Sensor-based data-driven differentiation between Parkinson's tremor and essential tremor

Tanmoy Sil, Veronika Selzam, Robert L. Peach, Gertrúd Tamás, Günther Deuschl, Sebastian R. Schreglmann, Jens Volkmann, Martin M. Reich, Muthuraman Muthuraman

Angaben zur Veröffentlichung / Publication details:

Sil, Tanmoy, Veronika Selzam, Robert L. Peach, Gertrúd Tamás, Günther Deuschl, Sebastian R. Schreglmann, Jens Volkmann, Martin M. Reich, and Muthuraman Muthuraman. 2026. "Sensor-based data-driven differentiation between Parkinson's tremor and essential tremor." Expert Systems with Applications 299 (Part D): 130336. https://doi.org/10.1016/j.eswa.2025.130336.

Expert Systems With Applications 299 (2026) 130336

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Sensor-based data-driven differentiation between Parkinson's tremor and essential tremor

Tanmoy Sil [a], Veronika Selzam [a], Robert L. Peach [a,b,*], Gertrúd Tamás [c], Günther Deuschl [d], Sebastian R. Schreglmann [a], Jens Volkmann [a], Martin M. Reich [a], Muthuraman Muthuraman [a,e]

[a] Department of Neurology, University Hospital Würzburg, Würzburg, Germany
[b] Department of Brain Sciences, Imperial College London, London, UK
[c] Department of Neurology, Semmelweis University, Budapest, Hungary
[d] Department of Neurology, UKSH, Christian-Albrechts University Kiel, Kiel, Germany
[e] Chair of the Informatics for Medical Technology, University Augsburg, Augsburg, Germany

ARTICLE INFO

Keywords:
Parkinson's disease
Essential tremor
Tremor
Clinical decision support
External validation
Electromyography (EMG)
Tremor differentiation
XGBoost

ABSTRACT

Despite distinct pathophysiologies, tremor due to Parkinson's disease and essential tremor are commonly misdiagnosed due to overlapping symptoms, hindering clinical intervention. Digital phenotyping with machine learning applied to peripheral sensor-based signals has shown potential. However, it lacks validation using an independent dataset or commercial medical device recordings broadly available for routine clinical use. Here, we present a scalable and generalizable diagnostic approach using features engineered from frequency, power, and non-linear phase domains combined with tree-based classification algorithms. Using data from a single hospital, our XGBoost model achieved a diagnostic specificity of 0.92, recall of 1.0, and ROC-AUC (Receiver Operating Characteristic – Area Under the Curve) of 1.0 on a held-out test set. Applying our pre-trained model to an external dataset, recorded in a different clinical setting using other devices, demonstrated strong generalizability with specificity of 0.70, recall of 0.80, and ROC-AUC of 0.79. Our findings not only outperform previous studies in predictive accuracy but also deliver clinically validated metrics that can be implemented in practice. These results highlight the feasibility of deploying affordable, widely available sensor-based diagnostics to enhance clinical accuracy and inform adaptive therapeutic interventions.

1. Introduction

Tremor due to Parkinson's disease (PD) and essential tremor (ET) is the most common adult-onset movement disorder: the prevalence for PD ranges from 1 to 4 % in the population > 60 years of age, while, for ET, the prevalence is reported at 5.8 % above the age of 65 and increases by a factor of 1.74 with every decade (Ascherio & Schwarzschild, 2016; Louis & McCreary, 2021; Tysnes & Storstein, 2017). Clinical differentiation between PD and ET is particularly challenging as symptoms overlap with 20–25 % of subjects in ET and PD lacking rest tremor (Cohen et al., 2003; Elble & Koller, 1990; Helmich et al., 2013; Jankovic, 2008), leading to misdiagnoses in 25–50 % of cases (Jain et al., 2006; Jo et al., 2025; Louis et al., 2015; Schrag et al., 2000; Shahed & Jankovic, 2007; Thenganatt & Louis, 2012; Tolosa et al., 2006), underscoring the

Abbreviations: ACC, Accelerometer; ANN, Artificial Neural Network; AUC, Area Under the Curve; CI, Confidence Intervals; CWT, Continuous Wavelet Transform; DaTScan, Dopamine Transporter Scan; DBS, Deep Brain Stimulation; EEG, Electroencephalography; EMG, Electromyography; ET, Essential Tremor; FTM, Fahn-Tolosa-Marin (tremor rating scale); MDS-UPDRS III, Movement Disorder Society-Unified Parkinson's Disease Rating Scale Part III; PCA, Principal Component Analysis; PD, Parkinson's Disease; RFECV, Recursive Feature Elimination with Cross-Validation; ROC-AUC, Receiver Operating Characteristic – Area Under the Curve; SEM, Standard Error of Mean; SHAP, SHapley Additive exPlanations; STD, Standard Deviation; SVM, Support Vector Machine; TSI, Tremor Stability Index; UPDRS, Unified Parkinson's Disease Rating Scale.

* Corresponding author at: Department of Neurology, University Hospital Würzburg, Würzburg, Germany.
E-mail addresses: sil_t@ukw.de (T. Sil), veronika.selzam@stud-mail.uni-wuerzburg.de (V. Selzam), Peach_R@ukw.de (R.L. Peach), tamas.gertrud@semmelweis.hu (G. Tamás), g.deuschl@neurologie.uni-kiel.de (G. Deuschl), schreglman_s@ukw.de (S.R. Schreglmann), volkmann_j@ukw.de (J. Volkmann), reich_m1@ukw.de (M.M. Reich), muthuraman.muthuraman@uni-a.de (M. Muthuraman).

https://doi.org/10.1016/j.eswa.2025.130336
Received 28 May 2025; Received in revised form 14 October 2025; Accepted 4 November 2025
Available online 6 November 2025

continued clinical need for objective diagnostic tools. The definition of ET as a disease has evolved over time and still presents with marked heterogeneity and unclear underlying mechanisms (Jankovic, 2002; Louis, 2009). Still, accurate differentiation is crucial for optimizing treatment strategies, as PD and ET have distinct pathophysiologies, progression patterns, and responses to therapeutic interventions (Jankovic, 2012). Thus, there is a growing need for robust, generalizable expert systems that can provide objective diagnostic support to assist clinicians in effectively and accurately discriminating between these conditions.

To address this, various diagnostic strategies have been developed, ranging from simple digital sensor-based approaches, such as electromyography (EMG), and accelerometry (ACC), to more complex and expensive methods like electroencephalography (EEG), Dopamine Transporter Scan (DaTScan), and genetic testing (Thenganatt & Louis, 2012). While additional diagnostic aids (such as DaTScan or genetic testing) can be powerful, they are often lengthy, expensive, and scarcely available (Antonini et al., 2008). Digital sensor data combined with machine learning, particularly those utilizing EMG and accelerometer, have shown promise in tremor differentiation, but the results remain inconsistent across studies. Machine learning approaches are attractive because they can integrate diverse data modalities, identify complex non-linear patterns, and produce uncertainty estimates, thereby reducing misclassification. Various machine-learning approaches for tremor differentiation have been proposed (Balachandar et al., 2022; di Biase et al., 2017; Ghassemi et al., 2016; Hossen et al., 2010; Muthuraman et al., 2011; Shahtalebi et al., 2021; Xing et al., 2022), yet conflicting findings and lack of consensus across different datasets, as seen in De et al., highlight the variability and limitations in current metrics derived from EMG and accelerometer recordings (De et al., 2023). This is in part due to sensor modality, experimental conditions and small, single-center datasets (De et al., 2023). However, a critical limitation across existing intelligent diagnostic systems is the lack of external validation across different clinical centers and recording conditions, severely limiting their practical deployment potential.

In this study, we aimed to develop and validate a robust data-driven expert system for differentiating between tremor in PD and ET using peripheral sensor data with the following primary objectives:

i. determine which sensor modality (EMG or accelerometer) provides the most distinctive features for classification.
ii. identify which experimental condition (rest, posture, or posture with weight) yields the most discriminative signal characteristics.
iii. evaluate which feature domain (frequency, power, or phase) contribute most significantly to accurate classification.
iv. assess whether our model generalizes to external datasets collected at a different clinical center with different recording equipment.
v. evaluate the feasibility of classifying states within disease conditions for adaptive device applications.

To achieve the presented objectives, we analyze a comprehensive dataset comprising 124 PD and 272 ET patients across three experimental conditions using both EMG and accelerometer recordings. We employ tree-based machine learning algorithms and other classification algorithms such as SVM, ANN, and logistic classifier, on features engineered from time-frequency domain and nonlinear phase analyses to build a classification model. We tested this model using an independent external dataset collected at a different clinical center with different recording equipment. This approach addresses a critical clinical need for objective, scalable diagnostic tools that can assist clinicians in accurate differentiation between these commonly misdiagnosed movement disorders.

2. Related work

To address the diagnostic challenge, various intelligent diagnostic strategies for PD vs ET tremor differentiation have evolved, ranging from single sensor modalities using accelerometers or EMG sensors, to hybrid strategies employing sensor fusion approaches, to rule-based systems leveraging specific neurophysiological markers, and more advanced methodologies such as Dopamine Transporter Scan (DaTScan) and genetic testing (Thenganatt & Louis, 2012). This section examines existing expert system strategies and their limitations for practical implementation, organized by methodological approach and diagnostic modality.

2.1. Sensor-based diagnostic expert systems

Accelerometer-based approaches have formed the foundation of accessible diagnostic solutions. Ai et al. (Ai et al., 2011) used empirical mode decomposition with SVM classification on 25 patients, achieving balanced metrics. Limitations include small sample size and exclusive focus on postural conditions. Loaiza Duque et al. (Duque et al., 2020) implemented smartphone gyroscope-based classification with frequency domain features on 39 patients, achieving 77.8 % accuracy with 75.7 % sensitivity and 80.0 % specificity. Performance remained insufficient for clinical decision support. Barrantes et al. (Barrantes et al., 2017) developed smartphone accelerometer-based classification using novel discriminatory features on 33 patients, achieving 84.38 % accuracy and ROC-AUC of 89.8 %. Performance gaps persisted, and smartphone-grade sensors introduced reliability concerns.

EMG-based approaches explored physiological signal analysis. Tavakkoli et al. (Tavakkoli et al., 2014) used nonlinear analysis of wrist muscle EMG signals with comprehensive feature extraction on 40 patients, achieving 95.75 % accuracy. Limitations include reporting only accuracy without sensitivity and specificity metrics, which is clinically inadequate for diagnostic applications.

2.2. Multi-modal expert systems

Multi-modal sensor fusion approaches combined accelerometer and EMG signals. Xing et al. (Xing et al., 2022) developed the largest study using accelerometer and dual EMG recordings on 398 patients with XGBoost classification, achieving AUC 0.95, 85 % accuracy, 96 % recall, and 64 % specificity. Significant imbalance between recall and specificity indicates systematic bias. Hossen and colleagues (Hossen, 2013; Hossen et al., 2010) explored sophisticated signal processing approaches, achieving 85–87.5 % accuracy with more balanced performance metrics. These systems revealed EMG flexor signals as the most reliable discriminator but demonstrated conservative diagnostic approaches that may miss some cases. While these systems show improved performance on single-center datasets, their robustness across diverse clinical environments remains largely unvalidated.

2.3. Rule-based expert systems

Rule-based diagnostic approaches have developed specialized neurophysiological markers that provide interpretable, clinically meaningful diagnostic criteria for tremor differentiation.

Tremor Stability Index (TSI) systems measured cycle-by-cycle frequency variation. di Biase et al. (di Biase et al., 2017) developed TSI-based classification on 91 subjects, achieving 95 % sensitivity, 95 % specificity, and 92 % accuracy in test cohort, but degraded to 69 % sensitivity and 85 % accuracy in validation cohort. Contradictory findings by other researchers raise robustness concerns (Balachandar et al., 2022; Luft et al., 2019; Panyakaew et al., 2020). While the Mean Harmonic Power (MHP) system analyzed harmonic frequency content. Muthuraman et al. (Muthuraman et al., 2011) used spectral harmonic analysis on 80 patients, achieving 94 % accuracy on clinically definite

cases and 91.7 % on unclear cases. The system revealed distinct oscillatory mechanisms between PD and ET but relied on specific recording protocols that may limit broader applicability.

3. Research gaps and validation limitations

Critical gaps significantly limit practical clinical utility of existing expert systems:

i. External validation and real-world deployment deficiency: None of the existing studies have demonstrated successful validation on external datasets collected outside their original clinical centers, using different recording equipment, or in diverse clinical settings beyond controlled environments. This absence of cross-institutional and real-world validation severely undermines confidence in system performance across different clinical environments and resource-constrained settings.
ii. Sample size limitations: Most studies utilized datasets with fewer than 100 subjects, with typical cohorts ranging from 25 to 81 patients. Only Xing et al. (Xing et al., 2022) achieved a substantial dataset of 398 patients, yet even this study lacked external validation.

Our model addresses these critical gaps through three key innovations: (1) the multi-modal tremor differentiation framework validated across two different clinical centers with different recording equipment and sampling rates, demonstrating generalizability; (2) systematic optimization of sensor modalities, experimental conditions, and feature domains to maximize both diagnostic accuracy and practical deployment feasibility; and (3) a robust classification approach using the largest combined dataset in the field (396 training + 40 external test patients) that proves the feasibility of multi-center diagnostic validation. This work distinguishes itself from purely theoretical approaches by prioritizing practical deployment considerations and demonstrated real-world performance over laboratory-optimized metrics.

4. Methods

4.1. Dataset

Dataset 1 – Training and validation: The core dataset (Table 1) comprised time-series data from EMG and accelerometer recordings of 124 PD and 272 ET patients, diagnosed by an experienced neurologist (GD), and acquired during routine diagnostic evaluations at the Department of Neurology, University of Kiel, Germany. Both accelerometer and EMG data were collected under identical conditions across all subjects. Accelerometry was performed using two 2 g monoaxial accelerometers placed on the dorsum of each hand, positioned over the midsection of the third metacarpal bone. EMG signals were recorded using four sensors, with electrodes placed on the flexor carpi ulnaris (FCU) and extensor carpi ulnaris (ECU) muscles of each hand. Signals were sampled at 800 Hz, providing 30-second recordings (24000 time points per time series).

Dataset 2 – External testing: We performed external testing using a separate dataset (Table 2) of 20 PD and 20 ET patients collected using a commercial device (Kinesia, Great Lakes NeuroTechnologies, Inc., Cleveland, OH) as explained in Muthuraman et al. (Muthuraman et al., 2021). Similar to our training dataset, the diagnoses were performed by an experienced neurologist (GT). Signals were sampled at 128 Hz, and the recordings were performed as a part of the diagnostic review by the Department of Neurology, Semmelweis University. The only difference with Dataset 1, which used a monoaxial accelerometer, was the use of a triaxial accelerometer which was mitigated by using the vector norm of the three (3) axes.

4.2. Dataset conditions

Three experimental conditions were employed to capture different tremor types:

Condition 1: Rest — Subjects were seated comfortably with their forearms supported by arm rests in an armchair, designed to capture rest-tremor.

Condition 2: Posture — Subjects extended their arms in front of their chest in a 0° neutral position to record postural-tremor.

Condition 3: Posture + Weight — Subjects extended their arms in front of their chest in a 0° neutral position, while 1000-gram weight was placed on top of both hands, intended to elicit postural tremor under load.

4.3. Ethics

The training dataset study was approved by the local ethics committee at the University Hospital Schleswig Holstein, Kiel, Germany (reference number: A 143/13). The external test dataset was approved by the Regional and Institutional Committee of Science and Research Ethics, Semmelweis University (271/2013). For both datasets, the subjects signed an informed consent according to the Declaration of Helsinki.

4.4. Signal preprocessing

Minimal preprocessing was applied to the raw time-series data. For EMG signals, only full-wave rectification was performed prior to analysis (Muthuraman et al., 2010). No additional preprocessing steps such as filtering, smoothing, or trimming of signals were applied to either the EMG or accelerometer recordings.

4.5. Feature engineering

To best differentiate between ET and tremor in PD, we computed a wide set of features engineered from a continuous wavelet transform of the signals. Each feature was computed for both accelerometer signals and each of the four EMG signals respectively. At a high-level, these

Table 1
Demographic and clinical characteristics of the 124 PD and 272 ET patients included in the model training and validation dataset for the study. Values are presented as mean ± SEM with standard deviations reported separately.

Diagnosis	*Count*	*Age (Mean ± SEM, STD)*	*Score (Action/Postural Tremor subset of UPDRS III for PD,* max. *= 8; Postural Tremor subset FTM for ET,* max. *= 8) Mean ± SEM, STD*
PD	124 (51 Females)	68.7 ± 1.12, 12.48	4.94 ± 0.12, 1.29
ET	272 (110 Females)	67.22 ± 1.02, 16.74	6.79 ± 0.13, 2.07

SEM: Standard Error of Mean.
STD: Standard Deviation.

Table 2
Demographic and clinical characteristics of the 20 PD and 20 ET patients in the external test dataset. Values are presented as mean ± SEM with standard deviations reported separately.

Diagnosis	*Count*	*Age (Mean ± SEM, STD)*	*Score (MDS-UPDRS III for PD,* max. *= 132; FTM for ET,* max. *= 144;) Mean ± SEM, STD*
PD	20 (4 Females)	64.8 ± 1.8, 8.1	38 ± 4.46, 20
ET	20 (12 Females)	72.5 ± 2.47, 11.1	45.6 ± 2.95, 13.2

SEM: Standard Error of Mean.
STD: Standard Deviation.

features emanate from:

i. Distance correlation between fundamental frequency and the second harmonic for both accelerometer and EMG.
ii. Distance correlation between fundamental frequency and the third harmonic for both accelerometer and EMG.
iii. Summative distance correlation for all frequencies and their harmonics between 2 Hz and 20 Hz i.e., the average distance correlation between all integer valued frequencies between 2 Hz and 20 Hz and their second harmonics.
iv. Average distance correlation for all non-harmonics between all frequencies between 2 Hz and 20 Hz. Non-harmonics are defined as the second harmonic ± 1 i.e., if 5 Hz is the fundamental frequency, non-harmonics would be defined at 5*2 + 1 = 11 Hz or 5*2–1 = 9 Hz.
v. The fundamental frequency for time series from both sensors (accelerometer and EMG) in all conditions i.e., fundamental frequency for accelerometer in both left and right hand, for EMG in both flexors and both extensors for all conditions.
vi. Power of the fundamental frequency bin for all sensors and for all conditions.
vii. Mean power across all frequency bins for all sensors and conditions.

4.6. Fundamental frequency and power using continuous wavelet transform

Continuous Wavelet Transform (CWT) with a single scale Morlet wavelet (scale width = 10) was selected for spectral analysis, primarily for its computational efficiency with our large dataset. The analysis was confined to the frequency range between 2 to 20 Hz, divided into 1 Hz frequency bins (2 Hz, 3 Hz, 4 Hz, and so on up to 20 Hz). CWT was used to compute the scalogram within this range, with the magnitude of the scalogram squared to derive the wavelet power spectrum.

For each signal, we identified the frequency bin with maximum power in the wavelet power spectrum, which was designated as the fundamental frequency (or 1st harmonic) of that specific signal. Higher harmonics were then determined as integer multiples of this fundamental frequency. For example, if the bin with maximum power was at 5 Hz, the second harmonic would be identified at 10 Hz, and the third harmonic at 15 Hz.

4.7. Phase calculation

For each frequency bin of interest, we first obtain the protophase θ by calculating the angle of the wavelet spectrum for that bin. Since these protophases do not grow uniformly in time (resulting in non-constant instantaneous frequencies, ω, i.e., $\dot{\varphi} \neq \omega$), we transform them into true phases φ following the methodology described in Kralemann et al. 2008, (Kralemann et al., 2008). This transformation is necessary because proto-phase angles derived from the wavelet spectrum are influenced by measurement conditions and do not provide observer-independent results (i.e., $\dot{\theta} = \omega + g(\theta)$, where $g(\theta)$ depends on the used embedding and observables and do not carry any physical meaning). The transformation to true phases ensures that we obtain a consistent representation of the underlying oscillatory dynamics, independent of the specific measurement technique. These true phases are then used in subsequent analyses to characterize tremor properties and relationships between different frequency components using the distance correlation metric (Székely et al., 2007).

4.8. Harmonics

Classical approaches for identifying harmonics have predominantly focused on spectral power; however, this is limited as harmonic interactions are not always regular and can be irregular as in PD (Muthuraman et al., 2011; Raethjen et al., 2009). While harmonics are defined mathematically as integer multiples of the fundamental frequency (which we directly calculate from the wavelet transform), the physiological interaction between these frequency components can provide valuable diagnostic information (Haufler & Pare, 2019; Luff et al., 2024). In our analysis, we treat the instantaneous true phases (φ) calculated from each frequency bin as time series. These phase time series exhibit variations due to tremor characteristics and measurement conditions.

To quantify the strength of phase interactions between the fundamental frequency and its harmonics, we use distance correlation (Peach et al., 2024; Székely et al., 2007). We do not use distance correlation to identify harmonics – these are determined directly from the fundamental frequency. Rather, we apply distance correlation to measure the degree of phase coupling between the fundamental frequency and its harmonics. Distance correlation (*dCor*) can detect both linear and nonlinear associations, making it well-suited for capturing complex phase relationships that may differ between PD and ET.

Distance correlation between two vectors X and Y of length n is computed as:

$$dCor(X, Y) = \frac{dCov(X,Y)}{\sqrt{dVar(X) \cdot dVar(Y)}}$$

where *dCov* represents distance covariance and *dVar* represents distance variance (Székely et al., 2007).

As an example, if we have two variables X and Y with marginal distributions P_X and P_Y and joint distribution $P_{X,Y}$:

i. if X and Y are independent, i.e., $P_{X,Y} - P_X.P_Y = 0$, $dCor(X, Y) = 0$,
ii. if X and Y are dependent, i.e., $P_{X,Y} - P_X.P_Y > 0$, $dCor(X, Y) > 0$, and
iii. if X and Y are perfectly dependent, $dCor(X, Y) = 1$.

For our analysis, X and Y represent instantaneous phase time series (extracted by Hilbert Transform) of length 24,000 (800 Hz × 30 s) for the training dataset and 3,840 (128 Hz × 30 s) for the external test dataset. Distance correlation measure was performed using the algorithm provided in Chaudhuri and Hu, 2019 (Chaudhuri & Hu, 2019).

4.9. Classification

For the initial classification task, we employed and compared the performance of several tree-based machine learning models. The model with the best performance, as determined by predefined metrics, was selected for further analysis. Tree-based classifiers are often favored for their ability to handle non-linear data, their automatic feature importance assessment, robustness to outliers, insensitivity to data distribution, and computational efficiency. Tree-based models can also be used in ensembles, which not only reduce overfitting but also require minimal feature engineering. Bagging ensembles build multiple models independently with different subsets of training data i.e., each model receives different subsets of the entire dataset. Sampling in bagging ensembles is done with replacement. Gradient boosting algorithms build models sequentially and every new iteration is used to correct errors made by the previous model. Bagging ensemble framework based classifiers included: (a) Random Forest classifier (Hastie et al., 2001), (b) Extra Trees Regressor (Geurts et al., 2006). Gradient Boosting framework algorithm included: (a) XGBoost (Chen & Guestrin, 2016), (b) LightGBM (Ke et al., 2017).

4.10. Training and evaluation

The training dataset (Dataset 1) was divided into an 80–20 split for training and validation, respectively. Due to the imbalance between PD (n = 124) and ET (n = 272) participants, we implemented stratified sampling to maintain the same class distribution in both training and

test sets. This ensured that each fold had a representative distribution of both classes, preventing potential bias in model evaluation.

We implemented a nested cross-validation approach to avoid data leakage and potential overfitting. All models were trained using an outer 10-fold cross-validation strategy for performance evaluation, with an inner 5-fold cross-validation for feature selection and hyperparameter tuning. The total initial feature set comprised 126 features derived from both sensor modalities across all experimental conditions.

To refine our feature set, we applied Recursive Feature Elimination with 10-fold cross-validation (RFECV). In k-fold cross validation, the training data set is split into k-subsets (folds), where the model is trained on k-1 folds and validated on the k^{th} fold. RFECV is an automated feature selection tool that is particularly valuable given our large number of features, as it helps us retain only the most relevant features for classification. RFECV was wrapped around the classifiers to compute feature importance and independently guide feature selection.

In addition to RFECV, we employed Bayesian Optimization (Nogueira, 2014) for hyperparameter tuning, using a 5 fold, 5 repeat cross validation setup (yielding 25 evaluations per model) within the training data to fine-tune our model for better performance. Bayesian Optimization was initialized with 10 random points and executed over 50 optimization rounds, providing a more efficient search for optimal hyperparameters compared to traditional grid search methods. This approach allowed for fine-tuning of the models, enhancing their performance and generalizability.

To quantify the uncertainty in our model performance estimates and assess the variability due to random data splitting, we report the 95 % confidence intervals (CI) for all performance metrics in the results section. These CIs should not be interpreted as a measure of model robustness to signal noise or data artifacts, but rather as an indication of statistical uncertainty in our performance estimates. The reported performance metrics (accuracy, precision, recall, F1, ROC-AUC, balanced accuracy, and specificity) are calculated on the test set after averaging across the 10 outer cross-validation folds.

Feature reduction was consistently applied to all modeling approaches to ensure fair comparison between different classifiers. However, for comparative purposes, we also report the performance of models without feature selection in Table 3.

4.11. Software packages used

Fundamental frequency and power using continuous wavelet transform and distance correlation measures were computed using MATLAB (Release 2023b). All classification tasks were performed using Python 3.10, and scikit-learn (Pedregosa et al., 2011), XGBoost (Chen & Guestrin, 2016), LightGBM (Ke et al., 2017) APIs.

5. Results

Our initial objective was to determine the most effective machine learning model for distinguishing between PD and ET using the complete feature set. Among the tree-based classifiers tested—XGBoost, LightGBM, Random Forests, and Extra Trees—all showed promising classification accuracy. However, XGBoost consistently outperformed the other models, exhibiting the highest metrics (Table 3). Given its superior performance, we selected XGBoost for further analyses, employing recursive feature elimination (RFECV) and Bayesian optimization for feature selection and model tuning. To identify top features, we computed SHAP values (SHapley Additive exPlanations) algorithm (Lundberg & Lee, 2017) that measure the impact that a certain feature has on the predictions (Fig. 3). Notably the top three features are all power-related features from EMG and accelerometer.

5.1. External testing

To externally test our predictive models, we sourced an independent dataset (Dataset 2) from the Department of Neurology, Semmelweis University, Budapest (Muthuraman et al., 2021). We extracted the same

Table 3
Discrimination between PD and ET using tree-based classifiers (XGBoost, LightGBM, Extra Trees, Random Forests) with different feature set combinations. Features were selected using RFECV except for one instance of XGBoost which used PCA. Optimization was performed using Bayesian methods across all postural conditions (rest, posture, and posture + weight).

Classifier	*Feature*	*Accuracy*	*Precision*	*Recall*	*F1*	*ROC AUC (95 % CI)*	*Balanced Accuracy*	*Specificity*
XGBoost (Without feature selection)	Phase, Frequency, Power	0.91	0.94	0.93	0.94	0.97 (0.93–0.99)	0.90	0.88
XGBoost (RFECV)	Phase	0.78	0.82	0.85	0.84	0.79 (0.66–0.89)	0.73	0.60
	Frequency	0.88	0.89	0.93	0.91	0.88 (0.76–0.96)	0.85	0.76
	Power	0.91	0.93	0.95	0.94	0.96 (0.89–0.99)	0.89	0.84
	Phase, Frequency, Power	0.98	0.96	1.0	0.98	1.0 (1.0 – 1.0)	0.96	0.92
XGBoost (PCA w/ 95 % variance)	Phase, Frequency, Power	0.78	0.83	0.87	0.85	0.87 (0.70–0.92)	0.74	0.60
LightGBM (Without feature selection)	Phase, Frequency, Power	0.90	0.93	0.93	0.83	0.97 (0.93–0.99)	0.88	0.84
LightGBM (RFECV)	Phase, Frequency, Power	0.85	0.88	0.91	0.89	0.94 (0.89–0.98)	0.81	0.72
Random Forest (Without feature selection)	Phase, Frequency, Power	0.93	0.93	0.96	0.95	0.96 (0.91–0.99)	0.90	0.84
Random Forest (RFECV)	Phase, Frequency, Power	0.91	0.91	0.96	0.94	0.96 (0.91 – 1.0)	0.88	0.80
Extra Trees (Without feature selection)	Phase, Frequency, Power	0.86	0.87	0.96	0.91	0.93 (0.87–0.98)	0.82	0.68
Extra Trees (RFECV)	Phase, Frequency, Power	0.94	0.93	0.98	0.96	0.97 (0.91 – 1.0)	0.91	0.84
External Validation Dataset (XGBoost with RFECV model for Inference)	Phase, Frequency, Power	0.75	0.73	0.80	0.76	0.79 (0.65–0.91)	0.75	0.70

Table 4
Discrimination between PD and ET using other classifiers (Support Vector Machine (SVM), Logistic Classifier, and artificial neural network (ANN)). Features were selected using RFECV. Optimization was performed using Bayesian methods across all postural conditions (rest, posture, and posture + weight).

Classifier	*Feature*	*Accuracy*	*Precision*	*Recall*	*F1*	*ROC AUC (95 % CI)*	*Balanced Accuracy*	*Specificity*
SVM (RFECV)	Phase, Frequency, Power	0.81	0.81	0.94	0.87	0.88 (0.79 – 0.95)	0.73	0.52
Logistic Classifier (RFECV)	Phase, Frequency, Power	0.81	0.80	0.96	0.87	0.87 (0.76–0.96)	0.72	0.48
ANN (RFECV)	Phase, Frequency, Power	0.77	0.84	0.84	0.84	0.78 (0.65–0.90)	0.74	0.64

features as Dataset 1 and using the pre-trained model from our training dataset we made out-of-sample predictions of disease. The overall specificity for the test dataset was 0.70, recall 0.80, and ROC-AUC 0.79 (Table 3).

5.2. Feature domain contribution to classification

To explicitly determine which domain of features (phase, frequency, or power) was most important for classification, we trained the XGBoost model on subsets of each feature domain using all conditions. Features associated with tremor power were best at differentiating between tremor in PD and ET (specificity: 0.84, recall: 0.95, ROC-AUC: 0.96), while frequency domain (specificity: 0.76, recall: 0.93, ROC-AUC: 0.88) and phase domain features (specificity: 0.60, recall: 0.85, ROC-AUC: 0.79) exhibited progressively lower but still informative performance (Fig. 2d).

For comparison, we also applied principal component analysis (PCA) for dimensionality reduction, retaining components that explained up to 95 % of the variance (Fig. 2a). Although PCA reduced the feature space from 126 original features to 27 principal components, it did not perform as well as RFECV, with lower classification accuracy and specificity across all models (Table 3).

5.3. Experimental condition contribution to classification

Next, we investigated which experimental condition (rest, posture, or posture + weight) provided the most discriminative features by training models using condition-specific data. Our results showed that rest-state features were the most effective in differentiating between PD and ET (Table 5), yielding the highest classification accuracy and specificity (specificity: 0.80, recall: 0.91, ROC-AUC: 0.90). Models trained on posture (specificity: 0.36, recall: 0.89, ROC-AUC: 0.69) or posture + weight data (specificity: 0.40, recall: 0.87, ROC-AUC: 0.67), while still informative, exhibited lower performance, suggesting that tremor patterns at rest offer more distinct signals for disease classification (Fig. 2c).

5.4. Sensor-specific contribution to classification

We then evaluated which sensor-specific features (EMG or accelerometer) contributed most to classification performance. The results indicated that EMG-specific features consistently outperformed accelerometer features, yielding higher recall and specificity in disease classification regardless of the feature domain (Table 6). EMG features alone achieved a specificity of 0.84, recall of 0.96, and ROC-AUC of 0.95, while accelerometer features alone reached a specificity of 0.72, recall of 0.89, and ROC-AUC of 0.90 (Fig. 2b). This suggests that EMG data captures more granular information related to tremor characteristics, particularly in the rest condition, which may be less detectable by accelerometry alone.

5.5. Motor task condition state classification

We further investigated whether we could differentiate between motor task condition states, which might be important for adaptive deep brain stimulation optimization. A binary classification between rest and posture states achieved moderate specificity for PD (0.68) and high specificity for ET (0.91), moderate to high recall (0.92 for PD, 0.87 for ET), and ROC-AUC (0.76 for PD, 0.96 for ET). Even a three-way classification between rest, posture, and weighted-posture states has reasonably good ROC-AUC (0.76 for PD, 0.89 for ET). However, the accuracy of classification between condition states was better for ET patients than PD patients (Table 7), suggesting more consistent state-dependent tremor characteristics in ET compared to PD.

5.6. Classification based on tremor stability index (TSI)

We evaluated classification performance using only the Tremor Stability Index (TSI) as input, applying the same classifiers and optimization procedures used for the multi-feature analysis. TSI, which quantifies cycle-by-cycle frequency variation, has been previously proposed for tremor differentiation (di Biase et al., 2017). As the original TSI implementation code was not publicly available, we implemented TSI based on the methodology described in the published literature.

TSI-based classification yielded substantially lower performance across all models in our dataset (Table 8). XGBoost with TSI achieved ROC-AUC 0.59 (95 % CI: 0.51–0.67), specificity 0.09, and recall 0.96. Other classifiers showed similar results: ANN (ROC-AUC 0.66, specificity 0.09, recall 0.97), SVM (ROC-AUC 0.63, specificity 0.00, recall 1.00), and logistic regression (ROC-AUC 0.55, specificity 0.00, recall 1.00). All models exhibited low specificity (0.00–0.09), contrasting with the multi-feature approach which achieved ROC-AUC 1.0, specificity 0.92, and recall 1.0 (Table 3).

6. Discussion

This study presents a robust approach for distinguishing PD from ET using peripheral sensor data, based on accelerometer and EMG signals, without the need for extensive preprocessing. By engineering features from a continuous wavelet transform of the sensor data, we achieve near-perfect classification accuracy on a held-out test set. Crucially, we validated our approach using an independent external test dataset comprising 20 PD and 20 ET patients collected at Semmelweis University using different recording equipment (Kinesia device with triaxial accelerometers sampled at 128 Hz). Despite these substantial differences in recording conditions, our pre-trained model maintained robust performance. Our metrics on internal validation surpass previous studies (Balachandar et al., 2022; Barrantes et al., 2017; Ghassemi et al., 2016; Hossen et al., 2010; Povalej Bržan et al., 2017; Shahtalebi et al., 2021; Xing et al., 2022), but the significant advancement is demonstrating generalizability through external testing – to our knowledge, the first such validation in tremor classification. Although we observed an expected performance decline during cross-institutional testing, this external validation represents a critical step toward establishing real-

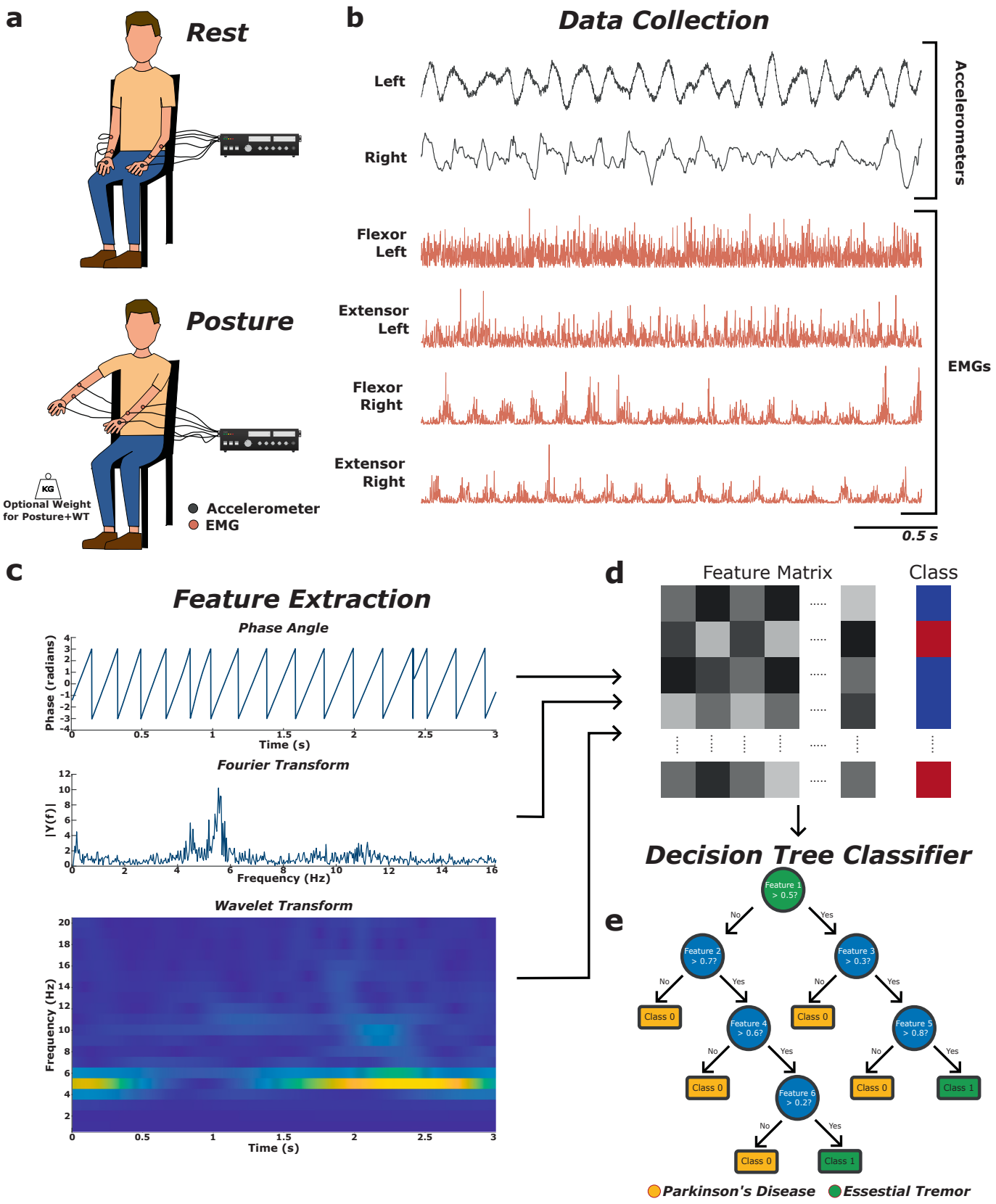

Fig. 1. Classification Model Workflow for Tremor Analysis. Workflow for the classification model using accelerometry and EMG signals from PD and ET patients. a Raw tremor accelerometry and EMG signals are collected in three states – rest, posture, posture + weight. b Exemplary accelerometer and EMG signals. c Time-series features are extracted from time and time–frequency domains. d Feature matrix created using collected features in step c. e A model is trained using tree-based classification algorithms.

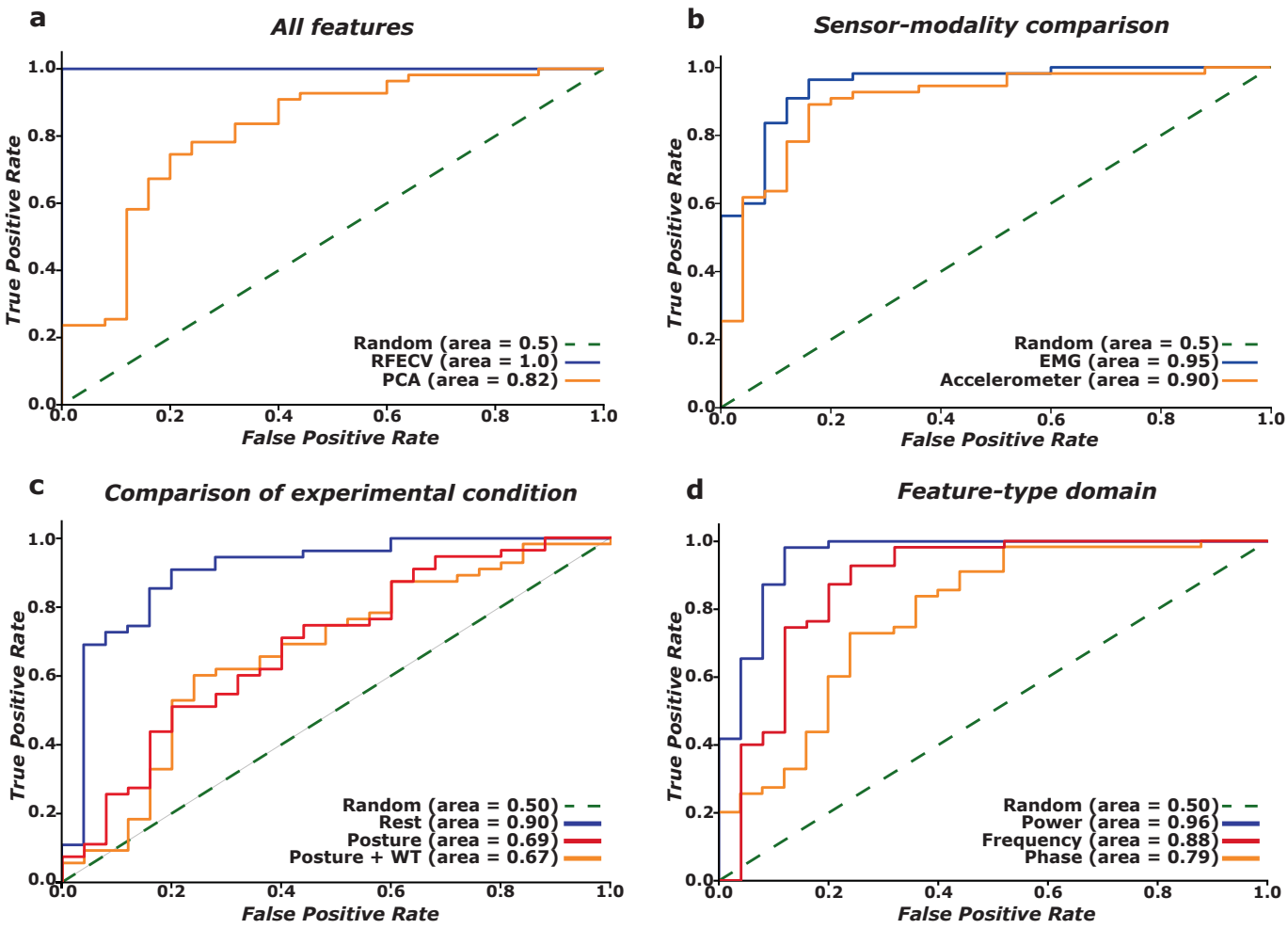

Fig. 2. ROC Curve Analysis of Classification Performance Factors. a-d. Receiver operating characteristics curve showing classification ratio between true positive rate (sensitivity) and false positive rate (1-specificity) for a. the choice of feature reduction approach, b. the sensor modalities (EMG, accelerometer), c. experimental conditions (rest, posture, posture + weight), d. the feature domain (power-, frequency-, phase-based features).

world clinical utility of sensor-based tremor classification.

Our study systematically compared multiple machine learning approaches, including tree-based models (XGBoost, LightGBM, Random Forests, Extra Trees), support vector machines, logistic regression, and artificial neural networks (Tables 3 and 4). Tree-based models, particularly XGBoost, consistently outperformed alternative approaches. This superior performance, combined with several practical advantages, guided our methodological choice. Tree-based models offer critical benefits for clinical deployment. First, they achieve strong performance with moderate-sized datasets (our 396 training participants), whereas deep learning approaches typically require thousands of samples to train robust networks. Prior studies in medical imaging have shown that deep learning models often require much larger sample sizes for stable generalization (Dhar et al., 2023; Wang et al., 2021; Woerner et al., 2025). Second, interpretability is crucial for clinical adoption and regulatory approval. Our tree-based approach enables transparent feature importance analysis (Fig. 3), allowing clinicians to understand which physiological characteristics such as EMG power at rest or fundamental frequency stability—drive specific classifications. This interpretability facilitates clinical trust. In contrast, deep learning models function as "black boxes," providing predictions without mechanistic insight. Third, computational efficiency enables broader deployment. Tree-based models require only milliseconds for inference on standard computing hardware, including smartphones, without requiring specialized GPU resources. This accessibility is particularly important for implementation in resource-constrained clinical settings and low- and middle-income countries where tremor disorders remain underdiagnosed. Deep learning models, conversely, demand substantial computational resources for both training and deployment. Finally, our feature engineering approach directly incorporates established domain knowledge about tremor physiology, including harmonic relationships between fundamental and higher frequencies, phase coupling patterns, and power spectral characteristics. Deep learning would need to learn these complex physiological patterns from scratch using raw spectrograms or time-series inputs, requiring substantially more data and computational resources.

We also compared our multi-feature approach with the Tremor Stability Index (TSI), an established single-metric method for tremor differentiation (di Biase et al., 2017). TSI quantifies cycle-by-cycle frequency variation and has shown promise in prior studies. We implemented TSI based on the published methodology, as the original code was not publicly available. We acknowledge that implementation differences, along with variations in dataset characteristics and recording protocols, may contribute to performance differences across studies. The lower specificity observed with TSI suggests that, when used as a standalone feature with our implementation and dataset, it tends to classify most patients as PD. This observation is consistent with previous reports of variability in TSI performance across different cohorts and recording conditions (Luft et al., 2019; Panyakaew et al., 2020). These differences may reflect the complexity and heterogeneity of tremor manifestations in both PD and ET populations, as well as potential sensitivity to specific implementation details and preprocessing steps.

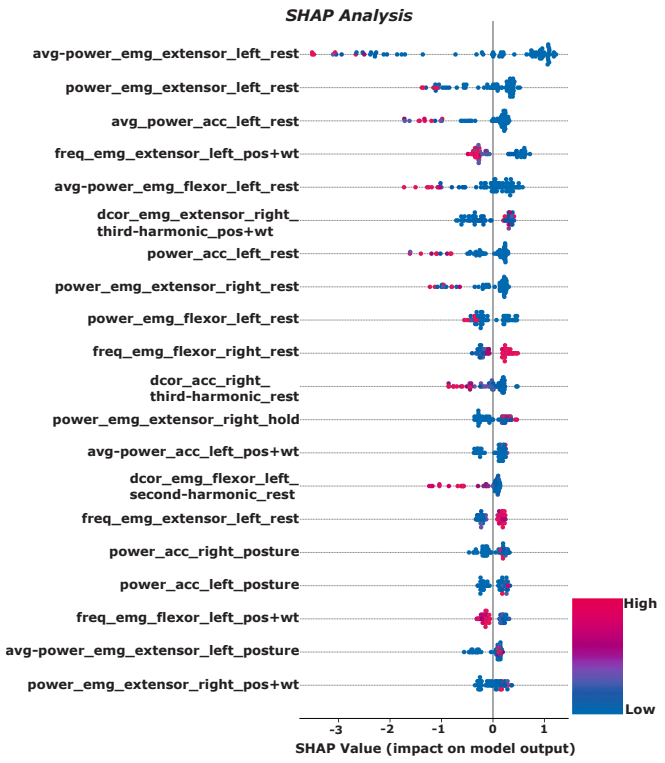

Fig. 3. SHAP Analysis of Predictive Features. This figure shows the SHAP values for the 20 most informative features. A positive SHAP value (further to the right on the graph) corresponds to a stronger prediction for the positive class (ET), whilst a negative SHAP value (further to the left on the graph) indicates a strong prediction for the negative class (PD). The color bar indicates the normalized feature value. As an example, a larger EMG power at rest (top row feature) leads to a strong prediction of PD. Abbr. acc – accelerometer, emg – electromyography, freq – frequency, dcor – distance correlation measure.

Our findings suggest that while single-metric approaches offer simplicity and clinical interpretability, integration of complementary features from multiple domains—including power spectral characteristics, frequency stability, and phase coupling patterns from both EMG and accelerometry—may provide more robust discrimination across diverse clinical settings and implementation environments. The enhanced performance of our multi-feature framework likely reflects its capacity to capture the multifaceted neurophysiological differences between these conditions. This does not diminish the conceptual value of focused metrics like TSI for understanding specific tremor characteristics but rather highlights the benefit of comprehensive feature integration for diagnostic classification tasks. Future collaborative efforts with access to original implementations could provide valuable insights into the optimal application of different methodological approaches.

Beyond methodological considerations, our findings provide insights into which sensor modalities and experimental conditions optimize tremor differentiation. Prior research has predominantly studied accelerometry due to its ease of access and low cost (Elble & McNames, 2016; Vescio et al., 2021). However, we show that features derived from EMG recordings demonstrated superior differentiability between ET and PD. This is expected for two reasons: First, EMG measures muscle activity which is more directly linked to the underlying neural activity, whilst accelerometry captures the physical movements following from muscle activity. The signal-to-noise ratio of relevant signal characteristics generated by disease will therefore be higher in EMG than accelerometer recordings, which are influenced by the weight of the person's hand as the patient maintains a posture. Secondly, EMG is independent of the axis of movement – whereas accelerometer sensors are either monoaxial that can lose information, or triaxial which requires processing to combine the separate axes. The benefits of combining EMG and accelerometry have been shown for differentiating multiple sclerosis from PD and ET (Hossen et al., 2022).

Consistent with several previous studies, our findings indicate that rest tremor is the most effective marker distinguishing PD from ET (Nisticò et al., 2011; Nisticò et al., 2022; Xing et al., 2022). Rest tremor is a primary characteristic of PD, present in approximately 90 % of PD cases, and typically subsides with movement onset. Although rest tremor is strongly indicative of PD, it can occasionally occur in ET cases, with prevalence rates reported between 20 % – 30 % in various studies (Cohen et al., 2003; Dotchin & Walker, 2008; Louis et al., 2011; Rajput et al., 2004). Consequently, upon addition of features from both postural and weighted-postural experimental conditions, we showed an improvement in classification accuracy. This finding is particularly significant as it suggests that while rest tremor provides the strongest single discriminative signal, a comprehensive assessment across multiple conditions yields the most accurate diagnosis.

The choice of features to differentiate between ET and PD is an ongoing debate (De et al., 2023). We show that traditional power related features capturing the amplitude of tremor are the strongest differentiators of disease – ET patients exhibit a lower tremor amplitude during rest but a larger tremor amplitude in postural position – agreeing with prior studies (Rahimi et al., 2011; Xing et al., 2022). Both frequency- and phase-based features show strong classification accuracy, but when combined with power-based features, the diagnosis is optimal. The importance of frequency-based and harmonic features was previously shown to differ between conditions, with irregular patterns for PD and regular patterns for ET, respectively (Hossen et al., 2010; Muthuraman et al., 2011; Raethjen et al., 2009; Sapir et al., 2003).

Our investigation into the classification of distinct motor states—specifically, differentiating between rest and postural conditions—yielded robust results. Given that rest and postural positions predominate in daily activities, the ability to discern between these states using peripheral signals as an additional feedback holds significant potential for the automatic adaptation of the deep brain stimulation (DBS) parameter space (Malekmohammadi et al., 2016; Swann et al., 2018; Yamamoto et al., 2013). Current closed-loop studies essentially close the loop with subthalamic beta band power as the feedback signal (Arlotti et al., 2018; Little et al., 2013; Nakajima et al., 2021; Velisar et al., 2019). The classification performance differed between binary vs. ternary classification and tremor diagnoses: Statistically, the addition of

Table 5
Discrimination between PD and ET using both sensors (EMG, accelerometer) based on a subset of conditions. Analysis performed with XGBoost using RFECV and Bayesian optimization.

Classifier	*Condition*	*Feature*	*Accuracy*	*Precision*	*Recall*	*F1*	*ROC AUC (95 % CI)*	*Balanced Accuracy*	*Specificity*
XGBoost (RFECV)	Rest (Condition 1)	Phase, Frequency, Power	0.88	0.91	0.91	0.91	0.90 (0.81–0.97)	0.85	0.80
	Posture (Condition 2)	Phase, Frequency, Power	0.73	0.75	0.89	0.82	0.69 (0.56–0.82)	0.63	0.36
	Posture+ Weight (Condition 3)	Phase, Frequency, Power	0.73	0.76	0.87	0.81	0.67 (0.53–0.80)	0.64	0.4

Table 6
Discrimination between PD and ET using EMG-only and Accelerometer-only features. Analysis performed with XGBoost using RFECV and Bayesian optimization.

Sensor	*Feature*	*Accuracy*	*Precision*	*Recall*	*F1*	*ROC AUC (95 % CI)*	*Balanced Accuracy*	*Specificity*
EMG	Phase	0.71	0.75	0.87	0.80	0.71 (0.57–0.84)	0.62	0.36
	Frequency	0.75	0.80	0.85	0.82	0.83 (0.73–0.92)	0.69	0.52
	Power	0.88	0.88	0.95	0.91	0.96 (0.91–0.99)	0.83	0.72
	Combined	0.93	0.93	0.96	0.95	0.95 (0.90–0.99)	0.90	0.84
Accelero-meter	Phase	0.66	0.71	0.87	0.78	0.68 (0.57–0.79)	0.54	0.20
	Frequency	0.69	0.69	1.0	0.81	0.69 (0.55–0.81)	0.5	0
	Power	0.81	0.81	0.94	0.87	0.84 (0.72–0.93)	0.73	0.52
	Combined	0.85	0.89	0.89	0.89	0.90 (0.85–0.97)	0.83	0.72

Table 7
Discrimination between states for PD and ET patients. Binary classifier distinguishes between rest and posture states, while ternary classifier distinguishes between rest, posture, and posture + weight states. All possible features from the respective states were included, with RFECV and Bayesian optimization applied.

Sensor	*States*	*Accuracy*	*Precision*	*Recall*	*F1*	*ROC AUC (95 % CI)*	*Balanced Accuracy*	*Specificity*
PD	Binary	0.80	0.74	0.92	0.82	0.89 (0.79–0.97)	0.80	0.68
	Ternary	0.56	0.56	0.56	0.56	0.76 (0.67–0.84)	0.56	0.78
ET	Binary	0.89	0.90	0.87	0.89	0.96 (0.93–0.99)	0.89	0.91
	Ternary	0.75	0.75	0.75	0.75	0.89 (0.84–0.93)	0.75	0.88

Table 8
Discrimination between PD and ET using TSI (tremor stability index) as the feature using XGBoost, artificial neural network (ANN), support vector machine (SVM), and a logistic classifier. Optimization was performed using Bayesian methods across all postural conditions (rest, posture, and posture + weight).

Classifier	*Feature*	*Accuracy*	*Precision*	*Recall*	*F1*	*ROC AUC (95 % CI)*	*Balanced Accuracy*	*Specificity*
XGBoost	TSI	0.69	0.70	0.96	0.87	0.59 (0.51–0.67)	0.52	0.09
ANN	TSI	0.70	0.70	0.97	0.82	0.66 (0.58–0.74)	0.53	0.09
SVM	TSI	0.69	0.69	1.00	0.81	0.63 (0.55–0.70)	0.50	0.00
Logistic Classifier	TSI	0.69	0.69	1.00	0.81	0.55 (0.46–0.63)	0.50	0.00

a third class to a classification system automatically increases the likelihood of false classification. Our finding of higher three-way (rest vs. posture vs posture + weight) as well as binary (rest vs. posture) classification accuracy in ET than PD indicates larger and more consistent differences between movement states in ET than PD – this is in keeping with the hypothesis of a single core oscillating driver in ET, and multiple ones in PD (Muthuraman et al., 2018; Raethjen et al., 2007; Raethjen et al., 2009). However, we note that using a continuous wavelet transform also poses some difficulties for application in live adaptive settings since a wavelet demands a width in time which cannot be extended into unknown future time points.

7. Limitations

Despite our study providing unique insights into tremor differentiation between ET and PD, several limitations require acknowledgment. First, while we compared multiple machine learning approaches (tree-based models, SVM, logistic regression, and artificial neural networks), deep learning methods were not explored. Deep learning could potentially be investigated with substantially larger multi-center datasets, though such approaches would need to address interpretability requirements for clinical acceptance and computational resource demands for deployment in resource-constrained settings. Second, performance in our external test dataset decreased by approximately 20 % (specificity: 0.92 → 0.70, recall: 1.0 → 0.80, ROC-AUC: 1.0 → 0.79). This reduction is expected when applying algorithms to data collected with different devices, clinical settings, and sampling rates (800 Hz versus 128 Hz). Standardizing recording protocols could improve generalizability. Third, we recognize PD's heterogeneity, but our study was not powered to stratify by PD subtype. Future research should investigate classification accuracy across different phenotypes. Fourth, we used monoaxial accelerometers in Dataset 1 and triaxial accelerometers in Dataset 2, which may have contributed to reduced external test performance despite vector normalization. Fifth, the lack of detailed information regarding disease stage and tremor phenotypes (e.g., pure resting tremor vs. combined resting and postural tremor in PD patients). The clinical heterogeneity of both PD and ET is well-established, and different phenotypes may exhibit varying classification accuracy. Future

multi-center studies should systematically collect these clinical characteristics to enable subgroup analyses and better understand which patient populations benefit most from sensor-based classification. Such analyses could inform personalized diagnostic strategies and identify cases where additional clinical assessment is warranted. Finally, diagnostic inconsistencies across hospitals could affect ground truth labels, highlighting the need for multi-site data collection with standardized protocols.

8. Conclusion

In summary, our findings indicate that generalized features derived from peripheral signals—specifically accelerometry and electromyography data—collected during routine diagnostic evaluations demonstrate significant potential in discriminating between PD and ET. While our machine learning approach shows promising results, we acknowledge the need for validation across more diverse datasets before clinical implementation. Nevertheless, these results suggest that quantitative tremor analysis could serve as a valuable complement to traditional medical diagnostic procedures, potentially reducing misdiagnosis rates.

This has implications for the usability of accelerometers as they are present in everyday smartphones, which could be used for preliminary screening. Additionally, specialized data gloves using both accelerometers and EMG could bring better model inference and help with clinical diagnoses. Furthermore, our ability to distinguish between different motor states (rest versus posture) holds promise for implementation in adaptive deep brain stimulation devices, which might need state information to augment stimulation parameter decision-making, potentially facilitating a more refined optimization of the programming parameter space.

Future work should focus on several key directions: (1) standardizing recording protocols across clinical centers to improve generalizability; (2) expanding external testing to additional institutions and patient populations; (3) investigating how these features perform across different PD subtypes and disease stages; and (4) exploring deep learning approaches with substantially larger multi-center datasets while addressing the interpretability and computational challenges inherent to such methods for clinical deployment.

9. Code availability

The code associated with this research is available on GitHub: https://github.com/tanmoysil/ETPD_Paper.

Funding

Tanmoy Sil was supported by a PhD fellowship of the Deutscher Akademischer Austauschdienst (DAAD). Tanmoy Sil, Robert L. Peach, Martin M. Reich, Sebastian R. Schreglmann, Jens Volkmann, and Muthuraman Muthuraman by the German Research Foundation (DFG): Project-ID 424778381-TRR 295. Sebastian R. Schreglmann is a fellow of the Thiemann foundation. Günther Deuschl receives funding from the German Research Council (SFB 1261, T1) and private foundations.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Martin M. Reich reports grant support and honoraria for speaking from Medtronic and Boston Scientific, Sebastian R. Schreglmann serves as a scientific advisor to Elemind Inc., Jens Volkmann reports grants and personal fees from Medtronic Inc., grants and personal fees from Boston Scientific, and personal fees from Abbott, Günther Deuschl has served as a consultant for Boston Scientific, Cavion, Functional Neuromodulation, and Insightec, and receives royalties from Thieme publishers, outside the submitted work.

Acknowledgements

The authors thank Ms. Ghadir Abbas for her assistance in creating Fig. 1.

Data availability

The authors do not have permission to share data.

References

Ai, L., Wang, J., & Yao, R. (2011). Classification of parkinsonian and essential tremor using empirical mode decomposition and support vector machine. *Digital Signal Processing, 21*, 543–550.

Antonini, A., Berto, P., Lopatriello, S., Tamma, F., Annemans, L., & Chambers, M. (2008). Cost-effectiveness of 123I-FP-CIT SPECT in the differential diagnosis of essential tremor and Parkinson's disease in Italy. *Movement Disorders, 23*(15), 2202–2209. https://doi.org/10.1002/mds.22278

Arlotti, M., Marceglia, S., Foffani, G., Volkmann, J., Lozano, A. M., Moro, E., Cogiamanian, F., Prenassi, M., Bocci, T., Cortese, F., Rampini, P., Barbieri, S., & Priori, A. (2018). Eight-hours adaptive deep brain stimulation in patients with Parkinson disease. *Neurology, 90*(11), e971–e976. https://doi.org/10.1212/WNL.0000000000005121

Ascherio, A., & Schwarzschild, M. A. (2016). The epidemiology of Parkinson's disease: Risk factors and prevention. *Lancet Neurology, 15*(12), 1257–1272. https://doi.org/10.1016/S1474-4422(16)30230-7

Balachandar, A., Algarni, M., Oliveira, L., Marsili, L., Merola, A., Sturchio, A., Espay, A. J., Hutchison, W. D., Balasubramaniam, A., Rudzicz, F., & Fasano, A. (2022). Are smartphones and machine learning enough to diagnose tremor? *Journal of Neurology, 269*(11), 6104–6115. https://doi.org/10.1007/s00415-022-11293-7

Barrantes, S., Sánchez Egea, A. J., González Rojas, H. A., Martí, M. J., Compta, Y., Valldeoriola, F., Simo Mezquita, E., Tolosa, E., & Valls-Solè, J. (2017). Differential diagnosis between Parkinson's disease and essential tremor using the smartphone's accelerometer. *PLoS One1, 12*(8), Article e0183843. https://doi.org/10.1371/journal.pone.0183843

Chaudhuri, A., & Hu, W. H. (2019). A fast algorithm for computing distance correlation. *Computational Statistics & Data Analysis, 135*, 15–24. https://doi.org/10.1016/j.csda.2019.01.016

Chen, T., & Guestrin, C. (2016). *XGBoost* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Cohen, O., Pullman, S., Jurewicz, E., Watner, D., & Louis, E. D. (2003). Rest Tremor in patients with Essential Tremor: Prevalence, Clinical Correlates, and Electrophysiologic Characteristics. *Archives of Neurology, 60*(3), 405–410. https://doi.org/10.1001/archneur.60.3.405

De, A., Bhatia, K. P., Volkmann, J., Peach, R., & Schreglmann, S. R. (2023). Machine Learning in Tremor Analysis: Critique and Directions. *Movement Disorders, 38*(5), 717–731. https://doi.org/10.1002/mds.29376

Dhar, T., Dey, N., Borra, S., & Sherratt, R. S. (2023). Challenges of Deep Learning in Medical image Analysis—Improving Explainability and Trust. *IEEE Transactions on Technology and Society, 4*(1), 68–75. https://doi.org/10.1109/TTS.2023.3234203

di Biase, L., Brittain, J. S., Shah, S. A., Pedrosa, D. J., Cagnan, H., Mathy, A., Chen, C. C., Martin-Rodriguez, J. F., Mir, P., Timmerman, L., Schwingenschuh, P., Bhatia, K., Di Lazzaro, V., & Brown, P. (2017). Tremor stability index: A new tool for differential diagnosis in tremor syndromes. *Brain, 140*(7), 1977–1986. https://doi.org/10.1093/brain/awx104

Dotchin, C. L., & Walker, R. W. (2008). The prevalence of essential tremor in rural northern Tanzania. *Journal of Neurology, Neurosurgery & Psychiatry, 79*(10), 1107–1109. https://doi.org/10.1136/jnnp.2007.134304

Duque, J. D. L., Egea, A. J. S., Reeb, T., Rojas, H. A. G., & González-Vargas, A. M. (2020). Angular Velocity Analysis Boosted by Machine Learning for helping in the Differential Diagnosis of Parkinson's Disease and Essential Tremor. *IEEE Access, 8*, 88866–88875.

Elble, R. J., & Koller, W. C. (1990). *Tremor*. Johns Hopkins University Press. https://books.google.de/books?id=-LdrAAAAMAAJ.

Elble, R. J., & McNames, J. (2016). Using Portable Transducers to measure Tremor Severity. *Tremor Other Hyperkinet Mov (N Y), 6*, 375. https://doi.org/10.7916/D8DR2VCC

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning, 63*(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1

Ghassemi, N. H., Marxreiter, F., Pasluosta, C. F., Kugler, P., Schlachetzki, J., Schramm, A., Eskofier, B. M., & Klucken, J. (2016). Combined accelerometer and EMG analysis to differentiate essential tremor from Parkinson's disease. *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*.

Hastie, T., Tibshirani, R., & Freidman, J. (2001). *The Elements of Statistical Learning*. Springer New York, NY. https://doi.org/10.1007/978-0-387-21606-5.

Haufler, D., & Pare, D. (2019). Detection of multiway gamma coordination reveals how frequency mixing shapes neural dynamics. *Neuron, 101*(4), 603–614 e606. https://doi.org/10.1016/j.neuron.2018.12.028

Helmich, R. C., Toni, I., Deuschl, G., & Bloem, B. R. (2013). The pathophysiology of essential tremor and Parkinson's tremor. *Current Neurology and Neuroscience Reports, 13*, 1–10.

Hossen, A. (2013). A neural network approach for feature extraction and discrimination between Parkinsonian tremor and essential tremor. *Technology and Health Care, 21* (4), 345–356. https://doi.org/10.3233/thc-130735

Hossen, A., Anwar, A. R., Koirala, N., Ding, H., Budker, D., Wickenbrock, A., Heute, U., Deuschl, G., Groppa, S., & Muthuraman, M. (2022). Machine learning aided classification of tremor in multiple sclerosis. *eBioMedicine, 82*, Article 104152. https://doi.org/10.1016/j.ebiom.2022.104152

Hossen, A., Muthuraman, M., Raethjen, J., Deuschl, G., & Heute, U. (2010). Discrimination of Parkinsonian tremor from essential tremor by implementation of a wavelet-based soft-decision technique on EMG and accelerometer signals. *Biomedical Signal Processing and Control, 5*(3), 181–188. https://doi.org/10.1016/j.bspc.2010.02.005

Jain, S., Lo, S. E., & Louis, E. D. (2006). Common misdiagnosis of a common neurological disorder: How are we misdiagnosing essential tremor? *Archives of Neurology, 63*(8), 1100–1104. https://doi.org/10.1001/archneur.63.8.1100

Jankovic, J. (2002). Essential tremor: A heterogenous disorder. *Movement Disorders, 17* (4), 638–644. https://doi.org/10.1002/mds.10221

Jankovic, J. (2008). Parkinson's disease: Clinical features and diagnosis. *Journal of Neurology, Neurosurgery & Psychiatry, 79*(4), 368–376. https://doi.org/10.1136/jnnp.2007.131045

Jankovic, J. M. (2012). *Distinguishing Essential Tremor From Parkinson's Disease*. Retrieved October 29, 2024 from https://practicalneurology.com/articles/2012-nov-dec/distinguishing-essential-tremor-from-parkinsons-disease.

Jo, S., Suh, C. H., Lee, S., Lee, J., Yoon, M., Heo, H., Shim, W. H., Kim, S. J., Kim, E. Y., & Chung, S. J. (2025). Susceptibility map-weighted MRI can distinguish tremor-dominant Parkinson's disease from essential tremor. *Scientific Reports, 15*(1), 823. https://doi.org/10.1038/s41598-024-81089-4

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). *LightGBM: a highly efficient gradient boosting decision tree* Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA.

Kralemann, B., Cimponeriu, L., Rosenblum, M., Pikovsky, A., & Mrowka, R. (2008). Phase dynamics of coupled oscillators reconstructed from data. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics, 77*(6 Pt 2), Article 066205. https://doi.org/10.1103/PhysRevE.77.066205

Little, S., Pogosyan, A., Neal, S., Zavala, B., Zrinzo, L., Hariz, M., Foltynie, T., Limousin, P., Ashkan, K., FitzGerald, J., Green, A. L., Aziz, T. Z., & Brown, P. (2013). Adaptive deep brain stimulation in advanced Parkinson disease. *Annals of Neurology, 74*(3), 449–457. https://doi.org/10.1002/ana.23951

Louis, E. D. (2009). Essential tremors: A family of neurodegenerative disorders? *Archives of Neurology, 66*(10), 1202–1208. https://doi.org/10.1001/archneurol.2009.217

Louis, E. D., Asabere, N., Agnew, A., Moskowitz, C. B., Lawton, A., Cortes, E., Faust, P. L., & Vonsattel, J.-P.-G. (2011). Rest tremor in advanced essential tremor: A post-mortem study of nine cases. *Journal of Neurology, Neurosurgery & Psychiatry, 82*(3), 261–265. https://doi.org/10.1136/jnnp.2010.215681

Louis, E. D., Hernandez, N., & Michalec, M. (2015). Prevalence and correlates of rest tremor in essential tremor: Cross-sectional survey of 831 patients across four distinct cohorts. *European Journal of Neurology, 22*(6), 927–932. https://doi.org/10.1111/ene.12683

Louis, E. D., & McCreary, M. (2021). How Common is Essential Tremor? Update on the Worldwide Prevalence of Essential Tremor. *Tremor Other Hyperkinet Mov (N Y), 11*, 28. https://doi.org/10.5334/tohm.632

Luff, C. E., Peach, R., Mallas, E. J., Rhodes, E., Laumann, F., Boyden, E. S., Sharp, D. J., Barahona, M., & Grossman, N. (2024). The neuron mixer and its impact on human brain dynamics. *Cell Reports, 43*(6), Article 114274. https://doi.org/10.1016/j.celrep.2024.114274

Luft, F., Sharifi, S., Mugge, W., Schouten, A. C., Bour, L. J., van Rootselaar, A. F., Veltink, P. H., & Heida, T. (2019). A Power Spectral Density-based Method to Detect Tremor and Tremor Intermittency in Movement Disorders. *Sensors (Basel), 19*(19), 4301. https://doi.org/10.3390/s19194301

Lundberg, S. M., & Lee, S.-I. (2017, 2017). *A Unified Approach to Interpreting Model Predictions* Advances in Neural Information Processing Systems, https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Malekmohammadi, M., Herron, J., Velisar, A., Blumenfeld, Z., Trager, M. H., Chizeck, H. J., & Brontë-Stewart, H. (2016). Kinematic Adaptive Deep Brain Stimulation for Resting Tremor in Parkinson's Disease. *Movement Disorders, 31*(3), 426–428. https://doi.org/10.1002/mds.26482

Muthuraman, M., Galka, A., Deuschl, G., Heute, U., & Raethjen, J. (2010). Dynamical correlation of non-stationary signals in time domain—A comparative study. *Biomedical Signal Processing and Control, 5*(3), 205–213.

Muthuraman, M., Hossen, A., Heute, U., Deuschl, G., & Raethjen, J. (2011). A new diagnostic test to distinguish tremulous Parkinson's disease from advanced essential tremor. *Movement Disorders, 26*(8), 1548–1552. https://doi.org/10.1002/mds.23672

Muthuraman, M., Palotai, M., Jávor-Duray, B., Kelemen, A., Koirala, N., Halász, L., Erőss, L., Fekete, G., Bognár, L., Deuschl, G., & Tamás, G. (2021). Frequency-specific network activity predicts bradykinesia severity in Parkinson's disease. *Neuroimage Clin, 32*, Article 102857. https://doi.org/10.1016/j.nicl.2021.102857

Muthuraman, M., Raethjen, J., Koirala, N., Anwar, A. R., Mideksa, K. G., Elble, R., Groppa, S., & Deuschl, G. (2018). Cerebello-cortical network fingerprints differ between essential. *Parkinson's and mimicked tremors. Brain, 141*(6), 1770–1781. https://doi.org/10.1093/brain/awy098

Nakajima, A., Shimo, Y., Fuse, A., Tokugawa, J., Hishii, M., Iwamuro, H., Umemura, A., & Hattori, N. (2021). Case Report: Chronic Adaptive Deep Brain Stimulation Personalizing Therapy based on Parkinsonian State [Case Report]. *Frontiers in Human Neuroscience, 15*. https://doi.org/10.3389/fnhum.2021.702961

Nisticò, R., Pirritano, D., Salsone, M., Novellino, F., Del Giudice, F., Morelli, M., Trotta, M., Bilotti, G., Condino, F., Cherubini, A., Valentino, P., & Quattrone, A. (2011). Synchronous pattern distinguishes resting tremor associated with essential tremor from rest tremor of Parkinson's disease. *Parkinsonism & Related Disorders, 17* (1), 30–33. https://doi.org/10.1016/j.parkreldis.2010.10.006

Nisticò, R., Quattrone, A., Crasà, M., De Maria, M., Vescio, B., & Quattrone, A. (2022). Evaluation of rest tremor in different positions in Parkinson's disease and essential tremor plus. *Neurological Sciences, 43*(6), 3621–3627. https://doi.org/10.1007/s10072-022-05885-4

Nogueira, F. (2014). *Bayesian Optimization: Open source constrained global optimization tool for Python*. https://github.com/fmfn/BayesianOptimization.

Panyakaew, P., Cho, H. J., Lee, S. W., Wu, T., & Hallett, M. (2020). The pathophysiology of dystonic tremors and comparison with essential tremor. *The Journal of Neuroscience, 40*(48), 9317–9326. https://doi.org/10.1523/jneurosci.1181-20.2020

Peach, R., Friedrich, M., Fronemann, L., Muthuraman, M., Schreglmann, S. R., Zeller, D., Schrader, C., Krauss, J. K., Schnitzler, A., Wittstock, M., Helmers, A.-K., Paschen, S., Kühn, A., Skogseid, I. M., Eisner, W., Mueller, J., Matthies, C., Reich, M., Volkmann, J., & Ip, C. W. (2024). Head movement dynamics in dystonia: A multi-centre retrospective study using visual perceptive deep learning. *npj Digital Medicine, 7*(1), 160. https://doi.org/10.1038/s41746-024-01140-6

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*(85), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html.

Povalej Bržan, P., Gallego, J. A., Romero, J. P., Glaser, V., Rocon, E., Benito-León, J., Bermejo-Pareja, F., Posada, I. J., & Holobar, A. (2017). New Perspectives for Computer-aided Discrimination of Parkinson's Disease and Essential Tremor. *Complexity, 2017*(1), Article 4327175. https://doi.org/10.1155/2017/4327175.

Raethjen, J., Govindan, R. B., Kopper, F., Muthuraman, M., & Deuschl, G. (2007). Cortical Involvement in the Generation of Essential Tremor. *Journal of Neurophysiology, 97*(5), 3219–3228. https://doi.org/10.1152/jn.00477.2006

Raethjen, J., Govindan, R. B., Muthuraman, M., Kopper, F., Volkmann, J., & Deuschl, G. (2009). Cortical correlates of the basic and first harmonic frequency of Parkinsonian tremor. *Clinical Neurophysiology, 120*(10), 1866–1872. https://doi.org/10.1016/j.clinph.2009.06.028

Rahimi, F., Bee, C., South, A., Debicki, D., & Jog, M. (2011, 30 Aug.-3 Sept. 2011). Variability of hand tremor in rest and in posture — A pilot study. 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society.

Rajput, A., Robinson, C. A., & Rajput, A. H. (2004). Essential tremor course and disability. *Neurology, 62*(6), 932–936. https://doi.org/10.1212/01.WNL.0000115145.18830.1A

Sapir, N., Karasik, R., Havlin, S., Simon, E., & Hausdorff, J. M. (2003). Detecting scaling in the period dynamics of multimodal signals: Application to Parkinsonian tremor. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics, 67*(3 Pt 1), Article 031903. https://doi.org/10.1103/PhysRevE.67.031903

Schrag, A., Munchau, A., Bhatia, K. P., Quinn, N. P., & Marsden, C. D. (2000). Essential tremor: An overdiagnosed condition? *Journal of Neurology, 247*(12), 955–959. https://doi.org/10.1007/s004150070053

Shahed, J., & Jankovic, J. (2007). Exploring the relationship between essential tremor and Parkinson's disease. *Parkinsonism & Related Disorders, 13*(2), 67–76. https://doi.org/10.1016/j.parkreldis.2006.05.033

Shahtalebi, S., Atashzar, S. F., Patel, R. V., Jog, M. S., & Mohammadi, A. (2021). A deep explainable artificial intelligent framework for neurological disorders discrimination. *Scientific Reports, 11*(1), 9630. https://doi.org/10.1038/s41598-021-88919-9

Swann, N. C., de Hemptinne, C., Thompson, M. C., Miocinovic, S., Miller, A. M., Gilron, R., Ostrem, J. L., Chizeck, H. J., & Starr, P. A. (2018). Adaptive deep brain stimulation for Parkinson's disease using motor cortex sensing. *Journal of Neural Engineering, 15*(4), Article 046006. https://doi.org/10.1088/1741-2552/aabc9b

Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist., 35*(6). https://doi.org/10.1214/009053607000000505

Tavakkoli, P., Kobravi, H. R., Shoeibi, A., & Zakhirehdari, N. (2014). Discrimination of Parkinsonian Tremor from Essential Tremor Using the Nonlinear Analysis of Wrist Muscle EMG Signals.

Thenganatt, M. A., & Louis, E. D. (2012). Distinguishing essential tremor from Parkinson's disease: Bedside tests and laboratory evaluations. *Expert Review of Neurotherapeutics, 12*(6), 687–696. https://doi.org/10.1586/ern.12.49

Tolosa, E., Wenning, G., & Poewe, W. (2006). The diagnosis of Parkinson's disease. *Lancet Neurology, 5*(1), 75–86. https://doi.org/10.1016/S1474-4422(05)70285-4

Tysnes, O. B., & Storstein, A. (2017). Epidemiology of Parkinson's disease. *Journal of Neural Transmission (Vienna), 124*(8), 901–905. https://doi.org/10.1007/s00702-017-1686-y

Velisar, A., Syrkin-Nikolau, J., Blumenfeld, Z., Trager, M. H., Afzal, M. F., Prabhakar, V., & Bronte-Stewart, H. (2019). Dual threshold neural closed loop deep brain stimulation in Parkinson disease patients. *Brain Stimulation, 12*(4), 868–876. https://doi.org/10.1016/j.brs.2019.02.020

Vescio, B., Quattrone, A., Nisticò, R., Crasà, M., & Quattrone, A. (2021). Wearable Devices for Assessment of Tremor [Mini Review]. *Frontiers in Neurology, 12*. https://doi.org/10.3389/fneur.2021.680011

Wang, S., Li, C., Wang, R., Liu, Z., Wang, M., Tan, H., Wu, Y., Liu, X., Sun, H., Yang, R., Liu, X., Chen, J., Zhou, H., Ben Ayed, I., & Zheng, H. (2021). Annotation-efficient deep learning for automatic medical image segmentation. *Nature Communications, 12* (1), 5915. https://doi.org/10.1038/s41467-021-26216-9

Woerner, S., Jaques, A., & Baumgartner, C. F. (2025). A comprehensive and easy-to-use multi-domain multi-task medical imaging meta-dataset. *Scientific Data, 12*(1), 666. https://doi.org/10.1038/s41597-025-04866-4

Xing, X., Luo, N., Li, S., Zhou, L., Song, C., & Liu, J. (2022). Identification and Classification of Parkinsonian and Essential Tremors for Diagnosis using Machine Learning Algorithms [Original Research]. *Frontiers in Neuroscience, 16*, Article 701632. https://doi.org/10.3389/fnins.2022.701632

Yamamoto, T., Katayama, Y., Ushiba, J., Yoshino, H., Obuchi, T., Kobayashi, K., Oshima, H., & Fukaya, C. (2013). On-demand Control System for Deep Brain Stimulation for Treatment of Intention Tremor. *Neuromodulation: Technology at the Neural Interface, 16*(3), 230–235. https://doi.org/10.1111/j.1525-1403.2012.00521.x