 doi:10.20944/preprints202103.0542.v1

Article

Sensor Validation and Diagnostic Potential of Smartwatches in Movement Disorders

Julian Varghese [1*], Catharina Marie van Alen [2], Michael Fujarski [1], Tobias Warnecke [3], and Christine Thomas [2]

[1] Institute of Medical Informatics, University of Münster, Germany; julian.varghese@uni-muenster.de
[2] Institute of Geophysics, University of Münster, Germany; cthom_01@uni-muenster.de
[3] Department of Neurology, University Hospital Münster, Germany; tobias.warnecke@ukmuenster.de

* Correspondence: julian.varghese@uni-muenster.de;

Abstract: Smartwatches provide technology-based assessments in Parkinson's disease (PD). We present results for sensor validation and disease classification via Machine Learning (ML). A comparison setup was designed with two different series of Apple smartwatches, one Nanometrics seismometer and a high-precision shaker to measure tremor-like amplitudes and frequencies. Clinical smartwatch measurements were acquired from a prospective study including 450 participants with PD, differential diagnoses (DD) and healthy participants. All participants wore two smartwatches and within a 15-min examination. Symptoms and medical history were captured on the paired smartphone. A broad range of different ML classifiers were cross-validated. Amplitude and frequency differences between smartwatches and the seismometer were under the level of clinical significance. The most advanced task of distinguishing PD vs DD was evaluated with 74,1% balanced accuracy, 86,5% precision and 90,5% recall by Multilayer Perceptrons. Deep Learning architectures significantly underperformed in all classification tasks. Smartwatches are capable of capturing subtle-tremor signs with low noise. This study provided the largest PD sample size of two-hand smartwatch measurements and our preliminary ML-evaluation shows that such a system provides powerful means for diagnosis classification and new digital biomarkers but it remains challenging for distinguishing similar disorders.

Keywords: Smartwatches, Artificial Intelligence, Movement Disorders, Parkinson's Disease

1. Introduction

Parkinson's Disease (PD) is the second-most neurodegenerative disorder – following Alzheimer Dementia - and worldwide burden that has more than doubled over the last two decades [1]. Early and accurate diagnoses improve quality of life, reduce work losses, which is why missed diagnoses mean missed opportunities [2]. Currently, PD diagnosis is primarily based on clinical assessment, which is challenging and associated with overall misclassification rates of around 20%-30%; Rizzo et al. 2016 conducted a meta-analysis and reported pooled diagnostic accuracy of 73.8% for general practitioners or general neurologists with 95% Credible Interval (CRI) of 67.8%-79.6%.

Clinical assessment may not identify subtle changes in movement pathologies as e.g. weak tremor, its frequency or slowness of movements [3]. Regarding diagnostic accuracy and treatment monitoring, there is a strong need for new technological objective biomarkers, which are capable of capturing these subtleties with high precision and machine-readable [4]. In the era of digital transformation of healthcare, consumer wearables with multi-sensor technology provide a source of objective movement monitoring allowing for greater precision in recording subtle changes unlike current clinical rating scales in hospital routine [5]. Though there is an increasing number of such wearables and mobile apps or even mature medical devices as the Parkinson's KinetiGraph™ system, there is a low number of large-scale deployments [6].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 March 2021 doi:10.20944/preprints202103.0542.v1

Regarding PD, some systems have shown promising diagnostic potential when analyzing voice, hand movements, gait, facial expressions, eye movements and balance [7–12]. Most of these promising examples used Machine Learning approaches for disease classification. However, the reported accuracies need to be taken with high caution because the implemented models were trained and tested on low sample sizes regarding PD (n<<100), which carries a high risk of overfitting. Moreover, we could not find any approach that include similar movement disorders as an important control group for differential diagnoses. A simple classification model that only differentiates between PD and Healthy is of only limited clinical use as it was only trained and tested between those classes and thus might have only learned to identify general movement anomalies, which differ from the healthy population but do not represent Parkinson-specific features. Hence, such models could misclassify other movement disorders as Multiple Sclerosis or Essential Tremor. Also, in clinical reality, the Health Practitioner or the Neurologist cannot initially assume whether the patient is either healthy or has PD. Therefore, classification models for potential diagnosis should consider differential diagnoses.

Our research focusses on acceleration-based hand movement analyses using the Smart Device System (SDS) that is currently part of a prospective human subject trial [13]. The study has recruited and measured >350 participants and has generated one of the largest databases for PD, differential diagnoses and healthy subjects with acceleration data from a neurological examination including left and right side of the body and structured clinical data on non-motor symptoms (e.g. sleep disturbances, loss of smell, depression). The system includes simple consumer devices by Apple, utilizing smartwatches to capture acceleration and a paired smartphone for clinical data. To our knowledge, official information on the smartwatch raw measurement accuracy is not publicly available. Therefore, the devices were evaluated by a systematic comparison with a gold standard utilizing a broadband seismometer.

Apart from this sensor validation, the SDS is integrated into a neurological examination. It consists of ten steps to monitor and provoke specific movement characteristics such as tremor or slowness of movements. While she study is still running till the end of 2021 and includes further smart devices data as tablet-based drawing and voice analyses, this manuscript aims to focus on following research aims:

- Sensor validation to measure the precision of smartwatches regarding acceleration amplitudes and tremor frequencies. As goldstandard, we conducted a comparison experiment utilizing a seismometer and a high-precision shaker. As a result, we assessed the level of precision regarding the smartwatches. This is particularly useful in case of subtle tremors, which have acceleration amplitudes of <0.05 g and are hard to capture by human vision.
- Timeseries features were extracted based on expert-based feature engineering and literature data. A broad range of Machine Learning models was trained and cross-validated to assess classification performances. To complement the expert-based feature engineering by a pure automatic feature extraction method, a Deep-Learning Neural Networks with the raw time series data as input were trained and cross-validated as well.

2. Materials and Methods

2.1. Study setting

The prospective study started in 2018 and was extended till the end of 2021. It received approval by the ethical board of the University of Münster and the physician's chamber of Westphalia-Lippe (Reference number: 2018-328-f-S). It is being conducted at the outpatient clinic of Movement Disorders at the University Hospital Münster in Germany. The details of the study design and the protocol have been published previously [13]. Study registration ID on ClinicalTrials.gov: NCT03638479.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 March 2021 doi:10.20944/preprints202103.0542.v1

Table 1 lists participants population characteristics. Further information on demographics, differential diagnoses is provided for each sample in the supplement Patient-Population. All diagnoses were confirmed by neurologists and finally reviewed by one senior movement disorder expert.

Each participant wore two smartwatches, one on each wrist, while seated in an armchair and following a pre-defined neurological examination, which was instructed by a study-nurse. This examination was designed by movement disorder experts with the primary aim to establish a simple to follow examination in order to capture most-relevant acceleration characteristics. The data included smartwatch-acceleration data and further clinical data encompassing non-motor symptoms based on the Parkinson's Non-motor Symptoms Questionnaire [14] – on the paired smartphone . Each examination took 15 minutes per participant on average. Each assessment step is summarized is Table 2.

Table 1. Participant population. DD: Differential diagnoses including movement disorders other than PD as Essential Tremor, Atypical Parkinsonism, secondary causes of Parkinsonism and Dystonia, Multiple Sclerosis.

Disease Class	Sample Size	Average Age (SD)
PD	260	66.26 (9.61)
DD	101	60.82 (12.87)
Healthy	89	61.45 (10.63)

Table 2. Smartwatch-based Examination steps.

Step	Duration (s)	Description
1a	20	**Rest tremor.** Participant is seated with his eyes closed in resting position, positioning standardized to Zhang et al. [15]
1b	20	**Rest tremor in a stressful situation**. Participant starts from 100, subtracts 7, and stops after five answers.
2	10	**Lift and extend arms** according to Zhang et al. [15]
3	10	**Remain arms lifted.**
4	10	**Hold 1kg** weight in every hand for 5 seconds. Start with the right hand. Then, have the participant's arm rested again as in 1a.
5	10	**Finger pointing.** Participant should point with his index fingers to examiner's lifted hand. Start with participant's right index, then left, then repeat.
6	10	**Drink from glass.** Have the participant grasp an empty glass with his right hand as if he/she would drink from it. Then repeat with his/her left hand.
7	10	**Cross and extend both arms.**
8	10	**Bring both index fingers to each other.**
9	10	**Let participant's both index fingers tap his/her nose.** Start with the right, then with left index. Then extend the arms.
10	20	**Entrainment.** While leaving the arms extended, have the participant stamp with his/her right foot according to the stamp frequency of the examiner. Then have him/her repeat with the left foot.

2.2. Smartwatch sensor validation

A Seismometer is a device that captures weak ground motion caused by seismic sources, e.g., earthquakes, explosions or ambient noise [16]. These instruments generally have a large bandwidth and dynamic range [17]. The Nanometrics Trillium Compact is a

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 March 2021 doi:10.20944/preprints202103.0542.v1

triaxial seismometer, measuring ground velocity and classified as a broadband instrument with -3dB points at 120s and 108 Hz. The self noise level is below -140dB and the clip level at 26mm/s up to 10Hz and 0.17g above 10Hz [18]. We combined the Trillium Compact with a Taurus 24-bit Digital Recorder [19] , that digitizes the motion that the seismometer measures. This combinations allows for accurate measurements of ground motion [20] and is therefore considered as a Gold standard instrument for raw measurements of acceleration.

We conducted a shaker table experiment, where two Apple watches, Series 3 and 4, and the Trillium Compact seismometer were simultaneously accelerated by oscillatory motions with tremor-typical frequencies and amplitudes. Because tremor is an oscillatory movement, the use of a shaker table provides a means of testing accuracy of the method. The setup of the validation experiment is shown in Figure 1, where the seismometer, smartwatches were placed on a shaking table.

The watches were further attached with tape to prevent unwanted movement due to the slightly curved backside of the watches. The shaker table is placed on a decoupled platform to reduce ambient noise and oscillates vertically with a range of frequencies and amplitudes. Due to the experimental setup and since the vibration table moves in the vertical direction, only the z-axis of the watches and the seismometer was examined here. However, a significant difference in measurement accuracy between all three sensor components of the seismometer is not to be expected since the device records on three orthogonal axes U.V.W, which are then rotated into vertical and two horizontal components North and East [18].

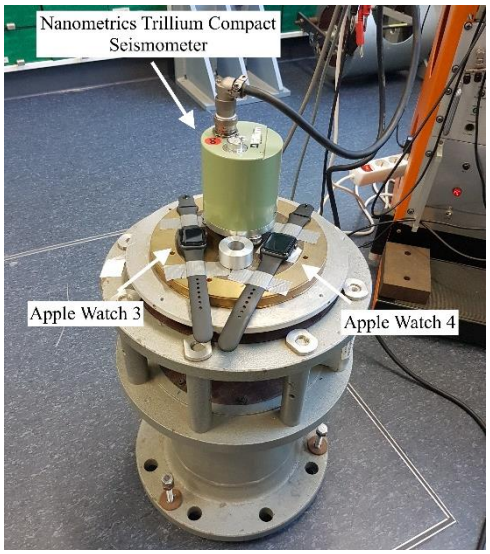

Figure 1. Experimental setup of the sensor validation experiment. Apple Watches Series 3 and 4 and a Nanometrics Trillium Compact seismometer are placed on a vertical vibration table. The table simultaneously accelerates the devices by oscillatory motions with tremor-typical frequencies and amplitudes. Both watches are connected to Apple iPhones (not in this figure) via Bluetooth, where the measurement data were stored. The seismometer data are collected on a digitizer (not in this figure), that the device is connected to.

The smartwatches are described to have a sampling rate of 100 Hz and we set the sapling rate of the seismometer to 100 Hz as well. A total of 43 measurements were performed on two different days. The duration of each measurement was set to 20 seconds for the watches, similar to the assessment steps performed with patients.

For each test, the table oscillated with a set amplitude and frequency that was kept constant during the measurement period. One test was carried out without vibration, to measure the difference in self noise of the watches and the seismometer. For the remaining tests we varied the frequency of the oscillation between 3Hz and 15Hz as this range covers tremor-typical frequencies [21]. The oscillation amplitude was varies between 0.002g and 0.1g, which is considered as high-resolution for tremor amplitudes as values < 0.01g are barely visible by human vision but still clinically relevant to measure subtle tremor in early disease.

The data had to be processed after the experiments: First, the data of the seismometer were deconvolved with the instrument response. During the deconvolution the counts per Volts scaling factor of the raw data and the frequency-dependent sensor response are

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 March 2021 doi:10.20944/preprints202103.0542.v1

removed. Since the seismometer records velocity while the watch records acceleration, the seismometer data were differentiated, scaled to SI units and divided by 9.81m/s^2, such that the output is in multiples of g, the Earth's acceleration.

To determine the oscillation frequency for each 20s measurement for both the seismometer and watches, the data were analyzed in the spectral domain, by applying the Fast Fourier Transform (FFT). The dominant frequency of each dataset was identified and compared. Prior to the FFT, all data were zero-padded to reach a frequency bin spacing of 0.01Hz because the frequency scale of the shaker table were graduated in 0.01 Hz steps.

The oscillation amplitude was calculated in the time domain on the pre-processed datasets. For 20 consecutive periods, the maxima and minima of the signal were identified and used to calculate the peak-to-peak amplitudes. The resulting 20 peak-to-peak amplitudes were averaged and divided by 2. Subsequently, the results of the watches were compared to those of the seismometer in order to assesses the accuracy of the watches.

2.3. Machine Learning Pipeline and Features

Three relevant classification tasks were trained and cross-validated:

1. PD vs Healthy
2. Movement Disorders (PD+DD) vs Healthy
3. PD vs DD

It is assumed, that the first two tasks are of lower classification difficulty as the system only needs to be trained for non-healthy characteristics. Such a system could still be helpful in home-based settings or at general practices e.g. to indicate whether certain abnormal movement characteristics (e.g. hand tremor) is pathologic or still normal (e.g. physiological tremor). The third one requires more advanced and differential features analyses in order to distinguish movement disorders with similar phenotypical characteristics from each other.

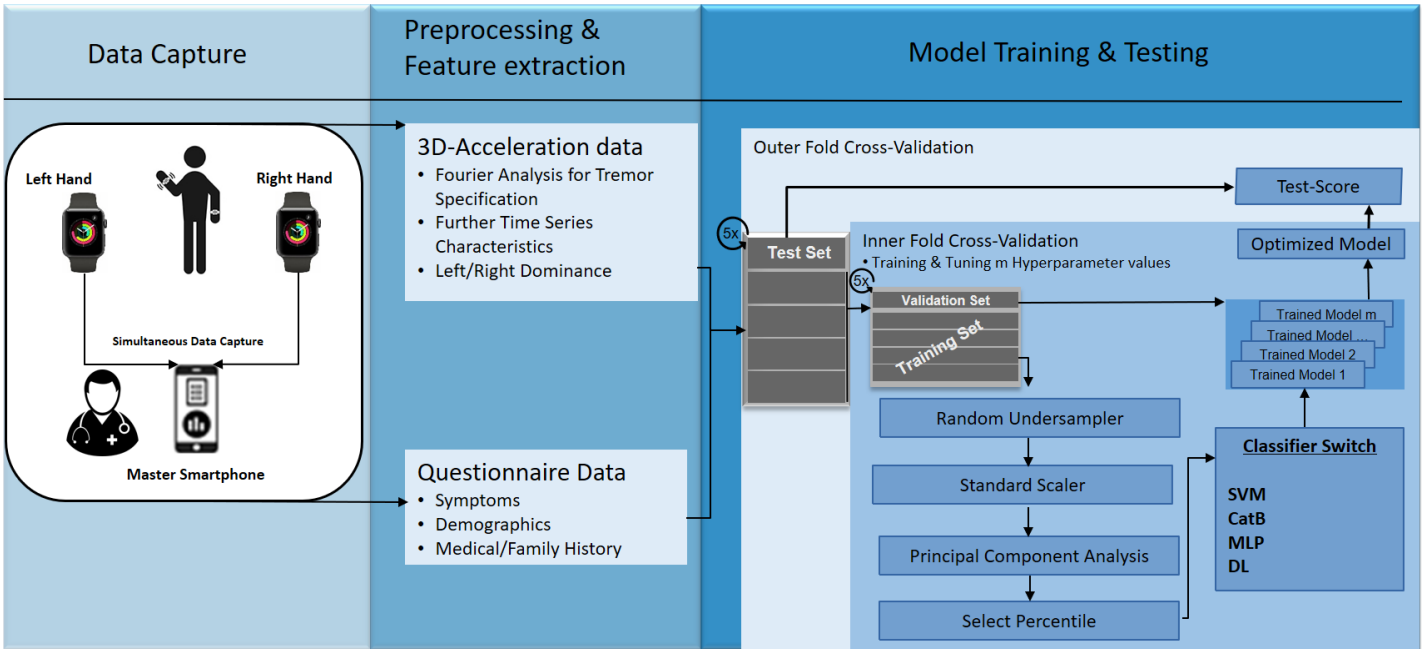

Figure 2. Overview of data analytics pipeline. SVM= Support Vector Machine with radial basis function. CatB = CatBoost, MLP = Multi Layer Perceptrons with two hidden layers, DL = Deep Learning Architecture.

Table 3. Machine Learning Features.

Feature	Average Age (SD)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 March 2021 doi:10.20944/preprints202103.0542.v1

Medical History Questionnaire	Information on demographics, height, weight, family history. Further details provided in Varghese et al. [13]. Medication is captured but not used as training-feature as it is too closely linked to the target classes.
Symptoms-Questionnaire	The number of items answered with 'yes' in the Parkinson's Disease Non-Motor Scale by the Movement Disorder Society.
Amplitude Distribution	Create an Amplitude-Histogram of acceleration data. Apply Euclidean norm on all three axes to generate 1-dimensional time-series vector, and pick the 30th to 70th percentile in 5 percent steps. Applied for all assessment steps.
Tremor Side Dominance	Use the 90th percentile of the left and right arm acceleration and calculate the ratio. Applied for all assessment steps.
Standard Deviation of Acceleration	Calculate the Standard Deviation of the acceleration data. Applied for all assessment steps.
Fast Fourier Transformation	Calculate the 3-dimensional FFT for the assessment step and use a polynomial regression to reduce the output dimensinoality. Polynomials of degree 3 are used. Applied for all assessment steps.

The extracted features are listed in Table 3. We provide further details and pseudo-code of feature extraction in the Machine-Learning Supplement. A previously developed Python-based data analytics pipeline is re-utilized [22]. The entire analytics process is summarized and illustrated in Figure 2. The different Machine Learning classifiers were Support Vector Machines (SVM), a modern gradient boosting decision-tree model called CatBoost (CatB) [23] and Multilayer Perceptrons (MLP), which are a classical type of Artificial Neural Networks. These were trained and validated within the framework of nested cross-validation [24] using 5 outer and 5 nested inner data folds to ensure unbiased training and testing, as well as unbiased optimization of hyperparameters. While the inner folds are used to train each model and to optimize its hyperparameters in a grid-search (m different hyperparameter values results in m different models configurations), the outer folds evaluates the test performance of trained and hyperparameter-optimized models. Before each inner fold model training, we apply the random undersampler from Scikit Learn 0.21.3 [25] in order to remove the bias towards the majority class by randomly removing samples of that set. Moreover, the standard scaler from Scikit Learn subtracts the mean and scales to unit variance for every feature. The Principal Component Analysis (PCA) reduces the dimensionality, the Scikit Learn-based 'Select Percentile' step randomly selects a subset of features, which are then used for training the classifier. We optimize the hyperparameters for the PCA, the Select Percentile and the specific classifiers.

The Multi-Layer Perceptron and the Deep Learning architecture is implemented using the KERAS 2.2.4 package and Google's Tensorflow 1.3.1, which provides full GPU support [26]. We considered various state-of-the-art architectures including Convolutional Neural Networks in ResNets and Long-Short-Term Memories (LSTM) [27]. Detailed architectures are provided in the Machine-Learning Supplement.

To evaluate their performance for automatic time-series feature extraction from acceleration data, they only received the raw acceleration data and the questionnaire data (medical history + symptoms) as input, but not the engineered time-series features in table 3.

Test performances for all three classification tasks are reported as mean values for Precision, Recall and F1-measure based on the outer fold validations including standard deviations. Due to the imbalance of the three disease classes, balanced accuracies [28,29]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 March 2021 doi:10.20944/preprints202103.0542.v1

are provided as well. As such, the baseline performance of all binary classification tasks is 50%, which corresponds to random guessing.

3. Results

3.1. Smartwatch sensor validation

Figure 3a shows the differences between dominant frequencies of the seismometer (used as the gold standard device) and Apple Watches Series 3 and 4 data (consumer grade device). Overall, Apple watches Series 3 and 4 seemed to measure higher frequencies than the seismometer, however, deviations were in the low milli-Hertz range (up to 10 mHz). With increasing frequencies, there was an increase in frequency deviation for both watches and for all experiments.

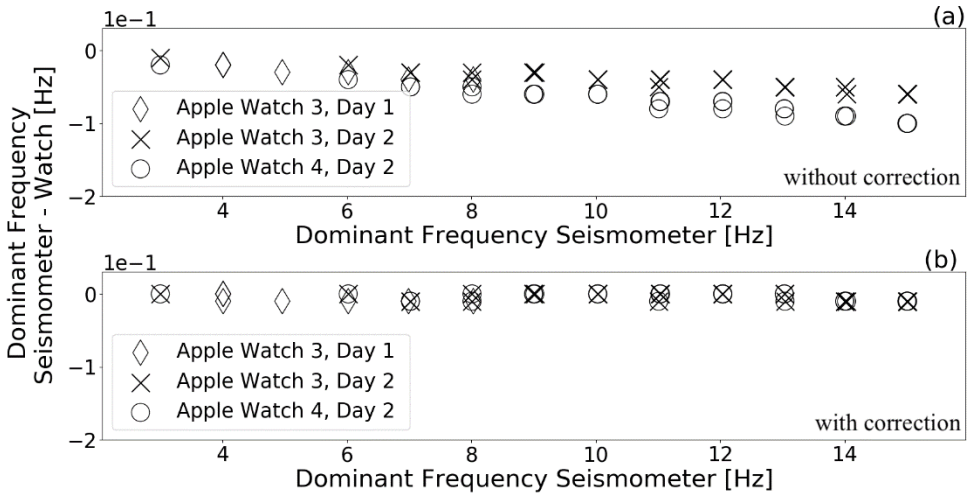

Figure 3. Differences between the dominant frequency measured by the Trillium Compact Seismometer and Apple Smartwatches Series 3 and 4 in a shaker table experiment. The experiment was conducted on two different days with the Apple Watch Series 3. The figure shows the difference in dominant frequency (a) using the pre-defined watches sample rate and (b) using the watches actual sample rate (calculated with watch-specific time vectors) for spectral calculations

As mentioned above, the watches' sampling rates were set to 100 Hz. When calculating the watches actual sampling rate using the watch-specific time vectors, we found that the sampling rates of the watches were non-uniform and up to 0.6 Hz smaller than specified 100 Hz. The increasing deviations with increasing frequency therefore resulted from the differences in programmed sampling rate of 100 Hz and the actual sampling rate for spectral calculations. Figure 3b shows corrected data based on the actual sampling rates of the watches. In the considered range, no clear increase in deviation with increasing frequency is recognizable anymore. Approximately 55% of the Series 3 and 59% of Series 4 dominant frequencies did not deviate from the seismometer up to the second decimal place. The remaining measurements deviated by up to 0.01Hz for both Series 3 and 4, which still provides high precision of tremor frequency capture, as clinical tremor documentation is documented in the rage of 4-18 Hz and step sizes of full Hz units [21].

We measured the self noise of the seismometer and the watches on the non-vibrating table. The results are depicted in Figure 4 and show that the watches had a higher noise compared with the seismometer but the RMS self-noise level was still below 0.001g for both watches. The 0g-offset was found to be below 2*10^-4 g. The power spectral density shows, that the noise of the smartwatches had a similar intensity at different frequencies.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 March 2021 doi:10.20944/preprints202103.0542.v1

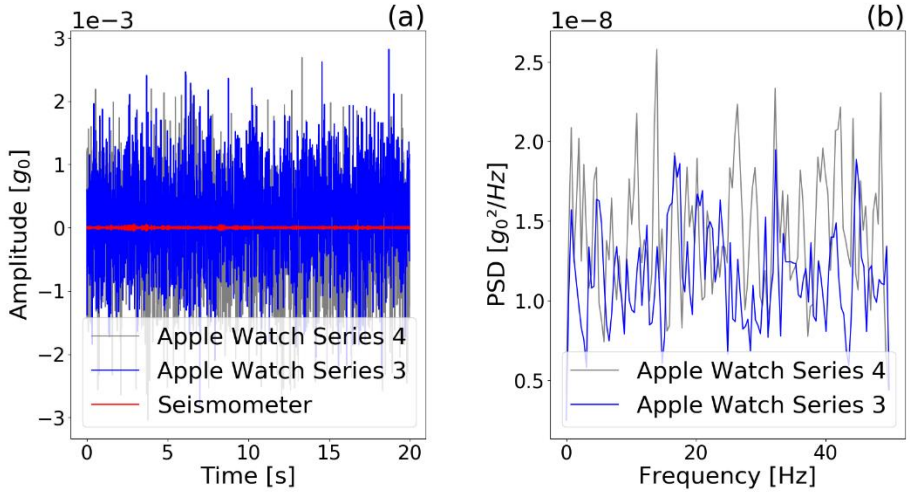

Figure 4. (a) Self Noise of watches and seismometer and (b) Power Spectral Density (PSD) of watches, captured during a 20 sec period without vibration of the shaker table.

Figure 5 depicts the difference in measured oscillation amplitude for the seismometer and the smartwatches. For all measurements, smartwatch Series 3 and 4 measured higher amplitudes than the seismometer. Up to 0.04g oscillation amplitudes, the amplitude differences between the watches and the seismometer showed no trend and were below 0.002g. Oscillation amplitudes > 0.05g lead to a larger deviations for both Series 3 and 4 and a trend is visible. We found the maximum deviation of 0.005g. The amplitude measurements of the watches and seismometer agree within their corresponding standard deviations.

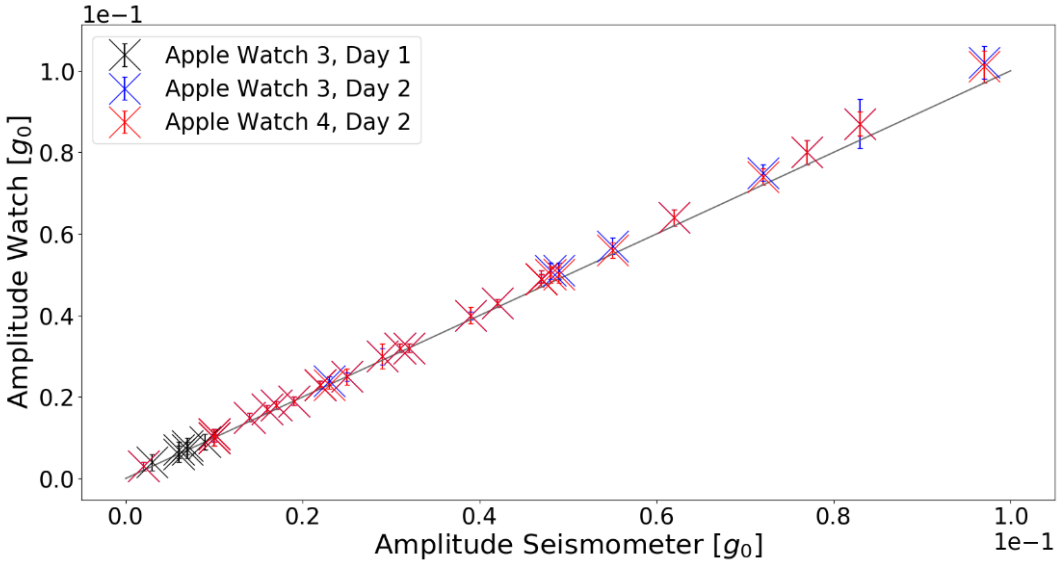

Figure 5. Measured oscillation amplitude of the seismometer and the watches are plotted against each other. The standard deviations of the amplitude mean values (of the watches) are plotted as error bars. The grey line corresponds to a perfect agreement between the oscillation amplitude measured by the watches and the seismometer.

3.2. Classification Performances and Feature Importance

Tables 4-6 list model performances for all three classification tasks. Apart from the Deep Learning model, the other three classical machine learning models performed similar in respect to their standard deviations, with balanced accuracies above 80%, precision and recall above 90% in the two simpler classification task. Regarding the most difficult task, which required separation of Parkinson's disease from similar movement disorders,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 March 2021 doi:10.20944/preprints202103.0542.v1

all three models performed lower with balanced accuracies between 67% and 74%. The MLP performed best in two of three tasks (PD+DD vs Healthy, PD vs DD) in terms of balanced accuracies.

Table 4. Performances for Classification Task 1, PD vs Healthy.

Estimator	Accuracy	Balanced Accuracy	Precision	Recall	F1
MLP	0.864 (0.03)	0.815 (0.05)	0.907 (0.03)	0.913 (0.03)	0.909 (0.02)
SVM - rbf	0.870 (0.02)	0.827 (0.01)	0.913 (0.01)	0.913 (0.03)	0.913 (0.01)
CatBoost	0.887 (0.02)	0.819 (0.04)	0.901 (0.03)	0.956 (0.03)	0.927 (0.01)
Deep Learning	0.768 (0.06)	0.591 (0.07)	0.782 (0.03)	0.954 (0.06)	0.859 (0.04)

Table 5. Performances for Classification Task 2, Movement Disorders (PD+DD) vs Healthy

Estimator	Accuracy	Balanced Accuracy	Precision	Recall	F1
MLP	0.856 (0.04)	0.772 (0.05)	0.907 (0.02)	0.914 (0.03)	0.910 (0.02)
SVM - rbf	0.838 (0.02)	0.750 (0.03)	0.901 (0.02)	0.897 (0.06)	0.897 (0.02)
CatBoost	0.882 (0.03)	0.757 (0.06)	0.895 (0.02)	0.968 (0.03)	0.929 (0.01)
Deep Learning 5	0.791 (0.03)	0.551 (0.06)	0.814 (0.01)	0.956 (0.03)	0.879 (0.02)

Table 6. Performances for Classification Task 3, PD vs DD.

Estimator	Accuracy	Balanced Accuracy	Precision	Recall	F1
MLP	0.823 (0.01)	0.741 (0.03)	0.865 (0.01)	0.905 (0.00)	0.885 (0.00)
SVM - rbf	0.800 (0.02)	0.682 (0.04)	0.831 (0.02)	0.921 (0.01)	0.873 (0.01)
CatBoost	0.817 (0.02)	0.678 (0.03)	0.826 (0.01)	0.956 (0.03)	0.887 (0.01)
Deep Learning 5	0.735 (0.01)	0.512 (0.01)	0.751 (0.01)	0.965 (0.04)	0.844 (0.01)

Among the different combinations of DL architectures, the best performing architecture included a light ResNet that could only reach balanced accuracies lower than 60%. Noteworthy, the inclusion of LSTMs consistently weakened the classification performance and therefore did not participate in our final DL architecture. Though the DL components underperformed in this complex task of diagnosis classification, a post-hoc analyses was conducted to figure out whether DL can correctly classify the assessment steps (e.g. does time-series belong to assessment step 6, “drinking glass” ?). This tasks would represent a typical time-series classification problem without requiring disease-dependent feature engineering. Here, the best DL model performed with an accuracy of 78,6% with the ResNet. The same tasks reduced to the assessment steps 'drink glass' and 'point finger' even performed with an accuracy of 94,6% using DL architecture with simple Dense Neural Networks. The detailed architecture for DL models and their performances is provided in the Machine-Learning Supplement.

4. Discussion

The SDS is an app-based mobile system that connects consumer devices for high-resolution monitoring of acceleration characteristics in different neurological disorders and questionnaire-based data capture of patient symptoms.

The seismological sensor validation showed high agreement between the smartwatches and the gold standard setting. While clinical tremor documentation ranges between 4-18 Hz with step sizes of 0.1 - 1Hz, the watches differed slightly from the gold standard at around 0.01 Hz. While human tremor amplitude threshold can be estimated at < 0.01-0.05g [3], the smartwatch amplitude-deviations were within the range of 0.001g and 0.005g. This shows the watches are capable of measuring movement subtleties or hand-tremor amplitudes and frequencies with much greater precision than clinical documentation or even human vision does. We reproduced these findings with multiple measurements and two Apple-based smartwatch models of different build years.

When integrating two smartwatches and paired smartphone to the SDS coupled with different AI-based classifiers we could show high diagnostic accuracies, above 80%, partially with precision and recall above 90% for simple classification tasks. Related work shows even higher performances, consistently above 90% accuracy when using other data modalities, e.g. voice-analyses [8]. However, while these findings doubtlessly show some diagnostic potential, they have to be interpreted with high caution as we believe these results are easily overestimated due to three key reasons: First, the overall sample size of almost all related studies were limited (n<<100). Second, model hyperparameters were not optimized in a separate nested set. Third, multiple measures of the same individual were taken, which can lead to identity confounding. To address these frequent drawbacks and provide higher degree of generalizability, we have generated – to the best of our knowledge - the largest database on this topic with more than 400 individually measured participants using nested cross-validation for all models and hyperparameters. In addition, we included the important control group of differential diagnoses. As expected, the most difficult task to separate PD from similar Movement Disorders was evaluated with much lower balanced accuracies of around 70%. This shows that further feature engineering and further integration of other promising modalities (acceleration, speech, voice or finger-tapping are needed. All these data modalities were studied in isolation with promising findings [4,8,30] and could be integrated within one system consisting of consumer devices. The results of our deep-learning architecture clearly show that automatic feature extraction is underperforming in this sample size dimension (n < 1000) and there is a strong need for engineering clinically relevant features in raw acceleration data.

A common limitation with related work, which is also not addressed by this study is the missing evaluation of real predictive capabilities for early diagnosis as we can only include patients, which have already been diagnosed or healthy participants, for which we don't know if they will develop a disease condition. Our study included a broad range of different disease progress states according to Hoehn & Yahr [31] or years from disease onset, but an observational epidemiological study with healthy to PD transformation data would be ideal to test disease prediction. Nevertheless, our work can provide potential features and methods, which need to be further studied in future study designs to evaluate on prediction performance. Moreover, our work contributes to new digital and objective biomarkers, which have the potential for disease stratification or disease monitoring of PD patients to provide personalized care and treatment optimization.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Figure S1: title, Table S1: title, Video S1: title. Patient-Population Supplement. Machine-Learning Supplement.

Author Contributions: Conceptualization, J.V. and C.T.; methodology, J.V and C.T.; software, M.F.; validation, J.V., C.T.; formal analysis, C.M.A and M.F.; investigation, J.V.; resources, C.T., T.W.; data curation, C.M.A.; writing—original draft preparation, J.V.; writing—review and editing, J.V.,C.M.A., M.F., T.W., C.T.; visualization, J.V. and C.M.A.; supervision, C.T.; project administration, J.V. and T.W.; funding acquisition, J.V. All authors have read and agreed to the published version of the manuscript.

Funding: This work is funded by the Innovative Medical Research Fund (Innovative Medizinische Forschung, I-VA111809) of the University of Münster..

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Ethics Committee of the University of Münster and the physician's chamber of Westphalia-Lippe (Reference num-ber: 2018-328-f-S, Approval date: Sept 5th, 2018).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study

Data Availability Statement: In this section, please provide details regarding where data supporting reported results can be found, including links to publicly archived datasets analyzed or generated during the study. Please refer to suggested Data Availability Statements in section "MDPI Research Data Policies" at https://www.mdpi.com/ethics. You might choose to exclude this statement if the study did not report any data.

Acknowledgments: In this section, you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

The appendix is an optional section that can contain details and data supplemental to the main text—for example, explanations of experimental details that would disrupt the flow of the main text but nonetheless remain crucial to understanding and reproducing the research shown; figures of replicates for experiments of which representative data is shown in the main text can be added here if brief, or as Supplementary data. Mathematical proofs of results not central to the paper can be added as an appendix.

Appendix B

All appendix sections must be cited in the main text. In the appendices, Figures, Tables, etc. should be labeled starting with "A"—e.g., Figure A1, Figure A2, etc.

References

1. Rocca WA. The burden of Parkinson's disease. The Lancet Neurology. 2018;17(11):928–9.
2. Postuma RB. Prodromal Parkinson disease. Nature reviews Neurology. 2019;15(8):437–8.
3. Varghese J, Niewöhner S, Fujarski M, Soto-Rey I, Schwake A-L, Warnecke T, et al. Smartwatch-based Examination of Movement Disorders: Early Implementation and Measurement Accuracy. EGMS [Internet]. 2019 DOI: 3205/19GMDS136; Available from: https://doi.org/10.3205/19GMDS136
4. Espay AJ, Bonato P, Nahab FB, Maetzler W, Dean JM, Klucken J, et al. Technology in Parkinson's disease. Movement disorders : official journal of the Movement Disorder Society. 2016;31(9):1272–82.
5. Heldman DA, Espay AJ, LeWitt PA, Giuffrida JP. Clinician versus machine. Parkinsonism & related disorders. 2014;20(6):590–5.
6. Silva de Lima AL, Hahn T, Evers LJW, Vries NM, Cohen E, Afek M, et al. Feasibility of large-scale deployment of multiple wearable sensors in Parkinson's disease. PloS one. 2017;12(12):e0189161.
7. Rusz J, Bonnet C, Klempíř J, Tykalová T, Baborová E, Novotný M, et al. Speech disorders reflect differing pathophysiology in Parkinson's disease, progressive supranuclear palsy and multiple system atrophy. J Neurol. 2015;262(4):992–1001.
8. Haq AU, Li JP, Memon MH, Malik A, Ahmad T, Ali A, et al. Feature Selection Based on L1-Norm Support Vector Machine and Effective Recognition System for Parkinson's Disease Using Voice Recordings. IEEE Access. 2019;7:37718–34.
9. Klucken J, Barth J, Maertens K, Eskofier B, Kugler P, Steidl R, et al. Mobile biometrische Ganganalyse. Der Nervenarzt. 2011;82(12):1604–11.
10. Klucken J, Gladow T, Hilgert JG, Stamminger M, Weigand C, Eskofier B. „Wearables" in der Behandlung neurologischer Erkrankungen – wo stehen wir heute? Der Nervenarzt. 2019;90(8):787–95.
11. Srulijes K, Mack DJ, Klenk J, Schwickert L, Ihlen EAF, Schwenk M, et al. Association between vestibulo-ocular reflex suppression, balance, gait, and fall risk in ageing and neurodegenerative disease. BMC Neurology [Internet]. 2015;15. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4600299/pdf/12883_2015_Article_447.pdf
12. Bandini A, Orlandi S, Escalante HJ, Giovannelli F, Cincotta M, Reyes-Garcia CA, et al. Analysis of facial expressions in parkinson's disease through video-based automatic methods. Journal of Neuroscience Methods. 2017;281:7–20.
13. Varghese J, Niewöhner S, Soto-Rey I, Schipmann-Miletić S, Warneke N, Warnecke T, et al. A Smart Device System to Identify New Phenotypical Characteristics in Movement Disorders. Frontiers in neurology. 2019;10:48.
14. Non-Motor Symptoms Questionnaire (NMSQ) by the International Parkinson and Movement Disorder Society [Internet]. [cited 2021 Jan 22]. Available from: https://www.movementdisorders.org/MDS/MDS-Rating-Scales/Non-Motor-Symptoms-Questionnaire.htm

15. Zhang B, Huang F, Liu J, Zhang D. A Novel Posture for Better Differentiation Between Parkinson's Tremor and Essential Tremor. Frontiers in neuroscience. 2018;12:317.
16. Routine Data Processing in Earthquake Seismology | SpringerLink [Internet]. [cited 2021 Feb 17]. Available from: https://link.springer.com/book/10.1007/978-90-481-8697-6
17. Havskov J, Ottemöller L, Trnkoczy A, Bormann P. Seismic Networks. New Manual of Seismological Observatory Practice 2 (NMSOP-2). 2012;1–65.
18. Nanometrics Trillium Compact Manual [Internet]. [cited 2021 Feb 17]. Available from: https://www.nanometrics.ca/sites/default/files/2018-04/trillium_compact_data_sheet.pdf
19. Taurus Portable Seismograph User Guide. :222.
20. Havskov J, Alguacil G. Instrumentation in Earthquake Seismology [Internet]. Springer Netherlands; 2004 [cited 2021 Feb 26]. (Modern Approaches in Geophysics). Available from: https://www.springer.com/gp/book/9789401751131
21. Bhatia KP, Bain P, Bajaj N, Elble RJ, Hallett M, Louis ED, et al. Consensus Statement on the classification of tremors. from the task force on tremor of the International Parkinson and Movement Disorder Society. Movement Disorders. 2018;33(1):75–87.
22. Varghese J, Fujarski M, Hahn T, Dugas M, Warnecke T. The Smart Device System for Movement Disorders: Preliminary Evaluation of Diagnostic Accuracy in a Prospective Study. Stud Health Technol Inform. 2020 Jun 16;270:889–93.
23. Dorogush AV, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support. arXiv:181011363 [cs, stat] [Internet]. 2018 Oct 24 [cited 2021 Mar 9]; Available from: http://arxiv.org/abs/1810.11363
24. Deisenroth MP, Faisal AA, Ong CS. Mathematics for Machine Learning. Cambridge University Press; 2020. 391 p.
25. Hackeling G. Mastering Machine Learning with scikit-learn. Packt Publishing Ltd; 2017. 249 p.
26. Keras: the Python deep learning API [Internet]. [cited 2021 Jan 28]. Available from: https://keras.io./
27. Wang Z, Yan W, Oates T. Time series classification from scratch with deep neural networks: A strong baseline. In: 2017 International Joint Conference on Neural Networks (IJCNN). 2017. p. 1578–85.
28. Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The balanced accuracy and its posterior distribution. In 2010. p. 3121–4.
29. Sklearn Metrics Balanced Accuracy Score — scikit-learn 0.24.1 documentation [Internet]. [cited 2021 Mar 8]. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.balanced_accuracy_score.html
30. Zham P, Arjunan S, Raghav S, Kumar DK. Efficacy of guided spiral drawing in the classification of Parkinson's Disease. IEEE journal of biomedical and health informatics [Internet]. 2017; Available from: https://doi.org/10.1109/JBHI.2017.2762008
31. Bhidayasiri R, Tarsy D. Parkinson's Disease: Hoehn and Yahr Scale. In: Bhidayasiri R, Tarsy D, editors. Movement Disorders: A Video Atlas: A Video Atlas [Internet]. Totowa, NJ: Humana Press; 2012 [cited 2021 Mar 9]. p. 4–5. (Current Clinical Neurology). Available from: https://doi.org/10.1007/978-1-60327-426-5_2