Machine-learning pipelines for classification of pathological tremor patients

A proof-of-concept

by

Alvaro Assis de Souza

to obtain the degree of Master of Science in Biomedical Engineering
at the Delft University of Technology,
to be defended publicly on Tuesday September 21, 2021 at 11:00 AM.

Student number:	5001005	
Project duration:	February 1, 2021 – September 21, 2021	
Thesis committee:	Dr. ir. W. Mugge,	TU Delft, supervisor
	Dr. M. Kok,	TU Delft, co-supervisor
	Prof. Dr. F.C.T. van der Helm,	TU Delft
	Dr. O. Scharenborg,	TU Delft

This thesis is confidential and cannot be made public until September 21, 2023.

TU Delft

MSc Thesis
MSc Biomedical Engineering

Machine-learning pipelines for classification of pathological tremor patients: a proof-of-concept

Alvaro Assis de Souza[1]*

Abstract

Background: Parkinson's Disease (PD), Essential tremor (ET), and dystonia are movement disorders often misdiagnosed as one another and commonly present tremor as one of their motor symptoms. Although similar tremor behaviors between the mentioned disorders lead to substantial misdiagnosis rates and, consequently, subpar care, tremorous signal acquired via wearable sensors can be used to discriminate between PD, ET, and dystonia patients. This study aims to develop three proofs-of-concept, accelerometer-based diagnostic assistance algorithms.

Methods: Hand and arm accelerometer data of eleven dystonia, ten ET, and seven PD patients, measured during the performance of standard clinical tremor evaluation motor tasks, is used to model three binary tremor classification pipelines. Principal Component Analysis reduces the dimensionality of the data of each sensor. A power spectral density-based tremor detection method (developed in a previously published study) identifies tremor and non-tremor windows from the time-series data. The windows are used as data sources for feature extraction. A feature matrix consisting of time and frequency-domain predictors is supplied to model-building pipelines to predict the probability of the patients belonging to each possible class. A nested cross-validation scheme selects and evaluates their performance. Logistic regression, Balanced Random Forest, and a voting ensemble of the two are used as classifiers. The pipeline that yields the lowest Brier score for each classification scenario is selected to develop the final pipeline.

Results: The Balanced Random Forest pipeline outperformed the other pipelines in all classification scenarios (PD x no-PD, ET x no-ET, and DT x no-DT). For the PD x no-PD case, a Brier score of 0.188 (0.128) and ROC-AUC of 0.84 were obtained. For an optimal decision threshold of 42% to classify a patient as PD, the pipeline achieved an accuracy of 75%, specificity of 67%, and sensitivity of 100%. A label permutation test (n=1000) was performed to assess the final pipeline's score significance for each classification case. Only the PD x no-PD presented a p-value < 0.05. The features extracted from rest, postural, and kinetic motor tasks had the largest influence on the classifier's predictions.

Conclusions: Based on hand and arm accelerometer measurements, PD patients are more easily differentiated from other pathological tremor patients than ET and DT patients. The promising results achieved by the proof-of-principle pipeline encourages further development of assistive diagnostic technologies in clinical practice.

Keywords
accelerometers — tremor classification — machine-learning — Parkinson's disease — Essential tremor — dystonic tremor — diagnostic assistance — proof-of-concept

[1] *Department of Biomechanical Engineering, Delft University of Technology, Delft, The Netherlands*
***Corresponding author**: a.assisdesouza@student.tudelft.nl

Contents

1. Introduction

Misdiagnosis of the three most common movement disorders, Essential tremor (ET), Parkinson's disease (PD), and dystonia, is a recurrent problem that leads to sub-optimal treatment and incorrect prognosis of millions of patients worldwide [1], [2]. Moreover, misclassification of the correct disorder may negatively affect the inclusion of the proper patients in clinical trials. Tremor, an involuntary and oscillatory movement, is a common symptom of these disorders, and its similar clinical presentation among patients often is a misleading factor for medical doctors in charge of their diagnoses [1]. However, tremor can also be used as a source of information that helps to discriminate between ET, PD, and dystonia.

There are no gold standard diagnostic tests for ET and dystonia, and current diagnostic procedures are based on clinical criteria and on the patients' medical records [3], [4]. Rates of misdiagnosis between 30 and 50% of ET patients have been reported, where dystonia and PD are the most common missed diagnoses [2], [5]. Additionally, up to 50% of dystonia cases are misdiagnosed/under-diagnosed at their first encounter [3]. Misdiagnosis rates up to 34% are reported for PD [6]. Furthermore, PD patients presenting tremors are more likely to be misclassified, especially if the diagnosis is made by a non-specialist neurologist [7]. Neuroimaging techniques are also available as Parkinson's disease diagnostic tools. Nevertheless, their adoption in clinical practice is debatable because, as in the case of DaTSCAN imaging, the diagnostic accuracy may be similar to the accuracy of a clinical diagnosis [8], with the drawbacks of being more invasive procedures and requiring high implementation costs [8], [9].

During clinical evaluation, it is common practice for doctors to assess tremor visually by asking ET, PD, and dystonic tremor (DT) patients to perform several standard motor tasks, as tremors are expected under specific circumstances, such as during rest, postural, and kinetic motor tasks, depending on the patient's diagnosis [10], [11]. The subjective nature of the current tremor assessment procedure adds uncertainty to the clinical evaluation of tremor [12]. An even more challenging diagnostic situation concerns the evaluation of the disorders when non-motor symptoms are absent, as is the case for early-stage Parkinson's disease [13].

Studies focusing on differentiating movement disorders from analysis of electromyography (EMG) and motion sensors data (e.g., accelerometers and gyroscopes) acquired during the performance of tremor evaluation motor tasks have been reported, presenting promising results [14]–[17]. These diagnostic approaches rely on the differentiation between the disorders based on several tremor features (e.g., amplitude and dominant frequency) extracted from the sensor measurements.

Nonetheless, key points are missing in current research: first, there is no validated demonstration of how tremor is confirmed in the sensors' recordings. As shown in [18], motion parameters calculated based on tremorless data differ significantly from those derived from tremorous data. Second, most of the previous works cover only the differentiation between essential and parkinsonian tremor patients, excluding people who suffer from DT, which consists of a representative portion of the misdiagnosed cases of ET and PD. Additionally, there is a need for studies involving probabilistic machine-learning for healthcare [19]. Probabilistic outputs from diagnostic models enable assessing the uncertainty associated with the predictions, which is particularly beneficial to assist medical doctors with decision-making in diagnostic scenarios.

To tackle the issue of tremor detection, Luft et al. [18], in 2019, proposed a power spectral density-based method to detect tremor windows (TW) and non-tremor windows (NTW) from accelerometer and EMG data recorded during the performance of clinical tremor evaluation tasks. A tremor window detection accuracy of 90% was achieved when only the acceleration information was used. The reporting of probabilistic outputs from diagnostic models was addressed by Ghassemi et al. [14] for the differentiation between ET and PD patients. However, the authors classification accuracy as the evaluation metric, which does not measure the quality of probabilistic predictions.

The present work aims to develop and evaluate three proofs-of-concept machine-learning pipelines able to differentiate between the three possible binary classification cases of PD, ET, and DT patients from accelerometer measurements when performing standard clinical evaluation motor tasks (rest, postural, kinetic, distraction, and entrainment tasks). The tremor detection technique developed by Luft et al. [18] will split the accelerometer data into TW and NTW for further feature extraction. The models will provide as output the probability of the patients belonging to each possible class. The assessment of their performance will be based on the Brier score; a *strictly proper* scoring rule. A scoring rule is considered *proper* if it is minimized as the probabilistic predictions of a classifier, instead of its binary outputs, approach the true probability outcome of the event being predicted. It is considered *strictly proper* if the minimum is unique [20]. Additionally, the area under the Receiver Operating Characteristic (ROC) curve, along with the ROC plot, will be assessed for supplementary information about the models' quality. The ROC curve, recurrently used to assess the performance of clinical tests, demonstrates the classifiers' ability to differentiate between the classes for different values of false positive (FP) and false-negative (FN) rates [21].

Regarding model selection and evaluation, a nested cross-validation (CV) scheme will be performed. The inner-loop cross-validation will host the search for the best set of features, classifier hyperparameters, and data transformations, while the outer-loop procedures will assess the model generalization performance. This scheme assures that the reported generalization performance of a model is not assessed against the same test set used for selecting the best model, i.e., classifier with final sets of hyperparameters, data transformations, and features. It is important to state that if some procedures on data, such as feature selection, a common step in classification modeling pipelines, are done before the split of the dataset into training and test sets, the estimated performance metrics of the classifier are likely to be overoptimistic due to data leakage, i.e., the unintended use of information of data from the test set to during model training. This has been reported as one of the most common issues in developing and reporting machine learning models in biomedical research [22]. This work will use *Imblearn* pipelines [23] to streamline the different procedures on the data and the classifiers to be tested. *ScikitLearn* library [24] provided most of the remaining machine-learning tools (see appendix A).

In the end, three final pipelines, one for each classification case, will be defined following the respective model-building strategy considered as best for each case. The statistical significance of their performance is checked through a permutation test.

Medical doctors could benefit from extra assistance when diagnosing tremor patients, primarily for unclear cases, and when the professionals work in a non-ideal resource setting and are non-neurologists or non-movement disorder experts. In addition, improvements in the correct inclusion of patients in clinical trials can be envisioned with such technology. Therefore, the main research questions to be answered at the end of this study are:

1. How reliably can essential, parkinsonian, and dystonic tremor patients be differentiated from each other based on accelerometer measurements?

2. What motor tasks are the most discriminative for the included tremor patients?

Regarding the first question, we expect our methodology to yield superior results for the PD x no-PD classification scenario, following what is most common in clinical practice. As discussed in [5], most misdiagnosed ET patients have dystonia assigned as their movement disorder. Therefore, differentiating between ET and DT should be more challenging than differentiating between PD and ET/DT. In what concerns the second research question, we foresee that tremor information retrieved during distraction motor tasks, in conjunction with rest, postural, and kinetic tasks, will assist in classifying the types of tremor. As discussed in [25], attention level affects tremor in PD patients, in contrast to what has been found for ET patients [26]. Entrainment tasks, however, are usually helpful for the differentiation between Psychogenic tremor and PD, ET, and DT [27]. Thus, these tasks are not expected to provide highly discriminative tremor information.

2. Materials and Methods

2.1 TIM-Tremor dataset

The classification experiments are conducted using the Technology in Motion Tremor (TIM-Tremor) dataset [28]. It aims to develop and implement motor function evaluation technologies in clinical practice, both in diagnosis and treatment.

The dataset contains videos and accelerometer measurements of tremor patients recruited from the LUMC Department of Clinical Neurophysiology outpatient clinic, performing up to 21 standard tremor clinical evaluation motor tasks, along with comma-separated values file with tremor severity scores and diagnosis for each study participant. The study experimenter assessed tremor intensity, and it followed the Bain and Findley clinical rating scale (Bain et al., 1993), ranging from 0 to 10 (0: no tremor, 1-3: mild tremor, 4-6: moderate tremor, 7-9: severe tremor, 10: very severe tremor). Two scores were given for each patient, one for each arm. No task-specific tremor scores are present in the dataset.

A neurologist, specialist in movement disorders, gave the diagnostic labels for each patient, taking into account the available medical information (e.g., history, anamnesis, neurological examination, MRI or DAT scans) and the tremor severity results provided by the study experimenter. The diagnostic labels are no (convincing), parkinsonian, essential, dystonic, and functional tremor. The neurologists provided additional comments about the diagnosis of patients with the label *other*. Some examples include patients diagnosed with myoclonus.

Patient inclusion

Patients with neurological disorders were recruited between May 2016 and October 2017. The authors included patients aged 18 years or older, with command of the Dutch language, who (according to their medical records) have had a hand or arm tremor assessment appointment. Patients unable to perform the motor tasks, either by physical or cognitive and communicative limitations, were excluded. Of all initially recruited 90 patients, 61 agreed to participate in the study, with a final amount of 55 included in the dataset. All patients gave prior informed consent according to the Declaration of Helsinki. The LUMC committee approved the study's protocol.

The patient inclusion flowchart is shown in Figure 1. From the 47 patients with sufficient data for diagnosis (according to the study neurologists), those with essential, parkinsonian, or dystonic tremors (28) were selected to develop the present work further.

Data acquisition

During several standard clinical tremor evaluation tasks, participants remained seated on a chair upright with their feet supported on the ground. Two tri-axial ACL300 accelerometers (Biometrics Ltd, Newport, UK) were taped to the forearm and to the back of the patient's hand most affected by tremors (approximately 6 cm proximal and distal to the wrist joint, respectively). The positioning of the sensors followed: z-axis

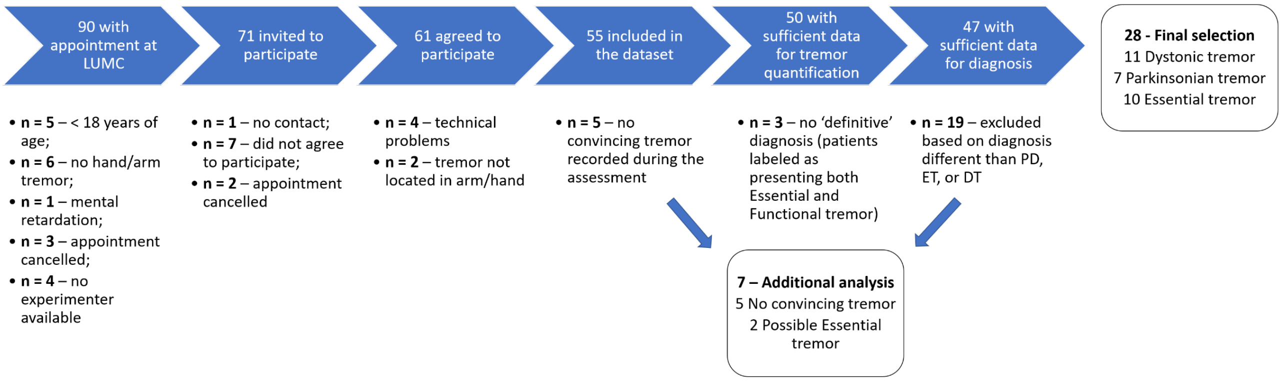
90 with appointment at LUMC
71 invited to participate
61 agreed to participate
55 included in the dataset
50 with sufficient data for tremor quantification
47 with sufficient data for diagnosis
28 - Final selection
11 Dystonic tremor
7 Parkinsonian tremor
10 Essential tremor
• n = 5 – < 18 years of age;
• n = 6 – no hand/arm tremor;
• n = 1 – mental retardation;
• n = 3 – appointment cancelled;
• n = 4 – no experimenter available
• n = 1 – no contact;
• n = 7 – did not agree to participate;
• n = 2 – appointment cancelled
• n = 4 – technical problems
• n = 2 – tremor not located in arm/hand
• n = 5 – no convincing tremor recorded during the assessment
• n = 3 – no 'definitive' diagnosis (patients labeled as presenting both Essential and Functional tremor)
• n = 19 – excluded based on diagnosis different than PD, ET, or DT
7 – Additional analysis
5 No convincing tremor
2 Possible Essential tremor

Figure 1. Patient inclusion flowchart and diagnostic distribution. The final selection set corresponds to the patients' data used for the development of the pipelines. The additional analysis set will be used to evaluate patients with inconclusive diagnoses.

normal to the skin surface, and the y-axis pointing towards the fingers. The devices have a range of +- 10g, each one with a mass of 10 grams, sensitivity of 100mV/g, and they measured acceleration at a 1000Hz sampling rate.

The sensors were positioned on the patients at the beginning of the recording and not removed until all clinical motor tasks were realized. The recordings took place in one go per patient without any breaks. All data was saved in the same folder for all the tasks performed. The data was then manually segmented into the different tasks' measurements. The recording duration of some motor tasks was not consistent among all patients; however, most of them took approximately 30s.

Motor tasks

Depending on the diagnostic, some patients may not display tremors for specific motor tasks and display significant tremors for others: e.g., a patient could not show tremors at rest but has a postural tremor when sustaining his/her arms in front of the chest. Therefore, the patient's performance of various motor tasks is required for the standard clinical evaluation of tremors.

The TIM-Tremor dataset contains measurements from rest, posture, action, distraction, and entrainment tasks, as they cover distinct situations on which different pathological tremors are elucidated, thus being good candidates to provide meaningful tremor information. A detailed description of the performed tasks is presented in appendix B. It is important to state that not every patient performed all motor tasks. Figure 2 presents the distribution of the available accelerometer recordings among the tasks.

2.2 Methodology

A three-blocks methodology was adopted in the present work: data preparation, model-building pipeline, and model selection and evaluation. It follows the simplified schematic diagram depicted in Figure 3. The Python programming language (Python Software Foundation, https://www.python.org/), along with standard Python libraries (described in appendix A), was used for the entirety of the methodology.

In the present work, three classification scenarios are assessed in a one-vs-all approach: PD x no-PD, ET x no-ET, and DT x no-DT. Table 1 depicts the classes distribution in all three cases.

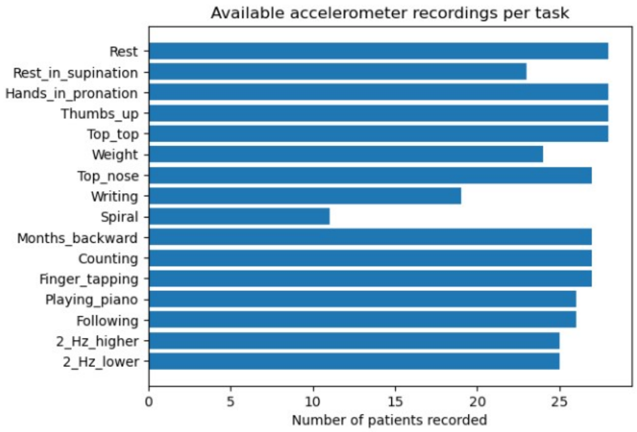

Figure 2. Available data in TIM-Tremor dataset. Varying amounts of patients recorded each motor task.

Table 1. Three binary classification scenarios and classes distribution.

Scenario	Class distribution
PD x no-PD	7 PD — 21 no-PD
ET x no-ET	10 ET — 18 no-ET
DT x no-DT	11 PD — 17 no-DT

The procedures adopted in each one of the blocks of the diagram shown in detail in Figures 4, 5, and 6, are now investigated.

Data preparation

We refer to data preparation as the processing and extraction of motion features from the two accelerometers placed on the participants that could discriminate between parkinsonian, essential, and dystonic tremor patients. Raw accelerometer recordings from the 28 patients performing the selected set of

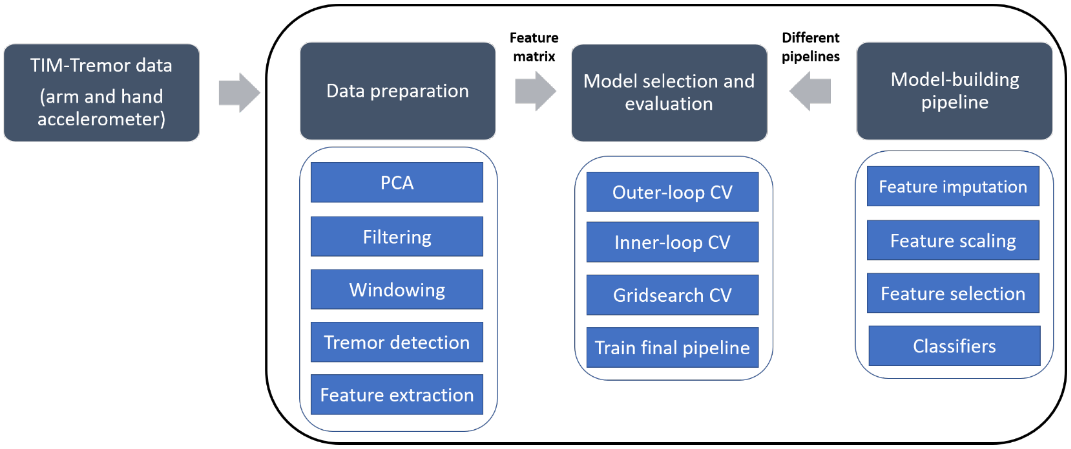

Figure 3. Simplified diagram of the presented three-blocks methodology.

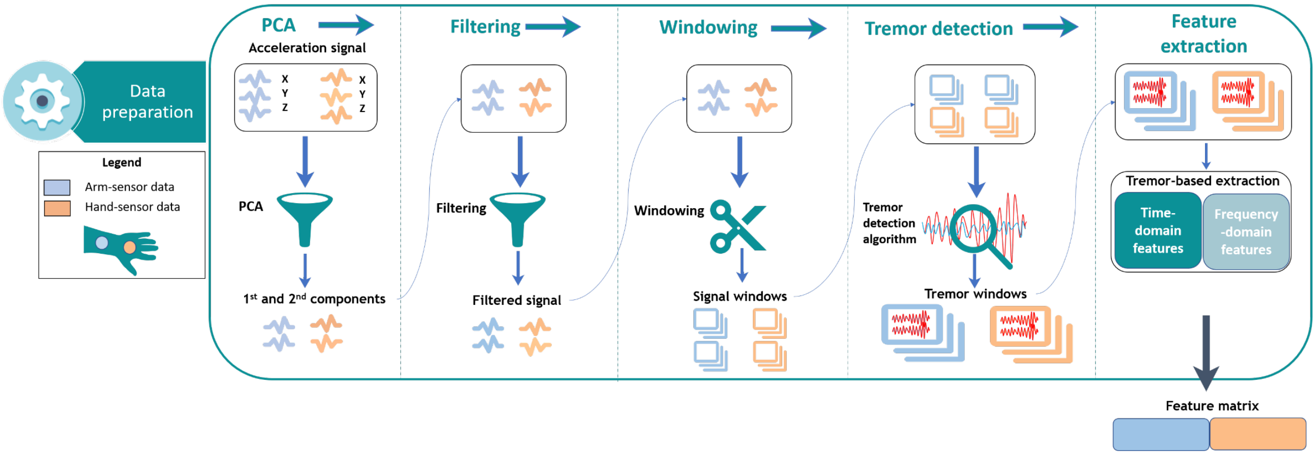

Figure 4. Data preparation diagram. Two accelerometers on the patients' hand and arm retrieve motion signals. Principal Component Analysis (PCA) reduces the dimensionality of the data. A band-pass filter removes voluntary movement and frequency content outside the pathological tremor range. The signal is windowed and passed to the power spectral density-based tremor detection algorithm. Lastly, features in the time and frequency domain are extracted, and a feature matrix is formed. It will be the input for the machine-learning algorithms.

motor tasks were used as input to the data preparation block, which consists of the following steps:

1. Principal Component Analysis (PCA): the first and second principal components from each 3D accelerometer data were extracted. This step reduces the dimensionality of the data while maintaining components that explain the signal variance. Besides removing noise, PCA prevents redundant information since the readings from the three sensor axes are expected to be highly correlated. The extraction of the dominant tremor axis (first principal component) through PCA is a required step in the tremor detection technique developed by Luft et al. [18]. The present work includes the second principal component of the signal to avoid discarding potential tremorous information during the tremor detection step.

2. Filtering: hand and arm accelerometer data were band-pass filtered (non-causal, zero-phase, 3–12 Hz, 2nd-order Butterworth filter). This frequency band covers most of the range of parkinsonian, essential, and dystonic tremors [29] and is suitable to exclude voluntary movement and physiological tremor information from the signal [30].

3. Windowing: data was split into 3s windows with 1.5s overlap, as in [18], and zero padding. The analysis of overlapping windows decreases the chance of the tremor detection algorithm to miss tremorous periods.

4. Tremor detection: following [18], each window was classified either as a tremor window (TW) or a non-tremor window (NTW). The classification relies on the power distribution within the tremor frequency band of the signal, acquired via the signal's power spectral density (PSD). In our case, a window is considered as TW if the power within the frequency band of dominant

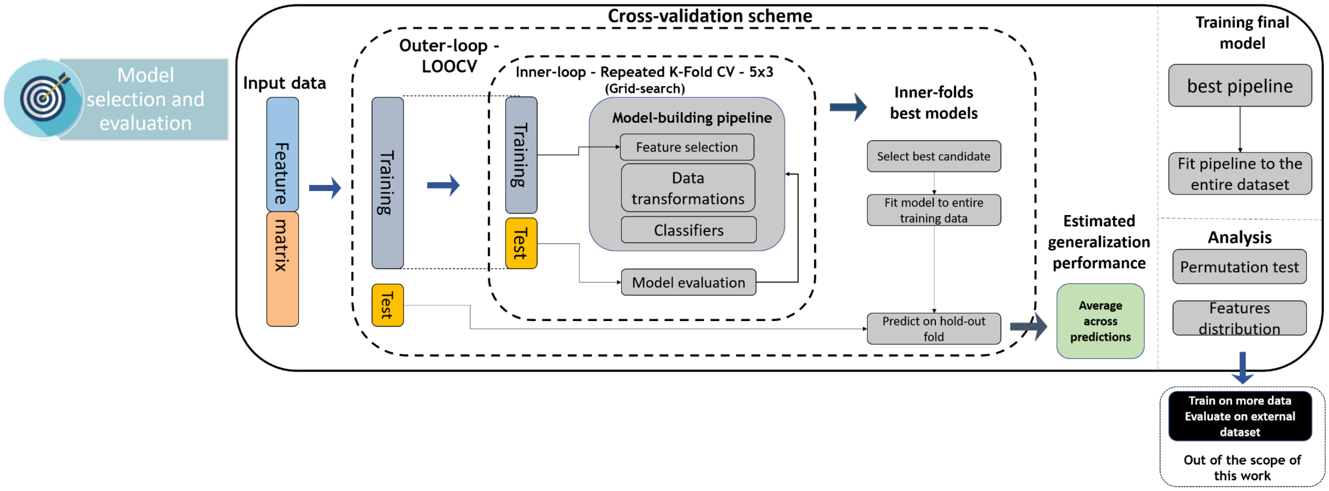

Figure 5. Model selection and evaluation via nested cross-validation. The outer-loop evaluates the generalization performance of pipelines, and the inner-loop select the best settings. After the pipeline with the lowest Brier score is identified, a final pipeline is built based on the distribution of the inner-loop results. A permutation test(n=1000) assesses the significance of the scores of the final pipeline.

frequency +- 0.5Hz represents 60% or more than the power within the 3 - 12 Hz band (see appendix C). In Luft et al. [18], by using a threshold of 40%, the technique incorrectly identified TW in healthy controls, both in the training and validation sets. The values of 45% and 50% were only tried in the training set. Therefore, a restrictive value of 60% was chosen to minimize the false positive rate of TW classification for the PD, ET, and DT patients.

5. Feature extraction: The presence of tremor can be task-dependent. Therefore, the strategy used for feature extraction in the time and frequency domain was as follows for each patient: if the presence of at least two TW was confirmed for the acceleration signal under analysis, the patient features were calculated as the mean of the values found for each TW. If less than two TW were detected, the features were computed as the mean of the values found for each NTW. As explained in 2.1, the patients' measurements were manually split into the corresponding motor tasks. Thus, the choice of two TW as a minimum to acknowledge tremor in a given motor task signal was employed to avoid possible tremorous motion recorded during the transition between tasks. The averaging of the values found in TWs to represent the final features is similar to what was done by Talitckii et al. [17]. The authors, however, used a different strategy to identify useful motion data (in our case, TWs) from the entire signal.

   We attempted to extract features covering different signal aspects, such as its linear and nonlinear autocorrelation, temporal statistics, complexity, and frequency content. Features used in machine-learning studies not related to tremor but with reported discriminative power across different datasets were also included [31]. The extracted predictors were divided into two groups: numeric and categorical. The numeric group was further subdivided into time and frequency-domain-based features. Table 2 summarizes the predictors used in this work.

   The output of the data preparation step, and input to the model-building pipelines, is a feature matrix consisting of 16 (tasks) * 12 (features) * 2 (sensors) * 2 (PCA components) = 268 features and 28 observations.

Model-building pipelines

Besides automating routine processes, the use of pipelines for modeling also prevents most cases of data leakage [33]. Each step of the pipeline fits to and transforms the training data; and transforms the test data based on the information retrieved during its fitting step. Therefore, the chances of the model-building process accidentally leak information from the test set is reduced.

The built pipelines consist of steps to select features to be used as input to the classifier; impute, scale, and encode the features; and the classifier itself. The steps of the pipelines varied according to the classifier to be integrated into it. The schematic diagram shown in Figure 6 shows the overall procedure to build the models. The steps of the pipelines are now described.

- Feature selection: a custom transformer extracts subsets of features from the feature matrix and passes them down the pipeline. These subsets are defined according to the user and correspond to the desired groups of motor tasks to be evaluated. The feature selection step of the model-building pipeline served to determine which group of motor tasks were considered by the

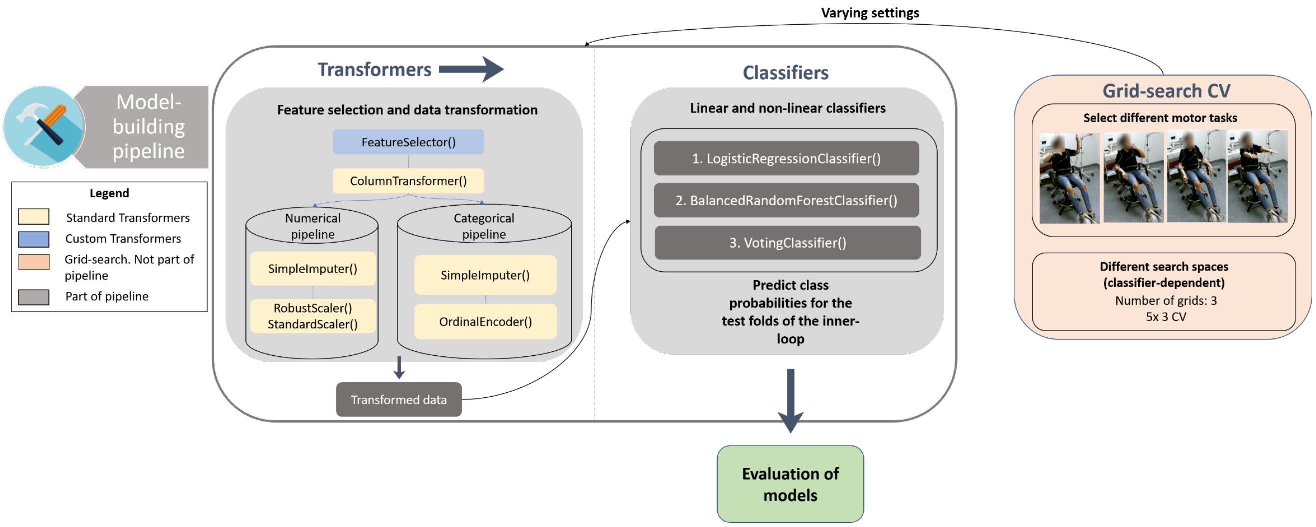

Figure 6. Model-building pipelines. The pipelines contain a custom feature selector transformer that selects subsets of the entire feature matrix to be passed down the pipeline. Predictors are divided into numerical and categorical, and appropriate data transformations are applied. The last step of the pipeline is feeding the transformed data to the classifiers to train and then predict on hold-out test folds. The grid-search CV is responsible for varying the combinations of features, scaling, and imputation methods. The transformers presented are available in the Scikit-learn library [24].

models as the most discriminative between essential, parkinsonian, and dystonic tremors.

- Column transformer: this transformer splits the features into numerical and categorical. It was a required step since appropriate modeling techniques depend on the feature data type. In our case, the features *presence of tremor* from the included tasks were grouped together, forming the categorical features. The remaining continuous features formed the numerical group.

  Feature Imputation: some patients were missing values for the features of the tasks in which they did not participate. Two widely used imputation strategies are considered for the numerical predictors: to replace the missing values by the mean or by the median of the other patient's respective features [34]. The imputed category was 'no tremor' for the categorical ones (i.e., *presence of tremor*).

  Feature scaling: two scaling methods are attempted: standardization of the predictors by removing their mean and scaling to unit variance; and a method robust to outliers that removes the median of the variables and scales the data according to the interquartile range. Scaling features is required to improve performance for many machine learning algorithms, especially those based on distance measures, such as Euclidean distance. Tree-based algorithms, on the other hand, can use non-scaled input data without a decrease in performance. The choice of an adequate scaling method depends on the distribution of the features among the patients (e.g., Gaussian) and the presence of outliers. Since we are blind to the statistical information of the features to prevent data leakage, both methods were attempted.

  Feature encoding: many machine learning algorithms require only numerical values as input. A categorical feature encoder assigned integer values for each unique category value. The values of '1' and '0' were used to represent the presence and absence of tremor during the motor tasks, respectively.

- Classifiers: one linear and one non-linear classifiers commonly used in scientific fields are evaluated [35]. In addition, a voting ensemble (ENS) of both classifiers is assessed. The linear algorithm is the Logistic Regression (LR); the non-linear is the Random Forest with an under-sampling implementation or Balanced Random Forest (BRF) [23]. The BRF was chosen rather than the regular RF due to its built-in strategy to account for class imbalance during training.

Six groups of features (see Table 3) from the feature matrix were created. The sets are extracted based on the type of motor task the patients perform to cover the range of rest and action (postural and kinetic) tremors. A baseline set consisting of the tasks *rest*, *hands_in_pronation*, and *top_nose* is considered, along with variations of it, by adding or subtracting features from the set. The baseline set was defined based on tasks used in previous tremor studies [11], [18], [36]. It is also considered the use of the entire feature matrix, with information from all motor tasks.

The choice of features to represent each category was made firstly by the availability of data. In case of a tie, the

Table 2. Features extracted from the time signals and the corresponding description.

Feature	Description
Numerical - Time-domain	
Root Mean Square (RMS)	Measure of the signal's strength
CID	Time-series complexity measure [32]
Tremor Stability Index (TSI)	Neurophysiological measure [15]
CO_FirstMin_ac	First minimum of the autocorrelation function [31]
CO_trev_1_num	Time-reversibility statistic [31]
SB_BinaryStats_mean_longstretch1	Longest period of consecutive values above the mean [31]
Skewness	Standard statistical measure
Kurtosis	Standard statistical measure
Numerical - Frequency-domain	
Dominant frequency	Frequency with max power from the signal's power-spectral density
Relative tremor power	Power within the frequency band of dominant frequency +-0.5Hz
Total tremor power	Power within the considered tremor frequency band (3 - 12Hz)
Categorical	
Presence of tremor	Tremor is considered present if two tremor windows are found in a given patient measurement

selection was randomized. The reasons for not including a greater range of feature groups were two: first, it was assumed that features from the same group retrieve redundant information about tremor. Second, a grid-search algorithm (explained in the next topic below) takes these sets of features and different data transformations methods as input to find the best-performing models. A great variability in the search space increases the computational time, and aligned with small sample size, may result in models with subpar generalization performance for unseen data [37].

Model selection and evaluation

A nested cross-validation (CV) procedure (see Figure 5) performs the steps of model selection and evaluation. In nested CV, the outer loop evaluates the model performance, and the inner loop is used for model selection. This method ensures that no data used for model selection is also used to assess its performance. It is similar to what was done by Lee et al. [38], where such steps were taken to prevent overfitting of the model and to provide an honest estimation of its generalization performance. In contrast to what was done in [38], the present work does not use a leave-one-out cross-validation procedure (LOOCV) in the model selection loop. This was done to prevent the optimization of the pipeline variables based on a single data sample.

- Model evaluation: LOOCV assesses the performance of the proposed model-building pipelines. LOOCV is a suitable evaluation method for small datasets [38]. It is the cross-validation procedure that provides the largest amount of training data. Each one of the observations (28 patients) is used once as the test set to estimate the performance of the different models trained in the correspondent inner training folds. The results for the outer folds of the LOOCV are averaged to obtain the final performance metric of the pipeline. Notice that the predictions of the samples left out, one at a time, are done by different models built using the same pipeline steps.
- Model selection: a repeated stratified K-fold CV, in a 5x3 configuration (i.e., K=3, five repetitions), is used to evaluate the different models built in the inner folds. Two-thirds of the available data in the inner folds of the LOOCV scheme (18 samples) are used to train the inner models, and a third (9 samples) is used for testing (K=3). The choice of K is a trade-off between available data for training and testing. Stratification of the data is done based on the class label of the samples, i.e., the inner training and test sets seek to keep the same proportion of observations from positive and negative classes. Larger values of K were discarded to do not severely restrict the size of the test sets. For each inner fold, the cross-validation procedure is repeated five times with different partitioning of the data. For each classification scenario (PD x no-PD, ET x no-ET, DT x no-DT), the positive class corresponds to the minority class: PD, ET, and DT, respectively.
- Grid search cross-validation: a cross-validated grid search performs an exhaustive search for the set of features, imputation and scaling methods, and classifier hyperparameters (see appendix D) that yield the best performance. The search is done using each one of the LOOCV training folds, with the cross-validation scheme defined for model selection. In the present

work, we have 5x3 scores for each one of these folds. Therefore, the grid-search CV calculates the mean Brier score across these 15 scores for each possible combination of hyperparameters, data transformation methods, and feature sets. The combinations that result in the lowest Brier score for each inner fold are stored.

- Final pipeline: once the best performing pipeline (with LR, BRF, or ENS as classifier) is defined, the most frequently selected hyperparameters, data transformation methods, and feature sets in the model selection folds (28) are used in the training of the final estimator.

- Analysis: lastly, the final estimator's cross-validation scores are tested for significance by permuting the labels (1000 times), and the empirical p-value is calculated against the null hypothesis that features and tremor classes are independent.

- Scoring rule: the strictly proper scoring rule Brier score (1)

$$\frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2, \tag{1}$$

  assesses the quality of each model-building pipeline, both in the outer and inner-loops. It measures the accuracy of probabilistic predictions, and it is defined as the mean squared error between the predictions and their corresponding true probabilities (100 or 0%), where N is the number of events, and f_t and o_t are the predicted probability and the true outcome of the t^{th} event, respectively. A perfect Brier score is 0, and the worst possible score is 1. In a binary classification context, a classifier that always predicts 50% ($f_t = 0.5$) for both classes achieves a Brier score of 0.25. The final score for each pipeline is the mean of the scores obtained from using the predictions for each patient (LOOCV) and the true labels, as done in [38]. The *predict_proba* method of each classifier instance estimated the class probabilities, as in [3], [14].

- Supplementary score: the area under the receiver operating characteristic curve (ROC AUC) is calculated for each one of the model pipelines. It is a common score used to assess the overall accuracy of diagnostic tests [39]. The ROC curve is a plot of the true positive rate (TPR) in function of the false positive rate (FPR) for different decision thresholds used to discriminate between the classes, providing information about the ability of the assessed classifier to correctly rank its predictions [40]. In the present work, for all the classification cases, the correct prediction of both classes is considered equally important.

Table 3. Groups of features based on the performed motor-tasks.

Group ID	Tasks included
Group 1	rest, pro, nose
Group 2	rest, pro
Group 3	pro, nose
Group 4	rest, pro, nose, tapping
Group 5	rest, pro, nose, higher
Group 6	All motor tasks

3. Results

This section is divided into two parts: main results and additional analysis. The main results correspond to the ones obtained following the proposed methodology (section 2). The additional analysis presents the class probabilities assigned by the final pipeline for patients in TIM-Tremor dataset that were labeled as not presenting convincing tremor (NCT) or as possible ET (PET).

3.1 Main results

Brier score and ROC-AUC

The pipelines with BRF as the classifier outperformed the other pipelines relying on LR or ENS for all classification scenarios. The Brier score and ROC-AUC obtained by the BRF pipelines for each scenario is shown in Table 4, along with the best set of motor tasks, imputation method, and hyperparameters that yielded the best performances. The results for the remaining pipelines are found in appendix E. The ROC curves for all pipelines (including a hypothetical model with no skill) are presented in Figure 7.

Permutation test

A final pipeline was defined for each classification scenario based on the most selected hyperparameters, imputation method, and motor tasks from the inner-folds of the model-evaluation procedure (see section 2.2). The significances of the scores obtained by the final pipelines in a LOOCV procedure are presented in Figure 10.

Selection of motor tasks, hyperparameters, and data transformations

Figure 8 presents the distribution (aggregate of inner-folds selections) of the selected classifier hyperparameters, motor tasks, and data transformations for the PD x no-PD classification scenario and the BRF pipeline. The distribution of selected motor tasks by all three pipelines in the PD x no-PD case is shown in Figure 9. It is clearly observed the consistent selection of the Group 1 for all pipelines, especially for the BRF. The settings' distributions for the remaining classification scenarios are shown in appendix E.

Predicted probabilities

As discussed in section 1, the reporting of class probabilities for patients is essential in a decision-making setting. Thus, Table 5 depicts the predicted probabilities assigned for each patient by the model-building procedure with BRF as the classifier for the PD x no-PD scenario. However, if binary labels are desired, a threshold needs to be defined to map the predicted probabilities to binary classes. An optimal threshold that balances TPR and FPR can be found by maximizing the geometric mean of sensitivity and specificity, which is defined as the square root of the product between sensitivity and specificity [41]. In our case, 42% was found to be the optimal threshold that should be used to classify a patient as having PD. For this threshold, the BRF pipeline yields an accuracy of 75% (21 out of 28 correctly classified patients), sensitivity of 100%, and specificity of 67%. If the threshold is set to 50%, the BRF pipeline achieves an accuracy, sensitivity, and specificity of 71%.

Table 5. Predicted class probabilities of the 28 patients by the BRF pipeline - PD x no-PD.

Patient ID	Diagnosis	Predicted Probability (%)	
		PD	NoPD
T002	no-PD	41	59
T004	PD	62	38
T005	no-PD	33	67
T006	no-PD	60	40
T008	no-PD	68	32
T010	no-PD	54	46
T012	PD	60	40
T013	PD	48	52
T014	PD	42	58
T019	no-PD	39	61
T020	PD	68	32
T022	PD	66	34
T023	no-PD	48	52
T024	PD	74	26
T026	no-PD	29	71
T027	no-PD	17	83
T028	no-PD	30	70
T029	no-PD	35	65
T030	no-PD	35	65
T031	no-PD	35	65
T032	no-PD	75	25
T036	no-PD	36	64
T039	no-PD	28	72
T042	no-PD	57	43
T045	no-PD	22	78
T046	no-PD	57	43
T050	no-PD	33	67
T052	no-PD	39	61

3.2 Additional analysis

Predicted probabilities for inconclusive diagnoses

Data from seven patients in the TIM-Tremor dataset considered as with no (convincing) tremor (NCT) or with a diagnosis of possible ET (PET) were used as input to the PD x no-PD final pipeline which used the initial 28 patients' data for training. The predicted probabilities assigned by the model are shown in Table 6.

Table 6. Predicted class probabilities of inconclusive patients by the BRF pipeline - PD x no-PD

Patient ID	Diagnosis	Predicted Probability (%)	
		PD	NoPD
T001	NCT	36	64
T003	NCT	74	26
T007	NCT	41	59
T025	NCT	33	67
T049	NCT	37	63
T021	PET	44	56
T038	PET	45	55

4. Discussion

4.1 Main results

Brier score and ROC-AUC

For all three diagnostic scenarios, at least one pipeline performed better than a hypothetical classifier that assigns 50% of chance for both positive and negative classes (Brier score = 0.25). As displayed in Figure 7, the pipelines with a linear classifier (LR) achieved the lowest scores among all scenarios, while the ones with a non-linear classifier (BRF) obtained the best results. The pipelines with a voting ensemble classifier performed better than random in the PD x no-PD and ET x no-ET case. These results suggest that the classes are not linearly separable by using the presented methodology.

The evaluation scores for the PD x no-PD scenario were significantly higher than the ones corresponding to the remaining scenarios, confirming our expectations. A possible reason for the subpar performance in the remaining classification scenarios could be because dystonia is the most frequent differential diagnosis of ET [5]. Therefore, the classifiers could have simply not found the underlying structure in the data delineating between ET and DT. On top of that, the use of only accelerometers to record tremor motion restricts the acquisition of tremor information. As discussed in [10], [11], tremor disorders may differ from each other based on tremor directionality. For instance, the study of Sternberg et al. [11] suggests that ET patients manifest more wrist flexion-extension tremor than PD patients during sustained arm extension (task *pro* in the present work). Thus, the additional measurements of angular rate with an inertial measurement unit (IMU), for instance, could increase the performance of tremor classification models. Other possibilities include a potential lack of discriminative power of the features and motor tasks used as inputs to the pipelines, and an eventual mislabeling of some ET and DT patients in the TIM-Tremor dataset, either during data collection or due to diagnostic error.

Regarding PD x no-PD, it is interesting to notice that although the pipeline with BRF as classifier achieved the highest

Table 4. Pipelines' results. The Brier score and ROC-AUC represents the estimated generalization performance. The bottom-four rows are the selected parameters to build the final pipeline. These parameters were chosen based on the inner-fold results of the model-selection grid-search procedure.

Results	PD x no-PD	ET x no-ET	DT x no-DT
Brier score	0.188 (0.128)	0.235 (0.156)	0.230 (0.061)
ROC-AUC	0.84	0.70	0.69
Best set of motor tasks	Group 1	Group 4	Group 6
Imputation method	mean	mean	median
min_samples_split	4	4	2
min_samples_leaf	1	1	1

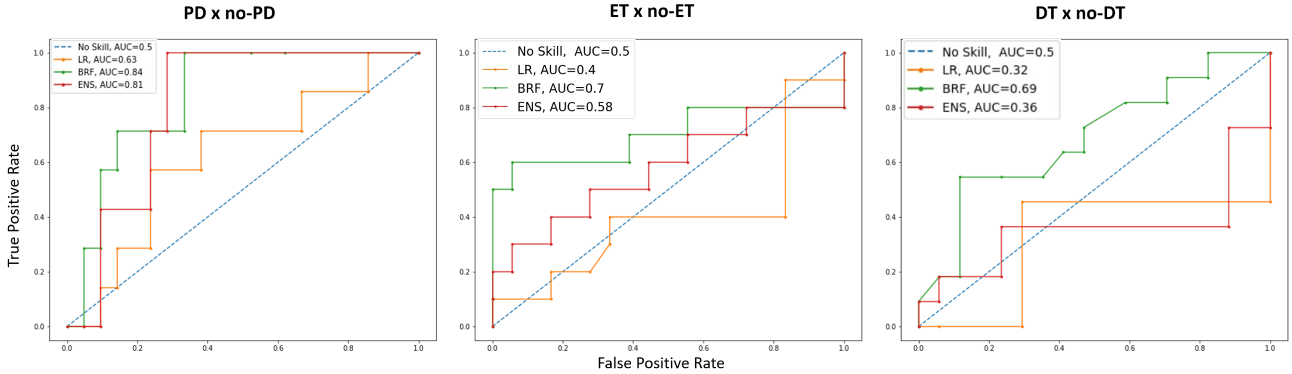

Figure 7. Receiver Operating Characteristic (ROC) curves and area under the curve (AUC) score for all pipelines, for each classification case.

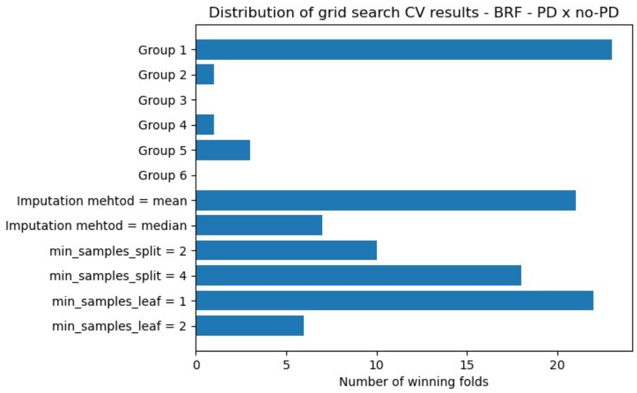

Figure 8. Distribution of selected parameters for the BRF - PD x no-PD. pipeline.

ROC AUC score, ENS performed better in part of the upper-left region of the ROC plot (Figure 7). This fact suggests that the linear model predictions contributed positively to some of the predictions of the BRF. Therefore, for a hypothetical diagnostic scenario using the developed models, the ENS-based pipeline should be considered the best if an operating point in that region (around 0.3% FPR) is desired.

Considering the ET x no-ET and the DT x no-DT cases, it is shown that, at some point, the ROC-AUC scores presented long horizontal lines. It means that even continuously increasing the FPR rate, no additional sample was assigned to the positive class. Additionally, it can be seen inclined segments in the plots. This type of phenomenon occurs when there is a tie, i.e., a region of uncertainty that appears when positive and negative samples are assigned to the same predicted probability [42].

Permutation test

The superior performance of the pipelines in classifying PD x no-PD patients when compared to the other diagnostic scenarios is further evidenced by the results of the permutation tests 10. It is observed that although the BRF pipelines obtained a Brier score below 0.25 and a ROC-AUC above 0.5 for all diagnostic cases (see Figure 4), statistical significance (p-value < 0.05) of the scores of the final BRF pipelines was only achieved for PD x no-PD, which was the case that presented the lowest Brier score.

Predicted probabilities

The report of assigned class probabilities allows patients and care providers to account for uncertainty in the decision of treatment plans and follow-up consultations, for instance [19]. It is observed from Table 5 that the predictions are rarely close to 100% or 0%. The reasons for this behavior, common to the Random Forest classifier, are explained in [43]. The BRF pipeline predictions for patients *T008* and *T032* are particularly interesting. Despite their diagnoses of no-PD and the discussed predictions behavior of the classifier, significant PD

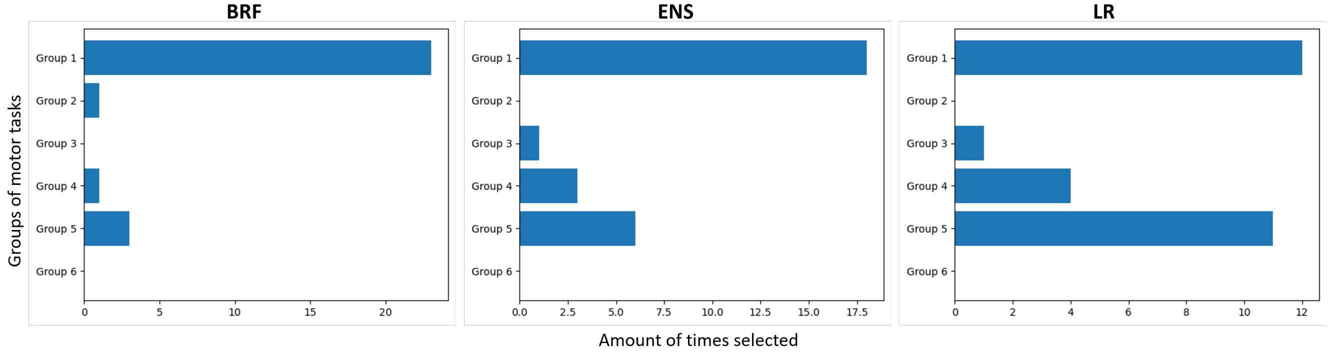

Figure 9. Frequency of selection for each group of features, corresponding to different motor tasks. Balanced Random Forest (BRF), voting ensemble (ENS), and logistic regression (LR) pipelines - PD x no-PD. Note that the scale of the plots are not the same.

probabilities were assigned to them.

An interesting discussion arises from the adjustment of the decision threshold used to classify the patients. If the optimal threshold of 42% for PD prediction is used, instead of 50%, patients *T013, T014* would be correctly classified.

Most discriminative set of motor tasks

Insights about the motor tasks considered by the estimators as the most important for their outputs will only be drawn from the PD x no-PD results since it is the case that yielded a final pipeline that had statistical significance for its performance.

As shown in Figure 9, the Group 1 of motor tasks was consistently chosen in the model selection loop (5x3 CV) for all the pipelines. This suggests that the inclusion of other motor tasks tends to add noisy, redundant, or irrelevant features to the feature space. For the BRF pipeline, Group 5, which adds features from the entrainment task *2_Hz_higher*, was selected for three out of the 28 leave-one-out cross-validation folds. Contrary to our expectations, Group 4, which adds the distraction task *tapping*, was selected only once. The same happened for the Group 2, which removes the kinetic task *top_nose*. These results suggest that for the diagnostic scenario of PD x no-PD, motor tasks other than the standard set of rest, posture, and kinetic tasks are not needed for the differentiation of tremor patients by using the methodology introduced in this study.

4.2 Additional analysis

Distribution of specific features

A particular interest arises over the patients identified with tremorous accelerometer measurements according to the tremor detection technique developed by Luft et al. [18]. Another point of discussion is the distribution of the neurophysiological measure TSI [15] among the patients due to its reported great performance in the differentiation of PD and ET patients. As done in Luft et al. [18], a discussion over the variability in values of TSI calculated based on TW and NTW takes place in this section. Figure 11 presents scatter plots of the analyzed features.

Some aspects observed in Figure 11 call our attention: first, tremor was detected in all PD patients during the *rest* task, based on the first principal component of the hand accelerometer signal. The manifestation of rest tremor in PD patients is a recurrently reported PD symptom [15], [18], [30], [36]. However, tremor was not detected for all PD patients when the second principal component was used as input to the tremor detection method used in the present work. It is interesting to see that at least two PD patients also presented tremor during the *hands_in_pronation* and *top_nose* tasks. Postural tremors may occur in PD, but, contrarily, kinetic tremors are rarely seen in such patients [11].

Concerning the features related to the TSI, it can be seen that higher TSI values were found for patients who did not have tremor detected during the performance of all tasks included in Figure 11 compared to those that presented tremor. This is in accordance with what was reported by Luft et al. [18]: TSI calculated based on NTWs presents higher values compared to TSI calculated in reliance on TWs.

Predicted probabilities for inconclusive diagnoses

One patient (T003) had a predicted probability of having PD higher than 50% (Table 6). In agreement with the treating neurologists' diagnosis, both PET patients presented a higher no-PD probability than PD.

An interesting point of discussion arises when the presence of tremor is assessed for those patients: tremor was detected in the *rest* task for two NCT patients (T003 and T025), and in the *top_nose* task for patient T025 (labeled as jaw tremor). As previously discussed, the occurrence of tremors in that task is common in PD. Perhaps that was one of the reasons for the higher assigned PD class probability for patient T003 compared to the remaining NCT patients. Figure 12 depicts the scatter plots of the features for the inconclusive diagnoses cases. The dependence on tremor occurrence for the low TSI values persisted in this analysis.

4.3 Limitations and recommendations

One limitation of this study is the amount of PD, ET, and DT patients included in the TIM-Tremor dataset. Additional

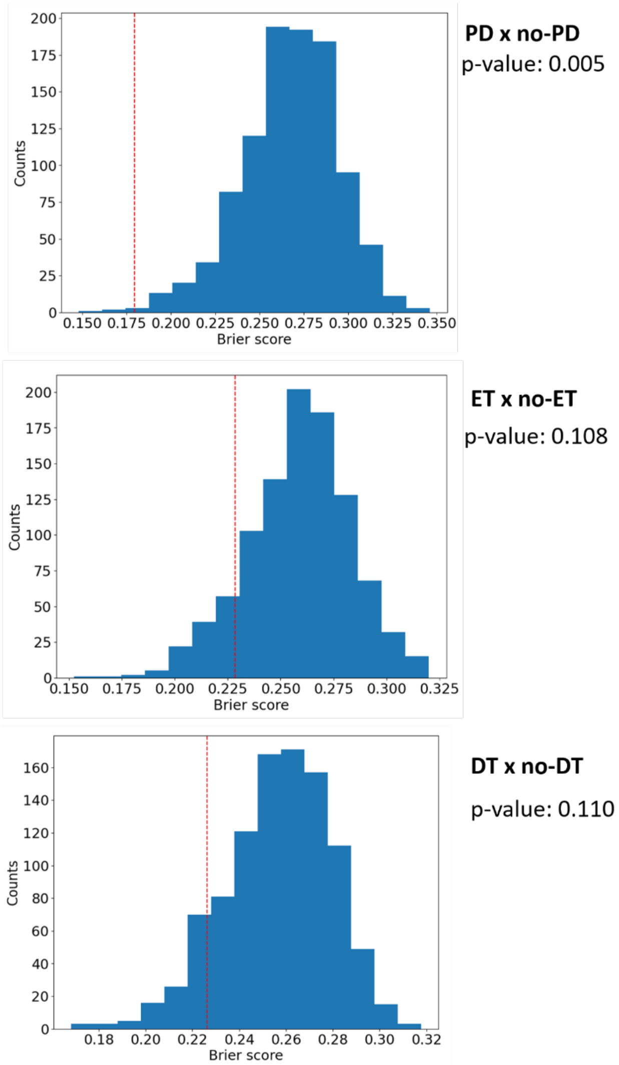

Figure 10. Permutation tests (n=1000) for the final BRF pipelines trained with fixed settings.

data is required to develop a more generalizable classification model, as more diverse patient characteristics will be used for model training.

A more in-depth analysis of the distributions of features among the patients is required to gain insights into the difference in performance among the three diagnostic scenarios. Since we were blinded to the feature distribution during this work's development (to avoid data leakage), this prior data analysis stage was not possible.

On top of that, the present work results are restricted by the ground-truth labels assigned to each patient. Even though neurologists specialists in movement disorders performed the diagnoses of the included patients, misdiagnosis can still occur. The possibility that some of the attributed labels were switched for one another during the dataset's creation cannot be discarded.

For future works in tremor classification, the use of sensors able to retrieve angular rate information of the patients is recommended. Also, the duration of the measurements of motor tasks should be set the same for all participants.

The adoption of machine-learning techniques in clinical settings is still in its early stages. More robust models developed and evaluated based on larger and distinct datasets are required. Nevertheless, models developed as proofs-of-concept are the initial step towards integrating novel diagnostic technologies in the medical field. Future works should focus first on acquiring quality data that are more representative of the problem being solved.

5. Conclusion

In this study, linear and non-linear machine-learning pipelines were developed to predict class probabilities of tremor patients diagnosed with PD, dystonia, and ET. The presented methodology showed to be suitable for the differentiation between PD and ET/DT patients. A set of one rest, one postural, and one kinetic motor task proved to contain the most discriminative group of features for the PD x no-PD diagnostic scenario. For an optimal decision threshold of 42% to classify a patient as PD, the pipeline achieved an accuracy of 75%, specificity of 67%, and sensitivity of 100%. In agreement with [18], TSI calculated based on TWs presented lower values when compared to TSI obtained from NTWs. In addition, patients labeled as NCT had tremor episodes detected during the performance of rest and postural motor tasks. Also, the predicted class probabilities assigned by a final model for patients labeled as PET conformed with their labels (higher chance of the no-PD class). The promising results achieved by the proof-of-principle pipeline encourages further development of assistive diagnostic technologies in clinical practice.

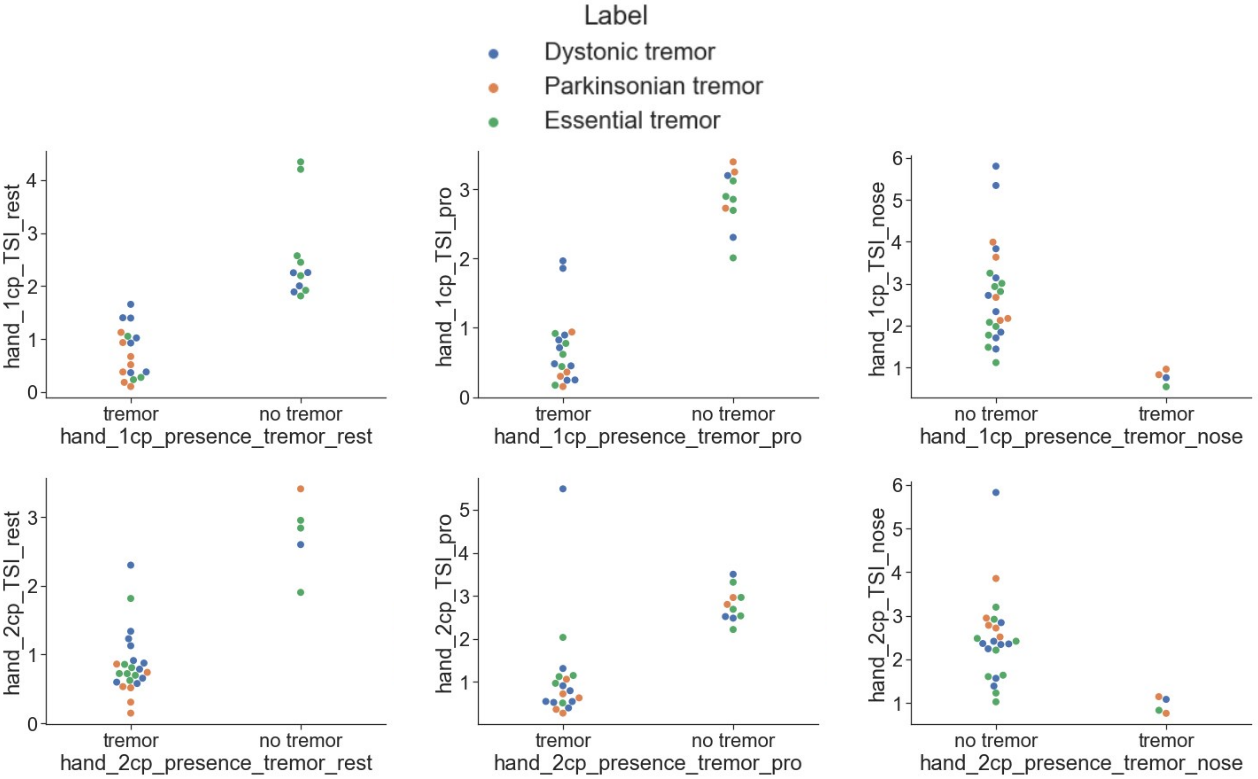

Figure 11. Scatter plots of TSI and presence of tremor features. The top three plots correspond to the features obtained from the first principal component of the hand accelerometer signal. The bottom three refer to the features obtained from the second principal component. The *rest*, *hands_in_pronation*, and *top_nose* tasks are analysed as they were considered the most discriminative set of motor tasks. The lower amount of data samples for the *top_nose* task is due to the absence of recordings for some patients.

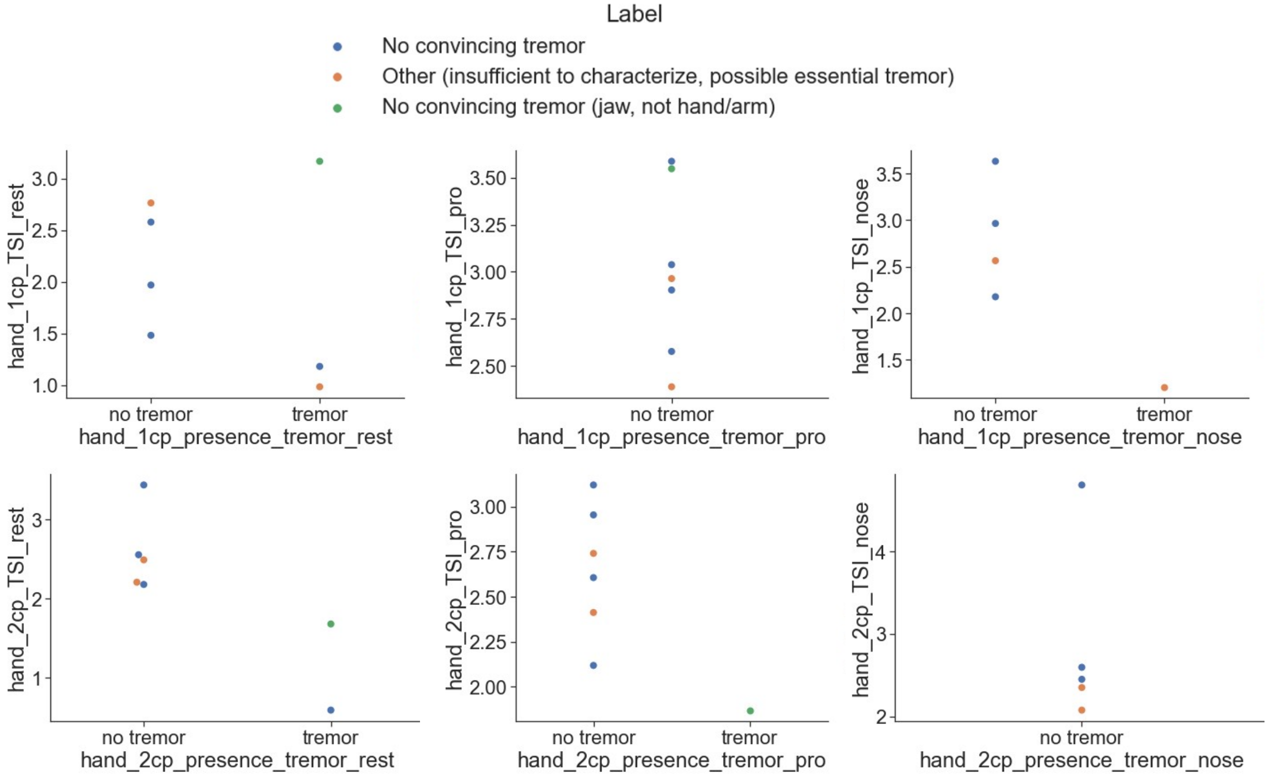

Figure 12. Scatter plots of TSI and presence of tremor features for the inconclusive diagnoses cases. The top three plots correspond to the features obtained from the first principal component of the hand accelerometer signal. The bottom three refer to the features obtained from the second principal component. Note that the patient with jaw tremor did not perform the *top_nose* task.

References

[1] J. L. Ostrem and N. B. Galifianakis, "Overview of common movement disorders.," eng, *Continuum (Minneapolis, Minn.)*, vol. 16, no. 1 Movement Disorders, pp. 13–48, Feb. 2010.

[2] S. Jain, S. E. Lo, and E. D. Louis, "Common Misdiagnosis of a Common Neurological Disorder: How Are We Misdiagnosing Essential Tremor?" *Archives of Neurology*, vol. 63, no. 8, pp. 1100–1104, Aug. 2006.

[3] D. Valeriani and K. Simonyan, "A microstructural neural network biomarker for dystonia diagnosis identified by a DystoniaNet deep learning platform," *Proc Natl Acad Sci U S A*, 2020.

[4] G. Deuschl, P. Bain, and M. Brin, "Consensus statement of the Movement Disorder Society on Tremor. Ad Hoc Scientific Committee.," eng, *Movement disorders : official journal of the Movement Disorder Society*, vol. 13 Suppl 3, pp. 2–23, 1998.

[5] C. J. Amlang, D. Trujillo Diaz, and E. D. Louis, "Essential Tremor as a "Waste Basket" Diagnosis: Diagnosing Essential Tremor Remains a Challenge.," eng, *Frontiers in neurology*, vol. 11, p. 172, 2020.

[6] M. Wang, G. Wenbo, D. Apthorp, and H. Suominen, "Robust Feature Engineering for Parkinson's Disease Diagnosis (Preprint)," *JMIR Biomedical Engineering*, vol. 5, 2019.

[7] M. Selikhova, P. Kempster, T. Revesz, J. Holton, and A. Lees, "Neuropathological findings in benign tremulous Parkinsonism: Benign Tremulous Parkinsonism," *Movement disorders : official journal of the Movement Disorder Society*, vol. 28, 2013.

[8] T. Xie, R. de la Fuente-Fernandez, P. Warnke, and U. J. Kang, "Role of DaTSCAN and clinical diagnosis in Parkinson diseaseAuthor Response:" *Neurology*, vol. 79, no. 16, p. 1744, 2012.

[9] K. Badiavas, E. Molyvda, I. Iakovou, M. Tsolaki, K. Psarrakos, and N. Karatzas, "SPECT imaging evaluation in movement disorders: far beyond visual assessment.," eng, *European journal of nuclear medicine and molecular imaging*, vol. 38, no. 4, pp. 764–773, Apr. 2011.

[10] A. C. Pigg, J. Thompson-Westra, K. Mente, C. W. Maurer, D. Haubenberger, M. Hallett, and S. K. Charles, "Distribution of tremor among the major degrees of freedom of the upper limb in subjects with Essential Tremor.," eng, *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, vol. 131, no. 11, pp. 2700–2712, Nov. 2020.

[11] E. J. Sternberg, R. N. Alcalay, O. A. Levy, and E. D. Louis, "Postural and Intention Tremors: A Detailed Clinical Study of Essential Tremor vs. Parkinson's Disease.," eng, *Frontiers in neurology*, vol. 4, p. 51, 2013.

[12] B. Post, M. P. Merkus, R. M. A. de Bie, R. J. de Haan, and J. D. Speelman, "Unified Parkinson's disease rating scale motor examination: are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable?" eng, *Movement disorders : official journal of the Movement Disorder Society*, vol. 20, no. 12, pp. 1577–1584, Dec. 2005.

[13] G. Rizzo, M. Copetti, S. Arcuti, D. Martino, A. Fontana, and G. Logroscino, "Accuracy of clinical diagnosis of Parkinson disease: A systematic review and meta-analysis.," eng, *Neurology*, vol. 86, no. 6, pp. 566–576, Feb. 2016.

[14] N. H. Ghassemi, F. Marxreiter, C. F. Pasluosta, P. Kugler, J. Schlachetzki, A. Schramm, B. M. Eskofier, and J. Klucken, "Combined accelerometer and EMG analysis to differentiate essential tremor from Parkinson's disease.," eng, *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2016, pp. 672–675, Aug. 2016.

[15] L. Di Biase, J.-S. Brittain, S. Shah, D. Pedrosa, H. Cagnan, A. Mathy, C. Chen, J. Martín-Rodríguez, P. Mir, L. Timmerman, P. Schwingenschuh, K. Bhatia, V. Di Lazzaro, and P. Brown, "Tremor stability index: A new tool for differential diagnosis in tremor syndromes," *Brain : a journal of neurology*, vol. 140, 2017.

[16] A. Hossen, M. Muthuraman, Z. Al-Hakim, J. Raethjen, G. Deuschl, and U. Heute, "Discrimination of Parkinsonian tremor from essential tremor using statistical signal characterization of the spectrum of accelerometer signal," English, *Bio-Medical Materials and Engineering*, vol. 23, no. 6, pp. 513–531, 2013.

[17] A. Talitckii, E. Kovalenko, A. Anikina, O. Zimniakova, M. Semenov, E. Bril, A. Shcherbak, D. V. Dylov, and A. Somov, "Avoiding Misdiagnosis of Parkinson's Disease With the Use of Wearable Sensors and Artificial Intelligence," *IEEE Sensors Journal*, vol. 21, no. 3, pp. 3738–3747, 2021.

[18] F. Luft, S. Sharifi, W. Mugge, A. Schouten, L. Bour, A.-F. van Rootselaar, P. Veltink, and T. Heida, "A power spectral density-based method to detect tremor and tremor intermittency in movement disorders," *Sensors (Switzerland)*, vol. 19, no. 19, 2019.

[19] I. Y. Chen, S. Joshi, M. Ghassemi, and R. Ranganath, "Probabilistic Machine Learning for Healthcare," *Annual Review of Biomedical Data Science*, vol. 4, no. 1, pp. 393–415, 2021.

[20] A. Dawid and M. Musio, “Theory and Applications of Proper Scoring Rules,” *METRON*, vol. 72, 2014.

[21] M. H. Zweig and G. Campbell, “Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine.,” eng, *Clinical chemistry*, vol. 39, no. 4, pp. 561–577, Apr. 1993.

[22] W. Luo, D. Phung, T. Tran, S. Gupta, S. Rana, C. Karmakar, A. Shilton, J. Yearwood, N. Dimitrova, B. Ho, S. Venkatesh, and M. Berk, “Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View,” *J Med Internet Res*, vol. 18, e323, 2016.

[23] G. Lemaitre, F. Nogueira, and C. K. Aridas, “Imbalanced-learn: {A} Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning,” *CoRR*, vol. abs/1609.0, 2016.

[24] 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,” *CoRR*, vol. abs/1201.0, 2012.

[25] A. S. B. Malling, B. M. Morberg, L. Wermuth, O. Gredal, P. Bech, and B. R. Jensen, “The influence of posture duration on hand tremor during tasks with attention-distraction in persons with Parkinson’s disease,” *Journal of NeuroEngineering and Rehabilitation*, vol. 16, no. 1, p. 61, 2019.

[26] C. Kenney, A. Diamond, N. Mejia, A. Davidson, C. Hunter, and J. Jankovic, “Distinguishing psychogenic and essential tremor.,” eng, *Journal of the neurological sciences*, vol. 263, no. 1-2, pp. 94–99, Dec. 2007.

[27] A. W. G. Buijink, M. F. Contarino, J. H. T. M. Koelman, J. D. Speelman, and A. F. van Rootselaar, “How to tackle tremor - systematic review of the literature and diagnostic work-up.,” eng, *Frontiers in neurology*, vol. 3, p. 146, 2012.

[28] P. Bank, J. Zheng, S. Pintea, P. Ouwehand, S. de Bot, and J. van Gemert, *Technology in Motion Tremor Dataset: TIM-Tremor*, 2019.

[29] C. W. Hess and S. L. Pullman, “Tremor: clinical phenomenology and assessment techniques.,” eng, *Tremor and other hyperkinetic movements (New York, N.Y.)*, vol. 2, 2012.

[30] T. Heida, E. C. Wentink, and E. Marani, “Power spectral density analysis of physiological, rest and action tremor in Parkinson’s disease patients treated with deep brain stimulation.,” eng, *Journal of neuroengineering and rehabilitation*, vol. 10, p. 70, Jul. 2013.

[31] C. H. Lubba, S. S. Sethi, P. Knaute, S. R. Schultz, B. D. Fulcher, and N. S. Jones, “catch22: CAnonical Time-series CHaracteristics,” *Data Mining and Knowledge Discovery*, vol. 33, no. 6, pp. 1821–1852, Jan. 2019.

[32] G. Batista, X. Wang, and E. Keogh, “A Complexity-Invariant Distance Measure for Time Series,” in *Proceedings of the 11th SIAM International Conference on Data Mining, SDM 2011*, 2011, pp. 699–710.

[33] M. Cearns, T. Hahn, and B. T. Baune, “Recommendations and future directions for supervised machine learning in psychiatry,” *Translational Psychiatry*, vol. 9, no. 1, p. 271, 2019.

[34] A. Jadhav, D. Pramod, and K. Ramanathan, “Comparison of Performance of Data Imputation Methods for Numeric Dataset,” *Applied Artificial Intelligence*, vol. 33, no. 10, pp. 913–933, 2019.

[35] R. Couronné, P. Probst, and A.-L. Boulesteix, “Random forest versus logistic regression: a large-scale benchmark experiment,” *BMC Bioinformatics*, vol. 19, no. 1, p. 270, 2018.

[36] J. C. van den Noort, R. Verhagen, K. J. van Dijk, P. H. Veltink, M. C. P. M. Vos, R. M. A. de Bie, L. J. Bour, and C. T. Heida, “Quantification of Hand Motor Symptoms in Parkinson’s Disease: A Proof-of-Principle Study Using Inertial and Force Sensors.,” eng, *Annals of biomedical engineering*, vol. 45, no. 10, pp. 2423–2436, Oct. 2017.

[37] J. Lever, M. Krzywinski, and N. Altman, “Model selection and overfitting,” *Nature Methods*, vol. 13, no. 9, pp. 703–704, 2016.

[38] Q. Y. Lee, S. J. Redmond, G. S. Chan, P. M. Middleton, E. Steel, P. Malouf, C. Critoph, G. Flynn, E. O’Lone, and N. H. Lovell, “Estimation of cardiac output and systemic vascular resistance using a multivariate regression model with features selected from the finger photoplethysmogram and routine cardiovascular measurements.,” eng, *Biomedical engineering online*, vol. 12, p. 19, Mar. 2013.

[39] J. N. Mandrekar, “Receiver operating characteristic curve in diagnostic test assessment.,” eng, *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer*, vol. 5, no. 9, pp. 1315–1316, Sep. 2010.

[40] H. He and Y. Ma, “Imbalanced learning: Foundations, algorithms, and applications,” *Imbalanced Learning: Foundations, Algorithms, and Applications*, pp. 1–210, Jan. 2013.

[41] J. Brownlee, *A Gentle Introduction to Threshold-Moving for Imbalanced Classification*, Feb. 2020.

[42] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[43] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning*, 2005, pp. 625–632.

[44] W. McKinney, "{D}ata {S}tructures for {S}tatistical {C}omputing in {P}ython," in *{P}roceedings of the 9th {P}ython in {S}cience {C}onference*, S. van der Walt and J. Millman, Eds., 2010, pp. 56–61.

[45] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Computing in Science \& Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[46] M. L. Waskom, "seaborn: statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021.

[47] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with {NumPy}," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020.

[48] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "{SciPy} 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[49] M. Christ, N. Braun, J. Neuffer, and A. W. Kempa-Liehr, "Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package)," *Neurocomputing*, vol. 307, pp. 72–77, Sep. 2018.

[50] B. D. Fulcher and N. S. Jones, "hctsa: A Computational Framework for Automated Time-Series Phenotyping Using Massive Feature Extraction," *Cell Systems*, vol. 5, no. 5, pp. 527–531, Nov. 2017.

A. Python libraries

The Python libraries used in the current study are described in Table 7.

Table 7. Python libraries

Library	Version	Description
Scikit-learn [24]	0.24.1	Machine-learning tools
Imblearn [23]	0.8.0	Machine-learning tools
Pandas [44]	1.2.4	Data analysis and manipulation
Matplotlib [45]	3.3.4	Data visualization
Seaborn [46]	0.11.1	Data visualization
Numpy [47]	1.20.1	Scientific computing
Scipy [48]	1.6.2	Scientific computing
tsfresh [49]	0.18.0	Time-series features
hctsa [50]	—	Time-series features

B. Description of motor tasks

Table 8 describes the motor tasks performed by the patients during data collection.

C. Tremor detection

Examples of TW and NTW detected using the method developed by Luft et al. [18] are presented in Figure 13. The detection does not depend on the intensity of the tremor, but on its power concentration around the peak tremor frequency (peak frequency +- 0.5Hz).

D. Classifiers' hyperparameters

The search for the best combination of classifier hyperparameters was performed by the grid-search cross-validation procedure discussed in section 2.2. According to [37], the variability in the search space should remain low when dealing with small datasets in order to prevent lack of generalization performance for unseen data. Table 9 shows the hyperparameters that were varied during grid-search.

Table 8. Motor tasks and description. Adapted from the TIM-Tremor [28] dataset documentation.

Task	Description
Rest	
Rest	Resting the arms on the chair handles.
Rest_in_supination	Resting the arms on the chair handles, hands in supination position.
Postural	
Hands_in_pronation	Both arms outstretched forward, hands in pronation position.
Top_top	Both hands in front of the chest with tips of the index fingers almost touching each other, elbows lifted sideways at approx. 90 degrees angle.
Thumbs_up	Holding the fingertips in front of each other, with the elbows lifted at 90 degrees angle.
Weight	The affected arm outstretched forward, with a weight attached to the wrist.
Extra_pose	Holding a pose proposed by the medical expert to better visualize the tremor.
Action	
Top_nose	Touching the top of the nose with the right/left index finger.
Writing	Writing a given sentence.
Spiral	Drawing a spiral.
Extra_writing	Extra writing task with a special pen, or diverging from the standard writing task.
Distraction	
Months_backward	Naming the months backwards, with the most affected arm outstretched forward.
Counting	Counting backwards from 100 in steps of 7, with the most affected arm outstretched forward.
Finger_tapping	Tapping using the index and thumb of the contralateral hand.
Playing_piano	Moving the thumb of the contralateral hand across all fingers from the index to the pinky finger, and back.
Following	Following a moving pointer, with the index finger of the contralateral hand.
Entrainment	
2_Hz_higher	Tapping with the contralateral hand in the rhythm of a flashing light, 2 Hz higher than the frequency estimated at rest.
2_Hz_lower	Tapping with the contralateral hand in the rhythm of a flashing light, 2 Hz lower than the frequency estimated at rest.

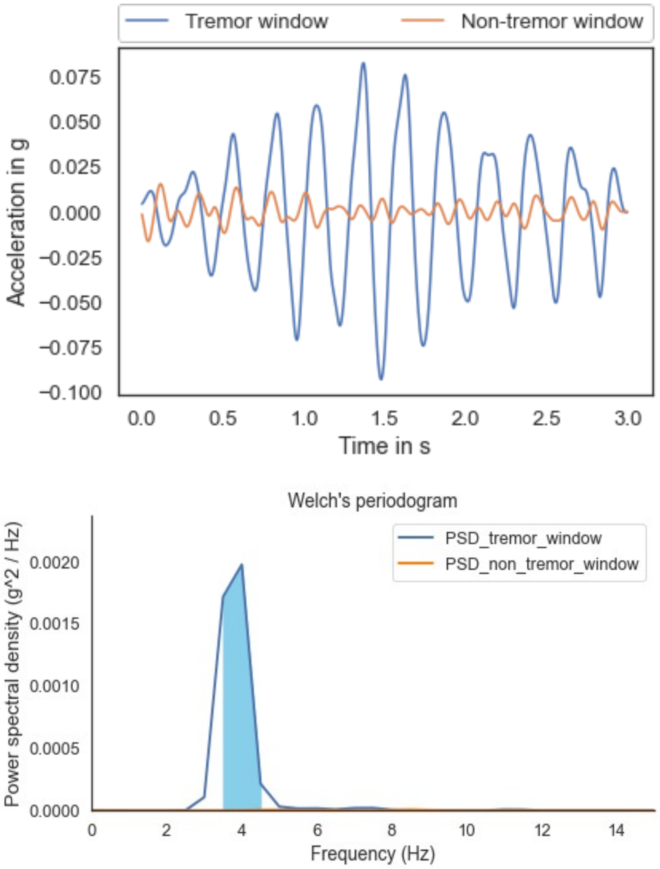

Figure 13. TW and NTW (and corresponding power spectral density) detected in the band-pass filtered first principal component of the hand accelerometer signal for the patient *T022*, during the performance of the *rest* motor task. The peak tremor frequency is 4Hz. The colored area represent the region around the tremor frequency.

Balanced Random Forest

- min_sample_split: determines the minimum number of samples required to slit an internal node of each decision tree in the BRF.
- min_sample_leaf: determines the number of samples required to be at a leaf node of each decision tree in the BRF.

Default values of the other hyperparameters were used. All hyperparameters and their explanation can be seen in the *Imblearn* library documentation [23].

Logistic regression

- C: it is the inverse of the regularization strength. Smaller values of *C* implies in stronger regularization. Regularization is related to lowering the variance of the model (increase in generalization power) by introducing a bias to it.
- penalty: term introduced to the loss function of the model. L1 or L2 regularization.

The solver selected for the LR classifier was the *liblinear*. The *class_weight* hyperparameter was set as 'balanced'. Default values were used for the remaining hyperparameters, according to the *Scikitlearn* library documentation [24].

Voting Ensemble

No hyperparameters were varied for the ENS classifier. The voting strategy selected was 'soft'. The classifier average the probabilities predicted by the LR and BRF to yield its final prediction. More information can be found in the *Scikitlearn* library documentation [24].

Table 9. Classifiers hyperparameter that are optimized in the inner-loop of the a nested cross-validation procedure.

Classifier	Hyperparameter	Values
Balanced Random Forest	min_sample_split	[2, 4]
Balanced Random Forest	min_sample_leaf	[1, 2]
Logistic regression	C	[0.01, 0.1, 1]
Logistic regression	penalty	['l1', 'l2']

E. Extra results

This section is divided in three parts: scores, predicted probabilities, and grid-search CV results. The results of all pipelines and for all classification scenarios are shown.

Scores

The Brier score and ROC-AUC achieved by all pipelines in each one of the classification scenarios are presented in Table 10. Although the ENS pipeline showed a Brier score lower lower than the BRF pipeline for the PD x no-PD case, the latter reached a higher ROC-AUC and a lower standard deviation of its Brier score. This suggests that the BRF pipeline has a higher class discrimination power. In addition, the BRF pipeline presented a higher stability in the selection of the group of motor tasks during grid-search CV compared to the ENS pipeline, consistently selecting the same group in 23 out of 28 folds, against 18 for the latter 14. In the other two diagnostic case, the BRF pipeline obtained the same Brier score as the ENS, but with higher ROC-AUC. Therefore, the BRF pipeline was considered the best for all classification scenarios.

Table 10. Pipelines' scores. The Brier score and ROC-AUC represents the estimated generalization performance.

Results	LR	BRF	ENS
PD x no-PD			
Brier score	0.229 (0.319)	0.188 (0.128)	0.180 (0.195)
ROC-AUC	0.63	0.84	0.81
ET x no-ET			
Brier score	0.330 (0.303)	0.235 (0.156)	0.235 (0.156)
ROC-AUC	0.40	0.70	0.58
DT x no-DT			
Brier score	0.333 (0.176)	0.230 (0.061)	0.230 (0.061)
ROC-AUC	0.32	0.69	0.36

Predicted probabilities

The assigned probabilities for the positive classes in the PD x no-PD, ET x no-ET, and DT x no-DT scenarios are shown in tables 11, 12, and 13, respectively. We can note the higher tendency of the LR pipeline in predicting probabilities closer to 100 and 0% than the BRF pipeline, as discussed in [43].

Overall, the DT x no-DT case yielded the lowest scores. It is interesting to note that for this scenario, the LR pipeline assigned probabilities of 50 or close to 50% for several patients, reassuring the difficulty in differentiating DT and ET patients [5].

Table 11. Predicted class probabilities of the 28 patients by all pipelines - PD x no-PD.

Patient ID	Diagnosis	Predicted PD Probability (%)		
		LR	BRF	ENS
T002	no-PD	15	41	20
T004	PD	2	62	31
T005	no-PD	73	33	65
T006	no-PD	42	60	59

Continued on next column

Continued Table 11

Patient ID	Diagnosis	Predicted PD Probability (%)		
		LR	BRF	ENS
T008	no-PD	100	68	80
T010	no-PD	2	54	27
T012	PD	23	60	42
T013	PD	39	48	43
T014	PD	87	42	71
T019	no-PD	19	39	22
T020	PD	58	68	71
T022	PD	38	66	72
T023	no-PD	24	48	51
T024	PD	8	74	30
T026	no-PD	8	29	15
T027	no-PD	2	17	14
T028	no-PD	42	30	38
T029	no-PD	6	35	24
T030	no-PD	37	35	22
T031	no-PD	26	35	21
T032	no-PD	88	75	81
T036	no-PD	10	36	27
T039	no-PD	0	28	15
T042	no-PD	8	57	28
T045	no-PD	0	22	11
T046	no-PD	6	57	26
T050	no-PD	4	33	20
T052	no-PD	16	39	20

Table 12. Predicted class probabilities of the 28 patients by the all pipelines - ET x no-ET.

Patient ID	Diagnosis	Predicted ET Probability (%)		
		LR	BRF	ENS
T002	no-ET	50	38	37
T004	no-ET	68	59	61
T005	ET	20	71	34
T006	ET	4	17	10
T008	ET	0	23	10
T010	no-ET	53	51	64
T012	no-ET	27	41	22
T013	no-ET	41	38	24
T014	no-ET	0	33	21
T019	no-ET	71	66	70
T020	no-ET	42	52	26
T022	no-ET	44	30	45
T023	ET	68	64	66
T024	no-ET	0	30	18
T026	ET	50	73	83
T027	ET	91	66	79
T028	ET	50	68	75
T029	no-ET	53	54	74

Continued on next column

Continued Table 12

Patient ID	Diagnosis	Predicted ET Probability (%)		
		LR	BRF	ENS
T030	no-ET	49	54	46
T031	no-ET	73	60	76
T032	no-ET	32	39	16
T036	no-ET	49	64	44
T039	no-ET	0	59	57
T042	no-ET	26	29	15
T045	ET	1	48	24
T046	no-ET	25	49	26
T050	ET	14	75	44
T052	ET	16	52	61

Table 13. Predicted class probabilities of the 28 patients by the all pipelines - DT x no-DT.

Patient ID	Diagnosis	Predicted DT Probability (%)		
		LR	BRF	ENS
T002	DT	53	53	66
T004	no-DT	50	38	56
T005	no-DT	50	49	30
T006	no-DT	50	36	38
T008	no-DT	50	44	72
T010	DT	38	45	22
T012	no-DT	50	44	35
T013	no-DT	50	41	32
T014	no-DT	50	56	77
T019	DT	41	48	24
T020	no-DT	50	47	26
T022	no-DT	50	48	37
T023	no-DT	50	49	39
T024	no-DT	50	47	47
T026	no-DT	99	50	66
T027	no-DT	82	51	71
T028	no-DT	50	45	26
T029	DT	19	62	28
T030	DT	50	66	79
T031	DT	50	47	70
T032	DT	21	52	27
T036	DT	50	49	25
T039	DT	50	52	76
T042	DT	49	56	28
T045	no-DT	57	62	37
T046	DT	50	42	26
T050	no-DT	54	51	48
T052	no-DT	57	45	75

Grid-search results

The distribution of the selected motor tasks, data transformation methods, and classifier hyperparameters are depicted in figures 14, 15, and 16 for PD x no-PD, ET x no-ET, and DT x no-DT, respectively.

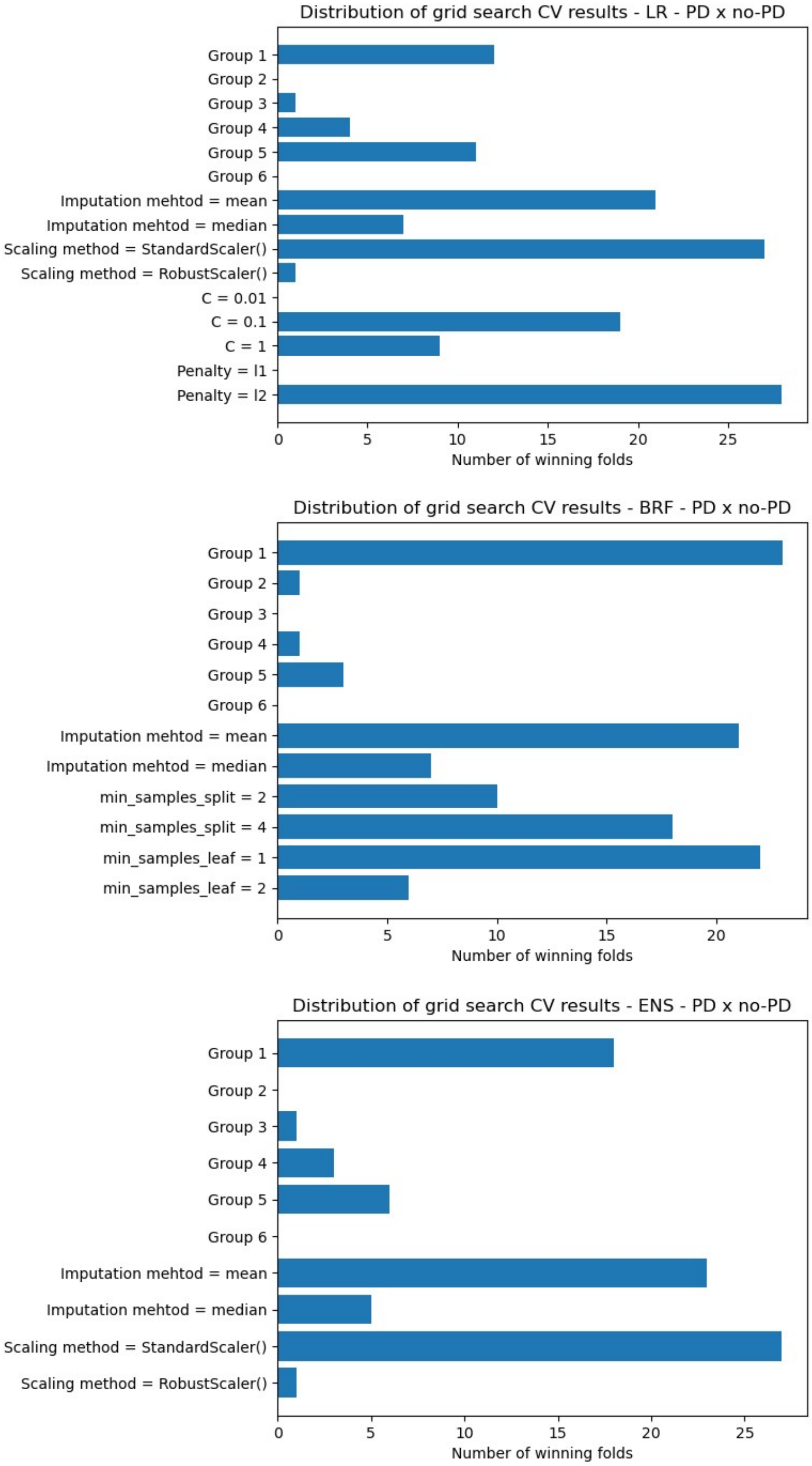

Figure 14. Selection distribution of motor tasks, data transformation methods, and hyperparameters - PD x no-PD.

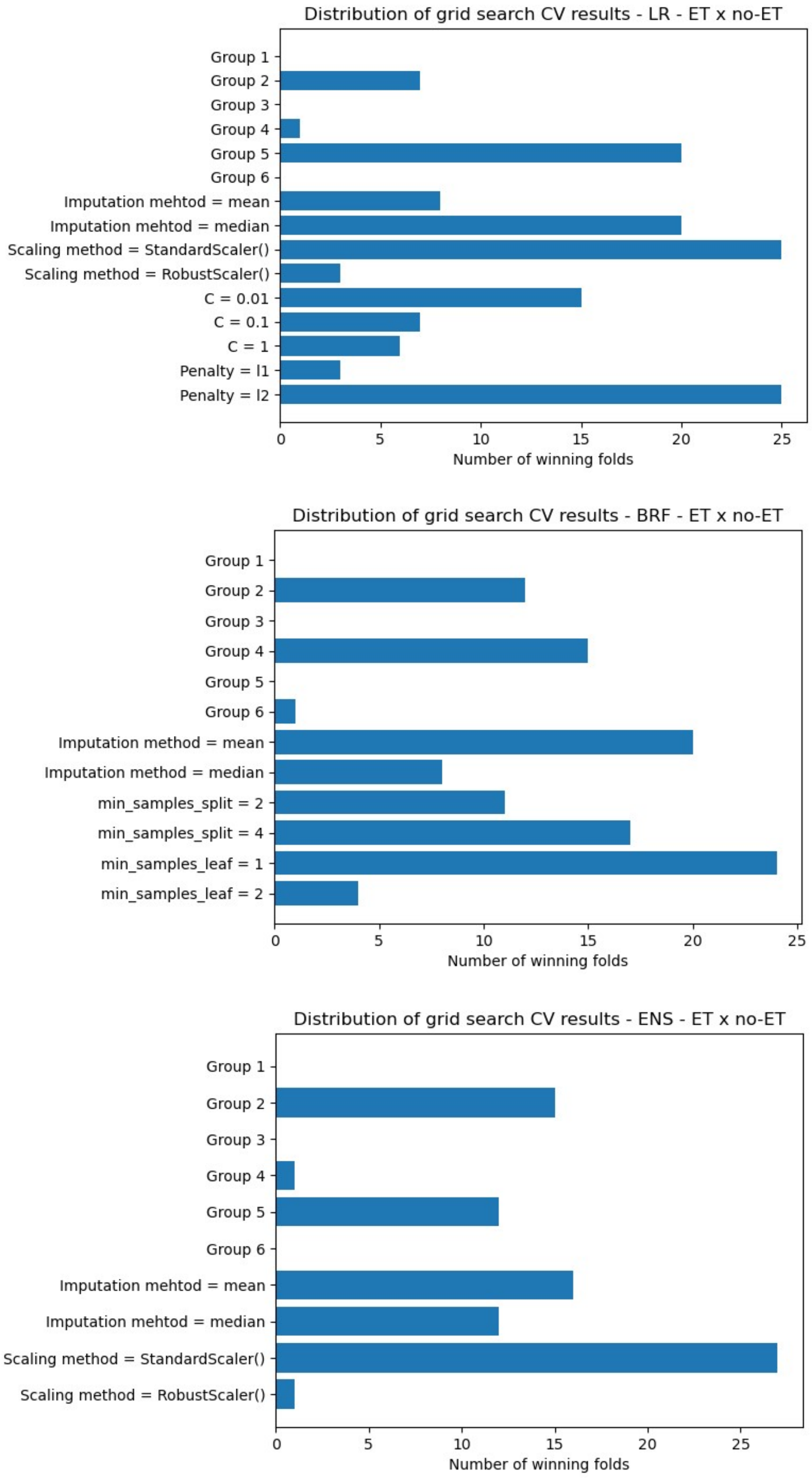

Figure 15. Selection distribution of motor tasks, data transformation methods, and hyperparameters - ET x no-ET.

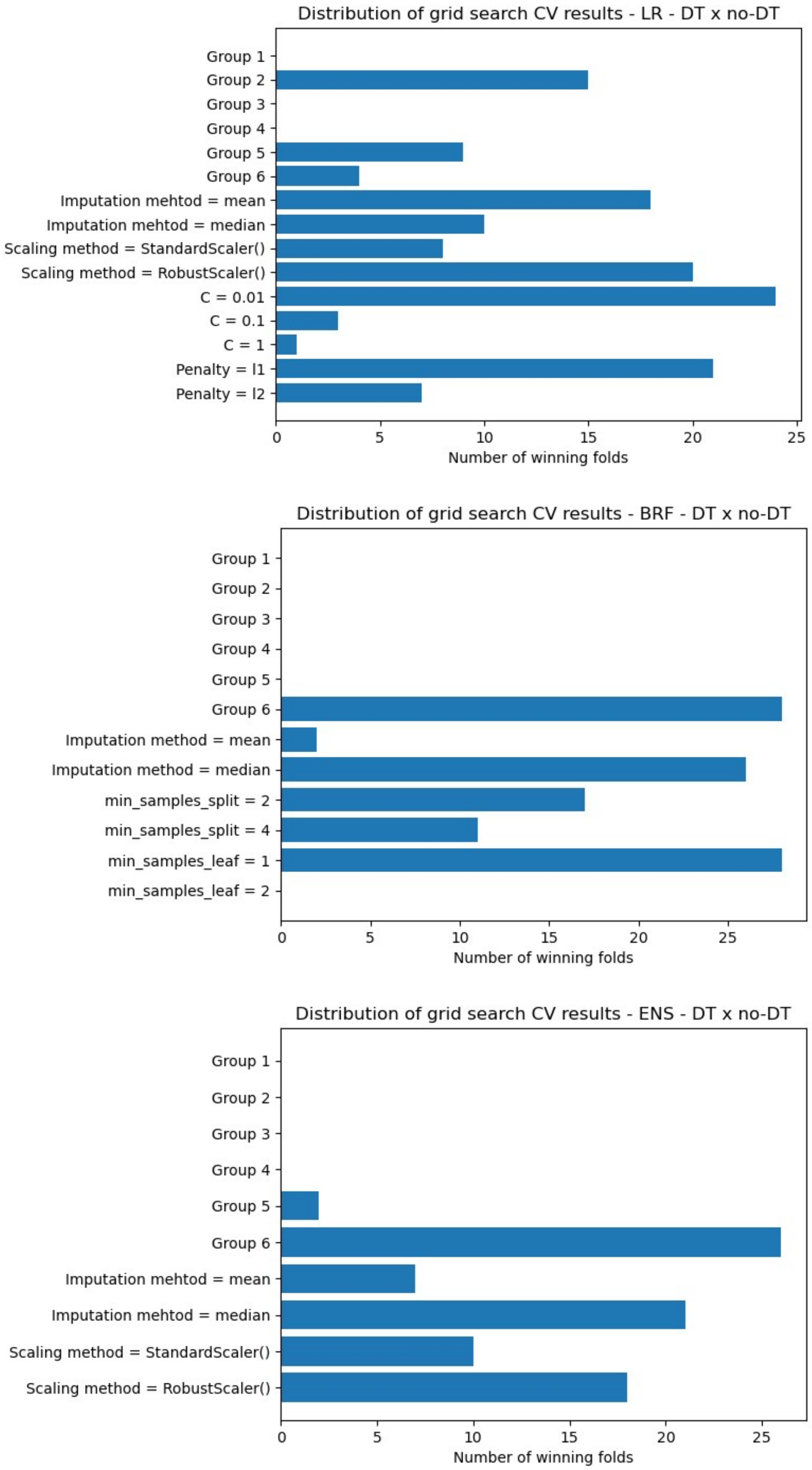

Figure 16. Selection distribution of motor tasks, data transformation methods, and hyperparameters - DT x no-DT.