Universiteit Leiden

Master Computer Science

A Machine Learning Approach to Identify Potential Biomarkers for Neuromuscular Diseases Using a Device That Simulates Regular Daily Upper Limb Activity

Name: M.A. Bax
Student ID: s1531298

Date: July 14, 2023

Specialisation: Data Science

1st supervisor: Dr. E.M. Bakker
2nd supervisor: Prof. Dr. M.S.K. Lew
Daily supervisors: Dr. R.J. Doll & I. van den Heuvel

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Biomarkers are an important tool to quantify the effect of a new drug in early-stage clinical drug research. In clinical research into neuromuscular diseases (NMDs), the most common methods that are used to measure these biomarkers, for example, using electromyography (sEMG), are often invasive, expensive, and/or labor-intensive. To mitigate this, a device that simulates a regular daily activity using the upper limbs like opening a jar, is introduced. It could potentially be used to collect data from which new biomarkers could be derived. At the Centre for Human Drug Research (CHDR), which specializes in early-stage clinical drug research, 75 test subjects performed a number of tasks using this device, while sEMG data was being recorded simultaneously. We built a classification model that retrieves features from the data, and is able to use these to differentiate between healthy test subjects and patients with NMDs. By experimenting with different features and classification algorithms, we identify features from both sEMG data and from data from the device that are influential in classification. Only using features derived from a task during which participants were instructed to hold open the device for as long as possible, a model was trained and evaluated using nested cross-validation and achieved an average F1 score of 0.77. The features that were found to be most influential were the time the device was held open, as well as a combination of features pertaining to the grip on the device by the other hand that was holding the device.

Contents

1 Introduction

Neuromuscular diseases (NMDs) comprise a wide range of diseases that affect the peripheral nervous system, skeletal muscles, or connections between the two. NMDs are often progressive and can cause muscle weakness, sensory loss, pain, fatigue, autonomic dysfunction, or postural abnormalities [14][4]. Most NMDs have no cure and this wide range of symptoms makes treatment and clinical research into new treatments difficult [17].

In early phase clinical trials, where potential new drugs are first tested on human test subjects, it can be important to be able to quantify disease severity and the possible effect a new drug is having on this severity. This is often done using biomarkers. Biomarkers are defined by the US National institute of Health as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention or other health care intervention" [7]. In other words, a biomarker is some measure that can be used to indicate the presence of a disease, its severity, and/or its reaction to treatment.

Ideally, biomarkers should be generally easy to measure, non-invasive, and time and cost effective. In early phase clinical trial, this could mean that measurements would not have to be performed by a medical doctor, but could be done by a research assistant or medical student. This would reduce operating costs and widen the pool of potential employees able to perform the measurement. Biomarkers retrieved from surface electromyography (sEMG) data are a common source of biomarkers for NMDs. Although collecting sEMG data is non-invasive, a trained medical professional and a relatively expensive setup are needed. Other commonly used biomarkers for NMDs are muscle MRI, nerve ultrasound, or muscle or nerve biopsy [1]. Biopsies are invasive and require further analysis of tissue in a laboratory setting, which increases necessary costs and effort. Muscle MRI and nerve ultrasound are less invasive, but are time consuming and require trained personnel and expensive equipment.

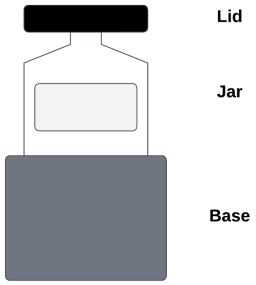


Figure 1: The Jar Device. One hand is to be placed on the black lid, while the other hand is placed around the white jar. The jar is secured onto the dark base.

The need for a more practical method of quantifying a loss in upper limb function due to NMDs, has led to the development of a new device which simulates daily upper limb activity, shown in Figure 1. This device simulates the daily task of opening a jar, which is a task that a patient with an NMD might have difficulties with. The device consists of a standing bottle with a lid that can be twisted. The torque required to twist the lid can be controlled. The

system measures the rotation of the lid and the gripping force on the jar itself. For simplicity and readability the device will be called the "Jar Device" in the remainder of this report
The rest of the report is structured as follows: The related work section outlines what research has been done with regards to NMD biomarkers, and using machine learning to find new candidates. The fundamentals section contains a list of definitions of the concepts used in the following sections. The methods section outlines the datasets that were used, the classification model, and feature importance analysis. The results section shows the performance of the models that were trained and the feature importance in these models. The discussion contains a review of the dataset, the methods, and an analysis of the results, as well as possible future research in this area. And finally, the conclusion includes a brief summary of the contributions of this report and some concluding remarks.

2 Related Research

Scotton et al. [17] lay out three different types of biomarkers in NMDs: genetic biomarkers, such as polymorphisms or allele variations in DNA and RNA, proteomic biomarkers, which can be measured from bodily fluids, and other biomarkers, such as MRI images of affected muscle tissue. The methods by which these biomarkers are discovered and validated are based on scientific or clinical knowledge, or "omic" technologies.

Using techniques like these a number of biomarkers for NMDs have been found. Barp et al. [1] discuss the application of biochemical biomarkers in the most prevalent NMDs, that can be used for diagnostics, prognostics, or therapeutics. Critical issues for the introduction of new biomarkers are also addressed. To facilitate easy and reliable collection of these biomarkers, they should ideally be low-cost, non-invasive, less time-consuming, and less operator dependent.

As mentioned above, one possible source of biomarkers for NMDs is sEMG signals. Flood et al. [6] study the sEMG signals in patients with Parkinson's disease (PD) and in healthy test subjects. They find that characteristics of the sEMG signals recorded from quadriceps and hamstring muscles during isometric leg extensions differ between healthy test subjects and patients with PD. Determinism (a measure of the degree of repeating non-linear patterns) was shown to be significantly higher, sample entropy (a measure of complexity) was shown to be significantly lower, and intermuscular coherence (a measure of correlation between EMG signals) is shown to be higher in PD patients. Besides sEMG signals, Emamzadeh et al. [5] outline a number of biomarkers for PD that can be found through other techniques. These include a variety of modern imaging techniques such as transcranial B-mode sonography (TCS) or magnetic resonance imaging (MRI), and neurochemical methods that use levels of certain proteins like Glial Fibrillary Acidic Protein (GFAP) or proteasomes.

The biomarkers mentioned in this section all need to be measured in a clinical setting and by trained medical staff. Youn et al. [20] give an overview of potential digital biomarkers that are measured using digital biosensors. A patient with an NMD or a participant in a clinical trial could wear or use these digital biosensors at home, which brings down cost and effort, while being able to measure the biomarker continuously for a longer period of time.

With the advent of machine learning and artificial intelligence in the past two decades, new methods of identifying potential biomarkers have been introduced. Other than differences between sEMG signals of healthy test subjects and patients with PD that were outlined above, Kugler et al. [9] also identify a number of specific biomarkers from sEMG signal for PD using machine learning. 12 channels of sEMG were recorded of ten participants, five with PD and five without PD, during standardized gait tests. Four time-domain and four frequency-domain features were extracted and used to train a Support Vector Machine classifier. Using leave-one-out cross-validation, specificity and sensitivity were at 0.9, and kurtosis and mean frequency were identified as the best features and most promising biomarkers.

Using a dataset of sEMG recordings of the upper limbs from four healthy test subjects and four patients with PD, Rezaee et al. [16] utilize three different pre-trained deep learning structures to generate feature vectors. The most promising features are then selected using a novel soft

ensembling of subset feature selection methods, and used to train an SVM. This approach achieved a specificity and sensitivity of 0.99. However because these features are created using deep learning, they are difficult to interpret as biomarkers.

Machine learning has been applied to identify new biomarkers for other diseases as well. Parmar et al. [12] use a dataset of CT images of 464 patients with lung cancer and extracts 440 features from the images. These images are fed into a range of feature selection algorithms and classifiers, of which the random forest and naïve bayes classifiers, in combination with Minimum redundancy maximum relevance, Mutual information feature selection, and Wilcoxon feature selection algorithms were most effective. Similarly, Abou Tabl et al. [18] study the application of machine learning to the identification of biomarkers for breast cancer. In this case genes function as features, and after feature selection, they are fed into a multi-class classifier. The classes are based on survivability and given therapy. Genes that are used in the model are identified as potential biomarkers and relevant literature confirms a relation between these biomarkers and breast cancer (survivability).

More sophisticated machine learning models have been used towards the identification of potential biomarkers as well. Putin et al. [15] use an ensemble of 21 deep neural networks (DNNs) to identify potential biomarkers of human aging (meaning the correlation between chronological and biological age). A relatively large dataset of 60,000 blood samples was used to successfully train these DNNs and five potential biomarkers for a humans chronological age were identified.

3 Fundamentals

This section contains descriptions of measures and concepts that are mentioned in the rest of the report, but are otherwise not defined. Concepts are arranged alphabetically.

ANOVA F-statistic The ANOVA F-statistic is the ratio between the between-group variance and the within-group variance:

$$F = \frac{\text{between-groups variance}}{\text{within-group variance}} \tag{1}$$

Here, the between-groups variance of a variable is the variance of the averages of the variable within each class, around the average over all classes. The within group variance is the variances within each class. Between-groups variance and within-group variance are sometimes called explained variance and unexplained variance, respectively. In this report, the F-statistic is used for feature selection by selecting the features (or variables) with the highest F-statistic.

Decision Tree Classifier The Decision Tree Classifier is a supervised machine learning model that can be adapted for classification or regression tasks. A decision tree consists of the root, nodes, branches, and leaves. A dataset enters the decision tree at the root and is split at the first node. This split consists of a feature and a cut-off point, where data points with a value for that feature on one side of the cut-off are sent to one branch and data points with a value on the other side are sent to the other branch. The branches that come from the first node are then split in new nodes and this process is repeated until the leaves are reached. The final classification is done according to the leave in which a data point ends up.

What features and cut-offs are used for the splits and when to stop splitting and create a leaf is determined by a number of user-specified hyperparameters. The quality of splits is determined by 1 of 2 measures: Gini impurity or entropy, where a minimal value indicates the optimal split. Gini impurity is the probability of misclassifying a data point in a branch if the data point would be randomly classified according to the class distribution in that branch. Entropy is a metric that describes the impurity of the datasets after a split. The max depth parameter regulates when the tree stops making splits and ends with a leaf, by giving the tree a maximum depth after which no more splits can be made.

The importance of each feature is defined as the cumulative loss in Gini impurity or entropy over all nodes that use this feature in the decision tree.

F1-Score The F1-score is a performance metric that can be used to measure the performance of a classifier. It is defined as the harmonic mean of precision and recall, which is calculated as follows:

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{2}$$

Here, P is the precision and R is the recall.

Fourier Transform The Fourier transform is a transform that converts a function of time to a function of frequency using the following formula:

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(t) e^{-i2\pi\xi t} \tag{3}$$

Here, $\hat{f}$ is the transformed function, ξ is the frequency, and t is the time. The output of this function can be a complex-valued function. However, because the functions that are used in this report are strictly real-valued, the Fourier transform results in real-valued functions. The Fourier transform gives the frequency components of a function. In our studies we will use the discrete fourier transform.

Nested Cross-Validation Nested cross-validation is an approach for optimizing hyper-parameters and evaluating performance of a model, while avoiding the issue of overfitting on the training dataset. A dataset is split into training and testing data k times. Here k is the number of cross-validation folds of the outer cross-validation loop. Then each training set is used to optimize the hyperparameters of the model, using a grid search with cross-validation. This results in a trained model with optimized hyperparameters in each outer fold and a corresponding test set to evaluate each model. The model can then be evaluated by analyzing the average performance over the outer cross validation folds. A visualisation of nested cross validation is part of the processing pipeline shown in Figure 2 on page 15

Precision Precision is a performance metric that can be used to measure the performance of a classifier. If a dataset consists of data points in 2 classes, positive and negative, the precision of a classification of this data is defined as the fraction of correctly classified data points among all data points classified as positive. It can be calculated using the following formula:

$$P = \frac{TP}{TP + FP} \tag{4}$$

Here, TP is the number of True Positive classifications and FP is the number of False Positive classification.

Random Forest Classifier A random forest is a model that consists of an ensemble of decision tree classifiers. Each decision tree makes a separate classification and the class with the most votes is chosen as the model's final prediction. The advantage of an ensemble of trees over an individual decision tree is that individual trees typically exhibit high variance and tend to over fit. By training each tree in a random forest on a different subset of the data, drawn with replacement, and not allowing trees to be too deep, variance and overfitting is reduced. Besides the same user-specified parameters as individual decision trees, the number of trees in the ensemble can also be specified by the user.

The importance of each feature is defined similarly to individual decision trees, where loss in Gini impurity or entropy is averaged over all trees in the random forest.

Recall Recall is a performance metric that can be used to measure the performance of a classifier. If a dataset consists of data points in 2 classes, positive and negative, the recall of a classification of this data is defined as the fraction of positive data points that are correctly classified. It can be calculated using the following formula:

$$R = \frac{TP}{TP + FN} \tag{5}$$

Here, TP is the number of True Positive classifications and FN is the number of False Negative classifications.

ROC-Curve The ROC curve is a plot that indicates the performance of a binary classifier. It is created by plotting the true positive rate against the false positive rate at various threshold settings. The area under the ROC curve is often used as a performance metric for a classifier. A random classifier's ROC curve is a straight diagonal line with an area under the curve of 0.5. A perfect classifier has a false positive rate of 0 and a true positive rate of 1, meaning its ROC curve starts at coordinates [0, 0] in the ROC space, to [0, 1], to [1, 1], resulting in an area under the curve of 1. This means that an area under the ROC curve that is close to 1 indicates a relatively strong classifier.

Support Vector Machine The support vector machine is a supervised machine learning model that can be adapted for classification or regression tasks. When using the model for a binary classification task, data points are viewed as points in n-dimensional space, where n is the dimension of the data points. Data points are then separated by creating a hyperplane that has the largest distance to the nearest training-data point of any class. New data points are then classified by which side of the hyperplane they fall on. The margin of an SVM classifier is the distance between the nearest training-data points of either class.

In reality, most datasets are not perfectly linearly separable. In these cases we adapt the model to allow for some misclassifications, by using a soft margin. This soft margin necessitates a user-specified parameter that regulates roughly how many misclassifications are allowed. This means that this user-specified parameter specifies the trade-off between the number of misclassifications and the size of the margin.

The importance of each feature in an SVM is indicated by the coordinates of the vector orthogonal to the hyperplane. A high value for a coordinate, corresponding to some feature, relative to other coordinates, indicates that this feature has a larger influence on the final classification.
SVMs are able make use of a number of kernels. Kernels are functions that transform the feature space, before a hyperplane is used make a seperation between classes. Common types of kernel functions are linear, polynomial, Radial Basis Function (RBF), or Sigmoid. Here, we only use the linear kernel, because the resulting separating hyperplane will still be in the same space as your input features. Therefore, its coefficients can be viewed as weights of the features, which are needed to analyze what features are most influential in the classification process.

4 Methods

4.1 Dataset

The data were collected by CHDR (Centre for Human Drug Research) and contain data from 60 healthy test subjects and 15 patients with neuromuscular diseases. Participants of the study performed 6 tasks on multiple occasions. Data was recorded using the Jar Device and 16 sEMG channels.

4.1.1 Study Protocol

To collect data, 60 healthy test subjects, and 15 patients with neuromuscular diseases (Parkinson's Disease, Myasthenia Gravis, or Inclusion Body Myositis) were recruited. Participants performed 6 tasks with the Jar Device, where 3 tasks were done at 3 different resistance settings, and three other tasks (ones that did not involve twisting the lid of the Jar Device) were done once, resulting in a total of 12 tasks. Healthy test subjects performed all 12 tasks on 4 occasions within 1 day. Patients with NMDs performed a set of tasks on 2 occasions, spread over 2 days, with 1 to 3 weeks in between occasions. sEMG data was not recorded on the first occasion of the patients with NMDs. The 3 different resistance settings were selected based on the subject's age. During the tasks, the angle of the lid of the Jar Device, the grip on the base of the Jar Device, as well as 16 channels of EMG data were recorded.

4.1.2 Healthy Subjects

Inclusion criteria for healthy test subjects were being between 21 and 80 years old and having a body mass index (BMI) between 18 and 35 kg/m2. Exclusion criteria were any diseases or conditions that impede upper limb function, not being able to understand the study requirements and lifestyle restrictions, or not being able to give informed consent. Lifestyle requirements for healthy test subjects were not taking in any medication that could affect upper limb function, not consuming any alcohol within 24 hours before the study day, not doing any strenuous physical activity within 48 hours before the study day, and not consuming more than 800 mg of caffeine per day within 7 days of the stud day. On Occasion 1 and Occasion 3, subjects were restricted in their hand and finger placement as well as in positioning of the arm. On Occasion 2 and Occasion 4, subjects were not restricted in the positioning of their arms.

4.1.3 Patients with NMDs

Inclusion criteria for patients were having a neuromuscular disease, and a self-reported weakness of the hands. Exclusion criteria were being unable to perform any of the tasks due to physical limitations, as well as not being able to understand the study requirements and lifestyle restrictions, or not being able to give informed consent. Patients were not restricted in the positioning of their arm.

4.1.4 Tasks

The 6 tasks that participants performed are defined as follows:

Task 1 The goal of this task is to twist the lid of the Jar Device 20 degrees and keep it at that angle. The participant is free to choose in what direction to twist the lid. The task is stopped when the lid reaches 15 degrees or when time runs out after 999 seconds. This task was done at 3 different resistance settings for a total of 3 tasks. These are called 1a, 1b, 1c, where 1a is the heaviest setting and 1c is the lightest.

Task 2 The goal of this task is to twist the lid of the Jar Device to 15 to 20 degrees, hold it there for 2 seconds, release it for 2 seconds, and repeat this for 30 seconds in total. The participant is free to choose in what direction to twist the lid. A metronome is used to assist the participant. This task was done at 3 different resistance settings for a total of 3 tasks. These are called 2a, 2b, 2c, where 2a is the heaviest setting and 2c is the lightest.

Task 3 The goal of this task is to grip the base of the Jar Device with maximum grip force, hold this for 2 seconds, release for 2 seconds, and repeat this for a total of 30 seconds. A metronome is used to assist the participant.

Task 4 The goal of this task is to grip the base of the Jar Device with maximum grip force, immediately release, and repeat this as often as possible within 30 seconds.

Task 5 The goal of this task is to twist the lid of the Jar Device to its maximum angle of 40 degrees, immediately release, and repeat this as often as possible within 30 seconds. The participant is free to choose in what direction to twist the lid. This task was done at 3 different resistance settings for a total of 3 tasks. These are called 5a, 5b, 5c, where 5a is the heaviest setting and 5c is the lightest.

Task 6 The goal of this task is to grip the base of the Jar Device with maximum grip force, and to maintain this grip as long as possible. The task is stopped when grip force reaches 75% of its own maximum or when time runs out after 999 seconds.

4.1.5 The Jar Device

The Jar Device is a device that simulates everyday tasks like twisting the lid of a jar and gripping an object. The Jar Device records the angle of the lid (in degrees) and the grip force on the base (in kg). The torque needed to open the Jar Device can be chosen manually and, as mentioned above, tasks that require the twisting of the lid of the Jar Device are performed at 3 different resistance levels. The resistance level is constant throughout a single task and does not change depending on the angle of the lid. The sampling frequency of the Jar Device is not constant throughout a single task.

4.1.6 sEMG

Concurrent with the recording of Jar Device data, 16 channels of sEMG data were also recorded. These recordings were made of 3 flexor muscles in both lower arms (musculus flexor carpi radialis (FCR), musculus flexor carpi ulnaris (FCU) and musculus flexor digitorum superficialis (FDS)), 3 extensor muscles in both lower arms (musculus extensor carpi radialis (ECR), musculus extensor carpi ulnaris (ECU) and musculus extensor digitorum (ED)), and 2 muscles in both hands (musculus abductor pollicis brevis (APB) and the first dorsal interosseous muscle (FDI)). These muscles were identified using an anatomy atlas for needle EMG and palpation.

The sEMG device has a sampling frequency of 125 Hz. It contains an analogue first-order low-pass filter with a -3 dB point at 4.8 kHz, as well as a third-order digital sinc filter with a cut-off frequency of 0.27 times the sampling frequency within the analogue-to-digital converter to remove unwanted noise.

4.1.7 Data Pre-Processing

The Jar Device data was irregularly sampled, and therefore required resampling and interpolation to a frequency of 11.11 Hz, corresponding to a sampling period of 90 ms. Spikes in the sEMG signal are, by their nature, random. This is because if 2 or more motor units fire at the same time and are located near an electrode, they produce a strong superposition spike [15]. To remove these non-reproducible random spikes in the sEMG data, smoothing algorithms were applied. The data was first passed through a band-pass Butterworth filter with a lower cut-off frequency of 5 Hz and a higher cut-off frequency of 50 Hz, a rectifier, and a low-pass Butterworth filter with a cut-off frequency of 5 Hz.

4.1.8 Final Datasets

The study, as outlined above, resulted in 1 dataset of multivariate time series data for each of the 12 tasks. These datasets include all instances where sEMG and Jar Device data were recorded. Most participants will appear multiple times because the tasks were repeated on multiple occasions: 4 for healthy subjects and 2 for patients, on one of which no sEMG data was recorded. Each instance is an 18-dimensional multivariate time series. The first 2 dimensions have a sampling frequency of 11.11 Hz and contain the grip on the base of the Jar Device and the angle of the lid of the Jar Device, respectively. The 16 subsequent dimensions have a sampling frequency of 125 and contain the EMG signals from the subject's dominant arm and non-dominant arm, respectively. Every measurment that was done done using the lowest resistance level for the participant performing the measurement is grouped together in one dataset. Because the lowest resistance level is chosen based on the assumed (based on age) relative strength of each participant, it can be assumed that it is equally difficult for each participant to twist the lid. The same was done for the middle and highest resistance settings.

4.2 Processing Pipeline

To build a machine learning model that can distinguish between healthy test subjects and patients with neuromuscular diseases, and to identify and analyze useful (combinations of) features or biomarkers from this model, a processing pipeline was set up. This processing pipeline takes in the raw multivariate time series data, transforms this into a feature set, and returns a model and the features this model uses for classification. It first calculates a set of feature sets, and then uses these to simultaneously train a data scaler, a feature selection algorithm, and a classifier. The data scaler, feature selection algorithm, and classifier need to be trained and tested. To prevent overfitting, we apply nested cross-validation (see Section 3. Fundamentals). Figure 2 shows an overview of this processing pipeline.

4.2.1 Feature Creation

Each instance in the raw datasets contains an 18-dimensional multivariate time series. Using this multivariate time series, features are created. These features can be split into 2 types:

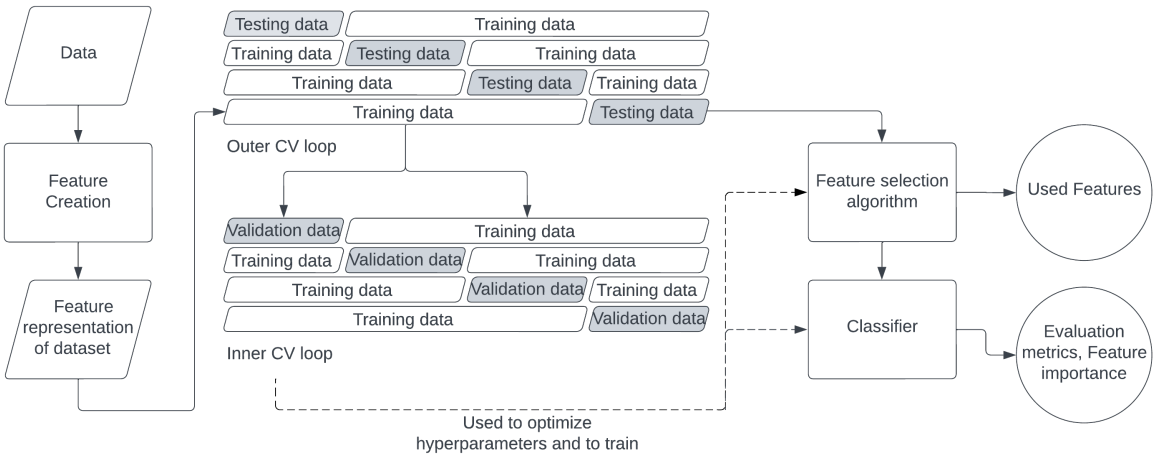


Figure 2: The processing pipeline that is used to identify which features are used for successful classification of the data. The raw data is turned into a feature set, which is used to train and test the models.

general features and specific features. Both types contain features from the time-domain and from the frequency domain for which the Fourier transform was used (see Section 3.).

Specific Features
Specific features are features that were specifically designed for the task at hand. An example of a specific feature would be the time a participant manages to keep the Jar Device opened during task 1a. The number of specific features and their precise definition differs per task. Appendix A contains all specific features and their exact definition, per task.

General Features
General features are calculated using functions not specifically designed for this task. This includes, among others, the minimum or maximum value throughout a time series, the average or standard deviation of a time series, or the average frequency. 12 general features are calculated for each dimension, leading to 192 features for each instance. Appendix A contains all general features and their exact definition.

Feature Sets
Table 1 contains a description of the number of patients and healthy test subjects, as well as the number of specific features based on Jar Device or sEMG data, for each feature set. Table 2 contains a lists of the specific features that are calculated for each task.

4.2.2 Data Scaler

After calculating features, each feature set is standardized using a data scaler. This entails that the mean of each feature's values in the training dataset is subtracted and the feature is scaled such that the training dataset has unit variance. This is done because features with higher variance, could inadvertently dominate the features with lower variance, which would then largely be ignored by the classifier [19].

Task	Patients	Healthy test subjects	Specific Jar Device features	Sepcific sEMG features	General Jar Device features	General sEMG features
1a	7	224	13	129	24	192
1b	14	226	13	129	24	192
1c	6	205	13	129	24	192
2a	7	225	11	34	24	192
2b	13	223	11	34	24	192
2c	16	213	11	34	24	192
3	16	237	9	34	24	192
4	16	237	6	33	24	192
5a	7	230	16	33	24	192
5b	13	234	16	33	24	192
5c	16	238	16	33	24	192
6	15	202	6	81	24	192

Table 1: The number of data points that were recorded, and the number of features that were created for each task.

Task 1	Task 2	Task 3	Task 4	Task 5	Task 6
Total Grip	Grip Frequency and Frequency Dominance	Grip Peak Polynomial Fit Parameters	Grip Frequency and Frequency Dominance	Grip Frequency and Frequency Dominance	Grip Hold Length
Grip Polynomial fit parameters	Angle-Grip Correlation	Grip Minimum Polynomial Fit Parameters	Grip Peak Polynomial fit Parameters	Angle-Grip Correlation	Time to Maximum Grip
Hold Length	Angle Peak Polynomial Fit Parameters	Grip Frequency and Frequency Dominance	Jamar-Jar Device Grip Difference	Total Grip	Grip Polynomial Fit Parameters
Angle Corrections	Angle Minimum Polynomial Fit Parameters	Jamar-Jar Device Grip Difference	AUC Difference Between Arms	Number of Openings	AUC Difference Between Arms
Angle Polynomial Fit Parameters	Angle Frequency and Frequency Dominance	Average Metronome Dominance	Frequencies and Frequency Dominances	Angle Peak Polynomial Fit Parameters	Frequencies and Frequency Dominances
Time to Maximum Angle	Average Metronome Dominance	AUC Difference Between Arms		Angle Minimum Polynomial Fit Parameters	Average rectified value slopes and intercepts
Frequencies and Frequency Dominances	AUC Difference Between Arms	Frequencies and Frequency Dominances		Angle Frequency and Frequency Dominance	Instantaneous median frequency slopes and intercepts
AUC Difference Between Arms	Frequencies and Frequency Dominances			Period Polynomial Fit Parameters	permutation entropy slopes and intercepts
Average Rectified Value Slopes and Intercepts				AUC Difference Between Arms	
Instantaneous Median Frequency Slopes and Intercepts				Frequencies and Frequency Dominances	
Permutation Entropy Slopes and Intercepts					

Table 2: The names of every specific feature that has been derived of each task. Descriptions of these features can be found in Appendix A.

4.2.3 Feature Selection

Each of the 12 feature sets, collected from the 12 corresponding tasks, contains relatively few instances and many features. A feature selection algorithm is applied to reduce the number

of features to a maximum of 10. This is done for 2 reasons: The first is that we want to keep the model, and which features it uses, interpretable. If classification were to be done using a combination of, for example, 100 features, the importance of each separate feature would be difficult to interpret. The second reason for applying a feature selection algorithm is to avoid the 'curse of dimensionality', which states that a high-dimensional dataset needs many instances to successfully train a classifier [2]. Because the number of positive cases is relatively low in each dataset, using a limited number of features helps in avoiding the curse of dimensionality.

Feature selection is based on the ANOVA (analysis of variance) F-value (see Section 3.), which estimates the degree of linear dependency between a feature and the target variable. The variables with the highest corresponding F-value are selected, with a maximum of 10 features being selected. If there are less than 10 features, all features are used, and no feature selection is applied.

4.2.4 Classifier

The selected, standardized features are then fed into a classifier. Here, we experiment with 3 different classifiers: a decision tree, a random forest, and a support vector machine (see section 3.). All 3 classifiers were implemented using sklearn [13], and are chosen over other popular classifiers because the relative importances of features after training the classifier are well defined. These importances can later be used to analyze which features are most influential.

4.2.5 Nested Cross-Validation

To evaluate the data scaler, feature selection algorithm, and classifier we use nested cross validation [11]. This is an approach to train models and optimize hyperparameters without risking overfitting. Nested cross validation uses 2 cross validation loops. In every outer fold, the dataset is split into a training dataset and a testing dataset. This training dataset is then split into a training dataset and a validation dataset in each inner fold. This inner training dataset and validation set are then used to optimize the models' hyperparameters using regular cross validation and a grid search of the hyperparameter space. Here, the F1-score (see Section 3.) is used to evaluate which hyperparameter setting performs best. Now, every outer cross-validation fold contains a model with optimized hyperparameters and these models can be evaluated using the test datasets from the outer folds. The hyperparameter spaces that were used are shown in Table 3.

In the outer nested cross validation loop, 20% of datapoints is set aside for testing and in the inner loop the same amount is set aside for validation. Data is sampled randomly using resampling with replacement, only ensuring a distribution of positive and negative cases like that of the original dataset. The outer cross-validation loop contains 100 folds and the inner cross-validation loop contains 25 folds.

4.2.6 Performance Metrics

As is shown in Table 1, the datasets that are used to train the models, are unbalanced, there are far more negative cases (healthy test subjects) than positive cases (patients with NMDs). Although accuracy is a popular and intuitive performance metric, it is not suited for models

Classifier	Hyperparameter	Possible values
Decision Tree	criterion	{"gini", "entropy"}
	max_depth	{3, 6, 9, None}
Random Forest	n_estimators	{50, 100}
	criterion	{"gini", "entropy"}
	max_depth	{1, 3, 5}
Support Vector Machine	C	{0.01, 0.1, 1, 10}

Table 3: The hyperparameter spaces of the three classifiers that were used during the grid search.

that are trained and tested on these unbalanced datasets, because if the model classifies every instance as negative, accuracy would still be high. 2 performance metrics that are of interest are recall and F1-score. Recall is the fraction of positive instances that are correctly classified. F1-score is the harmonic mean between recall and precision (see Section 3.). Using these 2 metrics ensures that more focus is placed on finding all positive instances, and that classifying every instance as negative is not incentivized. The average recall and F1-score are calculated to indicate the performance of the model, regardless of a specific train-test split. To further analyze the performance in the outer folds, confusion matrices and ROC-curves (see Section 3.) are created for every outer fold. Then the averages of these confusion matrices and ROC-curves are calculated to indicate the performance of the model, regardless of a specific train-test split.

4.2.7 Feature Importance

To interpret which of the selected features are actually used by the classifier and how influential they are relative to each other, we turn to feature importance. As mentioned earlier, feature importance is well defined for the 3 classifiers we are using. In the decision tree, a feature's importance is defined as the normalized cumulative reduction of the splitting criterion by that feature. It is defined similarly for a random forest, the only difference being that features can be used in multiple trees. In support vector machines weights are assigned to features. The sign of the dot product between a point in the feature space and a vector of these weights determines the final classification. These weights can be positive or negative and when the absolute value of one of these weights is relatively large, the corresponding feature can be interpreted as relatively important.

4.3 Experiments

Models using all possible combinations of data sources (Jar Device, sEMG, or both), types of features (specific, general, or both), and classifiers (decision tree, random forest, or SVM) are evaluated using the nested cross validation approach that was outlined previously. The average recall, average precision, average area under the ROC-curve, average confusion matrix, and average ROC-curve over the 100 outer cross-validation folds, are recorded. To further understand the role a particular feature plays in the classification process, the distribution of the feature within each of the 2 groups can be plotted.
To improve classification while only using the data obtained from the Jar Device, feature sets from multiple tasks are merged. This is done by concatenating any datapoints from different

tasks that have an identical subject number and occasion. The datasets from the 5 tasks that performed best when using only Jar Device features (1c, 2a, 4, 5a, and 5c) are chosen and any combination of these 5 feature set is used to train models. Again, the same 3 feature sets (specific, general, or both) and 3 classifiers (decision tree, random forest, or SVM) are studied.

5 Results

5.1 Single Task Experiments

To indicate the difference in performance between models that used only Jar Device data, sEMG data, or both, the best performing model per data source according to one or more numerical performance metrics (F1, Recall, or AUC ROC-curve) are shown in Table 4. To indicate how the choice of type of features impacts what model performs best, Table 5 shows the best performing model per type of feature.

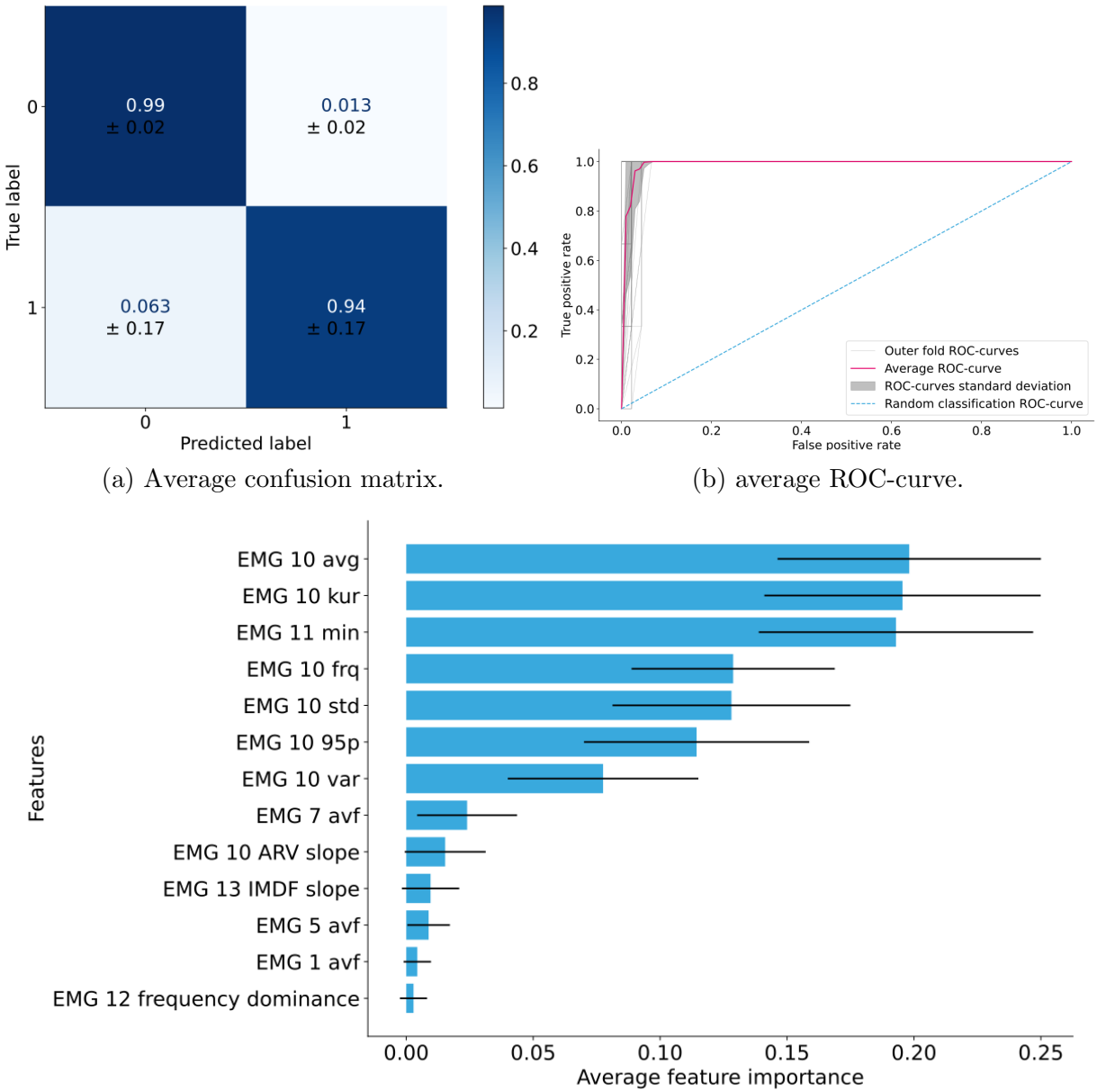
Data source	Task	Types of Features	Classifier	F1-score	Recall	AUC ROC-curve
Jar Device & sEMG	2b	General	Random Forest	0.865 ± 0.126	0.930 ± 0.151	0.991 ± 0.020
Jar Device & sEMG	4	General	Decision Tree	**0.886 ± 0.144**	0.870 ± 0.188	0.932 ± 0.094
Jar Device	1c	Specific	SVM	0.769 ± 0.241	0.695 ± 0.291	0.884 ± 0.210
Jar Device	1c	General & specific	Random Forest	0.326 ± 0.350	0.270 ± 0.303	0.989 ± 0.028
sEMG	1b	General & specific	Random Forest	0.878 ± 0.127	**0.937 ± 0.168**	**0.994 ± 0.009**
sEMG	4	General	Decision Tree	0.882 ± 0.149	0.867 ± 0.189	0.932 ± 0.095

Table 4: The models that performed best according to at least one of the three numerical performance metrics, per data source. The highest values for each of the 3 performance metrics are shown in bold type.

Data source	Task	Types of features	Classifier	F1-score	Recall	AUC ROC-curve
Jar Device & sEMG	2b	General	Random Forest	0.865 ±0.126	0.930 ±0.151	0.991 ±0.020
Jar Device & sEMG	4	General	Decision Tree	**0.886 ±0.144**	0.870 ±0.188	0.932 ±0.094
Jar Device & sEMG	5b	Specific	SVM	0.636 ±0.248	0.637 ±0.291	0.967 ±0.041
Jar Device	1c	Specific	SVM	0.769 ± 0.241	0.695 ± 0.291	0.884 ± 0.210
sEMG	1b	General & specific	Random Forest	0.878 ±0.127	**0.937 ±0.168**	**0.994 ±0.009**

Table 5: The models that performed best according to at least one of the three numerical performance metrics, per type of features that were used. The highest values for each of the 3 performance metrics are shown in bold type.

Figure 3 depicts the confusion matrix, ROC-curve and feature importance of the models that performed best over all the models that were tested, which used general and specific features derived from sEMG data from task 1b and a Random Forest classifier.



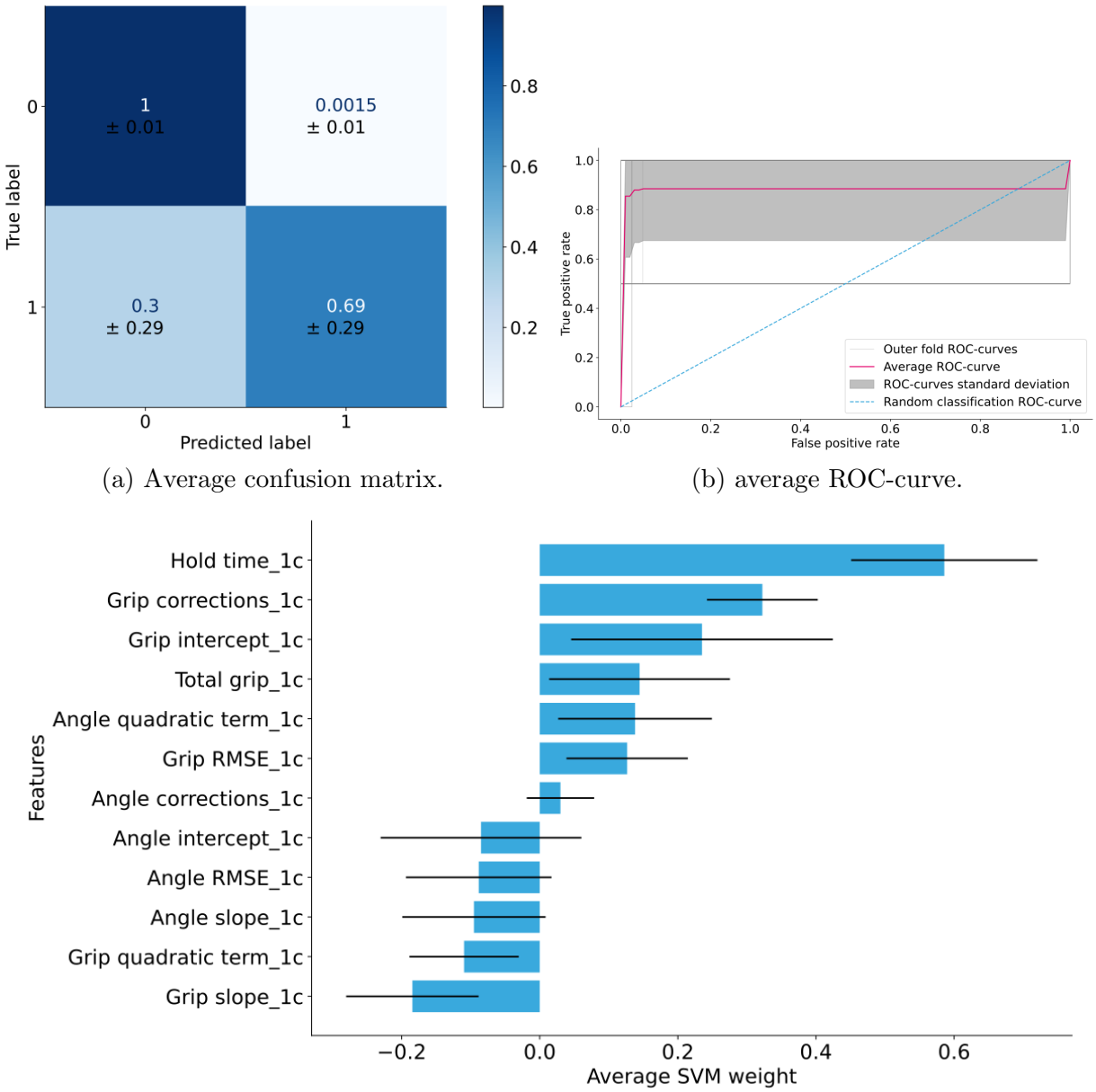
(a) Average confusion matrix.

(b) average ROC-curve.

(c) Average feature importances.

Figure 3: The average confusion matrix, average ROC-curve and average feature importances of the model that performed best over all the models that use features derived from Jar Device data, sEMG data, or both. This model used general features derived from sEMG data of task 1b and a Random Forest classifier.

Because the ideal biomarker would be derived only from Jar Device data, the model that performded best over all the models that only used Jar Device data is shown in Figure 4. Here the confusion matrix, ROC-curve and feature importance of this model, which used specific features from task 1c and an SVM classifier, are shown.To better understand the role the most influential features derived from Jar Device data play in the classification process of this model, the distribution of these features within the 2 groups are plotted. The distributions of the 5 features with the highest feature weights in the model shown in Figure 4, within each of the 2 groups are plotted in Figure 5.

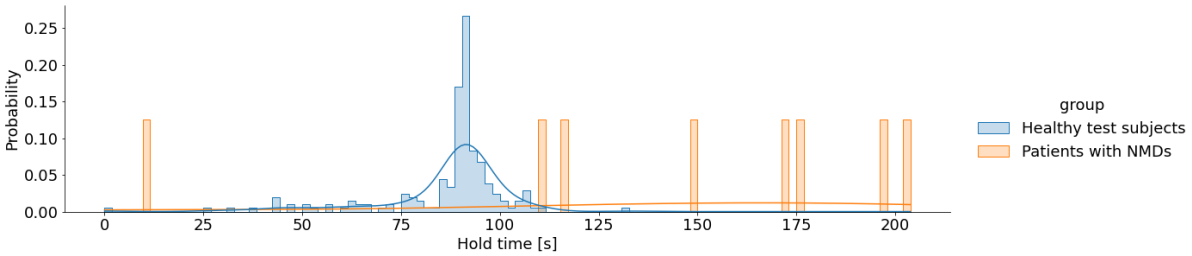


(a) Average confusion matrix.

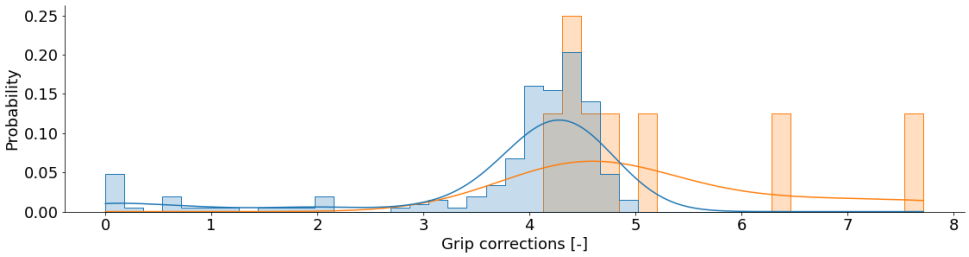
(b) average ROC-curve.

(c) Average feature weights.

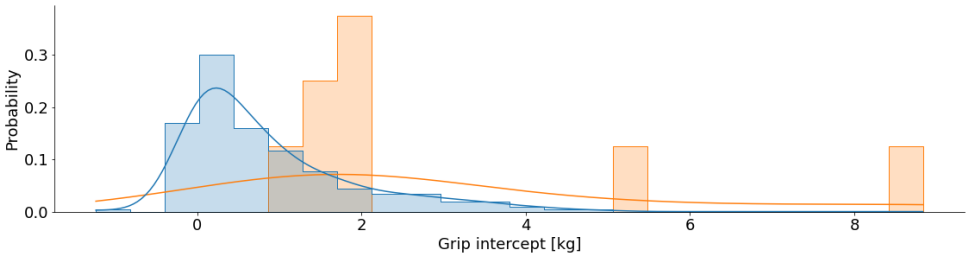
Figure 4: The average confusion matrix, average ROC-curve and average feature weights of the model that performed best over all the models that only use Jar Device data. This model used specific features derived from task 1c and an SVM classifier.



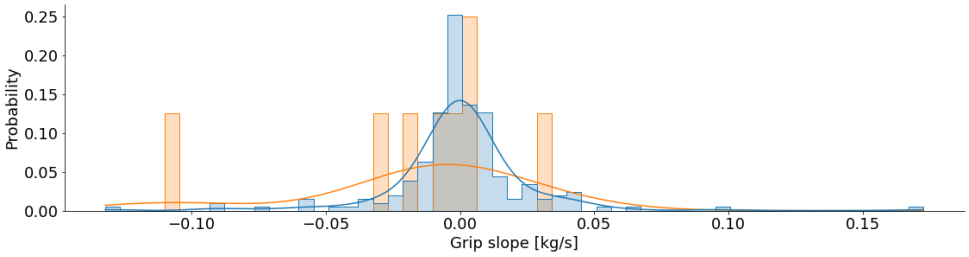
(a) Distribution of the hold time feature



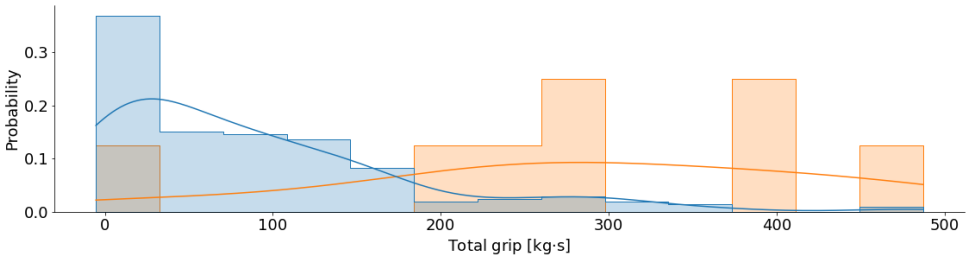
(b) Distribution of the grip corrections feature



(c) Distribution of the grip intercept feature



(d) Distribution of the grip slope feature



(e) Distribution of the total grip feature

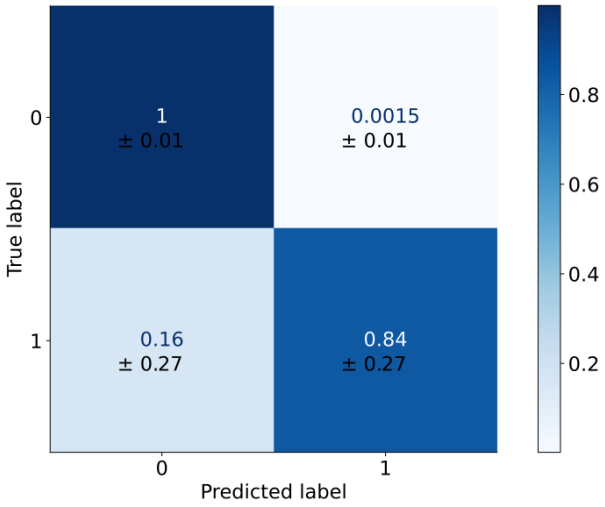
Figure 5: Distributions within each group of the five most influential features that were used in the model shown in Figure 4. This model was trained using specific features derived from Jar Device data of task 1c and an SVM classifier.

5.2 Experiments Using Merged Feature Sets

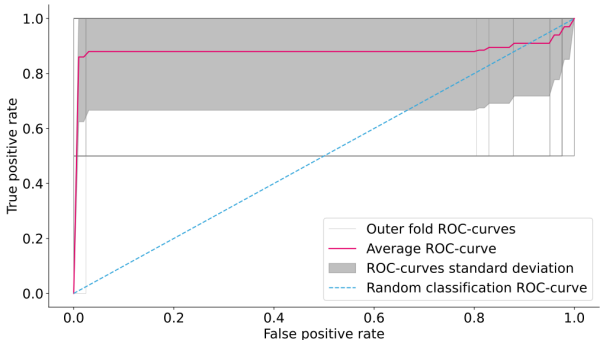
To attempt to train a better performing model using only Jar Device data, features from multiple datasets were used for training. Table 6 shows the model that outperformed all other models that were trained using features from multiple tasks in every performance measure. Figure 6 shows the confusion matrix, ROC-curve and feature importance of this model.

Tasks	Features	Classifier	F1-score	Recall	AUC ROC-curve
1c, 3	Specific features	SVM	**0.868 ± 0.230**	**0.840 ± 0.273**	**0.886 ± 0.202**

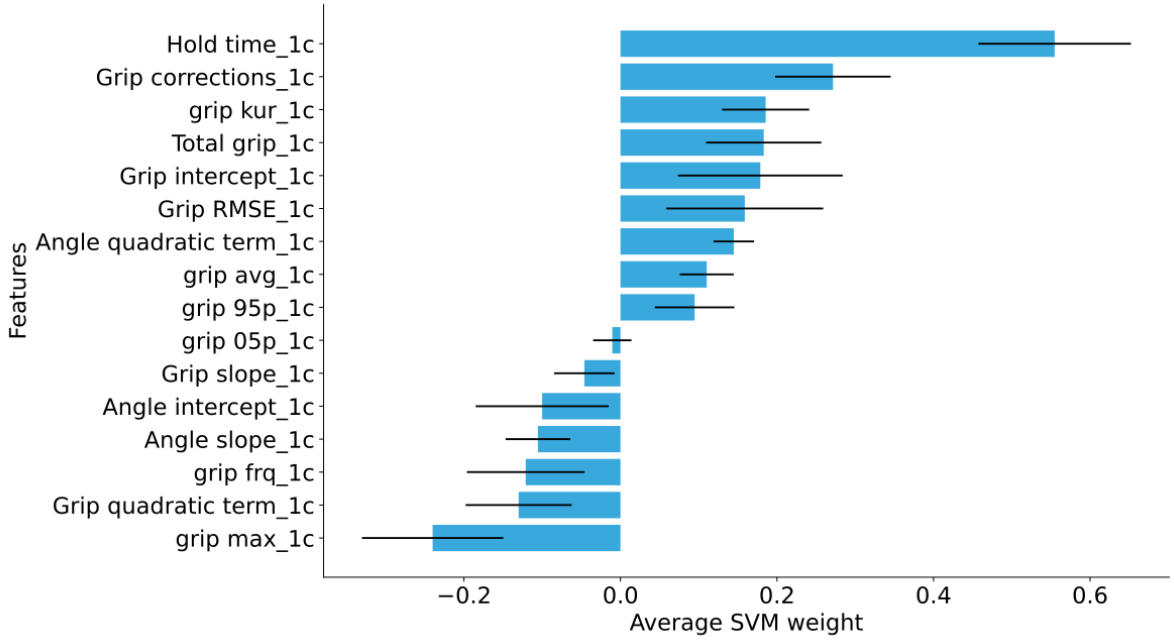
Table 6: The model, that makes use of features derived from Jar Device data of multiple tasks and performs best according to every performance metric



(a) Average confusion matrix.



(b) average ROC-curve.



(c) Average feature weights.

Figure 6: The average confusion matrix, average ROC-curve and average feature weights of the model that performed best over all the models that use features derived from Jar Device data of multiple multiple tasks. This model used specific features derived from tasks 1c and 3 and an SVM classifier.

6 Discussion

6.1 Dataset

The datasets that were collected for each of the 12 tasks were limited in size. Especially the amount of data points from patients with NMDs is relatively low, with datasets from tasks 1a and 5a even containing less than 10 complete (meaning both Jar Device data and sEMG data were recorded) data series from patients with NMDs. This was mainly caused by the fact that healthy test subjects were able to repeat all 12 tasks on 4 occasions on 1 day and patients with NMDs could only perform 1 set of tasks per day on 2 days. With a larger amount of data, especially from patients with NMDs, it is expected that the performance of the models would have improved, which, in turn, would have strengthened the potential of the identification of any biomarkers. With a large enough dataset, training more sophisticated machine learning models like the DNNs mentioned in the Related Reseach Section would also be possible.

When collecting data, it is important that the collection process is consistent for all participants to minimize unwanted noise in the data. When analyzing the datasets, some issues regarding this consistency arose. Healthy test subjects were restricted in their posture and hand placement during the first and third occasion. On the second and fourth occasion for the healthy test subjects and on both occasions for the patients with NMDs, posture and hand placement were not restricted, which may have led to inconsistencies in the sets.

When performing tasks that involve twisting the lid of the Jar Device, 3 different resistance levels were used. These levels were selected based on age, to compensate for the fact that younger people are generally stronger than older people. However, age is far from the only factor in a participant's strength and some participants may have performed these tasks on resistance levels that were lower or higher relative to their actual strength.

Lastly, because healthy test subjects performed 4 complete sets of tasks on 1 day, there might have been some fatigue that did not affect the patients with NMDs, who only performed 1 set of complete tasks per day, with at least a week in between. These issues have likely introduced unwanted experimental inconsistencies to the data. Without these inconsistencies, the models likely would have performed better and, like before, this would have strengthened the potential of any biomarkers that were identified.

Other possible inconsistencies in the datasets may have been introduced by apparent incorrect instructions or handling of the devices. For example, all datasets contain data points where data was recorded but the participant did not perform any task. Other examples of faulty data like this include data points where the participant starts the task before the recording has started, where the recording is ended before the participant finishes the task, or where the participant finishes a task and starts over, within the same recording. These data points were discarded, which largely negated these issues, although it did lead to a smaller amount of usable data. Issues like this could likely be avoided by automating larger parts of the recording process. For example, when a task started, the Jar Device recording and sEMG recording had to be started and stopped manually and separately. Starting and stopping these automatically and simultaneously would most likely have prevented some of these mistakes.

6.2 Methods

The results in Table 4 show that, generally, when using only sEMG data, using general features results in models that perform better. Of the specific features used by the model that uses general and specific features derived from sEMG data, only 3 are selected more than ten times and those features have relatively low feature importance. However, using specific features results in better performing models, when only using features from Jar Device data. A possible explanation for this is that Jar Device data is more intuitively interpretable than sEMG data, and this makes the creation of useful informed features easier. This, in turn, could mean that features derived from sEMG data could be improved with a deeper understanding of the data. Another possible explanation is that the low sampling frequency of the sEMG data resulted in data that contains a limited amount of information.

In the processing pipeline, we experimented with 3 different classifiers and optimized their hyperparameters in the inner cross-validation folds. In contrast, just 1 feature selection algorithm was experimented with, and its hyperparameters were not optimized. This could mean that some features that could potentially be used as biomarkers have been missed because the feature selection algorithm that was used was not optimized.

6.3 Results

Models trained using Jar Device & sEMG features
The best performing model over all the experiments is the model that uses general features derived from Jar Device and sEMG data from task 4. When looking at the average feature importances of this model, the feature with the highest feature importance is the average value of the 10th sEMG channel, which is collected from the flexor carpi ulnaris muscle in the subject's non-dominant hand. No single feature sticks out as a single deciding factor. However, of the features with an average importance of at least 0.05, all but one are characteristics of this same sEMG signal, indicating that this signal could contain potential biomarkers. Since participants were instructed to use their dominant hand (although patients with NMDs were allowed to switch hands, if they wanted to) and EMG channels 9 through 16 are derived from the subjects non-dominant arm, this sEMG channel is collected from the hand that is gripping the Jar Device. Although participants were not specifically instructed with regards to their grip for this task, this suggests that the way in which a subject grips the Jar Device, while keeping the lid at an angle, contains potential biomarkers. Other features are selected but have an average feature importance of (close to) 0. This indicates that the feature selection algorithm does not always select features that are useful for classification. In other words, features selected using the ANOVA F-statistic are not nececerally useful for classification.

Models trained using only Jar Device features
The ideal biomarker, however, would only use features derived from Jar Device data, since collecting data using only the Jar Device is more effective in terms of costs and effort. Models that only make use of features derived from Jar Device data did not perform as well as the other models, but were still able to correctly identify healthy test subjects in nearly all cases and patients with NMDs in a majority of cases. Since sEMG data is considered the state of the art when it comes to biomarkers for NMDs, this difference in performance between models using the two data sources is to be expected.

The model that performed best using only features derived from Jar Device data, used specific features derived from data of task 1c. When looking at the average feature importances, the most infuential feature is the time a participant is able to hold the Jar Device open. Holding the Jar Device open for as long as possible is the exact goal of the task, which indicates that this task was well designed. Other notable features with relatively high feature weights are the intercept, slope, quadratic term, and RMSE of the grip, which together indicate the shape of the grip curve, while the Jar Device is being held open. This further indicates that, even though no instructions were given with regards to gripping the Jar Device, the grip a participant needs to keep the lid twisted contains information that indicates the presence of an NMD.

When looking at the distributions of the most influential features in the model that was trained using only features dervide from Jar Device data, we see that patients with NMDs were generally able to hold the Jar Device open longer than healthy test subjects. This is unexpected because these are patients with self-reported weakness of the hands. An explanation could be that patients with NMDs are actually able to hold the Jar Device open for longer, but another possibility is a difference in the way in which the data was collected, such as a difference in encouragement from the person recording the data, or a choice in which hand to use, which healthy test subjects did not have. The number of patients in this dataset is too low to draw a definite conclusion. The other features that were most influential in this model all have to do with the grip on the Jar Device while it is being held open. We can see that patients generally correct their grip more times per second, have a higher grip intercept, and apply more total grip. The role which the slope with which the grip changes plays is less clear from this graph. The distributions of these 4 features show that patients with NMDs were more active with regards to their grip on the 4 to hold its lid open. This could be done unknowingly, or because the instructions that were given to the patients with NMDs were slightly different to those given to the healthy test subjects. Despite the reservations with regards to possible discrepancies within the data collection process, these are the most promising potential biomarkers derived from the Jar Device.

Models trained using merged feature sets
To attempt to improve performance while only using features derived from Jar Device data, features from multiple tasks were merged. The best performing model that utilizes data from multiple tasks, uses specific features derived from data from task 1c and 3, and an SVM classifier. However, only features derived from data from task 1c are selected in more than 10 outer folds. The model that uses the same features and classifier, but only data from task 1c, does not perform as well. This difference is likely caused by the exclusion of some datapoints in the merged dataset. Data points are only included in the merged dataset if the subject number and the occasion are present in both initial datasets. Such a difference in performance, caused by the exclusion of a small number of datapoints further indicates that a larger dataset is needed for more robust results.

The best performing model that utilized data from multiple tasks only uses features derived from data from task 1c, which seems to suggest that using features from multiple tasks does not necessarily improve results. Much like the model that only used data from task 1c, the hold time feature has the highest feature weight. Again, nearly all other influential features are derived from the subjects' grip on the Jar Device.

According to the characteristics of an ideal biomarker that were mentioned in the Related Research Section, biomarkers derived from Jar Device data should be prefered over those derived from sEMG data, given that they are equally informative, which remains the most important characteristic of any biomarker. The Jar device setup is cheaper, and less complicated than the sEMG system and, in contrast to the sEMG system, operating the Jar Device does not require a medical professional. Comparing the Jar Device biomarkers in this section to biomarkers for NMDs found in the literature that was mentioned in the Related Research section is difficult, because the Jar Device data from which potential biomarkers are derived has not yet been studied with regards to potential biomarkers.

6.4 Future Research

As was mentioned previously, a larger, more consistent dataset would most likely significantly improve the performance of the classifiers and make the results more robust. However, if the tasks were to be repeated to expand the dataset, it would suffice to only repeat those that seemed yielded the best performing models and the most promising biomarkers, like task 1b, 1c or 4. This would reduce the cost of repeating the same tasks somewhat.

Furthermore, the potential biomarkers that have been identified in this report would still need to be validated before they can be used in clinical research. This process of validation would need to ensure tolerability, difference between patients and controls, repeatability, detection of clinical events, and correlation with traditional biomarkers, as is described by Kruizinga et al. [8]. When biomarkers are technically validated, further validation in a clinical setting would be needed.

The methods that are described in this report have identified biomarkers for NMDs using the sEMG and Jar Device data. However, if a dataset containing time series data of patients with some disease and healthy test subjects is suspected to contain biomarkers for that disease, the same methods of extracting features, training and evaluating classification models, and analyzing the most influential features can be applied. When using only general features, the methods can even be copied exactly to identify potential biomarkers. Specific features would need to be handcrafted based on knowledge of the dataset and scientific knowledge of the disease at hand.

7 Conclusion

We have constructed a method of identifying potential biomarkers for NMDs using datasets recorded using the Jar Device and sEMG data. This data was recorded from a group of patients with NMDs and healthy test subjects. By building a classification model and analyzing which features played a role in the classification process, a number of potential biomarkers were identified. As was expected, classification using sEMG data was successful and some possible sources of biomarkers based on sEMG data were identified. Models that were trained using features derived from both Jar Device and sEMG data, were shown to achieve the highest performance metrics and the average feature importance of this model indicates that the sEMG signals from the arm that is gripping the Jar Device are most influential in the classification process and contain potential biomarkers.The best performing model that only used features derived from Jar Device data, used data of task 1c. The time a subject was able to hold the Jar Device open during this task and characteristics of the grip on the Jar Device during the task were identified as potential biomarkers that could be measured using only the Jar Device. These potential biomarkers are promising, but would need additional validation before being used in a clinical setting.

References

[1] Andrea Barp, Amanda Ferrero, Silvia Casagrande, Roberta Morini, and Riccardo Zuccarino. Circulating biomarkers in neuromuscular disorders: What is known, what is new. *Biomolecules*, 11(8):1246, 2021.

[2] Richard Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966.

[3] Edward A Clancy, Evelyn L Morin, and Roberto Merletti. Sampling, noise-reduction and amplitude estimation issues in surface electromyography. *Journal of electromyography and kinesiology*, 12(1):1–16, 2002.

[4] Edith H Cup, Allan J Pieterse, Jessica M ten Broek-Pastoor, Marten Munneke, Baziel G van Engelen, Henk T Hendricks, Gert J van der Wilt, and Rob A Oostendorp. Exercise therapy and other types of physical therapy for patients with neuromuscular diseases: a systematic review. *Archives of physical medicine and rehabilitation*, 88(11):1452–1464, 2007.

[5] Fatemeh N Emamzadeh and Andrei Surguchov. Parkinson's disease: biomarkers, treatment, and risk factors. *Frontiers in neuroscience*, 12:612, 2018.

[6] Matthew W Flood, Bente Rona Jensen, Anne-Sofie Malling, and Madeleine M Lowery. Increased emg intermuscular coherence and reduced signal complexity in parkinson's disease. *Clinical neurophysiology*, 130(2):259–269, 2019.

[7] Biomarkers Definitions Working Group, Arthur J Atkinson Jr, Wayne A Colburn, Victor G DeGruttola, David L DeMets, Gregory J Downing, Daniel F Hoth, John A Oates, Carl C Peck, Robert T Schooley, et al. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical pharmacology & therapeutics*, 69(3):89–95, 2001.

[8] M. D. Kruizinga, F. E. Stuurman, V. Exadaktylos, R. J. Doll, D. T. Stephenson, G. J. Groeneveld, G. J. A. Driessen, and A. F. Cohen. Development of novel, value-based, digital endpoints for clinical trials: A structured approach toward fit-for-purpose validation. *Pharmacological Reviews*, 72(4):899–909, 2020.

[9] Patrick Kugler, Christian Jaremenko, Johannes Schlachetzki, Juergen Winkler, Jochen Klucken, and Bjoern Eskofier. Automatic recognition of parkinson's disease using surface electromyography during standardized gait tests. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5781–5784. IEEE, 2013.

[10] Roberto Merletti and Dario Farina. *Surface electromyography: physiology, engineering, and applications*. John Wiley & Sons, 2016.

[11] Andreas C Müller and Sarah Guido. *Introduction to machine learning with Python: a guide for data scientists*. " O'Reilly Media, Inc.", 2016.

[12] Chintan Parmar, Patrick Grossmann, Johan Bussink, Philippe Lambin, and Hugo JWL Aerts. Machine learning methods for quantitative radiomic biomarkers. *Scientific reports*, 5(1):13087, 2015.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[14] Themistocles S Protopsaltis, Anthony J Boniello, and Frank J Schwab. Management of spinal deformity in adult patients with neuromuscular disease. *Journal of the American Academy of Orthopaedic Surgeons*, 24(9):634–644, 2016.

[15] Evgeny Putin, Polina Mamoshina, Alexander Aliper, Mikhail Korzinkin, Alexey Moskalev, Alexey Kolosov, Alexander Ostrovskiy, Charles Cantor, Jan Vijg, and Alex Zhavoronkov. Deep biomarkers of human aging: application of deep neural networks to biomarker development. *Aging (Albany NY)*, 8(5):1021, 2016.

[16] Khosro Rezaee, Somayeh Savarkar, Xiaofeng Yu, and Jingyu Zhang. A hybrid deep transfer learning-based approach for parkinson's disease classification in surface electromyography signals. *Biomedical Signal Processing and Control*, 71:103161, 2022.

[17] Chiara Scotton, Chiara Passarelli, Marcella Neri, and Alessandra Ferlini. Biomarkers in rare neuromuscular diseases. *Experimental Cell Research*, 325(1):44–49, 2014.

[18] Ashraf Abou Tabl, Abedalrhman Alkhateeb, Waguih ElMaraghy, Luis Rueda, and Alioune Ngom. A machine learning approach for identifying gene biomarkers guiding the treatment of breast cancer. *Frontiers in Genetics*, 10:256, 2019.

[19] Xing Wan. Influence of feature scaling on convergence of gradient iterative algorithm. In *Journal of physics: Conference series*, volume 1213, page 032021. IOP Publishing, 2019.

[20] Bo-Young Youn, Youme Ko, Seunghwan Moon, Jinhee Lee, Seung-Gyu Ko, and Jee-Young Kim. Digital biomarkers for neuromuscular disorders: a systematic scoping review. *Diagnostics*, 11(7):1275, 2021.

A Feature descriptions

This appendix contains a description of the specific features that were derived for each separate task, as well as a list of the general features that were used for all tasks.

A.1 Specific features

A.1.1 Task 1

Total Grip The total grip is defined as the area under the curve of the grip, over the period where the Jar Device is considered open. It is meant to capture the total amount of grip the subject needs to keep the Jar Device open. Grip Corrections Grip corrections is defined as the number of times the Jar Device goes from increasing to decreasing or from decreasing to increasing while the Jar Device is considered open, per second. It is meant to capture the number of times the subject has to re-adjust their grip to keep the Jar Device open.

Grip Polynomial fit parameters The grip polynomial fit parameters are defined as the 3 coefficients of a 2nd order polynomial fit of the grip, while the Jar Device is considered opened, as well as the RMSE of this fit. They are meant to capture the change in the grip on the Jar Device while the subject is keeping the Jar Device open.

Hold Length The hold length is defined as the amount of time the subject is able to hold the Jar Device open. The Jar Device is considered when the lid is twisted to 20 degrees in either direction and is considered closed when it reaches 15 degrees. They are meant to capture the change in angle while the subject is keeping the Jar Device open.

Angle Corrections Angle corrections is defined as the number of times the angle of the lid of the Jar Device goes from increasing to decreasing or from decreasing to increasing. It is meant to capture the number of corrections the subject needs to keep the lid of the Jar Device as close to 20 degrees as possible.

Angle Polynomial Fit Parameters The angle polynomial fit parameters are defined as the 3 coefficients of a 2nd order polynomial fit of the grip, while the Jar Device is considered opened, as well as the RMSE of this fit. It is meant to capture the change in the angle while the subject is keeping the Jar Device open.

Time to Maximum Angle The time to maximum angle is defined as the time between the start of the measurement and the moment the maximum angle over the entire measurement is reached. It is meant to capture how quickly the subject is able to twist the lid of the Jar Device.

Frequencies and Frequency Dominances The frequency and frequency dominance are defined as the frequency with the highest corresponding power spectral density in the frequency spectrum of an sEMG signal and the area under the curve of the sEMG signal within 0.25 Hz of its maximum divided by the area under the curve of the whole frequency spectrum, respectively. These features were calculated for all 16 sEMG channels. These features are meant to capture the most dominant frequency and its dominance in each of the sEMG signals.

AUC Difference Between Arms This AUC difference between arms is defined as the difference in average area under the curve of the sEMG signals of each of the 8 channels between the left and right arm. The sEMG signals are normalized using the maximum of the MVC measurements. These features are meant to capture the difference in sEMG activity between the subjects left and right arm.

Average Rectified Value Slopes and Intercepts The Average rectified value slopes and intercepts are defined as the slope and intercept fitted to the average rectified value (ARV) as described by Clancy et al. [3][10]. The ARV is calculated using windows of 1 second with 50% overlap between windows, resulting in a temporal resolution of 0.5 seconds. These features are meant to capture change in muscle fatigue, as muscle fatigue is associated with an increase in sEMG amplitude. The ARV slope and intercept were calculated for 16 sEMG channels.

Instantaneous Median Frequency Slopes and Intercepts The instantaneous median frequency slopes and intercepts are defined as the slope and intercept fitted to the instantaneous median frequency (IMDF). The IMDF is the frequency at which the area under the curve of the frequency spectrum is split in half. The IMDF is calculated using windows of 1 second with 50% overlap between windows, resulting in a temporal resolution of 0.5 seconds. These features are meant to capture the change in muscle fatigue during the measurement, as muscle fatigue is associated with a decrease in sEMG frequency.

Permutation Entropy Slopes and Intercepts The permutation entropy slopes and intercepts are defined as the slope and intercept fitted to the permutation entropy (PE). The PE is calculated by determining patterns in the signal using a moving window and calculating their relative probabilities within the entire signal. The PE is then defined as the entropy of these patterns. The PE is calculated using windows of 1 second with 50% overlap between windows, resulting in a temporal resolution of 0.5 seconds. These features are meant to capture the change in complexity of the sEMG signal.

A.1.2 Task 2

Grip Frequency and Frequency Dominance The grip frequency and frequency dominance are defined as the frequency with the highest corresponding power spectral density in the frequency spectrum of the grip and the area under the curve of the sEMG signal within 0.25 Hz of its maximum divided by the area under the curve of the whole frequency spectrum, respectively. They are meant to capture whether the subject grips the Jar Device with the same frequency as they twist its lid.

Angle-Grip Correlation The angle-grip correlation is defined as the correlation between the angle and the grip between the first and last opening. It is meant to capture whether the subject opens the Jar Device and applies pressure to it simultaneously.

Angle Peak Polynomial Fit Parameters The angle peak polynomial fit parameters are defined as the slope and intercept of the 2nd order polynomial fitted through the peaks (open angles) of the angle of the lid of the Jar Device, as well as the RMSE of these peaks

around the fit. They are meant to capture how the angle to which the subject opens the Jar Device changes.

Angle Minimum Polynomial Fit Parameters The angle minimum polynomial fit parameters are defined as the slope and intercept of the 2nd order polynomial fitted through the minima (closed angles) of the angle of the lid of the Jar Device, as well as the RMSE of these peaks around the fit. They are meant to capture how the angle to which the subject closes the Jar Device changes. This angle should always be 0, but subjects often did not fully close the Jar Device during this measurement.

Angle Frequency and Frequency Dominance The angle frequency and frequency dominance are defined as the frequency with the highest corresponding power spectral density in the frequency spectrum of the angle and the area under the curve of the sEMG signal within 0.25 Hz of its maximum divided by the area under the curve of the whole frequency spectrum, respectively. They are meant to capture how closely the subject is able to follow the metronome.

Average Metronome Dominance The average metronome dominance is defined as the average (over all sEMG channels) area under the curve of the frequency spectrum within 0.25 Hz of the frequency of the metronome, divided by the area under the curve of the entire frequency spectrum. It is meant to capture the presence of this frequency in the sEMG data.

AUC Difference Between Arms This feature has the same definition as the feature with the same name from task 1.

Frequencies and Frequency Dominances These features have the same definition as the features with the same names from task 1.

A.1.3 Task 3

Grip Peak Polynomial Fit Parameters The grip peak polynomial fit parameters are defined as the slope and intercept of the 2nd order polynomial fitted through the peaks (gripping moments) of the subjects' grip on the Jar Device, as well as the RMSE of these peaks around the fit. They are meant to capture how the grip the subject applies on the Jar Device changes during the measurement.

Grip Minimum Polynomial Fit Parameters The grip minimum polynomial fit parameters are defined as the slope and intercept of the 2nd order polynomial fitted through the minima (releasing moments) of the subjects grip on the Jar Device, as well as the RMSE of these peaks around the fit. They are meant to capture how the grip the subject applies to the Jar Device when they should release it changes. This grip should always be 0, but subjects often did not fully release the Jar Device during this measurement.

Grip Frequency and Frequency Dominance The grip frequency and frequency dominance are defined as the frequency with the highest corresponding power spectral density in the frequency spectrum of the grip and the area under the curve of the sEMG signal within 0.25 Hz of its maximum divided by the area under the curve of the whole frequency spectrum, respectively. They are meant to capture how closely the subject is able to follow the metronome.

Jamar-Jar Device Grip Difference The Jamar-Jar Device grip difference is defined as the difference between the maximum grip achieved during the maximum voluntary contraction measurement using the Jamar dynamometer and the maximum grip achieved during the measurement. It is meant to capture the difference between a subjects maximum grip on the 2 devices.

Average Metronome Dominance This feature has the same definition as the feature with the same name from task 2.

AUC Difference Between Arms This feature has the same definition as the feature with the same name from task 1.

Frequencies and Frequency Dominances These features have the same definition as the features with the same names from task 1.

A.1.4 Task 4

Grip Frequency and Frequency Dominance These features have the same definition as the features with the same names from task 3.

Grip Peak Polynomial fit Parameters These features have the same definition as the features with the same names from task 3.

Jamar-Jar Device Grip Difference This feature has the same definition as the feature with the same name from task 3.

AUC Difference Between Arms This feature has the same definition as the feature with the same name from task 1

Frequencies and Frequency Dominances These features have the same definition as the features with the same names from task 1

A.1.5 Task 5

Grip Frequency and Frequency Dominance These features have the same definition as the features with the same names from task 3.

Angle-Grip Correlation This feature has the same definition as the feature with the same name from task 2.

Total Grip The total grip is defined as the area under the curve of the grip, over the period in between the first and last gripping moment. It is meant to capture the total amount of grip the subject needs to open the Jar Device.

Number of Openings The number of openings is the number of times the subject is able to completely open the Jar Device during the measurement. Here, the Jar Device is considered completely opened when its lid is twisted more than 40 degrees in either direction. This feature is meant to capture how well the subject is able to perform the task, exactly as it is presented.

Angle Peak Polynomial Fit Parameters These features have the same definition as the features with the same names from task 2.

Angle Minimum Polynomial Fit Parameters These features have the same definition as the features with the same names from task 2.

Angle Frequency and Frequency Dominance These features have the same definition as the features with the same names from task 2.

Period Polynomial Fit Parameters The period polynomial fit parameters are defined as the slope and intercept of the 2nd order polynomial fitted through to the periods of time between consecutive openings, as well as the RMSE of these periods around the fit. These features are meant to capture the change in speed with which the subject is able to open and close the Jar Device.

AUC Difference Between Arms This feature has the same definition as the feature with the same name from task 1.

Frequencies and Frequency Dominances These features have the same definition as the features with the same names from task 1.

A.1.6 Task 6

Grip Hold Length The grip hold length is defined as the amount of time the subject is able to grip the Jar Device with sufficient force. The time is measured from the moment where the grip on the Jar Device is at its maximum value until it reaches 75% of this maximum value. This feature is meant to capture the time the subject is able to maintain maximum grip.

Time to Maximum Grip The time to maximum grip is defined as the time between the start of the measurement and the moment the maximum grip over the entire measurement is reached. It is meant to capture how quickly the subject is able to increase their grip on the Jar Device.

Grip Polynomial Fit Parameters The grip polynomial fit parameters are defined as the 3 coefficients of a 2nd order polynomial fit of the grip, from the time the subject reaches maximum grip, until it reaches 75% of this maximum value. They are meant to capture the change in the grip while the subject is trying to maintain maximum grip.

AUC Difference Between Arms This feature has the same definition as the feature with the same name from task 1.

Frequencies and Frequency Dominances These features have the same definition as the features with the same names from task 1.

Average rectified value slopes and intercepts These features have the same definition as the features with the same names from task 1.

Instantaneous median frequency slopes and intercepts These features have the same definition as the features with the same names from task 1.

permutation entropy slopes and intercepts These features have the same definition as the features with the same names from task 1.

A.2 General features

Minimum This feature is defined as the minimum value of the time series.

Maximum This feature is defined as the maximum value of the time series.

Fifth Percentile This feature is defined as the fifth percentile value of the time series.

Ninety-Fifth Percentile This feature is defined as the ninety-fifth percentile value of the time series.

Variance This feature is defined as the variance of the time series.

Average This feature is defined as the average value of the time series.

Standard Deviation This feature is defined as the Standard Deviation of the time series.

Kurtosis This feature is defined as the kurtosis of the time series.

Dominant Frequency This feature is defined as the frequency with the highest corresponding power spectral density in the frequency spectrum of the time series.

Dominant Frequency Power This feature is defined as the power spectral density of the dominant frequency of the time series.

Average Frequency This feature is defined as the sum of all frequencies in the frequency spectrum of the time series multiplied with their corresponding power spectral density, divided by the sum of all power spectral densities.

Instantaneous Median Frequency This feature is defined as the frequency at which the area under the curve of the frequency spectrum of the time series is split in half.