Smartphone-Based Digitized Neurological Examination Toolbox for Multi-test Neurological Abnormality Detection and Documentation

Trung-Hieu Hoang, Christopher Zallek, Minh N. Do *Fellow, IEEE*

***Abstract*— Understanding the efficacy of digital biomarkers in vision-based human motion analysis is essential, not only for interpreting the computer-aided exam results but also for advancing the next generation of digital health tool solutions. In this study, we extensively analyze digitized neurological examination (DNE) biomarkers for detecting and documenting exam features of Parkinson's disease (PD) and other neurological disorders (OD). Collected over 113 participants, *DNE-113, a multi-test DNE database* of finger tapping, finger to finger, forearm roll, stand-up and walk, and facial activation tests, covering a broader range of neurological abnormalities beyond PD is first proposed. Subsequently, DNE-113 is integrated into `pyDNE` - a convenient open-source toolbox, streamlining the creation and assessment of digital biomarkers. This toolbox empowers us to assess the quality of DNE biomarkers across diverse classification tasks. We showcase the discriminative potency of DNE biomarkers, successfully characterizing abnormal signals in neurological patients. Our findings highlight not only the potential use cases but also the persisting challenges in constructing digital biomarkers for computer-aided movement analysis on PD and OD patients.**

***Index Terms*— Digital biomarkers, Parkinson's disease, human motion analysis, human pose estimation, discriminant features analysis, machine learning.**

I. Introduction

Projecting to affect the quality of life of over 1.2 million people by 2030, Parkinson's disease (PD) [1] is one of the most common neurodegenerative disorders [2], [3]. While early diagnosis and regular check-ups with neurologists are the two ingredients to mitigate the situation and improve the patient's outcomes, the shortage of neurologists [4]–[6] creates a barrier for many people, especially individuals living in rural areas. At many clinics, the process of documenting patients' conditions after each visit creates a burden on the daily workload of clinicians [7]. Fortunately, information technology and engineering solutions in telehealth or digital clinical reports [8] are expected to address these challenges by allowing clinicians and patients to create a *digital record* of human motion from video-recorded neurological examinations [9]. The single-camera solutions, without requiring a complex setup or special equipment, have been receiving much attention recently due to their high accessibility [10], [11]. Moreover, computer-aided diagnosis can be effectively applied to videos recorded by commonly used devices like smartphones [12]–[15] or webcams [16]–[18] to empower automated analysis.

Describing the motor-control capability of PD patients, MDS-UPDRS-III, the largest part of the Unified Parkinson's Disease Rating Scales [19], assesses 27 neurological tests. Due to numerous obstacles in data collection, many studies are devoted to studying a *single neurological test* [16], [20]–[24]. Although the information from each MDS-UPDRS-III test varies across subjects, *multi-test* is still important for describing an individual's condition, especially in situations when the subject is unable to perform some tests (e.g., gait assessment on patients with high fall risk). Besides, while satisfactory results have been shown to distinguish PD patients from HC, it remains unclear whether these systems can *actually differentiate unique characteristics of PD patients, or just serve as a general neurological disorders detector*.

Towards a better understanding of these innovations in clinical care and set up a *testbed* for future development, we conduct a comprehensive *multi-test* digitized neurological examination (DNE) dataset collection, namely DNE-113. In this study, a total of 643 videos with *five* neurological examinations are recorded on a population of *113 subjects*. In contrast to earlier investigations which are primarily focused on HC and PD separation, our subject cohorts also include a group of subjects diagnosed with *other neurological disorders* (OD), besides PD, extending the coverage of a variety of abnormalities. Videos are collected using the *DNE Recorder* platform [12] installed on conventional smartphone/tablet devices.

Prospectively, the availability of DNE-113 facilitates the development of *digital biomarkers* that can extract clinically relevant features from the spatio-temporal information of human motion. We also introduce `pyDNE`, *a convenient open-source Python package* to streamline the regular workflow of designing such biomarkers. The goal is to construct a collection of multiple features and classifiers based on the clinician's examination techniques [12], [16], [25]. With high-quality biomarkers, it is sufficient for a simple machine learning (ML) model to predict the outcomes (e.g., normal versus abnormal) with comparable performance to large-scale

Trung-Hieu Hoang and Minh N. Do are with the Department of Electrical & Computer Engineering and Coordinated Science Laboratory at University of Illinois at Urbana-Champaign (UIUC), Urbana, IL, 61801, USA (e-mail: hthieu@illinois.edu, minhdo@illinois.edu). Minh N. Do is also with the VinUni-Illinois Smart Health Center, VinUniversity, Hanoi, Vietnam. Christopher Zallek is with OSF HealthCare Illinois Neurological Institute - Neurology, Peoria, IL, 61603, USA (e-mail: christopher.m.zallek@ini.org) *(Corresponding author: Trung-Hieu Hoang)*

black-box deep-learning models as shown in [12]. Besides the advantages of having a higher interpretability and requiring less computing resources, it is important to characterize the quality of each digital biomarker, and their contribution to the final prediction before utilizing ML models. To this end, our `pyDNE` provides a *unified platform* for biomarkers extraction on DNE-113, assessing the biomarkers quality, and subsequently incorporating them in classification models.

Finally, we showcase DNE-113 and `pyDNE` in action by performing a thorough validation of a set of DNE biomarkers (parts of them are described in [12]) *on real patients*. A *mixture of DNE classifiers* and a *hierarchical binary classifier* approach are introduced for handling the simultaneous availability of multiple neurological tests in the task of HC, OD, and PD classification. While purposely designed for differentiating the normal versus *simulated impairment* (SI) from HC subjects, our investigation demonstrates its satisfactory generalizability in differentiating HC and PD/OD subjects.

To sum up, this work provides a *comprehensive* study on the detection and documentation of PD patients using DNE digital biomarkers (described in **Sec.** IV) by proposing the DNE-113 (**Sec.** III) dataset and `pyDNE` toolbox (**Sec.** V-A). The quality of digital biomarkers in [12] are both assessed individually (**Sec.** VI-A) and collectively in the task of binary (**Sec.** VI-B) and multi-class (**Sec.** VI-C) classification.

II. Related Works and Context

Despite its widespread acceptance, the MDS-UPDRS-III is still subjected to physician-dependent [26]. With its guidelines, the majority of recent works have attempted to adopt an artificial intelligence (AI) model to analyze the neurological exams from the recorded videos [27]. The analysis typically aims to either distinguish PD patients from healthy controls (HC) [17], [28] or regress the MDS-UPDRS-III score [16], [29], [30]. Encouragingly, their capability can match the performance of many experienced human raters [16], [20], [31].

While the videos come from a variety of popular devices, e.g., smartphone [24], [32] or webcam [16], [17], a common pipeline is typically adopted across many vision-based automated PD assessments. Operating on the keypoints detected by an off-the-shelf human-pose estimator [33]–[35], the analysis phase can be clustered into two categories. The *data-driven*-like approach directly applies the deep-learning model to learn spatio-temporal relations of all keypoints to predict the clinical outcomes [20], [36], [37]. These approaches, however, are *lacking interpretability*, computationally expensive, or tend to overfit on small datasets which is critical for healthcare applications. With these limitations, black-box approaches cannot fully replace the roles of engineering clinical-relevant features as in other ML tasks. Meanwhile, *digital biomarkers*-based approaches focus on engineering discriminating features and adopt a simple ML model for the downstream tasks [12], [17], [23], [29]. This demonstrates the power of engineering solutions and highlights the importance of designing discriminant digital biomarkers in practice.

Furthermore, due to numerous obstacles in data collection, many studies are devoted to studying a single [16], [20]–[24], or with a limited collection of neurological tests (with a few participants, missing the HC group, or may not be publicly available [20], [30], [36]). With these shortcomings, this work provides a comprehensive data collection and toolbox for reasoning whether existing approaches can truly identify *unique PD characteristics* behind their promising results, and set a baseline for multi-test digitized neurological examination.

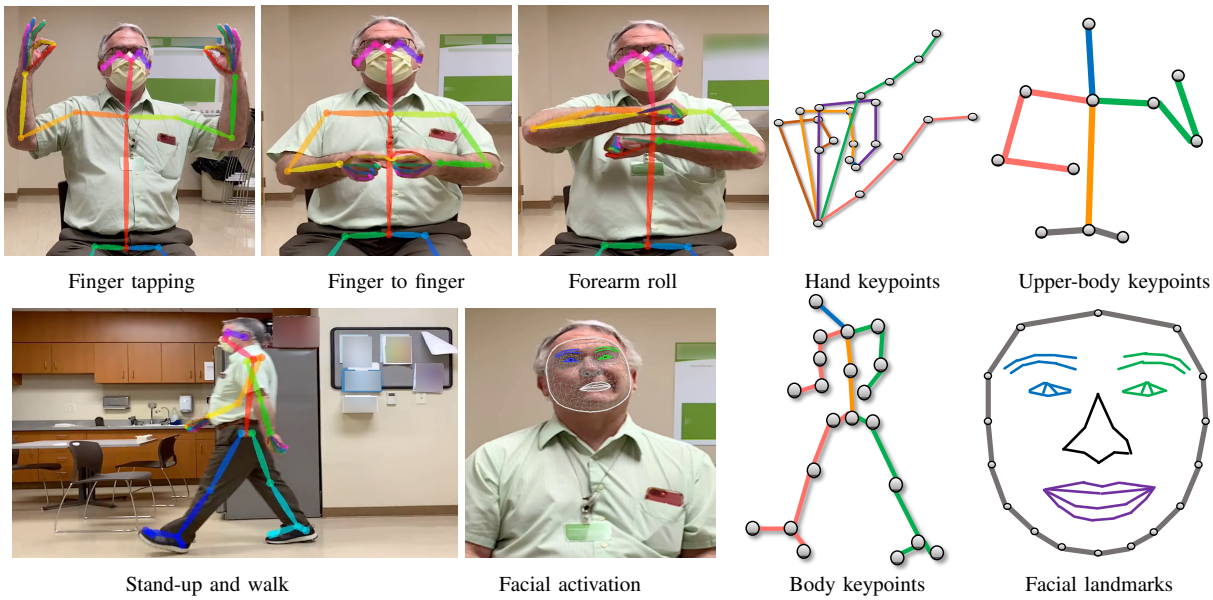
III. DNE-113 Dataset Collection

Our DNE-113 data collection protocol is IRB approved by the University of Illinois College of Medicine at Peoria Institute Review Board 1 (#IRB.1903797-9, approval date: 17/05/2022) and UICOMP Board 1 (#2110816-3, approval date: 27/10/2023). Of the 113 participants (55 males/58 females) in this study, 34 subjects are healthy control (HC), and 79 are patients with neurological disabilities. The majority of DNE-113 is collected at OSF Parkinson's Clinic, OSF Physical Therapy Clinic, Peoria Physical Therapy, Bradley Physical Therapy, and Saint Francis Inpatient Rehabilitation Center from 08/2022 to 11/2023. The duration of diagnosis ranges from several months to years (average of 2.2 ± 5.1 years).

Patients in DNE-113 were diagnosed with Parkinson's disease (PD) or at least one other neurological (OD) disorder, based on their clinical record. The OD includes subjects experiencing different conditions affecting various regions in the central/peripheral nervous system, or both. This includes prior strokes of different vasculature distributions, demyelinating disease (e.g. multiple sclerosis), muscular dystrophies, and motor neuron disease among others. We also denote the *neurological disease* (ND) category as a super-set of patients classified with neurological disorders, either PD or OD. In the HC group, 21 subjects were gathered from [12], with the remaining subjects being older individuals (age ≥ 60) in good neurological condition. The demographic of all participants grouped by cohorts is presented in **Tab.** I (Cols. 2-5).

Following the DNE protocol [12], each subject was examined with three fine-motor upper-limb tests: finger-tapping (FT), finger-to-finger (FTF), forearm roll (FR), one test of walking: stand-up and walk (SAW), and one facial activation (FA) test. The descriptions of five DNE tests investigated in this study are summarized as follows:

- **FT:** Participants are instructed to position their hands and touch each hand's thumb and index finger together. They would start tapping them as big open and close, and fast as they could for 15 seconds.
- **FTF:** Participants repetitively first point their index fingers towards the ceiling and then touch their fingers together out in front of their chests for 15 seconds.
- **FR:** Participants are asked to gently clench their hands, hold their forearms horizontally, and roll their hands around as fast as possible for 15 seconds.
- **SAW:** Participants stand from a seated position in a chair and walk back and forth 15 feet. The designated time for SAW test is 45 seconds.
- **FA:** Participants are asked to keep their heads still. They are instructed, in sequence, to lift their eyebrows as high as they can, make a big smile, relax their face, and then again lift their eyebrows as high as they can, make another big smile, and relax their face.



(a) Examples of DNE tests **(b)** Skeleton trees for motion analysis

Fig. 1. **DNE-113 dataset collection**. A complete digitized neurological examination (DNE) *record* contains five neurological exams: finger tapping, finger-to-finger, forearm roll, stand-up and walk, and facial activation. **(a)** Examples of RGB videos and human pose/facial landmark estimation results. **(b)** The skeleton trees of major keypoints and facial landmarks used for motion analysis. The hand and body keypoints are obtained with OpenPose [33], [38] while MediaPipe [34] provides facial landmarks. The authors have obtained consent to use the images in the publication.

TABLE I

DEMOGRAPHIC CHARACTERISTICS OF THE PARTICIPANTS AND THE STATISTICS OF OUR DNE-113 DATASET. DNE-113 CONTAINS 147 DNE RECORDS OF 113 SUBJECTS, PERFORMING FIVE NEUROLOGICAL EXAMINATIONS, INCLUDING THE FINGER-TAPPING (FT), FINGER-TO-FINGER (FTF), FOREARM-ROLL (FR), STAND-UP AND WALK (SAW), AND FACIAL ACTIVATION (FA). THE DIVERSITY OF DNE-113 IS EVIDENT IN PARTICIPANT COHORTS, THE NUMBER OF NEUROLOGICAL EXAMINATIONS, AND THE VOLUME OF RECORDS.

Cohort	Demographic				Number of	Number of Videos per Test				
	Size	Age	Male/Female	Disease Duration	Records	FT	FTF	FR	SAW	FA
Healthy Control (HC)	34	54.9 ± 13.6	8/26	-	65	52	53	60	32	53
Parkinson's Disease (PD)	33	73.0 ± 9.1	22/11	1.6 ± 3.5	35	35	35	35	30	35
Other Diseases (OD)	46	64.5 ± 14.0	25/21	2.6 ± 6.0	47	45	46	46	39	47
Total	113	64.1 ± 14.3	55/58	2.2 ± 5.1	147	132	134	141	101	135

Examples of DNE tests are provided in **Fig.** 1. Two subjects had a follow-up recording after several months. A collection of multiple DNE tests forms an *DNE record* with the total number provided in **Tab.** I (Col. 6). Note that not all records are complete e.g., for 36 records, the SAW test is unavailable due to the high fall risk of the participants (see **Appendix.** A for additional details). Some of the videos are removed due to the poor quality. **Tab.** I (Cols. 7-11) summarizes the statistics of the recorded videos. In total, a collection of 147 DNE records with 643 videos, from 113 subjects over all five tests forms the DNE-113 dataset. The pose estimation results are made publicly available for future development.

IV. DNE BIOMARKERS EXTRACTION

Regardless of the neurological test, DNE recordings follow a common data collection and processing pipeline for extracting DNE biomarkers. We briefly describe the four main steps.

A. DNE Record Acquisition

For all tests except SAW, the patient remained seated. A tripod with the iPhone/iPad was positioned with the subject in the field of view. The investigator briefly described and demonstrated the action to be performed for recording. The investigator used the *DNE Recorder* [12] app to record each exam individually while the investigator and *the app gave audio or visual instruction cues*. Participants are free to wear their everyday clothing or utilize walking aids as required. For SAW, the patient was recorded in an area to allow a sagittal view perspective of the standing and steps taken by the subject. The same system can be conveniently adopted in the at-home setting, either in self-recording or with support from a family member. Videos are recorded at 60 FPS with resolutions 1080×720 and uploaded to cloud storage for processing.

B. Automated Human Keypoints Detection.

Three off-the-shelf pose estimation solutions are adopted for our video analysis. For FT, FTF, FR, and SAW, we use OpenPose [33], [38] to estimate hand (21 keypoints) and body pose (25 keypoints). While the upper-limb movements are mostly located in parallel to the camera plane, this does not hold in the SAW test. Hence, for SAW, VideoPose3D [35] is used in addition to estimating 3D human pose location from detected 2D keypoints. For the FA test, MediaPipe [34] provides the prediction for 478 face landmark positions.

C. Pre-processing.

Due to several circumstances, the pose estimation results in a small number of frames are erroneous. Thus, a combination of median and Savitzky-Golay filtering [39] is also adopted for smoothing the signal filling in the missing values. Recordings are then truncated to segments that the subject performing the tests. Before further processing, the coordinates of all keypoints are normalized by a reference length. We use the median value of the length of the forearm for the FR, and FT test, the shoulder for the FTF, the trunk for the SAW, and the intercanthal distance for the FA test.

D. DNE Biomarkers Extraction

Upper-limb and SAW Tests. We adopt a full set of DNE biomarkers introduced in [12] for describing the upper-limb (FT, FR, FTF) and SAW tests. An incomplete list of some important biomarkers for each test is mentioned in **Tab.** II (with the full set of biomarkers provided in **Appendix.** B). In summary, quantifying the movement symmetry - left versus right (L/R) side, and the consistency across cycles for periodic movements are the most important. Therefore, the DNE biomarkers mainly focus on these aspects: L/R movement symmetry, angle, and the similarity to the trajectories in multiple repetitions. Additionally, observing the changes in keypoint positions also allows us to compute the amplitude, period, velocity, and acceleration of major joint locations. The discrete derivatives are used to compute speed and acceleration, local extrema are used to detect the starting and ending of each cycle. To avoid redundancy, [12] provides the full descriptions of all DNE biomarkers.

Facial Activation Test. Inspired by [40], we adopt a similar set of digital biomarkers for quantitatively characterizing the FA test. The feature construction is based on the distances between several groups of facial landmarks (detected by MediaPipe [34]). At each video frame, DNE performs in total a collection of 8 measurements, focusing on the opening of the mouth, eyes, and lifting of the eyebrows while the subject performs the test. From those measurements, 18 DNE biomarkers for the FA test are derived. These biomarkers can be classified into three primary categories. The first group calculates the standard deviation of both the measured distance and the corresponding instant speed for capturing the shuttle motions in the facial muscles. The second set of features evaluates the left and right symmetry of subject movement by computing the Pearson correlation coefficient between the changes on both sides. Finally, the remaining DNE biomarkers compute the normalized statistical interquartile range to quantify the range of motion during the FA test. Refer to **Appendix.** C for a detailed description.

V. DNE Biomarkers Analysis

A. `pyDNE` Tool Box:

We introduce `pyDNE`, a convenient, *open-source wrapper of common Python libraries* that streamlines the process of investigating and evaluating the quality of handcrafted digital biomarkers for general human motion analysis, given a dataset collection and interested classification tasks. The overview of `pyDNE` tool box is provided in **Fig.** 2. In this initial version, the DNE-113 is incorporated into `pyDNE`, unifying all DNE features from recorded RGB videos (via *DNE biomarkers extractor*, **Sec.** IV), and clinical annotation into a common data processing pipeline. `pyDNE` supports a wide range of operators for thorough digital biomarkers investigations.

First, the *individual DNE biomarker* quality analysis module separately performs statistical hypothesis tests and visualization for quantitatively and qualitatively assessing the variation and correlation of the same biomarker across different subject categories. Later, the effect of combining multiple features from each test can be analyzed using the *combined DNE biomarkers analysis*. Here, for a given classification model and evaluation metric, this module performs recursive feature elimination [41] to determine the most important biomarkers. Finally, towards the *downstream classification task*, both binary and multi-class classification tasks are supported. On DNE-113, three major binary classification problems: HC versus (v.s.) PD, HC v.s ND, PD v.s OD, and one three-class (HC, PD, OD) classification task are investigated. Users can freely implement their classification algorithm with `scikit-learn` [42] (for machine learning models) or `PyTorch` [43] (for deep-learning models) library. Common classification metrics (e.g., accuracy, precision, recall, specificity, AUC score) are reported for each classification task. For multi-class classifiers, the metrics are *macro-averaged* to summarize the overall performance. In short, the modular design of `pyDNE` allows easily (1) integrating new tests or other datasets, (2) designing digital biomarkers, or (3) customizing the data analysis pipeline.

The capability of `pyDNE` can be extended in the future, preferably in a community-driven fashion. Its modular design allows users to rapidly implement and integrate new functionalities, customizing classification models or adding extra statistical tests. See **Appendix.** D for an overview and **Appendix.** E for example code snippets, demonstrating the use of `pyDNE` in action.

B. Classification with DNE Biomarkers:

Mixture of DNE Tests Classifier. For classification at *DNE record level*, we propose a simple yet efficient ensembling approach, following the concept of mixture-of-experts [44], [45], namely *mixture of DNE tests classifier*. **Fig.** 3 (left) illustrates an inference of a DNE record using this model. Here, we compose a collection of multiple DNE test-specific estimators. Each estimator is trained individually, given all available samples of that test in the training set. At inference time, each estimator individually performs the classification task based on its expertise and the availability of DNE tests given in a record. For the aggregation scheme, a simple *majority voting* across all predictors is adopted in which the final prediction is aggregated by performing *majority voting* among all outputs. Noteworthy, while having a full set of DNE tests allows a better observation, the use of this ensembling strategy does not require a simultaneous availability of all DNE tests, which might not always hold in practice. The implementation of our mixture of DNE classifier is compatible with a regular `scikit-learn` API [42], [46] classifier.

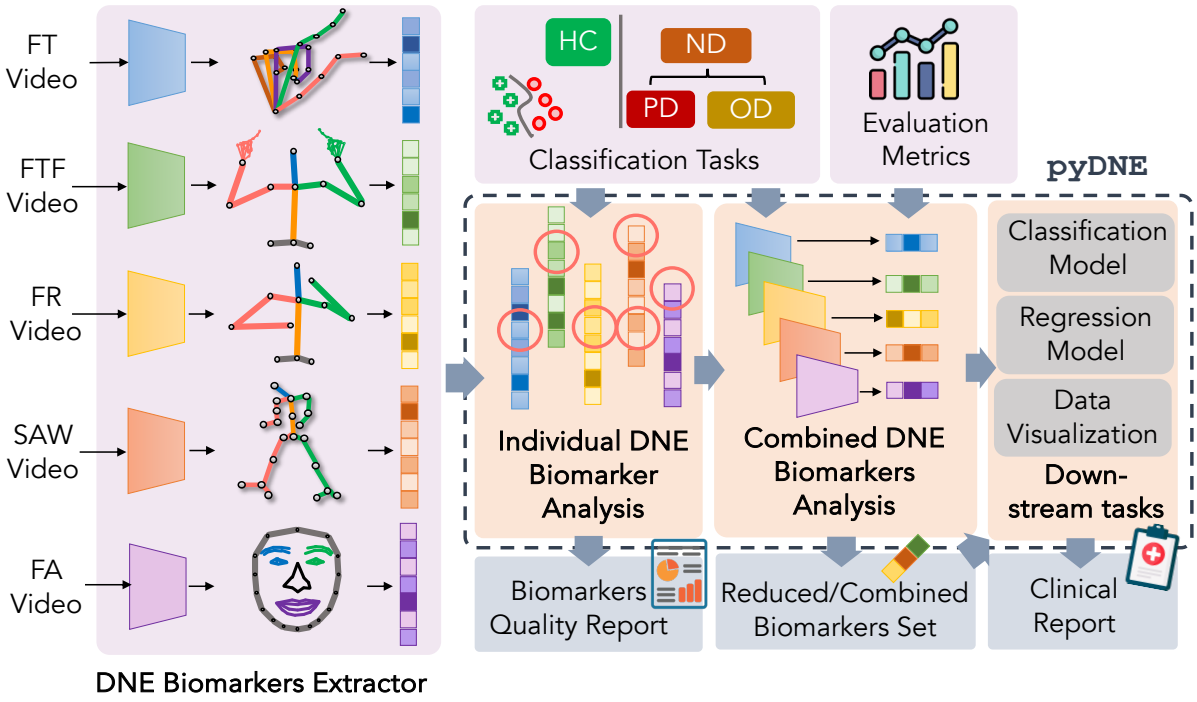


Fig. 2. **Overview of our open-source digital neurological examination (DNE) solution and the proposed `pyDNE` tool box**. Five neurological examinations, including finger tapping (FT), finger-to-finger (FTF), forearm roll (FR), stand-up and walk (SAW), and facial activation (FA) are recorded *(using DNE Recorder)* and digitized into DNE biomarkers *(DNE Biomarkers Extractor)* [12]. `pyDNE` represents those biomarkers into a unified platform for in-depth feature quality analysis, considered both separately *(Individual DNE Biomarker Analysis)* and collectively *(Combined DNE Biomarker Analysis)*. In addition, the digital biomarkers can be incorporated into several downstream binary and multi-class classification tasks. `pyDNE` here supports simple statistical tests, feature importance analysis, and deep-learning/machine learning models.

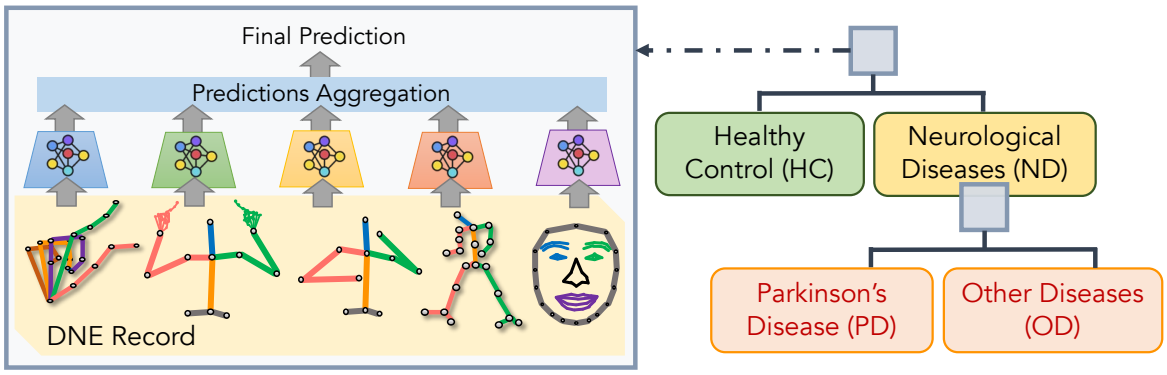


Fig. 3. **Illustration of the mixture of DNE tests classifier and the hierarchical classification task on DNE-113**. The mixture of DNE tests classifier aggregates the prediction results from multiple base estimators, each trained separately for classifying a single DNE test *(left)*. The categories in DNE-113 can be structured hierarchically *(right)*. With the square box denoting a binary classifier, the first classification level focuses on the task of general neurological disorders detection while the second one performs fine-grained differentiation of Parkinson's disease and others.

Hierarchical Classifier. Noticing the hierarchy structure of the HC, PD, and OD, we also explore the setup of a *hierarchical classification* model. Following the *local classifier per parent node strategy* [47], this model performs a two-step inference. Given an input sample, the first binary classifier serves as a general abnormality detection, deciding whether a given DNE test or record belongs to an ND (either OD or PD) or HC subject. In case ND is concluded for this sample, a second classifier is involved in separating PD and OD. A visualization of the category hierarchy and the representation of the classifier is shown in **Fig.** 3 (right). In the `pyDNE` framework, the hierarchical classification is accessible through the `DNEHiClassifier` object. We leverage the `HiClass` package [47] to conveniently structure the hierarchical classification tasks and implement this module. During training, at the first level, the data samples from the PD and OD groups are unified (forming the ND class) for training a base classifier, separating from the HC class. The second classifier follows similarly with PD and OD samples. While the training can be done in parallel, the inference at test time is done in a top-down style to avoid inconsistencies.

Experiment Setup. We perform 5-fold (stratified) cross-validation, i.e., 80%/20% of the videos will be used for train/test sets with preserved class distribution. Additionally, we randomly split the dataset 5 times. For all experiments, the average performance across 5 runs (each run corresponds to a random split) is reported. In all classification experiments in **Sec.** VI-B, VI-C, the choices of hyper-parameters for

all classifiers are simply selected by conducting grid-search with 5-fold cross-validation using the set of hyper-parameters suggested in [12]. Details regarding the parameters chosen for the grid search can be found in **Appendix.** H.1.

VI. Analysis Results

A. Individual Biomarker Quality Analysis

In our DNE, the recorded videos from the five DNE tests are condensed into a total of 75 digital biomarkers: FT (12), FTF (18), FR (12), SAW (15), and FA (18). **Appendix.** B presents a complete list of all biomarkers. As the most intuitive form of qualitative analysis, **Fig.** 4 compares the pair-wise mean and standard deviation of normalized DNE features grouped by subject cohorts. In contrast to other cohorts, the DNE biomarkers of the HC group exhibit the lowest variance across subjects. In several DNE tests, the mean of feature values measured in this group are also significantly different from those of the abnormal PD and OD cohorts (first and middle row, respectively). However, the biomarkers of PD and OD display a considerable similarity (last row), with a notable overlap between the two groups.

Following, we investigate a quantitative evaluation *for each DNE feature* when serving as a discriminant criterion for three binary classification tasks. The first task (HC v.s. PD) evaluates the ability to distinguish PD from a healthy population, while the second task (HC v.s. ND) assesses whether the system can serve as a general neurological movement disorders detector. The final one further observes the fine-grained differences between PD and OD subjects, collected in DNE-113. In this setup, the Mann-Whitney's U non-parametric test [48] is used to evaluate the difference in the distribution of estimated features from two separate groups of subjects (with the 2-sided p-value and two significance levels of $\alpha = 0.05$). **Tab.** II reports the top 5 features along with their corresponding p-values that are the most discriminate, between the two groups of subjects (refer to **Appendix.** G.1 for the results of other DNE biomarkers). Overall, the significance of each biomarker varies across tasks. The ranking of useful biomarkers is analogous in the HC v.s. PD or ND tasks (Cols. 1, 2), whereas the PD. v.s. OD tasks (Col. 3) prioritize different sets of biomarkers. Hence, deriving a complete set of digital biomarkers customized for the downstream task is important in application. Between HC and PD, or ND, a majority of the DNE features are statistically significant differences. Meanwhile, most of the features are struggling with separating the PD and OD classes, as reflected in the numbers of features with the $p-$value larger than the critical value. The observation is consistent with the qualitative visualization. DNE features, even though initially designed for identifying simulated impairment movements, demonstrate their effectiveness in capturing abnormalities of PD and OD patients. However, their capability only fits the purpose of general neurological abnormal movement detection. The ability to capture the unique, fine-grained PD motor phenotypes for separating PD requirements remains a challenge for this biomarker set. There is room for improvement: *enhancing the design of digital biomarkers to improve PD characterization*.

B. Binary Classification Tasks

Following, we assess the effect of combining multiple DNE biomarkers in binary classification tasks with simple machine learning (ML) classifiers. The investigation is conducted at both DNE *test level* (biomarkers extracted from a single test), and DNE *record level* (biomarkers from all DNE tests within a record, **Sec.** V-B). We benchmark the performance of several ML models by using the concatenated DNE biomarkers of each DNE test as input, for training and evaluation. The choices of ML models used in our experiments can be grouped into (1) tree-based models: Random Forest (RF), Gradient-Boosting Machine (GBM), XGBoost [49] and (2) parametric models trained using gradient-descent: Logistic Regression (LR), Support Vector Machine with radial basis function kernel (RSVM), Multi-layer Perceptron (MLP) with rectified linear unit (ReLU) activation. For reference, we also include two deep-learning (DL) models based on CNN [50], [51] and LSTM [52] architectures that operate directly on the detected 2D keypoints, without engineering the digital biomarkers.

In **Tab.** III, we provide the evaluation results of five tests (each corresponds to a row block), and at the record level (last row block). The evaluation results on three main tasks: HC v.s. PD, HC v.s. ND, and PD v.s. OD are organized in three blocks of columns. Overall, the combination of ML models and digital biomarkers achieves comparable accuracy with large-scale deep-learning models across most tasks while having a higher interpretability and computational efficiency. Simple ML models achieve higher accuracy on the FT, FR, and FA, while the SAW and FTF tests favor DL models. This is attributed to the high complexity of FTF, and SAW motion, which aligns with the strengths of DL models. Although the peak performance of ML models varies across different DNE tests, LR, GBM, and RF effectively combine the knowledge of DNE biomarkers in most tests. The SAW and FT are the most discriminate tests for HC v.s. PD/ND, followed by the FTF, FR, and FA tests. Meanwhile, the FTF is the most useful for PD v.s. OD, indicating the advantage of performing multi-test assessment. PD v.s. OD, with a notable gap, is still the most challenging task even when information from all features and tests are combined. In many tests, ML models outperform DL models on this task. Generally, the result aligns with the features quality analysis (**Sec.** VI-A), elaborating the role of constructing a high-quality set of DNE biomarkers.

Consequently, we examine the classification at the DNE record level, when information from multiple tests is consolidated. The benefits of forming a complete observation demonstrate a significant improvement in the HC v.s. ND/PD classification task. In PD v.s. OD, the ensemble model also surpasses the standalone models, except the FTF test. Note that, majority voting, the simplest scheme that considers the contribution of each test equally, requires all estimators to have a comparative performance. Further fine-tuning the aggregation schemes, and as more data is collected, the classification between PD and OD would be further enhanced.

C. Multi-class Classification

Finally, everything is put together into a three-class (HC, PD, and OD) classification problem. The setup at the record-

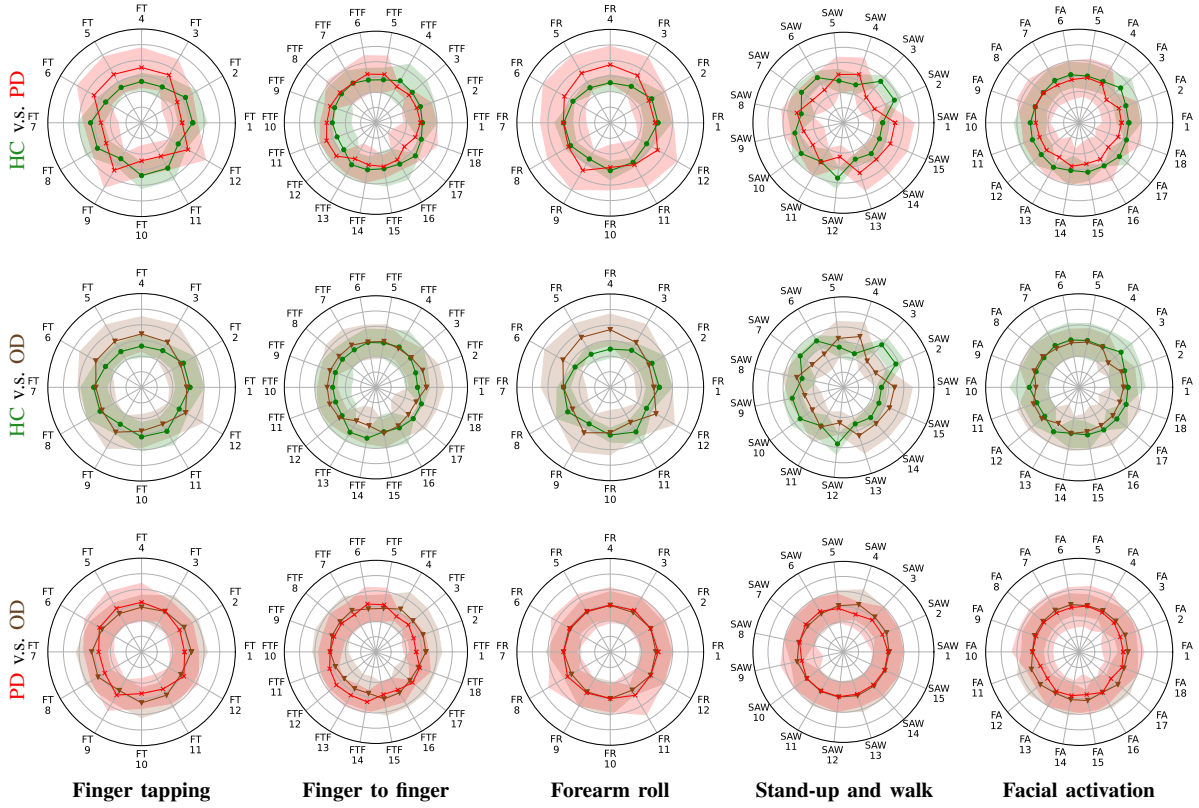


Fig. 4. **Qualitative pair-wise comparison between DNE biomarkers from different subject groups**. Mean *(solid line)* and variance *(shaded area)* of the normalized features from the Parkinson's disease (PD: —×—) and healthy control (HC: —●—) group *(first row)*. HC and other neurological diseases (OD: —▼—) group *(middle row)*. PD and OD group *(last row)*. Distinct variations are noted between HC and PD, while similarities exist between OD and PD. The mapping from abbreviations to the feature names is provided in **Appendix.** B. Best viewed in color.

TABLE II

MOST DISCRIMINATIVE BIOMARKERS AND THE CORRESPONDING p—VALUES SORTED IN DESCENDING ORDER. THE MANN-WHITNEY'S U TEST [48] COMPARES PAIRS OF TWO POPULATIONS, BETWEEN HEALTHY CONTROL (HC) AND NEUROLOGICAL DISEASES (ND) WHICH INCLUDE PARKINSON'S DISEASE (PD), AND OTHER NEUROLOGICAL DISORDERS (OD) GROUPS. MOST BIOMARKERS CAN INDICATE THE DIFFERENCES BETWEEN HC AND PD OR ND GROUPS IN GENERAL. THE UNDERLINE DENOTES THE FEATURES WITH THE p-VALUE GREATER THAN THE STATISTIC SIGNIFICANCE LEVEL OF 0.05. **APPENDIX.** G.1 PROVIDES THE FULL ASSESSMENT FOR ALL DNE BIOMARKERS.

Task	**HC v.s. PD**		**HC v.s. ND**		**PD v.s. OD**	
Test	**Statistic**	p-value	**Statistic**	p-value	**Statistic**	p-value
FT	Tapping period asymmetry	$2e^{-6}$	Tapping period asymmetry	$1e^{-7}$	Maximum tapping acceleration (R, L)	$(5e^{-3}, 5e^{-2})$
	Maximum tapping acceleration (R, L)	$(3e^{-6}, 1e^{-4})$	Tapping period (R, L)	$(1e^{-6}, 7e^{-6})$	Maximum tapping speed (R, L)	$(5e^{-2}, 3e^{-2})$
	Tapping period (R, L)	$(2e^{-5}, 2e^{-5})$	Maximum tapping acceleration (R, L)	$(2e^{-4}, 3e^{-4})$	Tapping amplitude (R, L)	$(3e^{-2}, \underline{6e^{-2}})$
	Tapping amplitude asymmetry	$1e^{-4}$	Tapping amplitude asymmetry	$5e^{-4}$	Maximum tapping speed asymmetry	$\underline{1e^{(-1)}}$
	Tapping amplitude (R, L)	$3e^{-4}$, $4e^{-4}$	Maximum tapping speed asymmetry	$1e^{-3}$	Tapping period (R, L)	$(\underline{3e^{-1}}, \underline{2e^{-1}})$
FTF	Horizontal finger symmetry	$5e^{-9}$	Horizontal finger symmetry	$2e^{-12}$	STD Period (R, L)	$(5e^{-4}, 8e^{-3})$
	STD path smoothness (R, L)	$(7e^{-3}, 9e^{-5})$	STD path smoothness (R, L)	$(1e^{-4}, \underline{8e^{-2}})$	Mean Period (R, L)	$(1e^{-3}, 2e^{-3})$
	STD period (R, L)	$(1e^{-2}, 5e^{-2})$	Mean velocity angle symmetry (R, L)	$(4e^{-3}, 2e^{-4})$	STD path smoothness (R, L)	$(8e^{-3}, \underline{1e^{-1}})$
	Vertical finger symmetry	$4e^{-2}$	Vertical finger symmetry	$2e^{-2}$	Mean velocity angle symmetry (R, L)	$(2e^{-2}, 2e^{-2})$
	Mean velocity angle symmetry (R, L)	$(\underline{9e^{-2}}, \underline{3e^{-1}})$	Mean period (R, L)	$(\underline{4e^{-1}}, \underline{2e^{-1}})$	STD speed (R, L)	$(\underline{3e^{-1}}, \underline{2e^{-1}})$
FR	Rolling period (R, L)	$(7e^{-10}, 8e^{-10})$	Rolling period (R, L)	$(2e^{-15}, 3e^{-15})$	Amplitude asymmetry	$\underline{1e^{-1}}$
	Rolling amplitude asymmetry	$2e^{-4}$	Maximum rolling acceleration (R, L)	$(3e^{-6}, 1e^{-6})$	Maximum speed asymmetry	$\underline{4e^{-1}}$
	Maximum rolling speed asymmetry	$4e^{-4}$	Rolling period asymmetry	$1e^{-5}$	Rolling period asymmetry	$\underline{4e^{-1}}$
	Maximum rolling acceleration (R, L)	$(7e^{-4}, 9e^{-4})$	Maximum acceleration asymmetry	$5e^{-5}$	Maximum acceleration asymmetry	$\underline{6e^{-1}}$
	Rolling period asymmetry	$3e^{-3}$	Maximum rolling speed asymmetry	$3e^{-4}$	Rolling amplitude (R, L)	$(\underline{6e^{-1}}, \underline{6e^{-1}})$
SAW	Median step length	$4e^{-9}$	Mean step length	$2e^{-12}$	Mean step width	$1e^{-2}$
	Mean step length	$5e^{-9}$	Median step length	$3e^{-12}$	Median knee angle symmetry	$\underline{3e^{-1}}$
	Mean walking speed	$1e^{-8}$	Mean walking speed	$1e^{-11}$	Mean knee angle symmetry	$\underline{3e^{-1}}$
	Mean step symmetry	$4e^{-5}$	STD step time	$1e^{-6}$	Mean step length	$\underline{3e^{-1}}$
	Mean turning time	$7e^{-5}$	Mean turning time	$1e^{-6}$	Mean step symmetry	$\underline{4e^{-1}}$
FA	STD mouth opening	$4e^{-8}$	STD mouth opening	$2e^{-10}$	Eyebrow lift speed (R, L)	$(1e^{-3}, 6e^{-3})$
	Mouth opening symmetry	$4e^{-6}$	Mouth opening symmetry	$8e^{-8}$	STD eye opening	$(1e^{-3}, 2e^{-3})$
	STD eyebrow lift speed (R, L)	$(1e^{-5}, 5e^{-5})$	Mouth opening speed symmetry	$1e^{-5}$	STD eye opening speed	$(2e^{-3}, 2e^{-3})$
	Eye opening speed symmetry	$2e^{-5}$	Normalized IQR Mouth opening	$5e^{-5}$	Normalized IQR eye opening (R, L)	$(3e^{-3}, 4e^{-3})$
	STD eye opening speed (R,L)	$(4e^{-5}, 3e^{-5})$	Eye opening speed symmetry	$4e^{-4}$	Eye opening speed symmetry	$4e^{-3}$

TABLE III

BINARY CLASSIFICATION PERFORMANCE ON DNE-113. EXPERIMENTS ARE DONE SEPARATELY AT THE DNE TEST LEVEL (FT, FTF, FR, SAW, AND FA) AND RECORD LEVEL (REC). SIX MACHINE LEARNING MODELS: RANDOM FOREST (RF), GRADIENT-BOOSTING MACHINE (GBM), XGBOOST, LOGISTIC REGRESSION (LR), SUPPORT VECTOR MACHINE WITH RBF KERNEL (RSVM), AND MULTI-LAYER PERCEPTRON (MLP). FOR TEST-LEVEL EXPERIMENTS, TWO DEEP-LEARNING-BASED MODELS (CNN AND LSTM) ARE INCLUDED FOR REFERENCE. THE BEST PERFORMANCE FROM EACH GROUP OF MODELS (MACHINE LEARNING/ DEEP-LEARNING) IS HIGHLIGHTED IN **BOLD**.

		HC v.s. PD					HC v.s. ND					PD v.s. OD				
Test	**Model**	**Acc**	**Precision**	**Recall**	**Specificity**	**AUC**	**Acc**	**Precision**	**Recall**	**Specificity**	**AUC**	**Acc**	**Precision**	**Recall**	**Specificity**	**AUC**
FT	GBM	0.8046	0.7682	0.7371	0.8500	0.7936	0.7561	0.8024	0.7925	0.7000	0.7463	0.5650	0.5026	0.4457	0.6578	0.5517
	LR	**0.8437**	**0.8467**	0.7486	**0.9077**	**0.8281**	0.7727	0.8361	0.7775	0.7654	0.7714	**0.6000**	**0.5509**	**0.4686**	0.7022	**0.5854**
	MLP	0.8345	0.8086	**0.7714**	0.8769	0.8242	0.7758	0.8172	0.8125	0.7192	0.7659	0.5675	0.5144	0.3943	0.7022	0.5483
	RF	0.8046	0.7707	0.7314	0.8538	0.7926	0.7758	0.8268	0.7975	0.7423	0.7699	0.5975	0.5498	0.4400	0.7200	0.5800
	RSVM	0.8345	0.8408	0.7257	**0.9077**	0.8167	**0.7864**	**0.8440**	0.7950	**0.7731**	**0.7840**	0.5725	0.5210	0.2857	**0.7956**	0.5406
	XGBOOST	0.8161	0.7914	0.7371	0.8692	0.8032	0.7576	0.7893	**0.8200**	0.6615	0.7408	0.5625	0.4961	0.4514	0.6489	0.5502
	CNN	**0.8184**	**0.7894**	0.7486	**0.8654**	0.8070	**0.8364**	**0.8652**	0.8650	**0.7923**	**0.8287**	0.5125	0.4348	0.4000	0.6000	0.5000
	LSTM	**0.8184**	0.7707	**0.7829**	0.8423	**0.8126**	**0.8364**	0.8453	**0.8950**	0.7462	0.8206	**0.5525**	**0.4883**	**0.4057**	**0.6667**	**0.5362**
FTF	GBM	**0.8250**	**0.8145**	**0.7257**	**0.8906**	**0.8081**	**0.8030**	**0.8392**	0.8346	**0.7547**	**0.7946**	**0.7111**	0.6729	**0.6571**	0.7522	**0.7047**
	LR	0.7886	0.7529	0.6971	0.8491	0.7731	0.7567	0.7731	0.8469	0.6189	0.7329	0.6296	0.5788	0.5200	0.7130	0.6165
	MLP	0.7795	0.7445	0.6743	0.8491	0.7617	0.7731	0.8038	0.8321	0.6830	0.7576	0.6395	0.5758	0.6286	0.6478	0.6382
	RF	0.7932	0.7885	0.6571	0.8830	0.7701	0.7836	**0.8299**	0.8099	**0.7434**	0.7766	0.7012	**0.6748**	0.6000	**0.7783**	0.6891
	RSVM	0.7182	0.6839	0.5257	0.8453	0.6855	0.7955	0.7757	**0.9309**	0.5887	0.7598	0.7012	0.6547	0.6571	0.7348	0.6960
	XGBOOST	0.7659	0.7341	0.6457	0.8453	0.7455	0.7791	0.8077	0.8346	0.6943	0.7645	0.6173	0.5609	0.5371	0.6783	0.6077
	CNN	**0.8318**	**0.8068**	**0.7600**	**0.8792**	**0.8196**	**0.8478**	**0.8548**	**0.9037**	**0.7623**	**0.8330**	**0.6543**	**0.6214**	0.5257	**0.7522**	**0.6389**
	LSTM	0.7773	0.7794	0.6286	0.8755	0.7520	0.7746	0.8010	0.8346	0.6830	0.7588	0.6222	0.5603	**0.5771**	0.6565	0.6168
FR	GBM	0.7663	0.7009	0.6457	0.8367	0.7412	0.7660	0.7871	0.8123	0.7033	0.7578	0.4938	**0.4011**	**0.3714**	0.5870	0.4792
	LR	0.7937	**0.8967**	0.4971	**0.9667**	0.7319	0.8099	0.8558	0.8049	0.8167	0.8108	**0.5383**	0.3797	0.0800	0.8870	**0.4835**
	MLP	0.7768	0.7274	0.6343	0.8600	0.7471	0.8014	0.8308	0.8222	0.7733	0.7978	0.4000	0.3019	0.2971	0.4783	0.3877
	RF	**0.8042**	0.7713	**0.6686**	0.8833	**0.7760**	0.7901	0.8223	0.8099	0.7633	0.7866	0.4938	0.3899	0.2914	0.6478	0.4696
	RSVM	0.7958	0.8686	0.5257	0.9533	0.7395	**0.8255**	**0.8874**	0.7975	**0.8633**	**0.8304**	0.5358	0.2541	0.0629	**0.8957**	0.4793
	XGBOOST	0.7705	0.7190	0.6229	0.8567	0.7398	0.7887	0.8112	**0.8247**	0.7400	0.7823	0.4790	0.3835	0.3371	0.5870	0.4620
	CNN	**0.7768**	0.7190	**0.6514**	0.8500	**0.7507**	**0.7957**	0.8145	**0.8346**	0.7433	0.7890	0.5111	0.4241	0.3543	0.6304	0.4924
	LSTM	0.7705	**0.7542**	0.5600	**0.8933**	0.7267	0.7858	**0.8519**	0.7605	**0.8200**	**0.7902**	**0.5210**	**0.4433**	**0.3714**	**0.6348**	**0.5031**
SAW	GBM	**0.8419**	0.8606	0.8067	0.8750	**0.8408**	**0.8337**	0.8657	0.8957	0.7000	0.7978	0.5101	0.4302	0.3867	0.6051	0.4959
	LR	0.8226	0.8755	0.7400	0.9000	0.8200	0.8277	0.8432	**0.9188**	0.6312	0.7750	0.5217	0.3131	0.1067	0.8410	0.4738
	MLP	0.7903	0.8040	0.7600	0.8187	0.7894	0.8257	0.8771	0.8667	0.7375	0.8021	0.4638	0.3804	0.3733	0.5333	0.4533
	RF	0.8194	0.8338	0.7867	0.8500	0.8183	0.8297	**0.8805**	0.8696	**0.7438**	**0.8067**	0.5246	**0.4405**	0.3533	0.6564	**0.5049**
	RSVM	**0.8419**	**0.8922**	0.7667	**0.9125**	0.8396	0.8099	0.8550	0.8696	0.6813	0.7754	**0.5304**	0.1178	0.0267	**0.9179**	0.4723
	XGBOOST	0.8194	0.8103	**0.8267**	0.8125	0.8196	0.8317	0.8573	0.9043	0.6750	0.7897	0.5130	0.4362	**0.4133**	0.5897	0.5015
	CNN	**0.8806**	**0.8602**	**0.9000**	0.8625	**0.8812**	**0.8970**	**0.9176**	**0.9333**	**0.8187**	**0.8760**	**0.5565**	**0.4835**	**0.3333**	**0.7282**	**0.5308**
	LSTM	0.8548	0.8567	0.8400	**0.8688**	0.8544	0.8713	0.8977	0.9159	0.7750	0.8455	0.5130	0.4193	0.3133	0.6667	0.4900
FA	GBM	0.7818	0.7526	0.6800	0.8491	0.7645	0.7615	0.7922	0.8244	0.6642	0.7443	**0.6122**	**0.5487**	**0.5143**	0.6851	**0.5997**
	LR	0.7727	**0.8381**	0.5314	**0.9321**	0.7318	0.7763	0.7775	**0.8854**	0.6075	0.7465	0.5829	0.5094	0.3029	**0.7915**	0.5472
	MLP	**0.8068**	0.7833	**0.7143**	0.8679	**0.7911**	**0.7778**	**0.8306**	0.7976	**0.7472**	**0.7724**	0.5512	0.4720	0.4171	0.6511	0.5341
	RF	0.7727	0.7615	0.6229	0.8717	0.7473	0.7704	0.7961	0.8366	0.6679	0.7523	0.6073	0.5480	0.4629	0.7149	0.5889
	RSVM	0.7636	0.7560	0.6000	0.8717	0.7358	0.7689	0.7906	0.8439	0.6528	0.7484	0.5537	0.4697	0.3657	0.6936	0.5297
	XGBOOST	0.7750	0.7452	0.6571	0.8528	0.7550	0.7719	0.8114	0.8146	0.7057	0.7601	0.5829	0.5096	0.4743	0.6638	0.5691
	CNN	**0.7545**	**0.7247**	**0.6229**	**0.8415**	**0.7322**	0.7393	0.7887	**0.7805**	0.6755	0.7280	**0.6220**	**0.5671**	**0.4857**	**0.7234**	**0.6046**
	LSTM	0.7318	0.6967	0.5714	0.8377	0.7046	**0.7467**	**0.8116**	0.7610	**0.7245**	**0.7428**	0.6195	0.5670	0.4800	**0.7234**	0.6017
REC	GBM	0.8940	0.8717	0.8171	0.9354	0.8763	0.8612	0.8519	0.9098	0.8000	0.8549	0.6195	0.5726	**0.4286**	0.7617	0.5951
	LR	0.8440	**0.9905**	0.5600	**0.9969**	0.7785	**0.9034**	0.8646	**0.9805**	0.8062	0.8933	0.6171	**0.8243**	0.1314	**0.9787**	0.5551
	MLP	**0.9160**	0.9180	**0.8343**	0.9600	**0.8971**	0.8776	0.8847	0.8976	0.8523	0.8749	0.6122	0.5621	0.4229	0.7532	0.5880
	RF	0.9020	0.8986	0.8114	0.9508	0.8811	0.8857	0.8900	0.9073	**0.8585**	0.8829	**0.6415**	0.6288	0.3943	0.8255	**0.6099**
	RSVM	0.8820	0.9336	0.7143	0.9723	0.8433	0.9020	**0.8915**	0.9390	0.8554	**0.8972**	0.5854	0.7443	0.0686	0.9702	0.5194
	XGBOOST	0.9060	0.8997	0.8229	0.9508	0.8868	0.8898	0.8873	0.9195	0.8523	0.8859	0.6293	0.5928	0.4114	0.7915	0.6015

level generalizes similarly, using the aforementioned ensemble approach. The choice of a base estimator is simply adopted from the list of the binary ML classifiers in **Sec.** VI-B. In **Tab.** IV, the numerical evaluation results of the multi-class (left column block) and hierarchical classifier - **Sec.** V-B (right column block) are reported. The classification performance also varies across different choices of base estimators at each DNE test. The hierarchical classifier performing at the record level (lower right corner block) establishes the *highest overall performance*, utilizing the information from both multiple DNE tests and the categories structure. **Fig.** 5 (top) showcases the confusion matrix of the best hierarchical models reported in **Tab.** IV. Unsurprisingly, we observe that most of the incorrect predictions come within the PD and OD groups.

For a further analysis of the multi-class setting, the *combined DNE biomarker analysis* performs the recursive feature elimination [41] (using GBM classifier, and accuracy as the evaluation metric). The surviving biomarkers are concatenated for visualization with t-SNE [53] in **Fig.** 5. See **Appendix.** G.2 for a complete result. While varying across tests, the t-SNE plots indicate a satisfactory separation between the HC v.s. ND, while this is not the case for PD v.s. OD. Qualitatively, the SAW test achieves the best separation. The visualization here provides insights for reasoning the classification results, based on the discriminative of combined biomarkers.

VII. DISCUSSION, CHALLENGES AND USE CASES

We summarize the key significance and discuss existing challenges and relevant applications in this work.

A. Discussion

The DNE-113 Dataset. The comparison of our DNE-113 with similar datasets is summarized in **Tab.** V. For the purpose of this study, only vision-based datasets, captured by a single camera, and having a primary focus on classifying PD subjects or regressing the MDS-UPDRS-III score are included. Acknowledging the privacy concerns, a dataset is determined as *available* when the 2D/3D pose estimation results from one human pose estimator are released, as a minimum requirement.

TABLE IV

MULTI-CLASS CLASSIFICATION PERFORMANCE ON DNE-113. EXPERIMENTS ARE DONE SEPARATELY AT THE DNE TEST LEVEL (FT, FTF, FR, SAW, AND FA) AND RECORD LEVEL (REC) FROM THE DNE-113 DATASET. WE REPORT THE EVALUATION RESULTS OF TWO POSSIBLE CHOICES OF CLASSIFIER: THE 3-CLASS CLASSIFICATION MODEL AND THE HIERARCHY CLASSIFICATION MODEL. THE BEST ONE IS HIGHLIGHTED IN **BOLD**.

		3-class Classifier					Hierarchical Classifier				
Test	**Model**	**Acc**	**Precision**	**Recall**	**Specificity**	**AUC**	**Acc**	**Precision**	**Recall**	**Specificity**	**AUC**
FT	GBM	0.5121	0.4880	0.4902	0.7533	0.6217	0.5576	0.5426	0.5422	0.7793	0.6608
	LR	**0.5742**	**0.5446**	**0.5489**	**0.7835**	**0.6662**	**0.5833**	0.5618	**0.5638**	**0.7885**	**0.6762**
	MLP	0.5485	0.5297	0.5267	0.7703	0.6485	0.5348	0.5174	0.5191	0.7664	0.6427
	RF	0.5409	0.5140	0.5187	0.7668	0.6428	0.5439	0.5399	0.5267	0.7704	0.6485
	RSVM	0.5682	0.5432	0.5369	0.7778	0.6574	0.5697	**0.5689**	0.5434	0.7797	0.6616
	XGBOOST	0.5303	0.5045	0.5058	0.7612	0.6335	0.5227	0.5151	0.5075	0.7612	0.6344
FTF	GBM	**0.6552**	0.6438	**0.6391**	**0.8252**	**0.7321**	0.6000	0.5922	0.5906	0.7983	0.6945
	LR	0.5478	0.5246	0.5258	0.7717	0.6488	0.5687	0.5682	0.5583	0.7886	0.6734
	MLP	0.5896	0.5724	0.5715	0.7937	0.6826	0.5985	0.5832	0.5812	0.7986	0.6899
	RF	**0.6552**	**0.6459**	0.6356	0.8240	0.7298	**0.6433**	0.6401	**0.6302**	**0.8198**	**0.7250**
	RSVM	0.5791	0.5675	0.5554	0.7836	0.6695	0.6164	**0.6430**	0.6186	0.8145	0.7166
	XGBOOST	0.6134	0.5959	0.5921	0.8043	0.6982	0.5970	0.5860	0.5839	0.7985	0.6912
FR	GBM	0.5872	0.5377	0.5414	0.7914	0.6664	0.5745	**0.5397**	0.5376	0.7887	0.6631
	LR	**0.6340**	**0.5573**	**0.5607**	0.8056	**0.6831**	0.6014	0.5289	0.5394	0.7942	0.6668
	MLP	0.5887	0.5213	0.5327	0.7926	0.6627	0.5418	0.4832	0.4903	0.7738	0.6321
	RF	0.5631	0.4863	0.5062	0.7790	0.6426	0.5830	0.5372	0.5409	0.7915	0.6662
	RSVM	0.6326	0.5449	0.5562	**0.8057**	0.6810	**0.6284**	0.4935	**0.5577**	**0.8071**	**0.6824**
	XGBOOST	0.5872	0.5276	0.5367	0.7921	0.6644	0.5518	0.5105	0.5107	0.7779	0.6443
SAW	GBM	0.5683	0.5644	0.5651	0.7807	0.6729	0.5802	0.5857	0.5802	0.7873	0.6838
	LR	**0.6040**	0.5842	0.5912	0.7966	0.6939	**0.5941**	0.5794	**0.5919**	**0.7946**	**0.6932**
	MLP	0.5762	0.5503	0.5643	0.7826	0.6735	0.5129	0.5161	0.5161	0.7541	0.6351
	RF	0.6020	0.5933	0.5996	**0.7992**	**0.6994**	0.5881	**0.6011**	0.5822	0.7893	0.6857
	RSVM	0.5644	0.5004	0.5426	0.7732	0.6579	0.5465	0.5663	0.5198	0.7608	0.6403
	XGBOOST	0.6000	**0.5999**	**0.6001**	0.7970	0.6986	0.5386	0.5555	0.5332	0.7650	0.6491
FA	GBM	0.5704	0.5591	0.5561	0.7815	0.6688	0.5585	0.5675	0.5494	0.7783	0.6639
	LR	0.5659	0.5359	0.5334	0.7771	0.6553	0.5570	0.5431	0.5378	0.7766	0.6572
	MLP	0.5556	0.5382	0.5359	0.7742	0.6551	0.5037	0.4965	0.4906	0.7493	0.6199
	RF	**0.5837**	**0.5648**	**0.5674**	**0.7888**	**0.6781**	0.5615	0.5561	0.5495	0.7787	0.6641
	RSVM	0.5333	0.5035	0.5095	0.7629	0.6362	0.4993	0.5077	0.4933	0.7501	0.6217
	XGBOOST	0.5719	0.5539	0.5561	0.7842	0.6702	**0.5778**	**0.5756**	**0.5648**	**0.7865**	**0.6757**
REC	GBM	0.6599	0.6226	0.6104	0.8225	0.7164	**0.7279**	0.7110	**0.6935**	**0.8668**	**0.7802**
	LR	0.6639	0.6455	0.5903	0.8159	0.7031	0.7224	0.7059	0.6724	0.8611	0.7667
	MLP	0.6381	0.5861	0.5857	0.8156	0.7007	0.6476	0.6085	0.5995	0.8271	0.7133
	RF	**0.6884**	**0.6682**	**0.6360**	**0.8336**	**0.7348**	0.7129	0.6980	0.6658	0.8583	0.7621
	RSVM	0.6599	0.6191	0.5802	0.8164	0.6983	0.7197	**0.7496**	0.6636	0.8613	0.7625
	XGBOOST	0.6707	0.6306	0.6186	0.8291	0.7239	0.6857	0.6588	0.6448	0.8453	0.7451

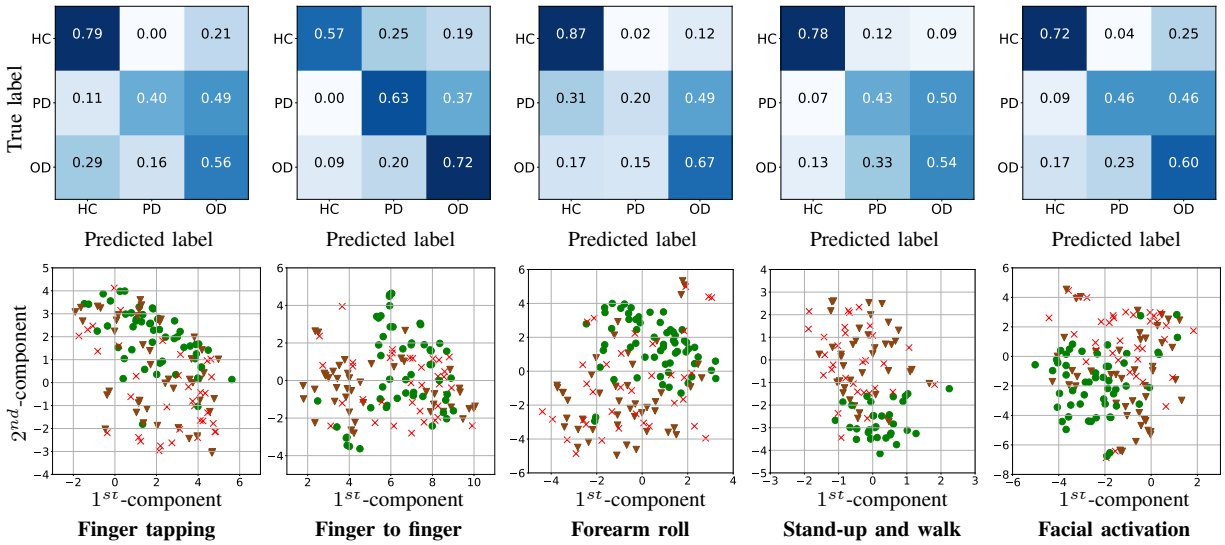


Fig. 5. **Multi-class classification on DNE-113.** *Top row*: Confusion matrix of the best model on the three-class (HC, PD, and OD) classification problem. For most DNE tests, a clear boundary between HC and PD/OD cohorts is observed but it is not the case within the PD and OD subgroups. This result is also reflected in the confusion matrices when using these biomarkers for machine-learning classification models. *Bottom row*: t-SNE [53] dimensionality reduction analysis of FT, FTF, FR, SAW, and FA tests. Green circle (●), red cross (×), and brown triangle (▼) stand for healthy control (HC), Parkinson's disease (PD), and other diseases (OD). Best viewed in color.

Our DNE-113 contains *multiple* neurological examinations, including upper limbs, SAW, and FA tests, collected by a simple smartphone device and covers a *wider* spectrum of neurological disorders compared to others. These advantages encourage the exploration of novel technologies to integrate information from multiple tests and address the *frontier challenges* of

TABLE V
A COMPARISON BETWEEN VISION-BASED, SINGLE-CAMERA PARKINSON'S DISEASE ASSESSMENT DATASETS. ONLY SINGLE-CAMERA (REGULAR CAMERA, SMARTPHONE, OR WEBCAM DEVICES), ARE INCLUDED. A DATASET IS DETERMINED AS AVAILABLE (✓) IF THE POSE ESTIMATION RESULTS ARE ACCESSIBLE. OUR DNE-113 DATASET IS A MULTI-TEST, SMARTPHONE-RECORDED, WITH A WIDER RANGE OF NEUROLOGICAL DISORDERS.

Approach	Cohort	Test(s)	Device	Available	Main Analysis Outcomes
Catherine *et al.* [28]	24 PD + 24 HC	Sit-to-stand	Camera	✓	PD/HC classification
Monje *et al.* [17]	22 PD + 20 HC	Upper limb	Webcam	✓	PD/HC classification
Vignoud *et al.* [29]	36 PD + 11 HC	Upper limb	Camera	✗	MDS-UPDRS-III score regression
Yin *et al.* [36]	39 PD	Gait + Upper limb	Camera	✗	Severe/mild PD classification
Shin *et al.* [21]	16 PD + 15 HC	Gait	Camera	✓	Gait parameters estimation
Rachneet *et al.* [22]	9 PD + 10 MS + 14 HC	Gait	Camera	✓	PD/MS/HC classification
Rupprechter *et al.* [54]	544 PD + 185 HC	Gait	Camera	✗	MDS-UPDRS-III score regression
Lu *et al.* [30]	55 PD	Gait, Finger tapping	Camera	✓	MDS-UPDRS-III score regression
Islam *et al.* [16]	172 PD + 78 HC	Finger tapping	Webcam	✗	MDS-UPDRS-III score regression
Li *et al.* [20]	157 PD	Finger tapping	Camera	✗	MDS-UPDRS-III score regression
Khan *et al.* [24]	13 PD + 6 HC	Finger tapping	Camera	✗	PD/HC classification
Williams *et al.* [55]	73 PD + 60 HC	Finger tapping	Smartphone	✗	MDS-UPDRS-III score regression
Morinan *et al.* [31]	1156 PD	Finger/toe tapping, hand movement, leg activity	Camera	✗	MDS-UPDRS-III score regression
Rajnoha *et al.* [56]	50 PD + 50 HC	Facial activation	Camera	✗	PD/HC classification
Grammatikopoulou *et al.* [32]	23 PD + 11 HC	Facial activation	Smartphone	✗	PD/HC classification
Gómez *et al.* [23]	30 PD + 24 HC	Facial Activation	Camera	✗	PD/HC classification
Hoang *et al. (ours)*	33 PD + 46 OD + 34 HC	Gait, upper limb, finger tapping, facial activation	Smartphone	✓	PD/OD/HC classification

PD: Parkinson's disease, OD: Other neurological disorders excluding PD, HC: Healthy control, MS: Multiple sclerosis.

characterizing PD/OD movements in clinical settings.

Inspection Tools for Digital Biomarkers. When analyzing recorded neurological examinations, or human motion analysis in a broader sense, *explainable* digital biomarkers take an important role in providing clinical-relevant explanations [57]. Doctors and bioengineers can manually design a set of useful digital biomarkers fitting their purposes. However, quantitatively evaluating the contributions of each feature should be carefully investigated. Besides, while the digital biomarkers are well designed with significant interpretability based on clinical-relevant knowledge, the process of making the final prediction is designated for a simple but still a black-box ML model. Hence, it is important to justify the benefits of these biomarkers. `pyDNE` serves as an inspection tool for *conveniently* and *systematically* analyzing their discriminate power that forms the basis for the success of existing works.

DNE Biomarkers and Clinical Relevance. The most meaningful DNE biomarkers in our analysis (**Tab.** II) are *in line* with the exam findings observed and subjectively judged (but not quantifiably measured) by clinicians evaluating *patients with PD/OD* [58], [59]. HC subjects activate and coordinate their motor movements well on both sides while this is not the case for PD/OD patients. As a result, asymmetry and/or the consistency of movements across cycles (especially on the upper limb tests, depending on the condition) reflect the distinction. In addition, the conditions affecting postural stability, such as due to Parkinson's syndrome, in the patient groups also reduce the step length, step width, turning time, and walking speed in the gait test. The measurement of these parameters, hence, is useful for characterizing the movements of PD/OD patients. When comparing the PD and OD, the features measuring acceleration, speed, or the period of movement are the most discriminating. The differences in those biomarkers signal the slowness of movement and speed (bradykinesia) in the PD cohort.

Digital Biomarkers Construction. Designing high-quality features is *crucial* for many applications [57]. Unfortunately, the accessibility to real-patient cohorts is highly constrained which limits the development and evaluation of digital biomarkers. In [12], a set of biomarkers for differentiating normal and simulated impairment (SI) movements is proposed. To acquire SI movements, subjects wear a wrist brace for FR, a rubber band for FT, a knee brace for SAW, perform clumsy movements in the FTF, and move one side of their face differently than the other in FA. Surprisingly, without any major modifications, the generalizability of those designs is demonstrated on DNE-113. As the distributions of *DNE biomarkers from SI group share many similar characteristics with the PD/OD group*, the features set fits the purpose of detecting ND patients, and moderately, separating PDs and ODs. Therefore, the use of SI groups or *synthesized* datasets during the development stage is beneficial, especially when it is costly for large-scale real-patient data collection.

The analysis results suggest that aggregating information from weak features will not improve the downstream tasks as polluting the estimator with noisily or irrelevant biomarkers can have negative impacts. Hence, it is important to examine the quality of constructed biomarkers thoroughly before adding them. The use of `pyDNE` streamlines this process.

B. Existing Challenges.

Undoubtedly, the DNE tests which are represented by high-quality biomarkers result in a higher classification performance. The FT, SAW or FTF tests are the *most discriminate*, depending on the task. Note that, this *does not* entirely imply the information from other tests is less significant, as the quality of the design of the biomarkers set is also a crucial factor for reflecting meaningful information captured in the test. Utilizing data-driven feature extraction, up to some extent, can empirically evaluate the gap between automated and feature-engineering digital biomarkers construction, suggesting room for improvement. In **Sec.** VI-B, DNE biomarkers with regular ML models is roughly at the same level as the two deep-learning architectures (CNN [50], [51] and LSTM [52]).

Leveraging a comprehensive, standardized data collection on a broader group of neurological disorders, this study reveals the potential and existing challenges of employing AI/ML for assessing PD severity using a smartphone device. While separating HC and ND groups proves relatively straightforward, providing a detailed justification (e.g., MDS-UPDRS-III score regression) within PD groups remains complex [16]. Here, we highlight the difficulties in distinguishing PD from OD. It is noteworthy to mention that the diagnosis of PD must include a comprehensive clinical history and non-motor examination (**Tab.** 5 of [58]) besides the physical tests assessed in these studies. In essence, the application of these technologies should be approached *cautiously*, serving initially as a tool to measure and document an abnormality but not associate the abnormalities with a specific type of neurological disease.

Differentiating PD versus OD can present similar challenges for humans and machines. Only seeing one or more exam features in isolation is often insufficient. In practice, neurologists combine information from the person's clinical history and the examination to formulate where the symptoms are originating in the nervous system and what are the diagnostic possibilities. Some people have classic symptom histories and exam findings that together easily establish a diagnosis clinically. For others, hands-on exams, additional testing, or monitoring is needed to establish a diagnosis as similarity or overlap in presenting symptoms can occur. For example, early symptoms of hand and arm clumsiness may be seen on an exam as asymmetry of FR or FT. This could be caused by a brain tumor, multiple sclerosis, Parkinson's syndrome, amyotrophic lateral sclerosis, cervical spinal cord compression, or even a mechanical/musculoskeletal condition. Applying the neurological method (history and hypothesis-driven exam) helps to narrow down localizations and possible diagnoses and reconsider alternative potential diagnoses (e.g., multiple system atrophy, essential tremor) [58], [60], [61].

C. Potential Use Cases of DNE.

An initial use case could be a *general neurological exam abnormality detection and documentation tool* that can instantly record and track changes of an individual, healthy aging or experiencing a neurological condition, over time. Individuals will establish their exam baselines and have their exam trends tracked to aid in their care. In the right context, the DNE, and other digital tools, can become for *force multipliers* for efficiently delivering care and performing research studies over a larger population, both in-person and remotely. They open avenues for new care models focused on wellness monitoring, prodromal screening of diseases, and monitoring of patients' conditions and responses to therapeutic interventions.

Clinicians and researchers will incorporate validated digital tools and information into clinical and research workflows as they find useful for specific situations. The DNE biomarkers, as a portion of the contactless screening neurological exam, are already familiar to clinicians. This familiarity, and the DNE measurements' ability to enhance current exam documentation, will facilitate the adoption of the DNE's information.

Finally, this work *validates* and *completes* the proof-of-concept of our comprehensive vision-based DNE solution. To this end, we introduce a common open-source software platform that streamlines the *entire process*, covering (smartphone-based) data acquisition, processing, and analysis. One benefit of this work is reducing the burden on researchers to set up their data collection platform, instead, concentrating more on their scientific questions. This development aligns with the recent momentum in the motion analysis community [15], [62], promoting an efficient and standardized community-driven data collection scheme for rapid data capturing, sharing, and automated analysis. We expect the DNE software can serve as a tool for accelerating the exploration of novel digital biomarkers, and serving a broader spectrum of use cases.

VIII. Conclusion and Future Work

Conclusion. This study empowers the DNE framework [12] with new capabilities (`pyDNE`) and carries out an extensive validation study on *real patients* (DNE-113). Collectively, we demonstrate and evaluate the ability of our DNE framework as a *comprehensive* solution for documenting neurological abnormalities, serving from the distributed data collection, preprocessing, and conducting detailed analysis.

Future Work. While covering a range of disorders, most subjects in DNE-113 only have a single record. Hence, we intend to *longitudinally* expand DNE-113 by periodically collecting DNE records from subjects over several months. This opens the opportunity for tracking the condition of patients over a prolonged duration. Furthermore, it is also necessary to extend the dataset collection to other age, ethnicity, BMI, and sex subgroups, for studying possible biases across population groups. Besides, there is room for enhancement in constructing biomarkers that can sufficiently capture the unique characteristics of PD. Also, novel ML-based data fusion approaches can be developed to aggregate information from multiple tests or explore information from alternative tests in circumstances of receiving incomplete DNE records. This will enable digital tools to evolve better classifying functions.

Acknowledgement

We would like to acknowledge George Heintz for his insightful discussions. This project has been funded by the Jump ARCHES endowment through the Health Care Engineering Systems Center, and the Innovation for Health (IFH) Collaboration with Bradley University and OSF HealthCare.

References

[1] J. Parkinson, "An Essay on the Shaking Palsy," *The Journal of Neuropsychiatry and Clinical Neurosciences*, vol. 14, no. 2, pp. 223–236, May 2002.
[2] C. Marras *et al.*, "Prevalence of Parkinson's disease across North America," *npj Parkinson's Disease*, vol. 4, no. 1, 2018.
[3] A. W. Willis *et al.*, "Incidence of Parkinson disease in North America," *npj Parkinson's Disease*, vol. 8, pp. 1–7, Dec. 2022.
[4] T. Dall *et al.*, "Supply and demand analysis of the current and future us neurology workforce." *Neurology*, vol. 81, no. 5, pp. 470–478, 2013.
[5] J. J. Majersik *et al.*, "A shortage of neurologists – we must act now," *Neurology*, vol. 96, no. 24, pp. 1122–1134, 2021.
[6] N. Kissani *et al.*, "Why does Africa have the lowest number of neurologists and how to cover the gap?" *Journal of the Neurological Sciences*, vol. 434, p. 120119, 2022.
[7] A. Gaffney *et al.*, "Medical Documentation Burden Among US Office-Based Physicians in 2019: A National Study," *JAMA Internal Medicine*, vol. 182, no. 5, pp. 564–566, 05 2022.
[8] J. McConville *et al.*, "A Neurology Advanced Referral Management System (NARMS) Reduces Face-to-Face Consultations By Over Sixty Percent," *The Ulster Medical Journal*, vol. 92, no. 1, pp. 19–23, 2023.

[9] A. Billnitzer *et al.*, "The Clinical Value of Patient Home Videos in Movement Disorders," *Tremor and Other Hyperkinetic Movements*, vol. 11, p. 37, 2021.

[10] M. Al Hussona *et al.*, "The Virtual Neurologic Exam: Instructional Videos and Guidance for the COVID-19 Era," *The Canadian Journal of Neurological Sciences. Le Journal Canadien Des Sciences Neurologiques*, pp. 1–6, 2020.

[11] A. Cohen *et al.*, "The Digital Neurologic Examination," *Digital Biomarkers*, vol. 5, no. 1, pp. 114–126, 04 2021.

[12] T.-H. Hoang *et al.*, "Towards a comprehensive solution for a vision-based digitized neurological examination," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 4020–4031, 2022.

[13] M. Grobe-Einsler *et al.*, "Development of sarahome, a new video-based tool for the assessment of ataxia at home," *Movement Disorders*, vol. 36, no. 5, pp. 1242–1246, 2021.

[14] M. A. Boswell *et al.*, "Smartphone videos of the sit-to-stand test predict osteoarthritis and health outcomes in a nationwide study," *npj Digital Medicine*, vol. 6, no. 1, pp. 1–7, Mar. 2023.

[15] S. D. Uhlrich *et al.*, "OpenCap: Human movement dynamics from smartphone videos," *PLOS Computational Biology*, vol. 19, no. 10, p. e1011462, 2023.

[16] M. S. Islam *et al.*, "Using AI to measure Parkinson's disease severity at home," *npj Digital Medicine*, 2023.

[17] M. H. G. Monje *et al.*, "Remote evaluation of Parkinson's disease using a conventional webcam and artificial intelligence," *Frontiers in Neurology*, vol. 12, 2021.

[18] P. Paruchuri, "ParkinSense: A novel approach to remote idiopathic Parkinson's disease diagnosis, severity profiling, and telemonitoring via ensemble learning and multimodal data fusion on webcam-derived digital biomarkers," *2022 7th International Conference on Intelligent Informatics and Biomedical Science*, vol. 7, pp. 359–366, 2022.

[19] C. G. Goetz *et al.*, "Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): Process, format, and clinimetric testing plan," *Movement Disorders*, vol. 22, no. 1, pp. 41–47, 2007.

[20] H. Li *et al.*, "Automated assessment of parkinsonian finger-tapping tests through a vision-based fine-grained classification model," *Neurocomputing*, vol. 441, pp. 260–271, 2021.

[21] J. H. Shin *et al.*, "Quantitative Gait Analysis Using a Pose-Estimation Algorithm with a Single 2D-Video of Parkinson's Disease Patients," *Journal of Parkinson's Disease*, vol. 11, no. 3, pp. 1271–1283, 2021.

[22] R. Kaur *et al.*, "A vision-based framework for predicting multiple sclerosis and Parkinson's disease gait dysfunctions—a deep learning approach," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 1, pp. 190–201, 2023.

[23] L. F. Gomez *et al.*, "Exploring facial expressions and action unit domains for Parkinson detection," *PLOS ONE*, vol. 18, no. 2, p. e0281248, 2023.

[24] T. Khan *et al.*, "A computer vision framework for finger-tapping evaluation in Parkinson's disease," *Artificial Intelligence in Medicine*, vol. 60, no. 1, pp. 27–40, 2014.

[25] K. Sato *et al.*, "Quantifying normal and Parkinsonian gait features from home movies: Practical application of a deep learning–based 2D pose estimator," *PLOS ONE*, vol. 14, no. 11, p. e0223549, Nov. 2019.

[26] Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease, "The unified Parkinson's disease rating scale (UPDRS): Status and recommendations," *Movement Disorders*, vol. 18, no. 7, pp. 738–750, 2003.

[27] K. G. Sibley *et al.*, "Video-Based Analyses of Parkinson's Disease Severity: A Brief Review," *Journal of Parkinson's Disease*, vol. 11, no. s1, pp. S83–S93, Jul. 2021.

[28] C. Morgan *et al.*, "Automated Real-World Video Analysis of Sit-to-Stand Transitions Predicts Parkinson's Disease Severity," *Digital Biomarkers*, vol. 7, no. 1, pp. 92–103, 08 2023.

[29] G. Vignoud *et al.*, "Video-Based Automated Assessment of Movement Parameters Consistent with MDS-UPDRS III in Parkinson's Disease," *Journal of Parkinson's Disease*, vol. 12, no. 7, pp. 2211–2222, 2022.

[30] M. Lu *et al.*, "Vision-based Estimation of MDS-UPDRS Gait Scores for Assessing Parkinson's Disease Motor Severity," *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, vol. 12263, pp. 637–647, Oct. 2020.

[31] G. Morinan *et al.*, "Computer vision quantification of whole-body Parkinsonian bradykinesia using a large multi-site population," *npj Parkinson's Disease*, vol. 9, no. 1, pp. 1–12, Jan. 2023.

[32] A. Grammatikopoulou *et al.*, "Detecting hypomimia symptoms by selfie photo analysis: For early parkinson disease detection," *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, p. 517–522, 2019.

[33] Z. Cao *et al.*, "Openpose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[34] C. Lugaresi *et al.*, "Mediapipe: A framework for perceiving and processing reality," *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*, 2019.

[35] D. Pavllo *et al.*, "3D human pose estimation in video with temporal convolutions and semi-supervised training," *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[36] Z. Yin *et al.*, "Assessment of Parkinson's disease severity from videos using deep architectures," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 3, pp. 1164–1176, 2022.

[37] Ł. Kidziński *et al.*, "Automatic real-time gait event detection in children using deep neural networks," *PloS one*, vol. 14, no. 1, p. e0211466, 2019.

[38] T. Simon *et al.*, "Hand keypoint detection in single images using multiview bootstrapping," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017.

[39] A. Savitzky *et al.*, "Smoothing and differentiation of data by simplified least squares procedures." *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[40] L. F. Gomez *et al.*, "Improving parkinson detection using dynamic features from evoked expressions in video," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1562–1570, 2021.

[41] I. Guyon *et al.*, "Gene Selection for Cancer Classification using Support Vector Machines," *Machine Learning*, vol. 46, no. 1, pp. 389–422, 2002.

[42] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[43] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," H. Wallach *et al.*, Eds., vol. 32. Curran Associates, Inc., 2019.

[44] S. Masoudnia *et al.*, "Mixture of experts: A literature survey," *Artificial Intelligence Review*, vol. 42, 08 2014.

[45] C. Zhang *et al.*, Eds., *Ensemble Machine Learning: Methods and Applications*. New York, NY: Springer, 2012.

[46] L. Buitinck *et al.*, "API design for machine learning software: experiences from the scikit-learn project," *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122, 2013.

[47] F. M. Miranda *et al.*, "Hiclass: a python library for local hierarchical classification compatible with scikit-learn," *Journal of Machine Learning Research*, vol. 24, no. 29, pp. 1–17, 2023.

[48] H. B. Mann *et al.*, "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other," *The Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50 – 60, 1947.

[49] T. Chen *et al.*, "XGBoost: A scalable tree boosting system," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

[50] A. Krizhevsky *et al.*, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira *et al.*, Eds., vol. 25. Curran Associates, Inc., 2012.

[51] L. Kidziński *et al.*, "Deep neural networks enable quantitative movement analysis using single-camera videos," *Nature Communications*, vol. 11, no. 1, p. 4054, Dec. 2020.

[52] S. Hochreiter *et al.*, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, p. 1735–1780, nov 1997.

[53] L. van der Maaten *et al.*, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.

[54] S. Rupprechter *et al.*, "A clinically interpretable computer-vision based method for quantifying gait in parkinson's disease," *Sensors*, vol. 21, no. 16, 2021.

[55] S. Williams *et al.*, "The discerning eye of computer vision: Can it measure Parkinson's finger tap bradykinesia?" *Journal of the Neurological Sciences*, vol. 416, p. 117003, 2020.

[56] M. Rajnoha *et al.*, "Towards identification of hypomimia in parkinson's disease based on face recognition methods," *2018 10th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, pp. 1–4, 2018.

[57] H. Fröhlich *et al.*, "Leveraging the potential of digital technology for better individualized treatment of parkinson's disease," *Frontiers in Neurology*, vol. 13, 2022.

[58] G. DeMaagd *et al.*, "Parkinson's Disease and Its Management," *Pharmacy and Therapeutics*, vol. 40, no. 8, pp. 504–532, Aug. 2015.

[59] W. Campbell *et al.*, *DeJong's The Neurologic Examination*. Lippincott Williams & Wilkins, 2019.

[60] F. Cardoso, "Difficult Diagnoses in Hyperkinetic Disorders – A Focused Review," *Frontiers in Neurology*, vol. 3, p. 151, Oct. 2012.

[61] M. D. Rajesh Pahwa *et al.*, "Early Diagnosis of Parkinson's Disease: Recommendations From Diagnostic Clinical Guidelines," *The American Journal of Managed Care*, vol. 16, Mar. 2010.
[62] K. Werling *et al.*, "Addbiomechanics: Automating model scaling, inverse kinematics, and inverse dynamics from human motion data through sequential optimization," *bioRxiv*, 2023.
[63] D. P. Kingma *et al.*, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

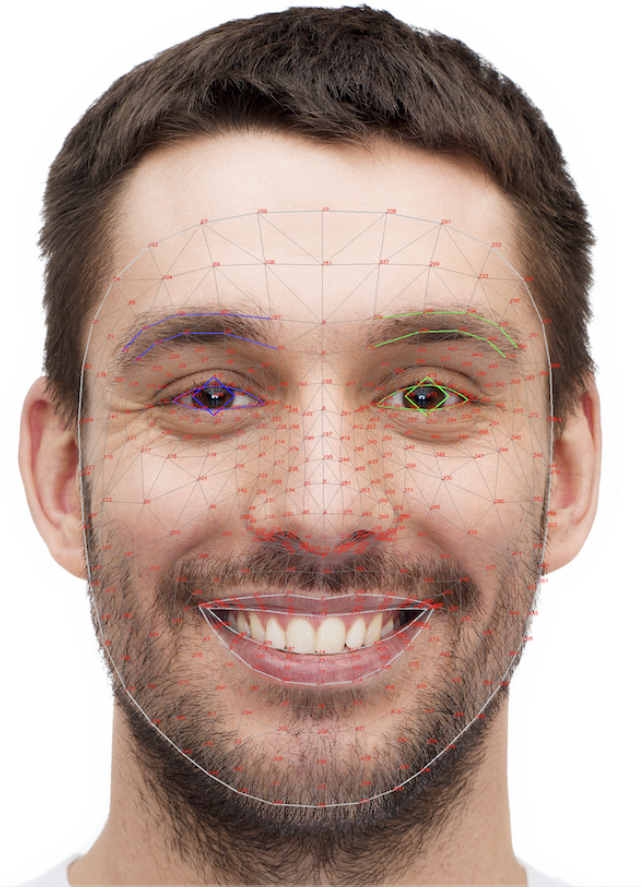


Fig. 6. **Visualization of facial landmark positions in F_3**. MediaPipe [34] introduces the estimation of all landmark positions in 3D. The figure is adopted from [1] visit the attached URL link for the full-size image.

APPENDIX

A. Statistics of DNE Records

In **Tab.** VI, we present a comprehensive overview of the statistical distribution of DNE tests within the dataset records. The predominant number of records feature 1 (8.84%), 4 (25.85%), or 5 (64.63%) DNE tests. Notably, among the DNE records with 4 tests, the stand-up and walk (SAW) test is most frequently absent (36 out of 38) since many subjects in our DNE-113 collection have walking difficulties.

B. A Complete List of DNE Biomarkers

In **Tab.** VII, VIII, IX, X, XI, we provide a complete list of all DNE biomarkers used in the finger tapping (FT), finger-to-finger (FTF), forearm-roll (FR), stand-up and walk (SAW) and facial activation (FA) DNE tests. The full descriptions of biomarkers in the first 4 tests is provided in [12] while the following section introduces the biomarkers of the FA test.

C. The Facial Activation Test

In this section, we introduce the set of DNE digital biomarkers for the FA test, the only DNE test was not formally included in [12] (the biomarkers' name are provided in **Tab.** XI). Following [12], we adopt the notation $\boldsymbol{s}_{k,F_3}[i] \in \mathbb{R}^3$ denoting the k-th keypoint at frame i in the facial landmarks tree F_3 [1] of MediaPipe [34], $k \in \mathcal{K}_{F_3}$ ($\mathcal{K}_{F_3}$ is the indices set of all landmarks). **Fig.** 6 visualizes the location of all F_3 landmarks. To analyze the facial activation tests, we

[1] https://developers.google.com/mediapipe/solutions/vision/face_landmarker

TABLE VI

STATISTICS OF DNE RECORDS IN DNE-113 DATASET. WE PROVIDE THE TOTAL NUMBER OF DNE RECORDS (*DNE record size*) HAVING 1, 2, 3, 4, AND 5 DNE TESTS, RESPECTIVELY. A DETAILED BREAKDOWN OF THE NUMBER OF TESTS IN EACH GROUP IS ALSO INCLUDED. IN DNE-113, A MAJORITY OF DNE RECORDS HAVE EITHER 4 OR 5 TESTS.

DNE record size	Number of DNE records	Number of DNE Tests				
		FT	FTF	FR	SAW	FA
1	13 (8.84%)	1/13	0/13	7/13	4/13	1/13
2	0 (0.00%)	0/0	0/0	0/0	0/0	0/0
3	1 (0.68%)	0/1	1/1	1/1	0/1	1/1
4	38 (25.85%)	36/38	38/38	38/38	2/38	38/38
5	95 (64.63%)	95/95	95/95	95/95	95/95	95/95

TABLE VII

A COMPLETE LIST OF DNE BIOMARKERS FOR THE FT TEST. REFER TO [12] FOR A DETAILED DESCRIPTION.

Index	DNE Biomarker
FT1	Tapping amplitude (R)
FT2	Tapping amplitude (L)
FT3	Tapping amplitude asymmetry
FT4	Tapping period (R)
FT5	Tapping period (L)
FT6	Tapping period asymmetry
FT7	Maximum tapping speed (R)
FT8	Maximum tapping speed (L)
FT9	Maximum tapping speed asymmetry
FT10	Maximum tapping acceleration (R)
FT11	Maximum tapping acceleration (L)
FT12	Maximum tapping acceleration asymmetry

TABLE VIII

A COMPLETE LIST OF DNE BIOMARKERS FOR THE FTF TEST. REFER TO [12] FOR A DETAILED DESCRIPTION.

Index	DNE Biomarker
FTF1	Mean period (L)
FTF2	Mean period (R)
FTF3	STD period (L)
FTF4	STD period (R)
FTF5	Mean speed (L)
FTF6	Mean speed (R)
FTF7	STD speed (L)
FTF8	STD speed (R)
FTF9	Mean path smoothness (L)
FTF10	Mean path smoothness (R)
FTF11	STD path smoothness (L)
FTF12	STD path smoothness (R)
FTF13	Mean velocity angle symmetry (L)
FTF14	Mean velocity angle symmetry (R)
FTF15	STD velocity angle symmetry (L)
FTF16	STD velocity angle symmetry (R)
FTF17	Horizontal finger symmetry
FTF18	Vertical finger symmetry

TABLE IX

A COMPLETE LIST OF DNE BIOMARKERS FOR THE FR TEST. REFER TO [12] FOR A DETAILED DESCRIPTION.

Index	DNE Biomarker
FR1	Rolling amplitude (R)
FR2	Rolling amplitude (L)
FR3	Rolling amplitude asymmetry
FR4	Rolling period (R)
FR5	Rolling period (L)
FR6	Rolling period asymmetry
FR7	Maximum rolling speed (R)
FR8	Maximum rolling speed (L)
FR9	Maximum rolling speed asymmetry
FR10	Maximum rolling acceleration (R)
FR11	Maximum rolling acceleration (L)
FR12	Maximum rolling acceleration asymmetry

TABLE X

A COMPLETE LIST OF DNE BIOMARKERS FOR THE SAW TEST. REFER TO [12] FOR A DETAILED DESCRIPTION.

Index	DNE Biomarker
SAW1	STD step time
SAW2	Mean step length
SAW3	Median step length
SAW4	Mean step width
SAW5	STD step width
SAW6	Median knee angle symmetry
SAW7	Mean knee angle symmetry
SAW8	STD knee angle symmetry
SAW9	Mean step symmetry
SAW10	Mean cadence
SAW11	STD cadence
SAW12	Mean walking speed
SAW13	Mean turning time
SAW14	STD turning time
SAW15	Mean step time

first group neighboring landmarks based on their anatomical position. After being normalized and centralized (in the pre-processing step), the L2 distance is computed between selective groups of keypoints. Let $\boldsymbol{\kappa}, \boldsymbol{\kappa'}$ be the two index vectors of the same size D, $\boldsymbol{\kappa}, \boldsymbol{\kappa}' \in \mathcal{K}_{F_3}^D$, the distance between two groups of landmarks (indexed by $\boldsymbol{\kappa}$ and $\boldsymbol{\kappa}'$) is defined as:

$$d(\boldsymbol{\kappa}, \boldsymbol{\kappa}')[i] = \sqrt{\sum_{d=1}^{D} \|\boldsymbol{s}_{\boldsymbol{\kappa}[d],F_3}[i] - \boldsymbol{s}_{\boldsymbol{\kappa}'[d],F_3}[i]\|_2^2}. \quad (1)$$

We are first interested in studying several distances d_* between two groups of landmarks (**Eq.** 1), listed in **Tab.** XII. For distances that are measured on both the left and right sides, a superscript r and l is used, for clarity. We compute the standard deviation (STD) of the distances $d^l_{\text{oe}}, d^r_{\text{oe}}$ (FA1, FA2), and $d^l_{\text{om}}, d^r_{\text{om}}, d_{\text{mw}}, d_{\text{ja}}$ (FA3, FA4, FA5) and report as DNE features. For FA5, we simply take the average between the values of the left and right sides.

Following, the speed of R/L eyes opening ($d^l_{\text{oe}}, d^r_{\text{oe}}$), eyebrow lifting ($d^l_{\text{ebh}}, d^r_{\text{ebh}}$), and mouth opening $\frac{1}{2}\left(d^l_{\text{om}} + d_{\text{om}}\right)$ are computed by taking the discrete derivative of the corresponding distance $v_*^{l/r} = d_*^{l/r}[i] - d_*^{l/r}[i-1]$. The STD of the computed speeds results in the DNE FA6-FA9 biomarkers.

We next focus on quantifying the range of motion, mathematically defined by the normalized interquartile range (NIQR). Here, we

TABLE XI
A COMPLETE LIST OF DNE BIOMARKERS FOR THE FA TEST. REFER TO **APPENDIX.** C FOR A DETAILED DESCRIPTION.

Index	DNE Biomarker
FA1	STD Eye Opening (R)
FA2	STD Eye Opening (L)
FA3	STD Mouth Opening
FA4	STD Mouth Width
FA5	STD Jaw Opening
FA6	STD Eye Opening Speed (R)
FA7	STD Eye Opening Speed (L)
FA8	STD Eyebrow Lift Speed (R)
FA9	STD Eyebrow Lift Speed (L)
FA10	STD Mouth Opening Speed
FA11	Normalized IQR Eye Opening (R)
FA12	Normalized IQR Eye Opening (L)
FA13	Normalized IQR Mouth Opening
FA14	Normalized IQR Mouth Width
FA15	Eyes Opening Symmetry
FA16	Mouth Opening Symmetry
FA17	Eyes Opening Speed Symmetry
FA18	Mouth Opening Speed Symmetry

introduce the NIQR value of a 1D discrete time series $\boldsymbol{x}$ as:

$$\mathrm{NIQR} = \frac{\mathrm{CDF}^{-1}(0.9; \boldsymbol{x}) - \mathrm{CDF}^{-1}(0.1; \boldsymbol{x})}{\mathrm{CDF}^{-1}(0.9; \boldsymbol{x})},$$

where $\mathrm{CDF}^{-1}(p; \boldsymbol{x})$ is the quartile function, computing the p-th percentile of the series $\boldsymbol{x}$. The use of quartiles is more robust against outliers. In FA tests, the difference between the 0.9 and 0.1 percentile is computed before normalizing by the value of the 0.9 percentile of the given signal. DNE biomarker FA11-FA14 measure the value of NIQR for the distance of two eyes, and mouth (in the vertical/horizontal directions) opening.

Finally, we evaluate the movement symmetry between the left and right sides when opening/closing the mouth, and two eyes (FA15-FA18 DNE biomarkers) as:

$$\begin{aligned} S_{\mathrm{fa}}^{\mathrm{mouth}} &= \mathrm{CC}\left(d_{\mathrm{om}}^{l}, d_{\mathrm{om}}^{r}\right) \\ S_{\mathrm{fa}}^{\mathrm{eyes}} &= \mathrm{CC}\left(d_{\mathrm{oe}}^{l}, d_{\mathrm{oe}}^{r}\right) \\ S_{\mathrm{fa}}^{\mathrm{mouth,v}} &= \mathrm{CC}\left(v_{\mathrm{om}}^{l}, v_{\mathrm{om}}^{r}\right) \\ S_{\mathrm{fa}}^{\mathrm{eyes,v}} &= \mathrm{CC}\left(v_{\mathrm{oe}}^{l}, v_{\mathrm{oe}}^{r}\right), \end{aligned}$$

where CC is the Pearson correlation coefficient. In the case of 1D discrete time series $\boldsymbol{x}_1, \boldsymbol{x}_2$ signal, CC is defined as:

$$\mathrm{CC}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{(\boldsymbol{x}_1 - \bar{\boldsymbol{x}}_1)^T (\boldsymbol{x}_2 - \bar{\boldsymbol{x}}_2)}{\|\boldsymbol{x}_1 - \bar{\boldsymbol{x}}_1\|_2 \|\boldsymbol{x}_2 - \bar{\boldsymbol{x}}_2\|_2}$$

where $\bar{\cdot}$ and $\cdot^T$ denotes the mean and transpose operator. With all the ingredients, the combination of all aforementioned features introduces a complete set of DNE biomarkers for characterizing the FA test.

D. Overview of `pyDNE`

Firstly, all records of the DNE-113 dataset (clinical annotation, and extracted DNE biomarkers from **Sec.** IV) are structured into a common dataloader class, sharing across different analysis tasks, namely `DNEDataset`. An instance of this object allows slicing the data from specific cohorts and returning DNE biomarkers at both record and test levels (e.g., gathering all DNE records from the HC and PD groups). Consequently, the data can be analyzed with a variety of analysis objects. The object `DNEStatisticalTest` provides statistical analysis tests and `DNEFeatureImportanceAnalyzer` conducts recursive feature elimination for picking the most discriminant features. The two objects realize the individual and combined DNE biomarker analysis functionality of `pyDNE`. Several objects in `pyDNE` support the analysis at the downstream task. For example, `DNEClassifier`, and `DNEDeepNetClassifier` handle the classification with various choices of regular ML models and deep-learning (with `PyTorch` [43]) models on DNE-113 dataset, respectively. Users can easily specify the choice of classifiers, and initial parameters via a simple configuration file. The grid search for parameter tuning can also be used, optionally. Visualization utilities, such as plotting the features radial plot, t-SNE [53], or PCA are also available in this package.

E. Examples of `pyDNE` Usages

1) Extracting DNE Biomarkers: We first demonstrate the DNE biomarkers extraction process from the pose or facial landmarks estimation results. We assume the those results are available, stored in the input directory.

```
from pydne.dne_tests.finger_to_finger import
    FingerToFingerVideo
import pandas as pd
# Initialize the FTF Video object
ftf_rec = FingerToFingerVideo(root_dir = '/path-to-
    pose-estimation-folder/',
        video_id = "subj_001/finger_to_finger",
        skip_first_sec=0,
        skip_last_sec = 0,
        log_dir = "dne-logs/")
# Convert to Pandas DataFrame (for display)
df_ftf_rec = pd.DataFrame.from_dict(ftf_rec.features
    , orient="index",
                columns=["Value"])
print(df_ftf_rec)

Output:
                          Value
l_cycle_period_mean    2.078333
r_cycle_period_mean    2.075000
l_cycle_period_std     0.107251
r_cycle_period_std     0.104947
l_mean_speed_mean      3.607417
r_mean_speed_mean      3.685745
l_mean_speed_std       0.807749
r_mean_speed_std       0.766697
l_optimized_path_mean  0.800748
r_optimized_path_mean  0.786693
l_optimized_path_std   0.123934
r_optimized_path_std   0.130408
l_vel_pattern_mean     0.634839
r_vel_pattern_mean     0.668573
l_vel_pattern_std      0.182242
r_vel_pattern_std      0.167678
idx_fin_alig_x         0.966203
idx_fin_alig_y         0.988407
```

2) Accessing DNE Biomarkers: `pyDNE` uses the same presentation for accessing DNE-113 dataset. The following code snippet creates an `DNEDataset` object storing and providing the DNE biomarkers of the FTF test from the HC and PD cohorts (specified by the task `"HC_vs_PD"`; other available options are: `"HC_vs_ND"`, `"HC_vs_OD"`, or `"HC_OD_PD"`). Here, the data and target label are provided via the `dts.data` and `dts.target` objects, correspondingly. With this availability, users can entirely focus on developing their machine-learning models or analyzing the data.

```
from pydne.dataset import DNEDatasets
from pydne.utils import load_dne2_configs

cfgs = load_dne_configs("my_configs.yaml")
# Initialize the DNE dataset object
dts = DNEDataset(cfgs, test = "ftf", task="HC_vs_PD"
    , record_level=False)
# Print dataset statistics
print(dts)
# Get data & labels
```

TABLE XII
DISTANCE BETWEEN SELECTED GROUPS OF FACIAL LANDMARKS OF THE FA TEST. COMPUTED BY **EQ.** 1, REFER TO **FIG.** 6 OR THE DOCUMENTATION OF MEDIAPIPE [34] FOR A MAPPING FROM THE NUMERICAL INDEX OF LANDMARKS TO THE CORRESPONDING ANATOMICAL POSITION.

Distance	Name	κ	κ'
d^l_{oe}	Opening of the eye (L)	[466, 388, 387, 386, 385, 384, 398]	[249, 390, 373, 374, 380, 381, 382]
d^r_{oe}	Opening of the eye (R)	[7, 163, 144, 145, 153, 154, 155]	[246, 161, 160, 159, 158, 157, 173]
d^l_{ebh}	Eyebrow height (L)	[300, 293, 334, 296, 336]	[388, 387, 386, 385, 384]
d^r_{ebh}	Eyebrow height (R)	[70, 63, 105, 66, 107]	[161, 160, 159, 158, 157]
d^l_{om}	Opening of mouth (L)	[88, 178, 87]	[80, 81, 82]
d^r_{om}	Opening of mouth (R)	[317, 402, 318]	[312, 311, 310]
d_{mw}	Mouth width	[78]	[308]
d_{ja}	Opening of jaw	[80, 81, 82, 13, 312, 311, 310]	[170, 140, 171, 175, 396, 369, 395]

```
print("Shape of dts.data:", dts.data.shape)
print("Shape of dts.target:", dts.target.shape)

Output:
+---------------+-----------------+-----------+
|     Stat.     | healthy-control | parkinson |
+---------------+-----------------+-----------+
| No. instances |        41       |     35    |
|  No. subjects |        21       |     33    |
+---------------+-----------------+-----------+
Shape of dts.data: (76, 18)
Shape of dts.target: (76,)
```

3) Running Statistical Tests on DNE Biomarkers: The following code snippet illustrates the use of `DNEMannWhitneyuTest` for performing the Mann-Whitney's U [48] statistical test. A DNE dataset object `DNEDataset` introduced in the previous subsection is first created. Here we take the FTF test (`ftf`) test of HC and PD subjects (`"HC_vs_PD"`) as an example. The $p-$value of all DNE biomarkers is returned as the output.

```
from pydne.datasets import DNEDataset
from pydne.statistical_tests import
    DNEMannWhitneyuTest
from pydne.utils import load_dne2_configs

cfgs = load_dne_configs("my_configs.yaml")
# Initialize the DNE dataset object
dts = DNEDataset(cfgs, test = "ftf", mode="HC_vs_PD"
    , record_level=False)
# Perform the Mann-Whitney U test on this dataset
stat_test = DNEMannWhitneyuTest(dts)
stat_test_results = stat_test.run_test()
print(stat_test_results)

Output:
                    features       p-value  rank

                 Mean period (L)  3.4e-01    13
                 Mean period (R)  1.1e-01     8
                  STD period (L)  5.1e-02     6
                  STD period (R)  9.8e-03     4
                  Mean speed (L)  1.5e-01     9
                  Mean speed (R)  2.1e-01    10
                   STD speed (L)  5.7e-01    14
                   STD speed (R)  7.1e-01    16
        Mean path smoothness (L)  7.3e-01    17
        Mean path smoothness (R)  6.0e-01    15
         STD path smoothness (L)  9.1e-05     2
         STD path smoothness (R)  6.9e-03     3
Mean velocity angle symmetry (L)  9.3e-02     7
Mean velocity angle symmetry (R)  2.6e-01    11
 STD velocity angle symmetry (L)  9.1e-01    18
 STD velocity angle symmetry (R)  3.3e-01    12
      Horizontal finger symmetry  5.2e-09     1
        Vertical finger symmetry  3.6e-02     5
```

4) Classification with DNE Biomarkers: We introduce an example of using `DNEClassifier` for conducting binary classification experiments (linear regression `"LR"` and radial basis kernel support vector machine model `"RSVM"`) using DNE biomarkers from the FT test. The model parameters are also provided via `"model_params"`. Additionally, the training/testing split is loaded from an external `JSON` file for consistency across all experiments.

```
from pydne.utils import load_dne2_configs
from pydne.classifiers import DNEClassifier
import pandas as pd

cfgs = load_dne_configs("my_configs.yaml")
# Split file (for K-fold classification), containing
     recording ids
split_file = "splits/seed_0_5folds_ft_HC_vs_PD.json"
# Initialize DNE Classifier object
dneclsf = DNEClassifier(
        cfgs, seed=0, split_file=split_file, n_folds
    =5,
        model_params={
            "LR":{ "max_iter": 100, "C": 0.1},
            "RSVM": {"kernel": "rbf", "C": 0.1}
        },
        grid_search=False
        )

eval_results = {}
# Try Linear Regression (LR) and RBF Kernel Support
    Vector Machine (RSVM)
for clsf_name in ["LR", "RSVM"]:
    # Do training/ testing. Predictions on testing
    set is returned
    y, y_preds = dneclsf.do_experiment(
        clsf_name = clsf_name,
        test = "ft",
        mode = "HC_vs_PD",
        fold = 0,
        split_file = split_file
        )
    # Evaluation on the testing set
    eval_results[clsf_name] = dneclsf.do_evaluation(
    y_preds, y)

# Convert to Pandas DataFrame (for display)
eval_df = pd.DataFrame.from_dict(eval_results,
    orient="index")
print(eval_df)

Output:
      acc      prec    recall      spec       auc
LR   0.83333  0.85714  0.75000  0.90000  0.82500
RSVM 0.66667  0.57143  0.57143  0.72727  0.64935
```

F. Feature Importance Analysis with `pyDNE`

The example below demonstrates the use of `DNEFeatureImportanceAnalyzer` for conducting recursive feature elimination to find the most discriminant biomarkers.

```
from pydne.utils import load_dne_configs
from pydne.feature_selection import
    DNEFeatureImportanceAnalyzer
from pydne.datasets import DNEDataset

cfgs = load_dne_configs("my_configs.yaml")

# Initialize the DNE dataset object
dts = DNEDataset(cfgs, test="ft", mode="HC_PD_OD")

# Initialize DNE feature importance analyzer object
alzr = DNEFeatureImportanceAnalyzer(cfgs, dts,
    out_dir="out_dir")
selected_features, _, __ = alzr.
    run_feature_selection(
                    estimator="gbc",
                    criteria="accuracy",
                    verbose = True)

Output:
[Score]: 0.53  -> 0.545
[No. Features]: 12 -> 10
                                features
index
FT1               Tapping amplitude (R)
FT2               Tapping amplitude (L)
FT3         Tapping amplitude asymmetry
FT4                  Tapping period (R)
FT5                  Tapping period (L)
FT6            Tapping period asymmetry
FT7           Maximum tapping speed (R)
FT8           Maximum tapping speed (L)
FT9     Maximum tapping speed asymmetry
FT10   Maximum tapping acceleration (R)
```

G. Additional Experiment Results

1) Individual Biomarker Analysis: Following the setup introduced in the *individual biomarker quality analysis* (**Sec.** VI-A), we provide the full assessment over all DNE biomarkers in **Tab.** XIII, XIV, XV, XVI, XVII, each table contains the results of a single DNE test.

2) Feature Important Analysis: In **Tab.** XVIII, we provide the list of all selected DNE biomarkers after performing the recursive feature elimination analysis. The three-class (HC, PD, and OD) classification task is performed, using GBM as the base classifier, and classification accuracy is used as the scoring function. The biomarkers listed here are used for the t-SNE [53] visualization.

H. Implementation Details

1) Classification with ML Model: The `scikit-learn` [42] and `XGBoost` [1] library provide the implementation of all ML models. For hyper-parameters tuning, an exhaustive grid search over parameter values is performed. The list of all model hyper-parameters is provided in **Tab.** XIX.

2) Classification with Deep-learning Models: `PyTorch` is used for the implementation of CNN and LSTM models, and trained on a single RTX 3090 GPU. The detailed architecture of the two models is introduced in the Appendix of [12]. During the training process, we used binary cross entropy as the loss function, utilizing the Adam optimizer [63]. The learning rate is chosen from $\{1e^{-3}, 5e^{-3}\}$, and scheduled by a step learning rate decay with a factor of $\gamma = 0.5$. The time series of all major keypoints are segmented into a sequence of $W \in 100, 200$ frames, depending on the characteristic of the test. A batch size of 64 samples is applied for all experiments.

[1] `https://xgboost.readthedocs.io/en/stable/install.html`

TABLE XIII
INDIVIDUAL DISCRIMINATING POWER OF FT DNE BIOMARKERS IN THE BINARY CLASSIFICATION TASKS.

Index	DNE Biomarker	HC v.s. PD		HC v.s. ND		PD v.s. OD	
		p-value	Rank	p-value	Rank	p-value	Rank
FT1	Tapping amplitude (R)	$3.12e^{-04}$	7	$7.16e^{-03}$	9	$3.45e^{-02}$	3
FT2	Tapping amplitude (L)	$4.2e^{-04}$	8	$3.27e^{-03}$	8	$6.4e^{-02}$	6
FT3	Tapping amplitude asymmetry	$1.37e^{-04}$	6	$5.17e^{-04}$	6	$3.37e^{-01}$	11
FT4	Tapping period (R)	$1.94e^{-05}$	3	$1.32e^{-06}$	2	$3.62e^{-01}$	12
FT5	Tapping period (L)	$2.26e^{-05}$	4	$7.14e^{-06}$	3	$2.04e^{-01}$	8
FT6	Tapping period asymmetry	$2.05e^{-06}$	1	$1.27e^{-07}$	1	$3.37e^{-01}$	10
FT7	Maximum tapping speed (R)	$3.21e^{-03}$	10	$2.78e^{-02}$	11	$4.57e^{-02}$	4
FT8	Maximum tapping speed (L)	$4.00e^{-03}$	11	$3.67e^{-02}$	12	$3.13e^{-02}$	2
FT9	Maximum tapping speed asymmetry	$9.02e^{-04}$	9	$1.14e^{-03}$	7	$1.71e^{-01}$	7
FT10	Maximum tapping acceleration (R)	$2.89e^{-06}$	2	$2.41e^{-04}$	4	$4.63e^{-03}$	1
FT11	Maximum tapping acceleration (L)	$1.33e^{-04}$	5	$3.27e^{-04}$	5	$4.68e^{-02}$	5
FT12	Maximum tapping acceleration asymmetry	$1.19e^{-02}$	12	$2.08e^{-02}$	10	$3.04e^{-01}$	9

TABLE XIV
INDIVIDUAL DISCRIMINATING POWER OF FTF DNE BIOMARKERS IN THE BINARY CLASSIFICATION TASKS.

Index	DNE Biomarker	HC v.s. PD		HC v.s. ND		PD v.s. OD	
		p-value	Rank	p-value	Rank	p-value	Rank
FTF1	Mean period (L)	$3.4e^{-01}$	13	$1.54e^{-01}$	7	$2.39e^{-03}$	3
FTF2	Mean period (R)	$1.15e^{-01}$	8	$4.19e^{-01}$	11	$9.55e^{-04}$	2
FTF3	STD period (L)	$5.09e^{-02}$	6	$8.59e^{-01}$	18	$7.93e^{-03}$	4
FTF4	STD period (R)	$9.79e^{-03}$	4	$5.45e^{-01}$	15	$5.11e^{-04}$	1
FTF5	Mean speed (L)	$1.5e^{-01}$	9	$3.23e^{-01}$	9	$3.94e^{-01}$	15
FTF6	Mean speed (R)	$2.13e^{-01}$	10	$4.39e^{-01}$	12	$2.75e^{-01}$	11
FTF7	STD speed (L)	$5.68e^{-01}$	14	$7.02e^{-01}$	17	$1.68e^{-01}$	9
FTF8	STD speed (R)	$7.14e^{-01}$	16	$2.18e^{-01}$	8	$3.48e^{-01}$	13
FTF9	Mean path smoothness (L)	$7.33e^{-01}$	17	$6.3e^{-01}$	16	$1.87e^{-01}$	10
FTF10	Mean path smoothness (R)	$6.03e^{-01}$	15	$3.32e^{-01}$	10	$5.26e^{-01}$	16
FTF11	STD path smoothness (L)	$9.11e^{-05}$	2	$1.23e^{-04}$	2	$1.49e^{-01}$	8
FTF12	STD path smoothness (R)	$6.88e^{-03}$	3	$8.22e^{-02}$	6	$8.15e^{-03}$	5
FTF13	Mean velocity angle symmetry (L)	$9.30e^{-02}$	7	$2.2e^{-04}$	3	$5.01e^{-02}$	7
FTF14	Mean velocity angle symmetry (R)	$2.64e^{-01}$	11	$3.97e^{-03}$	4	$2.08e^{-02}$	6
FTF15	STD velocity angle symmetry (L)	$9.12e^{-01}$	18	$5.36e^{-01}$	14	$3.24e^{-01}$	12
FTF16	STD velocity angle symmetry (R)	$3.31e^{-01}$	12	$5.21e^{-01}$	13	$3.63e^{-01}$	14
FTF17	Horizontal finger symmetry	$5.22e^{-09}$	1	$1.87e^{-12}$	1	$9.51e^{-01}$	18
FTF18	Vertical finger symmetry	$3.6e^{-02}$	5	$1.91e^{-02}$	5	$8.08e^{-01}$	17

TABLE XV
INDIVIDUAL DISCRIMINATING POWER OF FR DNE BIOMARKERS IN THE BINARY CLASSIFICATION TASKS.

Index	DNE Biomarker	HC v.s. PD		HC v.s. ND		PD v.s. OD	
		p-value	Rank	p-value	Rank	p-value	Rank
FR1	Rolling amplitude (R)	$2.58e^{-01}$	12	$7.46e^{-02}$	12	$6.17e^{-01}$	5
FR2	Rolling amplitude (L)	$3.83e^{-02}$	10	$1.68e^{-02}$	11	$6.30e^{-01}$	6
FR3	Rolling amplitude asymmetry	$1.64e^{-04}$	3	$7.04e^{-04}$	8	$1.71e^{-01}$	1
FR4	Rolling period (R)	$7.25e^{-10}$	1	$1.93e^{-15}$	1	$6.64e^{-01}$	8
FR5	Rolling period (L)	$8.39e^{-10}$	2	$3.53e^{-15}$	2	$6.30e^{-01}$	7
FR6	Rolling period asymmetry	$3.45e^{-03}$	8	$1.08e^{-05}$	5	$4.43e^{-01}$	3
FR7	Maximum rolling speed (R)	$9.04e^{-02}$	11	$1.62e^{-02}$	10	$9.28e^{-01}$	9
FR8	Maximum rolling speed (L)	$2.76e^{-02}$	9	$4.31e^{-03}$	9	$9.73e^{-01}$	10
FR9	Maximum rolling speed asymmetry	$3.92e^{-04}$	4	$3.38e^{-04}$	7	$3.68e^{-01}$	2
FR10	Maximum rolling acceleration (R)	$6.77e^{-04}$	6	$2.86e^{-06}$	4	$9.81e^{-01}$	11
FR11	Maximum rolling acceleration (L)	$9.47e^{-04}$	7	$1.36e^{-06}$	3	$9.89e^{-01}$	12
FR12	Maximum rolling acceleration asymmetry	$5.71e^{-04}$	5	$4.58e^{-05}$	6	$5.58e^{-01}$	4

TABLE XVI
INDIVIDUAL DISCRIMINATING POWER OF SAW DNE BIOMARKERS IN THE BINARY CLASSIFICATION TASKS.

Index	DNE Biomarker	HC v.s. PD		HC v.s. ND		PD v.s. OD	
		p-value	Rank	p-value	Rank	p-value	Rank
SAW1	STD step time	$5.16e^{-04}$	9	$9.51e^{-07}$	4	$4.14e^{-01}$	6
SAW2	Mean step length	$4.84e^{-09}$	2	$2.24e^{-12}$	1	$3.24e^{-01}$	4
SAW3	Median step length	$4.44e^{-09}$	1	$2.91e^{-12}$	2	$4.14e^{-01}$	7
SAW4	Mean step width	$1.81e^{-03}$	11	$1.53e^{-06}$	8	$9.76e^{-03}$	1
SAW5	STD step width	$1.79e^{-01}$	13	$5.72e^{-02}$	14	$5.17e^{-01}$	9
SAW6	Median knee angle symmetry	$2.30e^{-04}$	7	$1.28e^{-06}$	7	$2.84e^{-01}$	2
SAW7	Mean knee angle symmetry	$3.37e^{-04}$	8	$1.23e^{-06}$	6	$3.06e^{-01}$	3
SAW8	STD knee angle symmetry	$2.4e^{-01}$	14	$4.51e^{-02}$	13	$6.24e^{-01}$	10
SAW9	Mean step symmetry	$3.79e^{-05}$	4	$1.58e^{-05}$	10	$3.61e^{-01}$	5
SAW10	Mean cadence	$1.84e^{-04}$	6	$3.50e^{-06}$	9	$6.76e^{-01}$	12
SAW11	STD cadence	$6.67e^{-01}$	15	$4.68e^{-01}$	15	$6.41e^{-01}$	11
SAW12	Mean walking speed	$1.32e^{-08}$	3	$1.35e^{-11}$	3	$9.95e^{-01}$	15
SAW13	Mean turning time	$6.51e^{-05}$	5	$1.14e^{-06}$	5	$9.28e^{-01}$	14
SAW14	STD turning time	$1.34e^{-01}$	12	$2.21e^{-02}$	12	$7.12e^{-01}$	13
SAW15	Mean step time	$1.49e^{-03}$	10	$2.26e^{-05}$	11	$4.86e^{-01}$	8

TABLE XVII
INDIVIDUAL DISCRIMINATING POWER OF FA DNE BIOMARKERS IN THE BINARY CLASSIFICATION TASKS.

Index	DNE Biomarker	HC v.s. PD		HC v.s. ND		PD v.s. OD	
		p-value	Rank	p-value	Rank	p-value	Rank
FA1	STD Eye Opening (R)	$5.93e^{-05}$	8	$4.83e^{-03}$	10	$1.58e^{-03}$	3
FA2	STD Eye Opening (L)	$6.15e^{-05}$	9	$5.86e^{-03}$	11	$1.39e^{-03}$	2
FA3	STD Mouth Opening	$4.21e^{-08}$	1	$1.55e^{-10}$	1	$5.36e^{-01}$	16
FA4	STD Mouth Width	$6.7e^{-04}$	11	$3.29e^{-02}$	15	$5.68e^{-03}$	9
FA5	STD Jaw Opening	$1.13e^{-03}$	12	$9.14e^{-03}$	12	$1.05e^{-01}$	13
FA6	STD Eye Opening Speed (R)	$3.55e^{-05}$	6	$3.53e^{-03}$	8	$1.63e^{-03}$	4
FA7	STD Eye Opening Speed (L)	$3.18e^{-05}$	5	$1.50e^{-03}$	6	$2.31e^{-03}$	5
FA8	STD Eyebrow Lift Speed (R)	$1.37e^{-05}$	3	$1.81e^{-03}$	7	$1.14e^{-03}$	1
FA9	STD Eyebrow Lift Speed (L)	$4.6e^{-05}$	7	$3.58e^{-03}$	9	$5.85e^{-03}$	10
FA10	STD Mouth Opening Speed	$1.19e^{-02}$	17	$2.47e^{-02}$	14	$2.34e^{-01}$	14
FA11	Normalized IQR Eye Opening (R)	$6.97e^{-03}$	15	$2.9e^{-01}$	17	$3.24e^{-03}$	6
FA12	Normalized IQR Eye Opening (L)	$2.69e^{-02}$	18	$6.11e^{-01}$	18	$4.e^{-03}$	8
FA13	Normalized IQR Mouth Opening	$1.49e^{-03}$	13	$5.20e^{-05}$	4	$7.79e^{-01}$	18
FA14	Normalized IQR Mouth Width	$4.35e^{-03}$	14	$2.4e^{-02}$	13	$1.86e^{-02}$	11
FA15	Eye Opening Symmetry	$7.81e^{-03}$	16	$4.19e^{-02}$	16	$9.33e^{-02}$	12
FA16	Mouth Opening Symmetry	$4.15e^{-06}$	2	$7.73e^{-08}$	2	$7.01e^{-01}$	17
FA17	Eye Opening Speed Symmetry	$1.48e^{-05}$	4	$3.56e^{-04}$	5	$3.77e^{-03}$	7
FA18	Mouth Opening Speed Symmetry	$9.43e^{-05}$	10	$1.30e^{-05}$	3	$3.02e^{-01}$	15

TABLE XVIII
LIST OF ALL SELECTED BIOMARKERS BY COMBINED FEATURE IMPORTANCE ANALYSIS.

Test	Selected Biomarkers
FT	FT1, FT2, FT3, FT4, FT5, FT6, FT7, FT8, FT9, FT10
FTF	FTF1, FTF2, FTF4, FTF5, FTF6, FTF7, FTF8, FTF9, FTF11, FTF12, FTF13, FTF14, FTF16, FTF17, FTF18
FR	FR1, FR3, FR4, FR5, FR6, FR9, FR10, FR11, FR12
SAW	SAW1, SAW2, SAW3, SAW4, SAW5, SAW6, SAW8, SAW11, SAW12, SAW13
FA	FA3, FA5, FA6, FA7, FA8, FA9, FA10, FA11, FA12, FA14, FA16, FA17, FA18

TABLE XIX
PARAMTERS FOR ML MODEL GRID-SEARCH

Test	Model	Parameters
FT	RF	class weight $\in \{'balanced', None\}$; max depth $\in \{3, 10, 20\}$; n estimators $\in \{10, 30, 100\}$
	GBM	learning rate $\in \{0.1, 0.2, 0.5\}$; max depth $\in \{2, 4, 5\}$; max features $\in \{'sqrt', 'log2', None\}$; min samples leaf $\in \{1, 3\}$; min samples split $\in \{2, 4\}$; n estimators $\in \{50, 150\}$
	XGBoost	colsample bytree $\in \{0.7, 1.0\}$; gamma $\in \{0.5, 2.0\}$; max depth $\in \{5, 8\}$; min child weight $\in \{1, 2\}$; subsample $\in \{0.5, 0.8\}$
	LR	C $\in \{1.0, 10.0\}$; max iter $\in \{1000, 2000\}$
	RSVM	C $\in \{0.1, 0.5, 3.0, 4.0\}$; gamma $\in \{'scale', 'auto'\}$; kernel $\in \{'rbf'\}$; max iter $\in \{100, 500\}$
	MLP	alpha $\in \{0.1, 1.0, 10.0\}$; hidden layer sizes $\in \{[5, 6], [10, 12], [20, 10]\}$; max iter $\in \{2000, 5000\}$
FTF	RF	class weight $\in \{'balanced', None\}$; max depth $\in \{5, 10, 20\}$; n estimators $\in \{25, 50, 100\}$
	GBM	learning rate $\in \{0.1, 0.5\}$; max depth $\in \{5, 20\}$; max features $\in \{'log2'\}$; min samples leaf $\in \{5, 10\}$; min samples split $\in \{5, 10\}$; n estimators $\in \{200, 600\}$
	XGBoost	colsample bytree $\in \{0.7, 1.0\}$; gamma $\in \{0.5, 2.0\}$; max depth $\in \{5, 8\}$; min child weight $\in \{1, 2\}$; subsample $\in \{0.5, 0.8\}$
	LR	C $\in \{0.01, 0.1, 1.0, 10.0\}$; max iter $\in \{1000, 2000\}$
	RSVM	C $\in \{0.1, 1.0, 10.0\}$; gamma $\in \{'scale', 'auto'\}$; kernel $\in \{'rbf'\}$; max iter $\in \{500\}$
	MLP	alpha $\in \{0.01, 0.1, 1.0\}$; hidden layer sizes $\in \{10, 20\}$; max iter $\in \{2000\}$
FR	RF	class weight $\in \{'balanced', None\}$; max depth $\in \{3, 10, 20\}$; n estimators $\in \{10, 30, 100\}$
	GBM	learning rate $\in \{0.1, 0.2, 0.5\}$; max depth $\in \{2, 4, 5\}$; max features $\in \{'sqrt', 'log2', None\}$; min samples leaf $\in \{1, 3\}$; min samples split $\in \{2, 4\}$; n estimators $\in \{50, 150\}$
	XGBoost	colsample bytree $\in \{0.7, 1.0\}$; gamma $\in \{0.5, 2.0\}$; max depth $\in \{5, 8\}$; min child weight $\in \{1, 2\}$; subsample $\in \{0.5, 0.8\}$
	LR	C $\in \{0.01, 0.05, 0.1\}$; max iter $\in \{100, 200, 300\}$
	RSVM	C $\in \{0.1, 0.5, 3.0, 4.0\}$; gamma $\in \{'scale', 'auto'\}$; kernel $\in \{'rbf'\}$; max iter $\in \{100, 500\}$
	MLP	alpha $\in \{0.001, 0.005, 0.01\}$; hidden layer sizes $\in \{[10, 12], [20, 10], [20, 50, 20], [80, 200, 80]\}$; max iter $\in \{500, 1000\}$
SAW	RF	class weight $\in \{'balanced', None\}$; max depth $\in \{3, 10, 20\}$; n estimators $\in \{10, 30, 100\}$
	GBM	learning rate $\in \{0.1, 0.01\}$; max depth $\in \{20\}$; max features $\in \{'log2', 'sqrt'\}$; min samples leaf $\in \{5, 10\}$; min samples split $\in \{5, 10\}$; n estimators $\in \{500\}$
	XGBoost	colsample bytree $\in \{0.7, 1.0\}$; gamma $\in \{0.5, 2.0\}$; max depth $\in \{5, 8\}$; min child weight $\in \{1, 2\}$; subsample $\in \{0.5, 0.8\}$
	LR	C $\in \{0.01, 0.1, 1.0, 10.0\}$; max iter $\in \{1000, 2000\}$
	RSVM	C $\in \{0.1, 1.0\}$; gamma $\in \{'scale', 'auto'\}$; kernel $\in \{'rbf'\}$; max iter $\in \{500\}$
	MLP	activation $\in \{'relu', 'tanh'\}$; alpha $\in \{1.0, 10.0\}$; hidden layer sizes $\in \{10, 20\}$; learning rate $\in \{'adaptive', 'constant'\}$; max iter $\in \{3000\}$; solver $\in \{'adam'\}$
FA	RF	class weight $\in \{'balanced', None\}$; max depth $\in \{3, 10, 20\}$; n estimators $\in \{10, 30, 100\}$
	GBM	learning rate $\in \{0.1, 0.2, 0.5\}$; max depth $\in \{2, 4, 5\}$; max features $\in \{'sqrt', 'log2', None\}$; min samples leaf $\in \{1, 3\}$; min samples split $\in \{2, 4\}$; n estimators $\in \{50, 150\}$
	XGBoost	colsample bytree $\in \{0.7, 1.0\}$; gamma $\in \{0.5, 2.0\}$; max depth $\in \{5, 8\}$; min child weight $\in \{1, 2\}$; subsample $\in \{0.5, 0.8\}$
	LR	C $\in \{0.01, 0.05, 0.1\}$; max iter $\in \{100, 200, 300\}$
	RSVM	C $\in \{0.1, 0.5, 3.0, 4.0\}$; gamma $\in \{'scale', 'auto'\}$; kernel $\in \{'rbf'\}$; max iter $\in \{100, 500\}$
	MLP	alpha $\in \{0.001, 0.005, 0.01\}$; hidden layer sizes $\in \{[10, 12], [20, 10], [20, 50, 20], [80, 200, 80]\}$; max iter $\in \{500, 1000\}$