, Van Schooten KS, Pijnappels M, Van De Port IG, Wittink H, Van Dieën JH, 2016a Characteristics of daily life gait in fall and non fallprone stroke survivors and controls. J. NeuroEng. Rehabil 13 10.1186/s12984-016-0176-z.

Punt M, Bruijn SM, van Schooten KS, Pijnappels M, van de Port IG, Wittink H, van Dieën JH, 2016b Characteristics of daily life gait in fall and non fall-prone stroke survivors and controls. J. NeuroEng. Rehabil. JNER 13, 1–7. 10.1186/s12984-016-0176-z. [PubMed: 26728632]

Punt M, Bruijn SM, Wittink H, van de Port IG, van Dieen JH, 2017 Do clinical assessments, steady-state or daily-life gait characteristics predict falls in ambulatory chronic stroke survivors? J. Rehabil. Med 49, 402–409. 10.2340/16501977-2234. [PubMed: 28475196]

Quisel T, Foschini L, Signorini A, Kale DC, 2017 Collecting and analyzing millions of mHealth data streams. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining ACM, pp. 1971–1980.

Rahmati H, Martens H, Aamo OM, Stavdahl O, Stoen R, Adde L, 2016 Frequency analysis and feature reduction method for prediction of cerebral palsy in young infants. IEEE Trans. Neural Syst. Rehabil. Eng. Publ. IEEE Eng. Med. Biol. Soc 10.1109/TNSRE.2016.2539390.

Rajagopal A, Dembia CL, DeMers MS, Delp DD, Hicks JL, Delp SL, 2016 Fullbody musculoskeletal model for muscle-driven simulation of human gait. IEEE Trans. Biomed. Eng 63, 2068–2079. 10.1109/TBME.2016.2586891. [PubMed: 27392337]

Raknim P, Lan K-C, 2016 Gait monitoring for early neurological disorder detection using sensors in a smartphone: validation and a case study of parkinsonism. Telemed. J. E-Health Off. J. Am. Telemed. Assoc 22, 75–81. 10.1089/tmj.2015.0005.

Ratner A, Bach SH, Ehrenberg H, Fries J, Wu S, Ré C, 2017 Snorkel: rapid training data creation with weak supervision. Proc. VLDB Endow 11, 269–282. 10.14778/3157794.3157797.

Redfield MT, Cagle JC, Hafner BJ, Sanders JE, 2013 Classifying prosthetic use via accelerometry in persons with transtibial amputations. J. Rehabil. Res. Dev 50, 1201–1212. 10.1682/JRRD.2012.12.0233. [PubMed: 24458961]

Ries AJ, Novacheck TF, Schwartz MH, 2014 A data driven model for optimal orthosis selection in children with cerebral palsy. Gait Post. 40, 539–544. 10.1016/j.gaitpost.2014.06.011.

Rissanen SM, Kankaanpaa M, Meigal A, Tarvainen MP, Nuutinen J, Tarkka IM, Airaksinen O, Karjalainen PA, 2008 Surface EMG and acceleration signals in Parkinson's disease: feature extraction and cluster analysis. Med. Biol. Eng. Comput 46, 849–858. 10.1007/s11517-008-0369-0. [PubMed: 18633662]

Rodriguez-Martin D, Sama A, Perez-Lopez C, Catala A, Moreno Arostegui JM, Cabestany J, Bayes A, Alcaine S, Mestre B, Prats A, Crespo MC, Counihan TJ, Browne P, Quinlan LR, OLaighin G, Sweeney D, Lewy H, Azuri J, Vainstein G, Annicchiarico R, Costa A, Rodriguez-Molinero A, 2017 Home detection of freezing of gait using support vector machines through a single waist-worn triaxial accelerometer. PloS One 12, e0171764 10.1371/journal.pone.0171764. [PubMed: 28199357]

Rousseeuw PJ, 1987 Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math 20, 53–65. 10.1016/0377-0427(87)90125-7.

Roy SH, Cheng MS, Chang SS, Moore J, De Luca G, Nawab SH, De Luca CJ, 2009 A combined sEMG and accelerometer system for monitoring functional activity in stroke. IEEE Trans. Neural Syst. Rehabil. Eng. Publ. IEEE Eng. Med. Biol. Soc 17, 585–594. 10.1109/TNSRE.2009.2036615.

Roy SH, Cole BT, Gilmore LD, De Luca CJ, Thomas CA, Saint-Hilaire MM, Nawab SH, 2013 High-resolution tracking of motor disorders in Parkinson's disease during unconstrained activity. Mov. Disord. Off. J. Mov. Disord. Soc 28, 1080–1087. 10.1002/mds.25391.

Rozumalski A, Schwartz MH, 2009 Crouch gait patterns defined using k-means cluster analysis are related to underlying clinical pathology. Gait Post. 30, 155–160. 10.1016/j.gaitpost.2009.05.010.

Ryan W, Harrison A, Hayes K, 2006 Functional data analysis of knee joint kinematics in the vertical jump. Sports Biomech. 5, 121–138. 10.1080/14763141.2006.9628228. [PubMed: 16521626]

Sagawa Y, Watelain E, De Coulon G, Kaelin A, Gorce P, Armand S, 2013 Are clinical measurements linked to the gait deviation index in cerebral palsy patients? Gait Post. 38, 276–280. 10.1016/j.gaitpost.2012.11.026.

Samà A, Pérez-López C, Rodríguez-Martín D, Català A, Moreno-Aróstegui JM, Cabestany J, de Mingo E, Rodríguez-Molinero A, 2017 Estimating bradykinesia severity in Parkinson's disease by analysing gait through a waist-worn sensor. Comput. Biol. Med 84, 114–123. 10.1016/j.compbiomed.2017.03.020. [PubMed: 28351715]

Schwartz MH, Rozumalski A, Novacheck TF, 2014 Femoral derotational osteotomy: surgical indications and outcomes in children with cerebral palsy. Gait Post. 39, 778–783. 10.1016/j.gaitpost.2013.10.016.

Schwartz MH, Rozumalski A, Truong W, Novacheck TF, 2013 Predicting the outcome of intramuscular psoas lengthening in children with cerebral palsy using preoperative gait data and the random forest algorithm. Gait Post. 37, 473–479. 10.1016/j.gaitpost.2012.08.016.

Schwarz G, 1978 Estimating the dimension of a model. Ann. Stat 6, 461–464. 10.1214/aos/1176344136.

Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, Summers RM, 2016 Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging 35, 1285–1298. 10.1109/TMI.2016.2528162. [PubMed: 26886976]

Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, Hubert T, Baker L, Lai M, Bolton A, Chen Y, Lillicrap T, Hui F, Sifre L, van den Driessche G, Graepel T, Hassabis D, 2017 Mastering the game of Go without human knowledge. Nature 550, 354–359. 10.1038/nature24270. [PubMed: 29052630]

Sutton RS, Barto AG, 1998 Reinforcement Learning I: Introduction

Tibshirani R, 1996 Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Methodol 58, 267–288.

Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K, 2005 Sparsity and smoothness via the fused lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 67, 91–108.

Trost SG, Fragala-Pinkham M, Lennon N, O'Neil ME, 2016 Decision trees for detection of activity intensity in youth with cerebral palsy. Med. Sci. Sports Exerc 48, 958–966. 10.1249/MSS.0000000000000842. [PubMed: 26673127]

Vakanski A, Ferguson JM, Lee S, 2016 Mathematical modeling and evaluation of human motions in physical therapy using mixture density neural networks. J. Physiother. Phys. Rehabil 1.

Wei T-S, Liu P-T, Chang L-W, Liu S-Y, 2017 Gait asymmetry, ankle spasticity, and depression as independent predictors of falls in ambulatory stroke patients. PLoS ONE 12 10.1371/journal.pone.0177136.

Yu L, Xiong D, Guo L, Wang J, 2016 A remote quantitative Fugl-Meyer assessment framework for stroke patients based on wearable sensor networks. Comput. Methods Prog. Biomed 128, 100–110. 10.1016/j.cmpb.2016.02.012.

Zhang Z, Fang Q, Gu X, 2016 Objective assessment of upper-limb mobility for poststroke rehabilitation. IEEE Trans. Biomed. Eng 63, 859–868. 10.1109/TBME.2015.2477095. [PubMed: 26357394]

Zwick EB, Leistritz L, Milleit B, Saraph V, Zwick G, Galicki M, Witte H, Steinwender G, 2004 Classification of equinus in ambulatory children with cerebral palsy-discrimination between dynamic tightness and fixed contracture. Gait Post. 20, 273–279. 10.1016/j.gaitpost.2003.10.002.

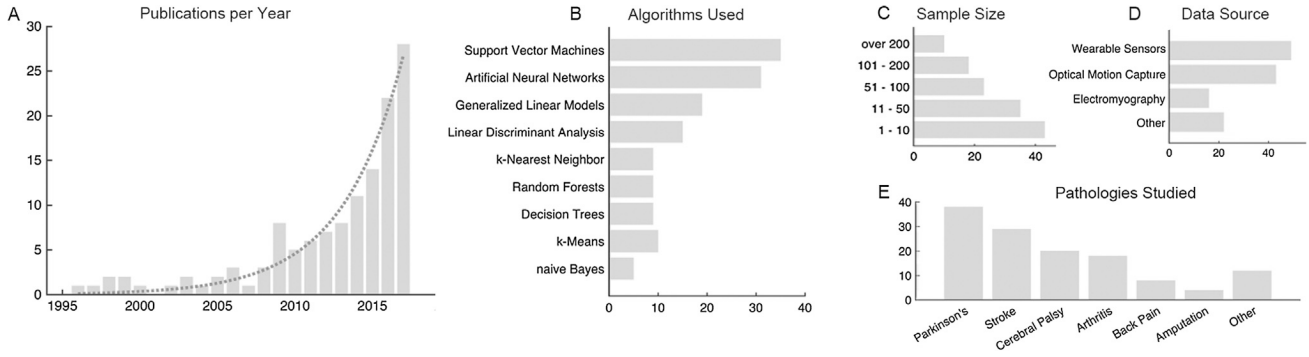

Fig. 1.

Study Characteristics. (A) The use of data science methods in human movement biomechanics studies that focus on neuromuscular and musculoskeletal pathologies has increased exponentially in recent years. (B) The mostly commonly used algorithms in the reviewed articles were support vector machines, neural networks, and generalized linear models, including linear and logistic regression. (C) The number of observations (subjects) in most studies was small to moderate. (D) Wearable sensors were the leading source of information, following by optical motion tracking. (E) The most commonly studied conditions were Parkinson's, stroke, cerebral palsy, and arthritis.

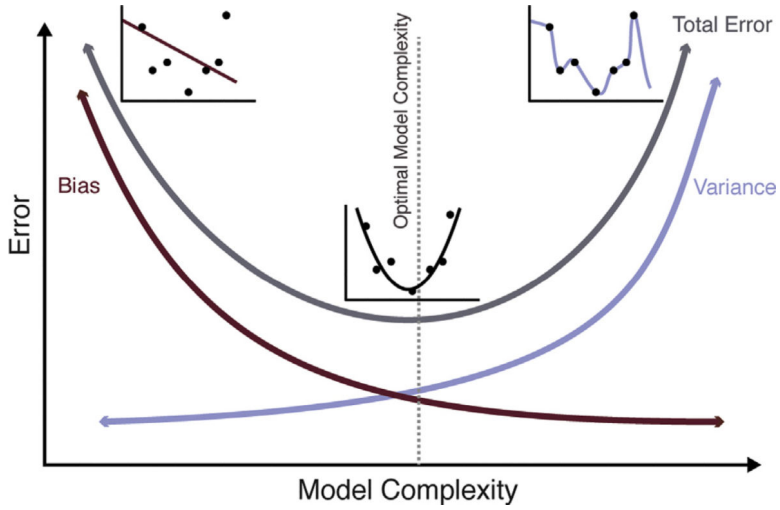

Fig. 2.

Model Complexity and the Bias-Variance Trade-off. In supervised learning tasks, bias is error that results from incorrect assumptions (e.g., fitting a linear model when the underlying relationship is not linear), causing algorithms to miss important relationships between features and labels, while variance is error that results from sensitivity to small variations, causing algorithms to model noise in the training data. The bias-variance tradeoff refers to the process of minimizing bias and variance in order to train models that can generalize well to non-training data. Simple models may be characterized by high bias and low variance (underfit to the training data), while complex models may be characterized by low bias and high variance (overfit to the training data).

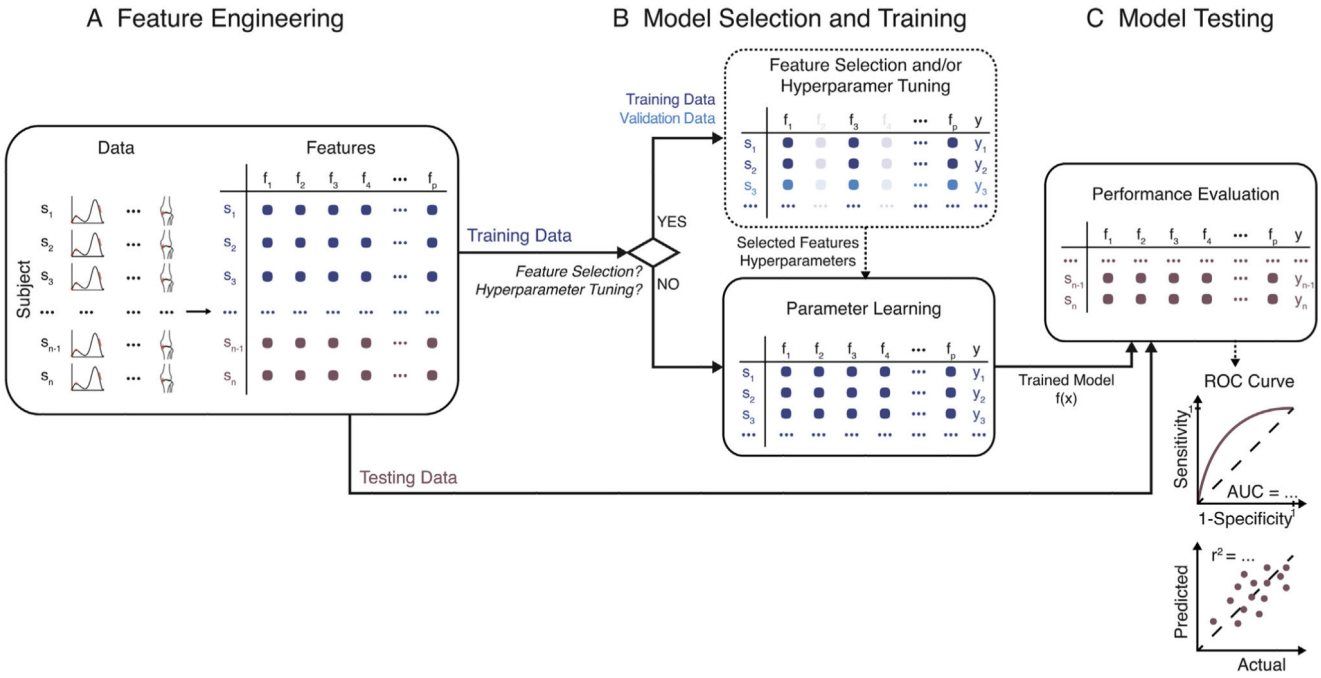

Fig. 3.

Framework for Building Robust Predictive Models. (A) When the data are high-dimensional, feature engineering is first used to derive lower-dimensional representations of the data. This can be accomplished through domain knowledge (i.e., extract important gait characteristics) or automated techniques, such as principal component analysis, fast Fourier transforms, and other approaches. (B) After this step, the data are split into training and test sets, with the test set aside for performance evaluation. The model selection and training step, which uses only training data, may focus solely on parameter learning (e.g., learn regression coefficients) or may include one or two additional steps: feature selection and hyper-parameter tuning. Feature selection is performed to ensure that only relevant features are included in the model. This often improves model performance. Hyper-parameter tuning is performed to select additional parameters in complex models (e.g., select the degree of the polynomial if not a linear model). To avoid overfitting, both feature selection and hyper-parameter tuning are carried out using training data alone, which are further split into training and validation sets. (C) Last, the trained model is tested on held-out test data and comprehensive performance metrics that are more meaningful than model accuracy are reported.

Table 1

Search terms used to identify studies that used machine learning methods to investigate movement biomechanics in populations with common musculoskeletal and neuromuscular diseases.

	Specific Terms
Machine learning method	support vector OR neural network* OR random forest* OR principal component OR linear discriminant OR lasso OR k means OR k-nn OR knn OR k-nearest* OR dimensionality reduction OR feature selection OR independent component analysis OR decision tree* OR regression model* OR logistic regression* OR reinforcement learning OR graphical model* OR Bayesian OR *fuzzy* AND
Neuromuscular or musculoskeletal condition affecting mobility	osteoarthritis OR osteoporosis OR osteonecrosis OR cerebral palsy OR running injur* OR running-related injur* OR stroke OR fracture OR Parkinson* OR polio OR carpal tunnel OR gout OR arthritis OR back pain OR fibromyalgia OR spinal cord injury OR amputee* OR joint implant OR joint replacement OR paralysis OR lumbar stenosis OR ACL tear OR meniscal tear OR muscular dystrophy OR muscular atrophy OR amyotrophic lateral sclerosis OR myopathy AND
Biomechanical outcome measure	gait OR motion OR movement OR kinematic* OR kinetic* OR accelerometer OR wearable OR IMU OR inertial measurement unit OR electromyogra* OR EMG OR force plate OR motor OR ground reaction force OR biomechanic*