applied sciences

MDPI

Article

Predicting Wearing-Off of Parkinson's Disease Patients Using a Wrist-Worn Fitness Tracker and a Smartphone: A Case Study

John Noel Victorino *, Yuko Shibata, Sozo Inoue and Tomohiro Shibata

Graduate School of Life Science & Systems Engineering, Kyushu Institute of Technology, Kitakyushu 808-0135, Japan; y_shibata@fukujo.ac.jp (Y.S.); sozo@brain.kyutech.ac.jp (S.I.); tom@brain.kyutech.ac.jp (T.S.)

* Correspondence: jnoelvictorino@gmail.com or victorino.john-noel783@mail.kyutech.jp

Abstract: Parkinson's disease (PD) patients experience varying symptoms related to their illness. Therefore, each patient needs a tailored treatment program from their doctors. One approach is the use of anti-PD medicines. However, a "wearing-off" phenomenon occurs when these medicines lose their effect. As a result, patients start to experience the symptoms again until their next medicine intake. In the long term, the duration of "wearing-off" begins to shorten. Thus, patients and doctors have to work together to manage PD symptoms effectively. This study aims to develop a prediction model that can determine the "wearing-off" of anti-PD medicine. We used fitness tracker data and self-reported symptoms from a smartphone application in a real-world environment. Two participants wore the fitness tracker for a month while reporting any symptoms using the Wearing-Off Questionnaire (WoQ-9) on a smartphone application. Then, we processed and combined the datasets for each participant's models. Our analysis produced prediction models for each participant. The average balanced accuracy with the best hyperparameters was at 70.0–71.7% for participant 1 and 76.1–76.9% for participant 2, suggesting that our approach would be helpful to manage the "wearing-off" of anti-PD medicine, motor fluctuations of PD patients, and customized treatment for PD patients.

Keywords: prediction; statistical model; wearing-off phenomenon

Citation: Victorino, J.N.; Shibata, Y.; Inoue, S.; Shibata, T. Predicting Wearing-Off of Parkinson's Disease Patients Using a Wrist-Worn Fitness Tracker and a Smartphone: A Case Study. *Appl. Sci.* **2021**, *11*, 7354. https://doi.org/10.3390/app11167354

Academic Editor: Syoji Kobashi

Received: 5 May 2021
Accepted: 2 August 2021
Published: 10 August 2021

1. Introduction

Parkinson's disease (PD) patients experience difficulties in managing the symptoms of their illness. Each PD patient experiences symptoms differently and with varying severity. Doctors and medical practitioners customize the treatment of PD for every patient depending on their experienced symptoms. In order to do that, patients have to monitor their symptoms actively and report to their doctor. Then, doctors provide different doses of the anti-PD drug depending on the reported symptoms. The "wearing-off" phenomenon [1–4] occurs when the anti-PD medicine loses its effect. Patients start to experience the symptoms again until their next intake. In the long term, the duration of wearing-off begins to shorten. Thus, patients and doctors have to work together to manage PD symptoms effectively. Patients have to track their symptoms and the wearing-off period. Meanwhile, doctors need to have this information to give a customized approach to the patients' treatment.

Previous works have used different wearable data to detect PD symptoms. They mainly worked with accelerometer, electromyography (EMG), and gyroscope data to detect tremors, freezing of gait (FoG), and "wearing-off" periods [5–7]. Some of these previous works collected data in a closed and controlled environment. In contrast, other studies showed the feasibility of medium to large-scale data collection using different wearable sensors [8,9]. The Parkinson's KinetiGraph (PKG) has been used in some of these studies [9,10]. Statistical analysis [11] and machine learning-based analysis have been conducted on gait features; in one of these studies, machine learning algorithms were applied to wearable accelerometer-based sensor data to detect "on" and "off" states for PD

patients [12]. In another study, machine learning classifiers were applied to PKG data to estimate the levodopa response from the Unified Parkinson's Disease Rating Scale Part III (UPDRS III) [10]. Smartwatches have also been employed to monitor motor fluctuations based on accelerometer data in a real-world environment [8,13]. Studies have usually employed either PKG or other medical-grade wearable sensors. The medium to large-scale data collection approach has yet to apply automated analysis using the machine learning approach to analyze and predict the wearing-off phenomenon. Data collection in a real-life environment has shown promise for further analysis and thus could help PD patients.

There is still untapped potential in the use of commercially available fitness trackers and smartwatches for PD management and for predicting the wearing-off phenomenon. Fitness trackers and smartwatches can now report a user's heart rate, sleep quality, stress, and even blood pressure, among many other data types. Although the reliability of fitness trackers and smartwatches is contestable [14], these data can be used to verify earlier studies. For example, the blood pressure and heart rate of PD patients have been investigated for the wearing-off phenomenon, and vlood pressure change was statistically significant among PD patients experiencing wearing-off [15].

PD patients have also reported effects on their sleep patterns. The associated risk factors of rapid eye movement (REM) sleep behavior disorder (RBD) in PD patients were examined [16,17]. Aside from RBD as a form of sleep disturbance, light, fragmented sleep due to increased muscle activity, disruption of biological rhythms during sleep, breathing difficulties, insomnia, and excessive daytime sleepiness were sleep disturbances that manifested in PD patients [18,19]. A study with 3075 PD patients showed an increase of PD symptoms during sleep disturbances for 32% of the patients, while depressive moods were found for 20% of the patients. This study reported that each PD patient had varying degrees of symptoms and a differing psychological stress structure [20].

Moreover, in a recent study, PD patients were shown with clinical evidence to be highly sensitive to the effects of stress. The prevalence of stress-related symptoms in PD patients was 30% to 40% for depression and 25% to 30% for anxiety. Furthermore, stress worsened tremors, FoG, and dyskinesia. Thus, the authors suggested further investigation using wearable sensors [21].

As more people are adopting wearable technologies, this study explores the use of other wearable datasets to predict the "wearing-off" of PD patients. This research aims to develop a prediction model to determine the "wearing-off" of anti-PD drugs from fitness tracker data in a real-world environment. This study seeks to answer the following research questions:

1. How do we collect and combine a fitness tracker dataset and wearing-off dataset?
2. Can we develop a prediction model to determine the wearing-off of PD patients?

This paper is organized as follows: Section 2 describes the participants, the data collection tools, and the collected datasets. It also presents the data collection, data processing, and model development. Then, Section 3 presents the summary of collected and combined datasets. The performance and the best configuration of the prediction models are also reported. Next, Section 4 elaborates the results of our analysis and prediction models. Finally, Section 5 concludes this paper.

2. Materials and Methods

This section describes the data collection, data processing, and model development, as summarized in Figure 1.

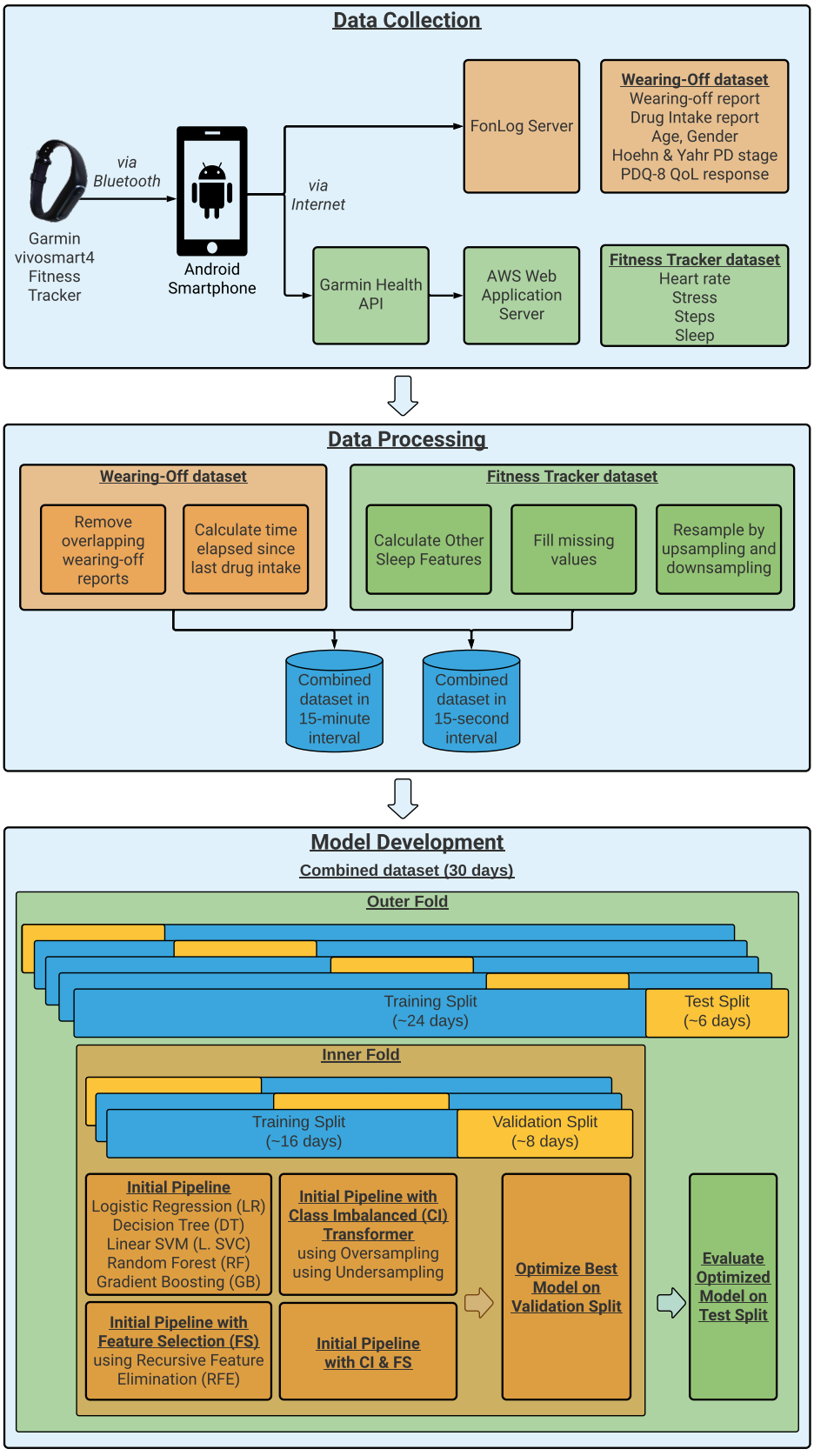


Figure 1. An overview of the study, describing the process from data collection to model development.

2.1. Data Collection

2.1.1. Data and Data Collection Tools

This study used two data collection tools that were distributed to the participants. First, a fitness tracker collected the participants' heart rate, step count, stress score, and sleep data. Second, a smartphone application enabled the participants to record wearing-off periods, drug intake time, and other one-time basic information such as age and gender.

2.1.2. Garmin Vivosmart4 Fitness Tracker

The Garmin vivosmart4 fitness tracker was chosen for this study. The participants preferred the vivosmart4 over smartwatches because it was sleek, lightweight, and waterproof. It weighs 16.5 g up to 17.1 g, with dimensions of 15 × 10.5 × 197 mm (https://buy.garmin.com/en-US/US/p/605739, accessed on 1 May 2021). The Garmin vivosmart4 contains an optical photoplethysmography (PPG) sensor and accelerometer sensor. Using light emitted into the skin, PPG estimates heart rate by monitoring the changes in the intensity of the reflected light due to the contraction and swelling of the arteries and arterioles caused by pulsating blood pressure [22,23]. Then, the Garmin vivosmart4 uses heart rate variability to estimate stress levels [24,25]. The combination of the heart rate, heart rate variability, and accelerometer reading is used to estimate sleep stages [23,26]. Finally, the Garmin vivosmart4 uses the accelerometer sensor to assess the number of steps taken [27].

End-users can monitor their data using the Garmin Connect smartphone application. On the other hand, developers can access heart rate, sleep, steps, and stress level data using the Garmin Health Application Programming Interface (Garmin Health API) (https://developer.garmin.com/gc-developer-program/health-api/, accessed on 1 May 2021). Upon gaining access, the Garmin Health API expected a web application server to receive the data via HTTP POST. We deployed the web application server with Amazon Web Services (AWS), and the acquired data were stored on that server. The data flow from the fitness tracker to data storage is described in Figure 1, while the data available via the Garmin Health API are summarized in Table 1.

Table 1. The Garmin vivosmart 4 fitness tracker datasets available via the Garmin Health API. This shows the initial time interval for each dataset, as well as the range of possible values reported by Garmin.

Data Type	Granularity	Description
Heart rate	15 s interval	Beats per minute
Steps	15 min interval	Cumulative count per interval, with a lowest value of 0
Stress score	3 min interval	Estimated stress score from 0 to 100 [28] • 100: the highest score, • −1: not enough data to detect stress, • −2: too much motion
Sleep classification and sleep period	Varying interval, with specific calendar date	Start and end time for each sleep classification [29] • Light sleep • Rapid eye movement (REM) sleep • Deep sleep

2.1.3. FonLog Smartphone Application

A smartphone application, called FonLog, was used as a data collection tool for human activity recognition in nursing services [30]. In this study, FonLog was customized specifically to collect wearing-off periods and drug intake time. To collect the data, we adapted the Wearing-Off Questionnaire (WoQ-9) using the Japanese translation [1,31]. The participants answered the remaining components of FonLog one time; these included age, gender, the Hoehn and Yahr Scale, and the Parkinson's Disease Questionnaire (PDQ-8). The Hoehn and

Yahr scale was used to determine the patient's PD stage [32,33]. Then, the PDQ-8 identified the self-reported quality of life (QoL) among the PD patients [34]. Table 2 summarized all the data recorded using FonLog.

Table 2. FonLog Data.

Data Type	Description
WoQ-9	Symptoms onset and drug intake time
Basic Information	Age and gender
Hoehn and Yahr Scale (H&Y) Japan Ministry of Health, Labor, and Welfare's classification of living dysfunction (JCLD)	Participant's PD stage
PDQ-8	Participant's QoL measurement specific to PD 0–100%, with 100% showing worst QoL

2.1.4. Participant Demographics

Two patients participated in this initial study with their informed consent. Both of our participants were female and were in their late 30s to early 40s. According to their Hoehn and Yahr Scale (H&Y) scores and the Japan Ministry of Health, Labor, and Welfare's classification of living dysfunction (JCLD), both participants were in a similar PD stage. However, their views on their QoL were different from each other. Table 3 summarized the demographics of the two participants.

Table 3. Participants' demographics.

	Participant 1	Participant 2
Age	43	38
Gender	Female	Female
H&Y	2: Bilateral or midline involvement without impairment of balance	3: Bilateral disease: mild to moderate disability with impaired postural reflexes; physically independent
JCLD	1: Little assistance is needed in daily life and outpatient visits	2: Partial assistance is required for daily life and outpatient visits
PDQ-8	37.5%	65.63%

The participants were subjected to the levodopa therapy, taking different medicines such as levodopa and carbidopa, dopamine agonist for D_2 receptors, and a selective monoamine oxidase type B (MAO-B) inhibitor drug. Participant 1 was taking 400 mg levodopa and carbidopa hydrate per day while participant 2 was taking 600 mg to 800 mg per day. Both participants were taking 1 tablet of 8 mg dopamine agonist medicine to stimulate D_2 receptors. Finally, participant 1 was taking 1 tablet of 2.5 mg MAO-B inhibitor while participant 2 was taking 0.5 mg per day.

2.1.5. Data Collection Method

The two PD patients received the data collection tools needed for our experiment. The data collection lasted for 30 days, with an additional one-day setup for the PD patients.

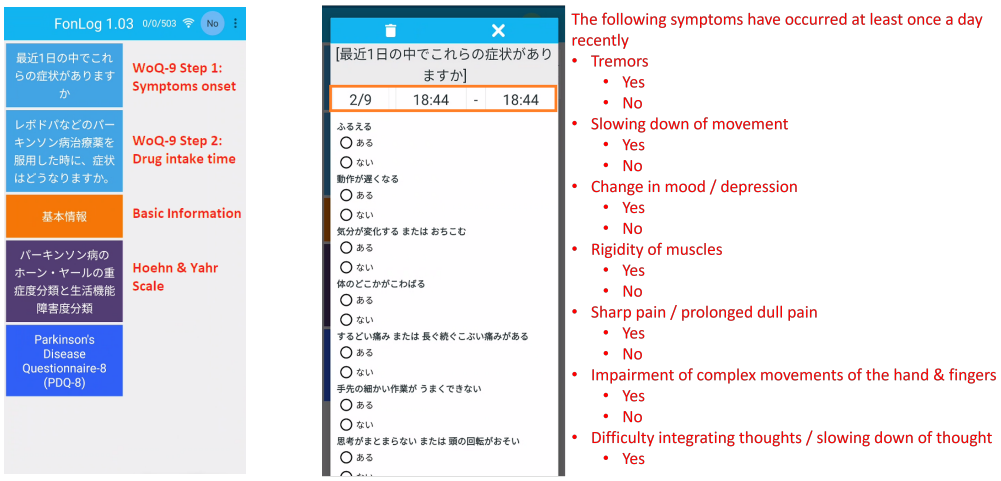
Each participant created their own Garmin Connect account to start the Garmin vivosmart4 fitness tracker's data collection process. Next, the participants signed into our web application server using their Garmin Connect account. At this point, they had control

over what information they would share with our web application server while reading the privacy policy. After all the steps, the Garmin Health API automatically transmitted available data to our AWS Web Application Server.

Throughout the 30-day data collection period, the participants were encouraged to always wear the Garmin vivosmart4, even during sleep. However, there were still cases when they forgot to wear it, such as after taking a shower or when doing household chores. In the worst case, they could not wear it due to some PD symptoms. These conditions were accepted to capture data that were as close as possible to the real-world scenario.

Using FonLog, the participants answered questions regarding (1) basic information, (2) the Hoehn and Yahr Scale, and (3) the PDQ-8 once. Then, they reported any of the following nine symptoms: tremors, slowing down of movement, change in mood or depression, rigidity of muscles, sharp pain or prolonged dull pain, impairment of complex movements of the hand and fingers, difficulty integrating thoughts or slowing down of thought, anxiety or panic attacks, and muscle spasms [1]. However, this study simplified the wearing-off label to either wearing-off (1) or not (0). Any recorded symptoms were considered to represent a wearing-off label.

The participants recorded the onset of any symptoms as accurately as possible. From the home screen presented in Figure 2a, the participants could click on the activity on the left side; i.e., activity enclosed in the red box. The home screen showed every recording on the right side. However, if a recording was difficult for the participants due to their symptoms, they could retroactively record data with their best estimation of the time. They could correct and review the onset of the symptoms, as shown in the top part of Figure 2b.



(**a**) (**b**)

Figure 2. The customized FonLog smartphone application. (**a**) The home screen shows each questionnaire: WoQ-9 step 1 for symptoms onset, WoQ-9 step 2 for drug intake time, basic information, H&Y, and PDQ-8, as discussed in Section 2.1.3. (**b**) A sample FonLog form presents the WoQ-9 step 1, asking if each symptom described in Section 2.1.5 has been experienced.

2.2. Data Processing

This subsection describes how we processed and combined the fitness tracker dataset and the wearing-off dataset.

First, the sleep dataset was aggregated according to each calendar date and the sleep classification, as listed in Table 1. The original sleep data (Table 4) were transformed into aggregated sleep data (Table 5). Furthermore, additional sleep features were calculated.

Table 4. The original sleep dataset provided by Garmin Health API. Each row shows the start and end time of each sleep stage.

Calendar Date	Start Time	End Time	Sleep Type
2021-02-23	2021-02-23 02:24:00	2021-02-23 02:32:00	Light
2021-02-23	2021-02-23 02:32:00	2021-02-23 02:33:00	Awake
2021-02-23	2021-02-23 02:33:00	2021-02-23 02:36:00	Light

Table 5. The sleep dataset aggregated by calendar date. Each sleep stage was converted as a feature, while the sleep duration for each sleep stage was calculated in minutes. Other sleep features were calculated.

Calendar Date	Awake	Deep	Light	REM	Total Non-REM	Total Sleep	Total Non-REM (% of Total Sleep)	Sleep Efficiency
2021-02-23	2.0	0.0	150.0	27.0	150.0	177.0	0.847	0.989
2021-02-25	0.0	150.0	66.0	54.0	216.0	270.0	0.800	1.000
2021-02-26	0.0	83.0	55.0	0.0	138.0	138.0	1.000	1.000

The total non-REM sleep duration was the sum of deep and light sleep duration (Equation (1)). Next, the total sleep duration was the sum of total non-REM sleep duration and REM sleep duration (Equation (2)). Then, the percentage of non-REM sleep was the ratio of the non-REM sleep duration to the total sleep duration (Equation (3)) [35]. Finally, we estimated the sleep efficiency as the ratio of the total sleep duration to the sum of the total sleep and total waking duration (Equation (3)) [36].

$$\text{Total non-REM duration} = \text{Deep sleep duration} + \text{Light sleep duration} \quad (1)$$

$$\text{Total sleep duration} = \text{Total non-REM duration} + \text{REM sleep duration} \quad (2)$$

$$\text{Total non-REM percentage} = \frac{\text{Total non-REM duration}}{\text{Total sleep duration}} \quad (3)$$

$$\text{Sleep efficiency} = \frac{\text{Total sleep duration}}{\text{Total sleep duration} + \text{Total awake duration}} \quad (4)$$

Second, the missing values were filled with "−1" before re-sampling. This was in accordance with how Garmin Health API reported stress scores when they could not estimate the stress level [28]. Replacing missing values with "−1" would also indicate that the participant was not wearing the fitness tracker.

Third, the fitness tracker datasets were re-sampled due to their varying granularity. We chose 15 s and 15 min intervals to represent the minimum and maximum granularity in the fitness tracker datasets, respectively. The resulting missing values due to re-sampling were filled by copying the last available data. This was the chosen fill method to replicate the streaming behavior of the fitness tracker dataset, as the future values would not be available.

Finally, the wearing-off dataset was cleaned by removing overlapping wearing-off data. Then, the wearing-off and the drug intake datasets were combined with the fitness tracker dataset. Each record was checked if it fell within the timestamp. If so, a value of "1" was assigned for the "y" or wearing-off variable; otherwise, a value of "0" was set. Similarly, if each drug intake record fell within the timestamp, the time that elapsed from the start of each drug intake record was computed. The time that elapsed for the succeeding drug intake record was set back to "0".

The raw available datasets from Garmin and the FonLog smartphone application and the combined dataset are provided in the Supplementary Materials.

2.3. Model Development

The prediction models for wearing-off were developed using the 15 s and 15 min datasets for each participant. For this initial case study, we built individual-level models because each participant experienced PD differently. Thus, we wanted to discover and optimize a prediction model for each individual to understand their PD situation fully.

In this subsection, the features and the different prediction models are discussed in detail (1) to understand how each model estimates the wearing-off phenomenon, and (2) to interpret how they use different features in the prediction. We also explain how we trained, evaluated, and optimized the models using the nested cross-validation approach. We used the Python programming language with PhotonAI [37] and Scikit-Learn [38] as the main libraries.

The following features extracted from the processed dataset were used to develop the prediction model:

- x_1: Heart rate (HR);
- x_2: Step count (Steps);
- x_3: Stress score (Steps);
- x_4: Awake duration during the estimated sleep period (Awake);
- x_5: Deep sleep duration (Deep);
- x_6: Light sleep duration (Light);
- x_7: REM sleep duration (REM);
- x_8: Total non-REM sleep duration (NonREMTotal);
- x_9: Total sleep duration (Total);
- x_{10}: Total non-REM sleep percentage (NonREMPercentage);
- x_{11}: Sleep efficiency (SleepEfficiency);
- x_{12}: Time elapsed from the last drug taken (DrugElapsed);
- y: Wearing-off.

In this study, five machine learning algorithms were applied to develop the prediction models. These algorithms were Logistic Regression (LR), Linear Support Vector Machine (L. SVM), Decision Tree (DT), Random Forest (RF), and Gradient Boosting (GB) classifiers. We used these machine learning models to interpret the importance of each feature to the wearing-off prediction.

2.3.1. Logistic Regression (LR)

In this study, the Logistic Regression model $p(X)$ was used to approximate the probability of wearing-off using the different data from the fitness tracker and drug intake data. During the training of a Logistic Regression model, it estimates the weights for each feature in relation to wearing-off. A larger weight for a feature shows that it contributes more to the probability of wearing-off than smaller weights. Mathematically, the Logistic Regression model is a sigmoid function of $f(X)$ such that

$$f(X) = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + w_5x_5 + w_6x_6 + w_7x_7 + w_8x_8 + w_9\ x_9 + w_{10}x_{10} + w_{11}x_{11} + w_{12}x_{12} + b, \quad (5)$$

where $x_1 \ldots x_12$ are the defined features, b refers to a bias, and $w_1 \ldots w_12$ are the weights for each feature.

2.3.2. Linear Support Vector Machine (L. SVM)

The Linear Support Vector Machine's key task is to find a hyperplane $w \in \mathbb{R}^p$ which has the largest distance (by $\min_{w,b} \frac{1}{2}w^Tw$) to the nearest training data $x_i \in \mathbb{R}^p$. It is defined as

$$\min_{w,b} \frac{1}{2}w^Tw + C\sum_{i=1} \max(0, y_i(w^T\phi(x_i) + b)) \quad (6)$$

where $y \in \{1, -1\}^n$ is the target, $\phi(x_i)$ is an identity function, b refers to a bias, and C refers to the strength of the penalty [38–40].

2.3.3. Decision Tree (DT)

The Decision Tree model finds the best split of data multiple times based on thresholds for each feature. Each subset of the data is assigned to a node of the tree with specific feature thresholds. Equation (7) encapsulates the Decision Tree model [41].

$$y = f(x) = \sum_{m=1}^{M} c_m I\{x \in R_m\} \tag{7}$$

where R_m is the subset of the data and $I\{x \in R_m\}$ is an identity function which returns "1" if x is in subset R_m—otherwise, the identity function returns "0"—and y is the predicted value, which is equal to the average of all training instances c_m.

2.3.4. Random Forest (RF)

Built using Decision Tress, the Random Forest model is an ensemble of smaller Decision Trees that produces a predictive model [38]. A Random Forest classification model follows these steps [42]:

1. Dataset D is bootstrapped b to produce samples of size n such that $D_b = \{(x_{1b}, y_{1b}), \ldots, (x_{nb}, y_{nb})\}$;
2. A Decision Tree f_b is trained with D_b to get $\hat{f}_b$;
3. The tree is grown to the largest extent;
4. Steps 1 to 3 are repeated to build bootstrapped Decision Trees $(\hat{f}_b \ldots \hat{f}_B)$;
5. The different Decision Trees are combined using majority voting $\hat{f} = mode(\hat{f}_1 \ldots \hat{f}_B)$.

2.3.5. Gradient Boosting (GB)

Unlike the Random Forest model, in which a majority vote occurs among Decision Trees, the Gradient Boosting model is an additive model that proceeds in a forward stage-wise fashion. The goal is to improve a weak classifier that starts from initialization. Then, the negative gradient is calculated and the error is minimized. The next step size is determined until the final model is achieved [38,40,43,44].

2.3.6. Analysis Pipeline

This section describes the nested cross-fold (Nested CV) approach for model selection, hyperparameter optimization, and performance evaluation using PhotonAI. PhotonAI is a high-level machine learning library for designing and optimizing machine learning models using pipelines [37]. PhotonAI wraps around other machine learning libraries such as Scikit-Learn for general machine learning tools [38], Scikit-Optimize for hyperparameter optimization [45], and Imbalanced-Learn for handling class imbalance [46].

Nested CV Approach

Furthermore, PhotonAI allows the implementation of the nested CV approach, which was used for small to medium datasets. The dataset was divided into training, validation, and test splits in this approach, depending on the outer and the inner fold data split technique. The outer fold handled the model evaluation while the inner fold chose the best model with the optimized hyperparameters. Over-fitting would be avoided because the hyperparameter optimization was only exposed to the subset of each outer fold. The expected output of a nested CV approach was a set of optimized best models [37,42,47–49]. The model development in Figure 1 showed the nested CV approach.

Stratified Validation Technique

A non-shuffled stratified validation technique was used to divide the dataset to preserve the time-based dependency and the wearing-off distribution percentage of the dataset.

Similar to *k*-fold cross-validation, the dataset was divided by *k* folds while maintaining the distribution of each class. The non-shuffled stratified validation split was used in this study due to the imbalanced nature of wearing-off reports. The dataset was divided into 5 outer folds, where each fold had approximately 24 days of training data and 6 days of test data. Then, each outer fold's training data were divided into 3 inner folds, with approximately 16 days of training data and 8 days of validation data. The depiction of the outer fold and inner fold of the model development in Figure 1 illustrates the stratified validation technique.

Hyperparameter Optimization Using Scikit-Optimize

The hyperparameter optimization was conducted using the Scikit-optimize Python library within the PhotonAI library. Given a hyperparameter search space, the balanced accuracy metric in our study was maximized. Then, the next hyperparameter was suggested until 30 configurations were generated and tested. This optimization search occurred in the nested CV's inner fold.

Pipelines for Model Selection and Hyperparameter Optimization

With PhotonAI, we compared different learning algorithms while each learning algorithm was optimized (initial pipeline). Next, the recursive feature elimination (RFE) algorithm was added for the feature selection pipeline (FS). The goal of RFE was to choose the features by repeatedly removing the least important feature for each iteration until the desired number of features was reached [38,41]. Then, a class imbalance pipeline (CI) handled the few wearing-off labels in our dataset by adding $ImbalanceDataTransformer$. This transformer was part of the Imbalanced-Learn [46] library in PhotonAI. The CI pipeline chose the best over and under-sampling technique that improved the model's performance. Finally, we combined FS and CI elements into one pipeline (CI + FS) for comparison with the previous pipelines. Table 6 summarizes the hyperparameter search space used in these pipelines.

Table 6. Hyperparameter search space for each learning algorithm and other pipeline elements specific to each pipeline.

Learning Algorithms and Other Pipeline Elements	Hyperparameter Range
Logistic Regression (LR)	C = Float([1,. . .,10])
	class_weight = 'balanced'
Decision Tree (TR)	min_samples_split = Integer([2,. . .,30])
	min_samples_leaf = Integer([2,. . .,30])
	criterion = 'gini'
Linear SVM (L. SVM)	C = Float([1,. . .,10])
	class_weight = 'balanced'
Random Forest (RF)	min_samples_split = Integer([2,. . .,30])
	max_features = ['auto', 'sqrt', 'log2']
	criterion = 'gini'
	bootstrap = True
Gradient Boosting (GB)	loss = ['deviance', 'exponential']
	learning_rate = Float([0.001,. . ., 1], 'logspace')
Recursive Feature Elimination (RFE)	n_features_to_select = Integer([2,. . .,12])
Imbalance Data Transformer	method_name = ['RandomUnderSampler', 'RandomOverSampler', 'SMOTE', 'BorderlineSMOTE']

Performance Metrics

In this study, the balanced accuracy was the chosen primary metric to handle the prediction model's bias towards the more frequent class in our dataset ("on" or "0"). The balanced accuracy was also chosen so that both a high true positive rate (or sensitivity, Sn) and a high true negative rate (or specificity, Sp) were achieved. The balanced accuracy was defined as the arithmetic mean of the sensitivity and specificity, as defined in Equations (8) and (9) [38,50].

$$\text{Sn} = \frac{TP}{TP + FN}, \text{Sp} = \frac{TN}{TN + FP}, \tag{8}$$

$$\text{Bal. Acc.} = \frac{1}{2}(\text{Sn} + \text{Sp}), \tag{9}$$

where TP is the true positive value, FN is the false negative value, FP is the false positive value, and TN is the true negative value, according to the confusion matrix in Table 7.

Table 7. Confusion matrix for wearing-off and on of anti-PD medicine.

	Predicted Wearing-Off (1)	**Predicted On (0)**
Actual Wearing-Off (1)	True positive (TP)	False negative (FN)
Actual On (0)	False positive (FP)	True negative (TN)

In this study, the precision and f1-score were also calculated, as shown in Equations (10) and (11). The precision and F1-score were considered because we wanted to avoid an "on" state being predicted as wearing-off. Using the balanced accuracy and F1-score gave further importance to false negative and false positive errors in our prediction model.

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{10}$$

$$\text{F1 Score} = 2 \cdot \left(\frac{Precision \cdot \text{Sn}}{Precision + \text{Sn}} \right) \tag{11}$$

3. Results

3.1. Wearing-Off Data Collection

Fitness tracker data and wearing-off data were collected for 30 days from 23 February until 24 March 2021. The participants recorded their wearing-off periods using a customized FonLog application. Participant 1 collected a total of 227 wearing-off periods, while participant 2 reported 44 wearing-off periods. The average duration of wearing-off for participant 1 was 87.432 min (σ = 77.943 min); on the other hand, participant 2 had an average duration of wearing-off of 25.295 min (σ = 34.512 min). Over the 30-day collection period, participant 1 reported at least one wearing-off every day. Meanwhile, participant 2's reporting was more sparse than participant 1; there was even no report on the third week from participant 2. Figures 3 and 4 present the distribution of wearing-off data for both participants.

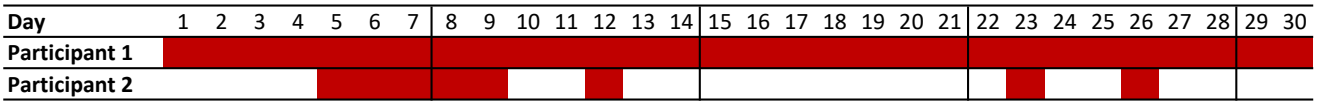


Figure 3. Distribution of reported wearing-off over the 30-day collection period.

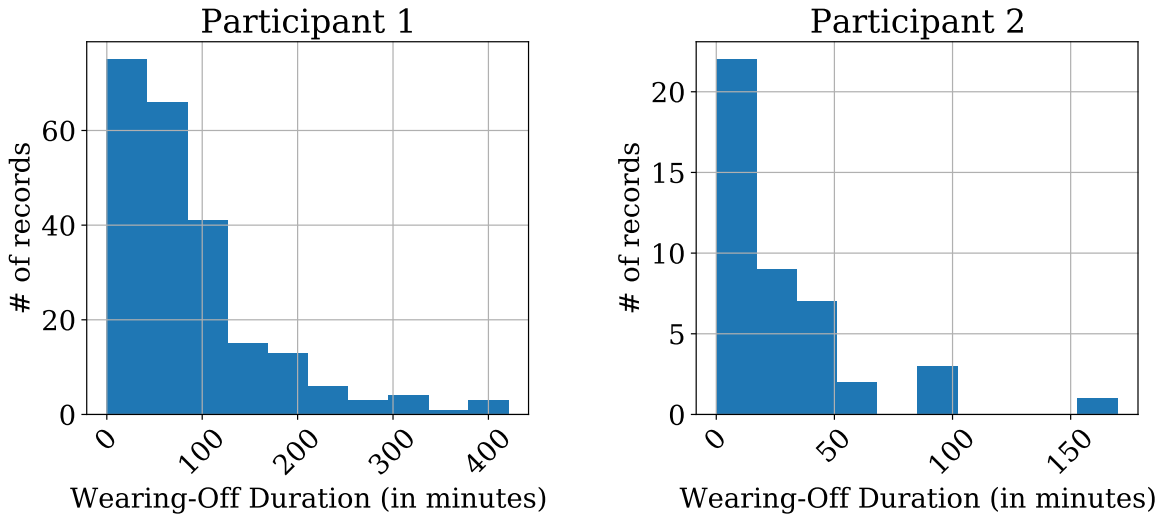


Figure 4. The collected wearing-off histogram. The x-axis represents the wearing-off duration in minutes, and the y-axis represents the number of records for a specific wearing-off duration.

3.2. Fitness Tracker Data Collection

Each participant was given a Garmin vivosmart4 fitness tracker. We asked the participants to wear the fitness tracker during the 30-day collection period. The collected data are summarized in Table 8.

Table 8. Summary of collected fitness tracker data . The expected n records is the number of records for 30 days specific to each dataset's time interval. The reported statistics were for the whole 30 days.

		Heart Rate	Steps	Stress Score	Sleep
	Expected *n* records	172,800.00	2880.00	14,400.00	30.00
Participant 1	n records	152,804.00	2880.00	14,400.00	29.00
	$\overline{x} \pm \sigma$	85.792 ± 15.927	34.116 ± 109.190	29.667 ± 32.514	265.793 ± 93.681
	$[min, max]$	[50, 168]	[0, 1549]	$[-2, 98]$	[85, 469]
Participant 2	n records	148,193.00	2654.00	13,135.00	25.00
	$\overline{x} \pm \sigma$	69.962 ± 15.880	109.275 ± 195.430	11.230 ± 20.752	237.48 ± 67.430
	$[min, max]$	[43, 186]	[0, 1392]	$[-2, 94]$	[138, 404]

Both participants missed at least one day of wearing the fitness tracker over 30 days. Still, participant 1 wore the fitness tracker for longer than participant 2, as shown by the lower number of n records in Table 8. Moreover, participant 1 had a higher average heart rate than participant 2 over 30 days, as shown in Figure 5.

Figure 6 shows the distribution of 15 min records for the number of steps. Both participants' distributions of step records presented a positively skewed distribution, where most records were under 100 steps. Specifically, for every 15 min record, the collected number of steps was mainly under 100 steps.

In terms of the reported stress score from the Garmin Health API, participant 1 had an average stress score (n = 14,400 of 3 min records) of 29.67 ± 32.51 over the 30-day collection period, or a low-stress classification based on Garmin's stress level classification [28]. On the other hand, participant 2 had an average stress score (n = 13,135 of 3 min records) of 11.23 ± 20.75 over the 30-day collection period, or a resting state classification. Figure 7 shows the stress scores for each participant.

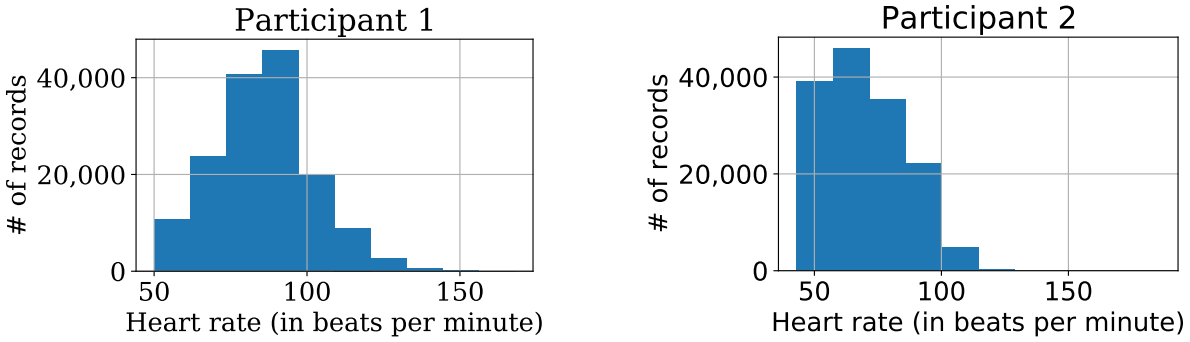


Figure 5. Participants' heart rate record distribution. The x-axis represents the heart rate (in bpm) for every 15 s record. The y-axis represents the number of 15 s records collected during the 30-day period.

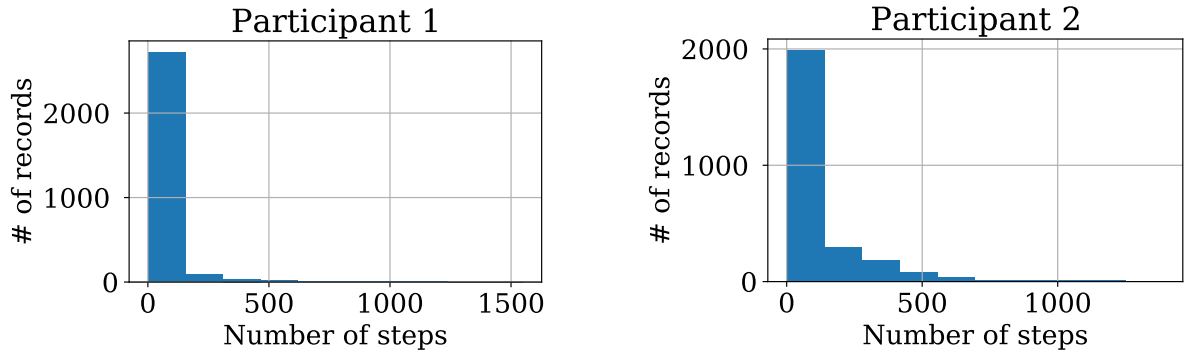


Figure 6. Participants' step count record distribution. The x-axis represents the number of steps for every 15 min record. The y-axis represents the number of 15 min records collected during the 30-day period.

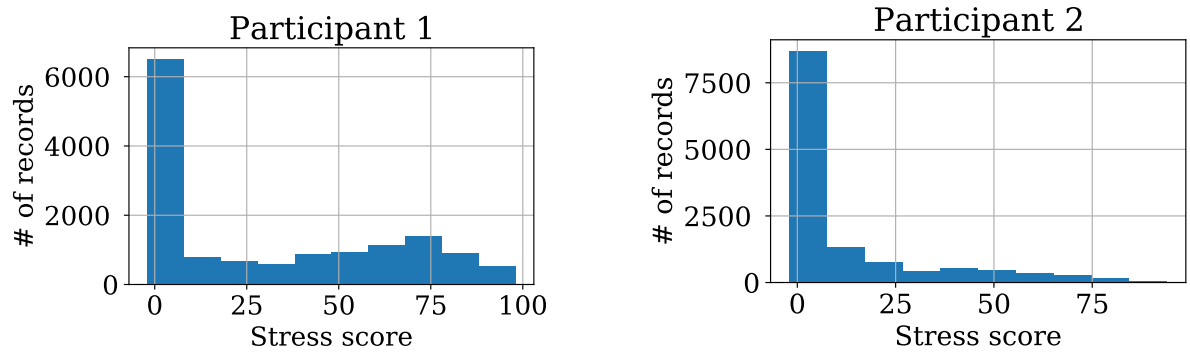


Figure 7. Participants' stress count record distribution. The x-axis represents the 3 min stress score reported by Garmin Health API. The y-axis represents the number of 3 min records collected during the 30-day period.

Finally, the duration of sleep was computed from the provided Garmin data. Participant 1 had an average sleep duration of 265.793 ± 93.681 min, or 4 h and 25 min, over 29 days. Meanwhile, participant 2 had an average sleep duration of 237.48 ± 67.430 min, or 3 h and 57 min, over a smaller number of days ($n = 25$). Figure 8 shows that participant 1 had longer light and REM sleep than participant 2. On the other hand, participant 2 had longer deep sleep than participant 1.

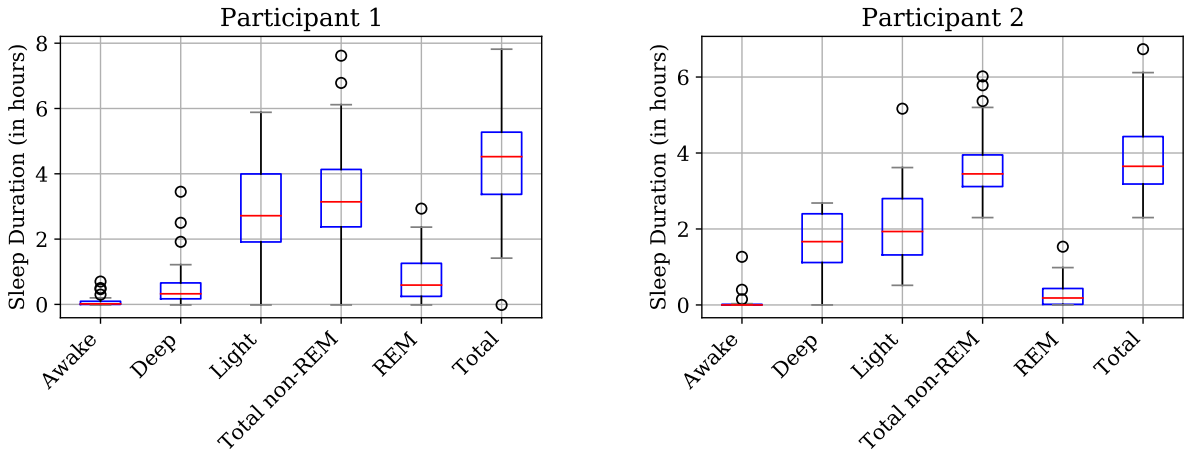


Figure 8. Participants' amount of sleep by sleep classification. The x-axis represents each sleep classification (awake, deep, light, and REM sleep) and aggregated sleep feature (total non-REM and total sleep). The y-axis represents the sleep duration in hours.

3.3. Difference between the Participants' PD Experience

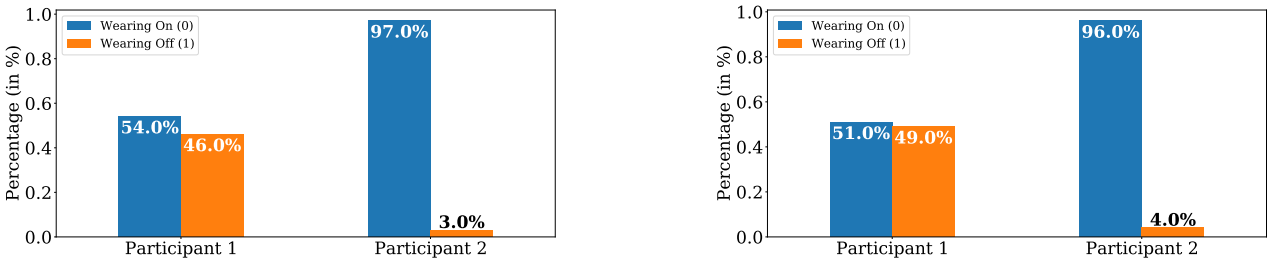
We also investigated our initial hypothesis that each patient would experience PD differently using a two-sample t-test with unequal variances, and effect sizes were reported using Cohen's d. There were significant differences between the participants' heart rates ($t(300{,}765) = 1.96$, $p < 0.001$, $d = 0.995$), stress scores ($t(17{,}043) = 1.96$, $p < 0.001$, $d = 0.688$), and numbers of steps ($t(1606) = 1.96$, $p < 0.001$, $d = 0.486$), where all three p values were below 0.001. However, there was no significant difference between the participants' daily amount of sleep, $t(52) = 2.01$, $p = 0.23$, $d = 0.328$.

3.4. Combined Datasets

The fitness tracker datasets, the wearing-off dataset, and the drug intake dataset were combined according to the method explained in Section 2.2. To recap, we prepared 15 s and 15 min datasets based on the 15 s heart rate interval and 15 min step interval, respectively. Then, these 15 s and 15 min datasets were used for each participant in our prediction model.

3.5. Wearing-Off Prediction Model

A series of pipelines for developing the wearing-off prediction model was applied to each participant's 15-s and 15-min datasets. We observed a class imbalance for the wearing-off y variable of participant 2 in both time frames. On the other hand, participant 1's dataset did not have a substantial class imbalance. Figure 9 presents the disparity in the classes for each participant.



(**a**) 15-s Dataset (**b**) 15-min Dataset

Figure 9. Wearing-Off ratio among participants.

3.5.1. Participant 1 Prediction Model

For participant 1, we analyzed the performance of each pipeline on the validation set. The CI pipeline produced the best balanced accuracy of 74.7% for the 15 min time frame and 73.5% for the 15 s time frame, as shown in Table 9.

Table 9. Best hyperparameter configuration performance on the validation set for participant 1. CI pipeline resulted as the best pipeline for both time frames, as shown in bold.

	15 min Interval				15 s Interval			
Metrics	**Initial**	**FS**	**CI**	**CI + FS**	**Initial**	**FS**	**CI**	**CI + FS**
Bal. Acc.	0.742	0.744	**0.747**	0.743	0.730	0.729	**0.735**	0.728
F1 Score	0.735	0.732	**0.745**	0.731	0.686	0.682	**0.704**	0.691
Acc.	0.742	0.744	**0.747**	0.743	0.738	0.738	**0.739**	0.734
Precision	0.749	0.761	**0.746**	0.760	0.764	0.768	**0.739**	0.743
Recall/Sn	0.722	0.706	**0.747**	0.713	0.633	0.624	**0.687**	0.656
Sp	0.761	0.781	**0.746**	0.773	0.827	0.833	**0.784**	0.800

Next, we used the CI pipeline to compare the performance of each learning algorithms using the validation set. In the 15 min time frame, the GB algorithm produced the best balanced accuracy of 74.3% and and an f1-score of 74.0% over other algorithms. The next best learning algorithm was DT, with a 73.9% balanced accuracy and a 74.4% f1-score. Similarly, the GB algorithm had the best balanced accuracy of 72.7% and an f1-score of 69.7% on the 15 s time frame. However, the L. SVM had the second best performance, with 71.4% balanced accuracy and an f1-score of 63.3%. Table 10 shows the performance of the other learning algorithms.

Table 10. Comparison of learning algorithm performance on the validation set using the CI pipeline for participant 1. DT, L. SVM, and GB learning algorithms produced the best results in 15 min and 15 s interval, as shown in bold.

	15 min Interval					15 s Interval				
Metrics	**LR**	**DT**	**L. SVM**	**RF**	**GB**	**LR**	**DT**	**L. SVM**	**RF**	**GB**
Bal. Acc.	0.712	**0.739**	0.655	0.691	**0.743**	0.698	0.659	**0.714**	0.668	**0.727**
F1 Score	0.713	**0.744**	0.563	0.671	**0.740**	0.682	0.617	**0.633**	0.609	**0.697**
Acc.	0.712	**0.739**	0.655	0.691	**0.742**	0.697	0.664	**0.731**	0.678	**0.731**
Precision	0.704	**0.723**	0.757	0.709	**0.744**	0.655	0.645	**0.839**	0.684	**0.720**
Recall/Sn	0.725	**0.771**	0.597	0.642	**0.739**	0.715	0.598	**0.519**	0.552	**0.679**
Sp	0.699	**0.707**	0.712	0.740	**0.746**	0.682	0.720	**0.909**	0.784	**0.775**

Thus, we evaluated the GB, DT, and L. SVM algorithms using the CI pipeline for model evaluation. In the end, the CI pipeline with the GB learning algorithm had the best balanced accuracy and f1-score in both time frames. The prediction model had an average balanced accuracy of 71.7% $\pm$ 0.035 for the 15 min time frame and 70.0% $\pm$ 0.032 for the 15 s time frame, as shown in Table 11. In addition, the permutation feature importance for the prediction models were computed to identify which feature contributed to the prediction model. The permutation feature importance measured the prediction model error after a feature was permuted [41]. DrugElapsed, Stress, and HR were the top features in both time frames for participant 1 (Table 12).

Table 11. Average performance of best models on the test set using the CI pipeline for participant 1. GB learning algorithm produced the best results in both intervals, as shown in bold.

	15 min Interval		15 s Interval	
Metrics	**GB**	**DT**	**GB**	**L. SVM**
Bal. Acc.	**0.717 ± 0.04**	0.675 ± 0.04	**0.700 ± 0.03**	0.563 ± 0.09
F1 Score	**0.700 ± 0.04**	0.664 ± 0.03	**0.676 ± 0.03**	0.633 ± 0.04
Acc.	**0.718 ± 0.03**	0.676 ± 0.04	**0.701 ± 0.04**	0.538 ± 0.10
Precision	**0.740 ± 0.05**	0.685 ± 0.04	**0.683 ± 0.07**	0.513 ± 0.08
Recall/Sn	**0.669 ± 0.07**	0.647 ± 0.05	**0.685 ± 0.09**	0.860 ± 0.12
Sp	**0.765 ± 0.07**	0.704 ± 0.07	**0.714 ± 0.13**	0.266 ± 0.26

Table 12. The permutation feature importance for the best models for participant 1 using the CI pipeline with GB. Values were balanced accuracy error when a feature was permuted. The larger the error value, the more important the feature was for the model.

15 min		15 s	
Features	**Feature Importance**	**Feature**	**Feature Importance**
DrugElapsed	0.190 ± 0.018	DrugElapsed	0.168 ± 0.002
Stress	0.023 ± 0.009	Stress	0.048 ± 0.002
HR	0.012 ± 0.009	HR	0.009 ± 0.001
Awake	0.008 ± 0.004	Steps	0.007 ± 0.001
Steps	0.006 ± 0.003	Deep	0.002 ± 0.000
SleepEfficiency	0.004 ± 0.002	Awake	0.000 ± 0.000
Total	0.003 ± 0.005	Total	0.000 ± 0.000
NonREMTotal	0.002 ± 0.001	NonREMPercentage	−0.000 ± 0.000
Light	−0.000 ± 0.002	NonREMTotal	−0.001 ± 0.000
Deep	−0.002 ± 0.003	SleepEfficiency	−0.001 ± 0.000
REM	−0.004 ± 0.003	Light	−0.002 ± 0.000
NonREMPercentage	−0.008 ± 0.002	REM	−0.002 ± 0.000

3.5.2. Participant 2 Prediction Model

The same process was conducted for participant 2's prediction model. First, we compared the performance of each pipeline on the validation set in both time frames. The best balanced accuracy for both time frame was the initial pipeline. However, upon further testing with the pipelines, and due to the lack of wearing-off reports from participant 2, we opted to use the CI pipeline to handle the class imbalance. The 15 min time frame had a balanced accuracy of 75.6%, while the 15 s time frame had a balanced accuracy of 74.8%. The other pipelines' performances on the validation set are presented in Table 13.

Table 13. Best hyperparameter configuration performance on the validation set for participant 2. CI pipeline was the chosen pipeline, as shown in bold.

	15 min Interval				15 s Interval			
Metrics	**Initial**	**FS**	**CI**	**CI + FS**	**Initial**	**FS**	**CI**	**CI + FS**
Bal. Acc.	0.770	0.763	**0.756**	0.758	0.745	0.734	**0.748**	0.754
F1 Score	0.265	0.233	**0.304**	0.235	0.242	0.174	**0.183**	0.183
Acc.	0.658	0.597	**0.573**	0.562	0.776	0.711	**0.567**	0.555
Precision	0.273	0.191	**0.275**	0.182	0.236	0.132	**0.147**	0.147
Recall/Sn	0.889	0.941	**0.953**	0.967	0.713	0.758	**0.938**	0.964
Sp	0.650	0.585	**0.560**	0.548	0.778	0.710	**0.558**	0.544

Second, the learning algorithms were compared using the CI pipeline for both time frames. For the 15 min time frame, the LR learning algorithm had the best balanced

accuracy of 73.4% on the validation set. On the other hand, GB had the best balanced accuracy of 70.7% for the 15 s time frame. In both time frames, the second best algorithm was L. SVM. The results of the other learning algorithms, as well as other metrics, are shown in Table 14.

Table 14. Comparison of learning algorithm performance on the validation set using the CI for participant 2. The learning algorithms (LR, L. SVM, and GB) with the best results for each interval were shown in bold.

	15 min Interval					15 s Interval				
Metrics	**LR**	**DT**	**L. SVM**	**RF**	**GB**	**LR**	**DT**	**L. SVM**	**RF**	**GB**
Bal. Acc.	**0.734**	0.695	**0.724**	0.649	0.697	0.672	0.665	**0.678**	0.682	**0.707**
F1 Score	**0.310**	0.204	**0.226**	0.195	0.202	0.141	0.093	**0.165**	0.092	**0.172**
Acc.	**0.541**	0.448	**0.485**	0.438	0.431	0.434	0.479	**0.415**	0.582	**0.475**
Precision	**0.295**	0.159	**0.186**	0.159	0.157	0.113	0.050	**0.135**	0.049	**0.141**
Recall/Sn	**0.940**	0.959	**0.980**	0.875	0.982	0.923	0.860	**0.956**	0.788	**0.952**
Sp	**0.527**	0.430	**0.468**	0.422	0.411	0.421	0.469	**0.400**	0.576	**0.462**

Third, the LR, GB, and L. SVM algorithms were evaluated on the test sets using the CI pipeline. In the 15 min time frame, LR had an average balanced accuracy of 76.9% $\pm$ 0.176. Meanwhile, GB had an average balanced accuracy of 76.1% $\pm$ 0.120 in the 15 s time frame (Table 15). The evaluated prediction models for participant 2 only showed DrugElapsed as the common top feature, while other features' importances varied, as shown in Table 16.

Table 15. Average performance of best models on the test set using the CI pipeline for participant 2. LR and GB learning algorithm produced the best result in 15 min interval and 15 s interval, respectively.

	15 min Interval		15 s Interval	
Metrics	**LR**	**L. SVM**	**GB**	**L. SVM**
Bal. Acc.	**0.769 $\pm$ 0.176**	0.571 $\pm$ 0.255	**0.761 $\pm$ 0.120**	0.493 $\pm$ 0.048
F1 Score	**0.486 $\pm$ 0.350**	0.272 $\pm$ 0.314	**0.425 $\pm$ 0.264**	0.066 $\pm$ 0.039
Acc.	**0.768 $\pm$ 0.344**	0.674 $\pm$ 0.329	**0.783 $\pm$ 0.324**	0.358 $\pm$ 0.405
Precision	**0.635 $\pm$ 0.449**	0.354 $\pm$ 0.423	**0.461 $\pm$ 0.367**	0.049 $\pm$ 0.050
Recall/Sn	**0.770 $\pm$ 0.316**	0.460 $\pm$ 0.413	**0.739 $\pm$ 0.227**	0.636 $\pm$ 0.380
Sp	**0.769 $\pm$ 0.361**	0.682 $\pm$ 0.341	**0.784 $\pm$ 0.337**	0.351 $\pm$ 0.425

Table 16. The permutation feature importance for the best models for participant 2 using the CI pipeline with LR and GB. Values were balanced accuracy error when a feature was permuted. The larger the error value, the more important the feature was for the model.

15 min		15 s	
Features	**Feature Importance**	**Feature**	**Feature Importance**
DrugElapsed	0.255 $\pm$ 0.025	DrugElapsed	0.254 $\pm$ 0.003
NonREMTotal	0.108 $\pm$ 0.023	Steps	0.039 $\pm$ 0.005
Total	0.079 $\pm$ 0.024	NonREMPercentage	0.013 $\pm$ 0.002
Awake	0.064 $\pm$ 0.028	REM	0.006 $\pm$ 0.001
Steps	0.013 $\pm$ 0.011	Light	0.003 $\pm$ 0.002
Light	0.008 $\pm$ 0.026	Awake	0.002 $\pm$ 0.000
HR	0.003 $\pm$ 0.011	Stress	0.001 $\pm$ 0.002
Stress	0.002 $\pm$ 0.007	Deep	$-$0.003 $\pm$ 0.001
Deep	$-$0.018 $\pm$ 0.022	SleepEfficiency	$-$0.006 $\pm$ 0.001
NonREMPercentage	$-$0.028 $\pm$ 0.006	Total	$-$0.008 $\pm$ 0.002
SleepEfficiency	$-$0.029 $\pm$ 0.005	HR	$-$0.019 $\pm$ 0.002
REM	$-$0.029 $\pm$ 0.022	NonREMTotal	$-$0.023 $\pm$ 0.001

4. Discussion

This study aimed to develop an individual-level prediction model for the wearing-off of anti-PD medicine using a readily available fitness tracker in a real-world environment. Our first goal was to collect and combine the related datasets needed for this study. Second, prediction models were built to predict the wearing-off period. Lastly, we analyzed the produced results to assess our primary goal.

In response to our first research question, there were two main challenges in data collection and processing. First, various datasets from different tools were collected in different time intervals. There were some significant gaps without available data. Before re-sampling, the datasets were filled with values of "−1", based on Garmin's data handling protocol for stress scores. After re-sampling the datasets, the last known values were copied to missing values. It would be convenient to have a uniform interval rather than to re-sample the data for future studies. The use of smartwatches can also be beneficial to read and record raw sensor data, without sacrificing the comfort of the participants. Second, there were few wearing-off reports for participant 2. The sparsity and the low number of reports affected the performance of the prediction models. Participant 1 had sufficient reports, while participant 2 did not have very many reports.

In this study, we have shown the feasibility of using a fitness tracker to predict the wearing-off phenomenon. Based on the feature importance from the models, the elapsed time since the last drug intake affected both participants' models regardless of the sampling interval. For participant 1, the stress score and heart rate were indicators of the wearing-off phenomenon.

For both participants, the 15 min time frame produced better results than the 15 s time frame. The difference in the time frame affected the prediction model due to the presence of large amounts of missing data. In the future, the sampling interval can be optimized since the differences between the two time frames were substantial.

In response to the second research question, our case study showed the possibility of utilizing the fitness tracker dataset to estimate wearing-off. The individual-level prediction models were built to personalize the estimation based on the participants' experiences of PD. It was highlighted in a previous study that each PD patient had varying degrees of symptoms [20]. In this study's analysis, the participants' heart rate, number of steps, and stress score were also shown have significant differences. Thus, this study was able to produce a prediction model with an accuracy of 70.0–71.7% for participant 1 and 76.1–76.9% for participant 2 with the use of a non-motor dataset; i.e., in contrast with the accelerometer and gyroscope datasets, which were based on motor functions (Table 17).

Table 17. Comparison of previous studies on predicting wearing-off and the current study.

Study	Aim	Data Used	Method	Result
Keijsers, 2006 [51]	Determine between "On" and "Off" based on daily activities using wearable data	Accelerometer	Unsupervised method using frequency-based method	Sensitivity: 97% Specificity: 97%
Hssayeni, 2019 [52]	Detect "On" and"Off" states	Gyroscope	SVM with fuzzy labeling	Accuracy: 90.5% Sensitivity: 94.2% Specificity: 85.4%
Aich, 2020 [12]	Detect "On" and "Off" using gait signals	Accelerometer	Random forest, SVM, kNN, Naive Bayes	Accuracy: 96.72% Sensitivity: 97.35%
Current study	Predict "wearing-off" on individual-level	Heart rate Stress score Sleep features Step count	Random Forest, Gradient Boosting, Logistic Regression, Linear SVM, Decision Tree	Accuracy P1: 70.0–71.7% P2: 76.1–76.9%

The prediction models could also be improved by reducing false positive and false negative predictions, as shown in Tables 18 and 19. It was possible that when the prediction model learned and discovered the wearing-off event, the participant has not reported the wearing-off. We learned this possibility from our short interview with the participants. Participant 2 often experienced mild to severe symptoms which resulted in fewer self-reports. Moreover, the participants forgot to replace their fitness tracker after doing household chores. Then, their symptoms began to manifest again. These events were backed up by the known information about the participants, such as fitness tracker usage (Table 8), PDQ-8, and H&Y profile (Table 3). In future work, these information will be helpful before data collection and would improve the study overall.

Table 18. Confusion matrix for participant 1's best model on the 15 min test set.

	Predicted Wearing-Off (1)	**Predicted Actual On (0)**
Actual Wearing-Off (1)	955 True positive (TP)	472 False negative (FN)
Actual On (0)	341 False positive (FP)	1112 True negative (TN)

Table 19. Confusion matrix for participant 2's best model on the 15 min test set.

	Predicted Wearing-Off (1)	**Predicted Actual On (0)**
Actual Wearing-Off (1)	76 True positive (TP)	23 False negative (FN)
Actual On (0)	644 False positive (FP)	2137 True negative (TN)

5. Conclusions

This study showed the development of prediction models to predict wearing-off for PD patients. A commercially available fitness tracker (Garmin vivosmart4) was used to collect data in a real-world environment. Moreover, the participants recorded their wearing-off periods and drug intake with a customized smartphone application (FonLog). Then, the datasets were combined to build our prediction models. Predictive models were built for each participant as there were significant differences between the participants' data, except for one feature. Developing predictive models for each participant in a challenging data collection environment resulted in a balanced accuracy of 70.0–71.7% for participant 1 and 76.1–76.9% for participant 2 on the test set. Both models utilized a class imbalanced transformer in their model development pipeline. Finally, only the time that elapsed after taking anti-PD medicine was a common predictor between the participants. Heart rates and stress scores contributed to the prediction model of participant 1, who presented more reports of wearing-off over the 30 days.

In the future, more participants will be part of the study to assess the differences among demographic groups. Moreover, these baseline models will be deployed to an early warning system for the participants. The reported feedback will be used to improve the prediction models and develop a symptom-based prediction model. We envision the utilization of these baseline models while collecting more data using a smartwatch.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/app11167354/s1: "data/garmin" contains the available raw Garmin dataset from Garmin Health API, "data/fonlog/records.xlsx" contains the raw FonLog dataset, "data/combined_data/combined_ data_participant*.xlsx" contains data for the processed and combined dataset for each participant in different re-sampled time interval, "Data Processing.ipynb" contains the data processing source code, "Model Development" contains the source code for model pipelines, and the "About Supplementary Files" contains the guide to the Supplementary Materials.

Author Contributions: Conceptualization, T.S. and S.I.; Data curation, J.N.V. and S.I.; Formal analysis, J.N.V.; Funding acquisition, T.S.; Investigation, J.N.V. and Y.S.; Methodology, J.N.V., T.S. and S.I.; Project administration, T.S. and J.N.V.; Resources, T.S. and S.I.; Software, J.N.V. and S.I.; Supervision, T.S. and S.I.; Validation, T.S. and Y.S.; Visualization, J.N.V.; Writing—original draft, J.N.V.; Writing—review and editing, Y.S., S.I. and T.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the JSPS KAKENHI Grant Number JP16H06534.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki.

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

Data Availability Statement: The data are made available in the Supplementary Materials.

Acknowledgments: The authors would like to thank the Japanese Ministry of Education, Culture, Sports, Science, and Technology (MEXT) and the participants of our study.

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

Sample Availability: Samples of the data and the analysis can be obtained with the corresponding author.

Abbreviations

The following abbreviations are used in this manuscript:

API	Application Programming Interface
AWS	Amazon Web Services
Bal. Acc.	Balanced accuracy
bpm	Beats per minute
CI	Class Imbalance transformer
CI + FS	Class Imbalance transformer with Feature Selection
DT	Decision Tree
EMG	Electromyography
FN	False negative
FoG	Freezing of gait
FP	False positive
FS	Feature Selection
GB	Gradient Boosting
H&Y	Hoehn and Yahr Scale
JCLD	Japan Ministry of Health, Labor & Welfare's classification of living dysfunction
L. SVM	Linear Support Vector Machine
LR	Logistic Regression
MAO-B	Monoamine oxidase type B
Nested CV	Nested cross-fold
PD	Parkinson's Disease
PDQ-8	Parkinson's Disease Questionnaire-8
PKG	Parkinson's KinetiGraph
PPG	Photoplethysmography
QoL	Quality of Life
RBD	REM sleep behavior disorder
REM	Rapid eye movement
RF	Random Forest
RFE	Recursive Feature Selection
SN	Sensitivity
SP	Specificity
TN	True negative
TP	True positive
UPDRS III	Unified Parkinson's Disease Rating Scale Part III
WoQ-9	Wearing-off Questionnaire-9

References

1. Antonini, A.; Martinez-Martin, P.; Chaudhuri, R.K.; Merello, M.; Hauser, R.; Katzenschlager, R.; Odin, P.; Stacy, M.; Stocchi, F.; Poewe, W.; et al. Wearing-off Scales in Parkinson's Disease: Critique and Recommendations: Scales to Assess Wearing-Off in PD. *Mov. Disord.* **2011**, *26*, 2169–2175. [CrossRef]
2. Colombo, D.; Abbruzzese, G.; Antonini, A.; Barone, P.; Bellia, G.; Franconi, F.; Simoni, L.; Attar, M.; Zagni, E.; Haggiag, S.; et al. The "Gender Factor" in Wearing-Off among Patients with Parkinson's Disease: A Post Hoc Analysis of DEEP Study. *Sci. World J.* **2015**, *2015*, 787451. [CrossRef]
3. Stacy, M.; Hauser, R.; Oertel, W.; Schapira, A.; Sethi, K.; Stocchi, F.; Tolosa, E. End-of-Dose Wearing off in Parkinson Disease: A 9-Question Survey Assessment. *Clin. Neuropharmacol.* **2006**, *29*, 312–321. [CrossRef]
4. Stocchi, F.; Antonini, A.; Barone, P.; Tinazzi, M.; Zappia, M.; Onofrj, M.; Ruggieri, S.; Morgante, L.; Bonuccelli, U.; Lopiano, L.; et al. Early DEtection of wEaring off in Parkinson Disease: The DEEP Study. *Park. Relat. Disord.* **2014**, *20*, 204–211. [CrossRef] [PubMed]
5. Jeon, H.; Lee, W.; Park, H.; Lee, H.J.; Kim, S.K.; Kim, H.B.; Jeon, B.; Park, K.S. Automatic Classification of Tremor Severity in Parkinson's Disease Using a Wearable Device. *Sensors* **2017**, *17*, 2067. [CrossRef] [PubMed]
6. Samà, A.; Pérez-López, C.; Rodríguez-Martín, D.; Català, A.; Moreno-Aróstegui, J.M.; Cabestany, J.; de Mingo, E.; Rodríguez-Molinero, A. Estimating Bradykinesia Severity in Parkinson's Disease by Analysing Gait through a Waist-Worn Sensor. *Comput. Biol. Med.* **2017**, *84*, 114–123. [CrossRef]
7. Naghavi, N.; Miller, A.; Wade, E. Towards Real-Time Prediction of Freezing of Gait in Patients With Parkinson's Disease: Addressing the Class Imbalance Problem. *Sensors* **2019**, *19*, 3898. [CrossRef]
8. Silva de Lima, A.L.; Hahn, T.; Evers, L.J.W.; de Vries, N.M.; Cohen, E.; Afek, M.; Bataille, L.; Daeschler, M.; Claes, K.; Boroojerdi, B.; et al. Feasibility of Large-Scale Deployment of Multiple Wearable Sensors in Parkinson's Disease. *PLoS ONE* **2017**, *12*, e0189161. [CrossRef]
9. Farzanehfar, P.; Woodrow, H.; Horne, M. Assessment of Wearing Off in Parkinson's Disease Using Objective Measurement. *J. Neurol.* **2021**, *268*, 914–922. [CrossRef] [PubMed]
10. Khodakarami, H.; Ricciardi, L.; Contarino, M.F.; Pahwa, R.; Lyons, K.; Geraedts, V.; Morgante, F.; Leake, A.; Paviour, D.; De Angelis, A.; et al. Prediction of the Levodopa Challenge Test in Parkinson's Disease Using Data from a Wrist-Worn Sensor. *Sensors* **2019**, *19*, 5153. [CrossRef] [PubMed]
11. Ossig, C.; Gandor, F.; Fauser, M.; Bosredon, C.; Churilov, L.; Reichmann, H.; Horne, M.K.; Ebersbach, G.; Storch, A. Correlation of Quantitative Motor State Assessment Using a Kinetograph and Patient Diaries in Advanced PD: Data from an Observational Study. *PLoS ONE* **2016**, *11*, e0161559. [CrossRef] [PubMed]
12. Aich, S.; Youn, J.; Chakraborty, S.; Pradhan, P.M.; Park, J.H.; Park, S.; Park, J. A Supervised Machine Learning Approach to Detect the On/Off State in Parkinson's Disease Using Wearable Based Gait Signals. *Diagnostics* **2020**, *10*, 421. [CrossRef] [PubMed]
13. Powers, R.; Etezadi-Amoli, M.; Arnold, E.M.; Kianian, S.; Mance, I.; Gibiansky, M.; Trietsch, D.; Alvarado, A.S.; Kretlow, J.D.; Herrington, T.M.; et al. Smartwatch Inertial Sensors Continuously Monitor Real-World Motor Fluctuations in Parkinson's Disease. *Sci. Transl. Med.* **2021**, *13*, eabd7865. [CrossRef]
14. Garmin. Activity Tracking and Fitness Metric Accuracy. 2021. Available online: https://www.garmin.com/en-US/legal/atdisclaimer/ (accessed on 1 May 2021).
15. Pursiainen, V.; Korpelainen, J.; Haapaniemi, T.; Sotaniemi, K.; Myllylä, V. Blood Pressure and Heart Rate in Parkinsonian Patients with and without Wearing-Off. *Eur. J. Neurol. Off. J. Eur. Fed. Neurol. Soc.* **2007**, *14*, 373–378. [CrossRef]
16. Chaudhuri, K.R.; Healy, D.G.; Schapira, A.H. Non-Motor Symptoms of Parkinson's Disease: Diagnosis and Management. *Lancet Neurol.* **2006**, *5*, 235–245. [CrossRef]
17. Sixel-Döring, F.; Trautmann, E.; Mollenhauer, B.; Trenkwalder, C. Associated Factors for REM Sleep Behavior Disorder in Parkinson Disease. *Neurology* **2011**, *77*, 1048–1054. [CrossRef]
18. Smith, M.C.; Ellgring, H.; Oertel, W.H. Sleep Disturbances in Parkinson's Disease Patients and Spouses. *J. Am. Geriatr. Soc.* **1997**, *45*, 194–199. [CrossRef]
19. Claassen, D.O.; Kutscher, S.J. Sleep Disturbances in Parkinson's Disease Patients and Management Options. *Nat. Sci. Sleep* **2011**, *3*, 125–133. [CrossRef]
20. Macht, M.; Schwarz, R.; Ellgring, H. Patterns of Psychological Problems in Parkinson's Disease. *Acta Neurol. Scand.* **2005**, *111*, 95–101. [CrossRef]
21. Van der Heide, A.; Meinders, M.J.; Speckens, A.E.M.; Peerbolte, T.F.; Bloem, B.R.; Helmich, R.C. Stress and Mindfulness in Parkinson's Disease: Clinical Effects and Potential Underlying Mechanisms. *Mov. Disord.* **2021**, *36*, 64–70. [CrossRef]
22. Collins, T.; Woolley, S.I.; Oniani, S.; Pires, I.M.; Garcia, N.M.; Ledger, S.J.; Pandyan, A. Version Reporting and Assessment Approaches for New and Updated Activity and Heart Rate Monitors. *Sensors* **2019**, *19*, 1705. [CrossRef]
23. Stevens, S.; Siengsukon, C. Commercially-Available Wearable Provides Valid Estimate of Sleep Stages (P3.6-042). 2019. Available online: https://n.neurology.org/content/92/15_Supplement/P3.6-042 (accessed on 3 July 2021).
24. Kim, H.G.; Cheon, E.J.; Bai, D.S.; Lee, Y.H.; Koo, B.H. Stress and Heart Rate Variability: A Meta-Analysis and Review of the Literature. *Psychiatry Investig.* **2018**, *15*, 235–245. [CrossRef] [PubMed]
25. Hehlmann, M.I.; Schwartz, B.; Lutz, T.; Gómez Penedo, J.M.; Rubel, J.A.; Lutz, W. The Use of Digitally Assessed Stress Levels to Model Change Processes in CBT - A Feasibility Study on Seven Case Examples. *Front. Psychiatry* **2021**, *12*, 258. [CrossRef]

26. Mouritzen, N.J.; Larsen, L.H.; Lauritzen, M.H.; Kjær, T.W. Assessing the Performance of a Commercial Multisensory Sleep Tracker. *PLoS ONE* **2020**, *15*, e0243214. [CrossRef]
27. Garmin. My Fitness Device Is Not Accurately Counting Steps. 2021. Available online: https://support.garmin.com/en-US/?faq=z1TfJCqajl8ZEZey72gg98 (accessed on 3 July 2021).
28. Garmin. Vívosmart 4—Heart Rate Variability and Stress Level. 2020. Available online: https://www8.garmin.com/manuals/webhelp/vivosmart4/EN-US/GUID-9282196F-D969-404D-B678-F48A13D8D0CB.html (accessed on 1 May 2021).
29. Garmin. Garmin Vivosmart 4. 2020. Available online: https://buy.garmin.com/en-US/US/p/605739 (accessed on 1 May 2021).
30. Mairittha, N.; Mairittha, T.; Inoue, S. A mobile app for nursing activity recognition. In Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, Singapore, 8–12 October 2018; pp. 400–403.
31. Fukae, J.; Higuchi, M.A.; Yanamoto, S.; Fukuhara, K.; Tsugawa, J.; Ouma, S.; Hatano, T.; Yoritaka, A.; Okuma, Y.; Kashihara, K.; et al. Utility of the Japanese Version of the 9-Item Wearing-off Questionnaire. *Clin. Neurol. Neurosurg.* **2015**, *134*, 110–115. [CrossRef]
32. Bhidayasiri, R.; Tarsy, D. Parkinson's disease: Hoehn and Yahr Scale. In *Movement Disorders: A Video Atlas*; Current Clinical Neurology; Humana: Totowa, NJ, USA, 2012; pp. 4–5._2 [CrossRef]
33. Kashiwara, K.; Takeda, A.; Maeda, T. *Learning Parkinson's Disease Together with Patients: Toward a Medical Practice that Works with Patients, with Q&A*; Nankodo: Tokyo, Japan, 2013. Available online: https://honto.jp/netstore/pd-book_25644244.html (accessed on 4 April 2021). (In Japanese)
34. Jenkinson, C.; Fitzpatrick, R.; Peto, V.; Greenhall, R.; Hyman, N. The PDQ-8: Development and validation of a short-form Parkinson's disease questionnaire. *Psychol. Health* **1997**, *12*, 805–814. [CrossRef]
35. Schroeder, L.A.; Rufra, O.; Sauvageot, N.; Fays, F.; Pieri, V.; Diederich, N.J. Reduced Rapid Eye Movement Density in Parkinson Disease: A Polysomnography-Based Case-Control Study. *Sleep* **2016**, *39*, 2133–2139. [CrossRef]
36. Reed, D.L.; Sacco, W.P. Measuring Sleep Efficiency: What Should the Denominator Be? *J. Clin. Sleep Med. JCSM Off. Publ. Am. Acad. Sleep Med.* **2016**, *12*, 263–266. [CrossRef]
37. Leenings, R.; Winter, N.R.; Plagwitz, L.; Holstein, V.; Ernsting, J.; Steenweg, J.; Gebker, J.; Sarink, K.; Emden, D.; Grotegerd, D.; et al. PHOTONAI—A Python API for Rapid Machine Learning Model Development. *PLoS ONE* **2021**, *16*, e0254062. [CrossRef] [PubMed]
38. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
39. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
40. Hastie, T.; Tibshirani, R.; Friedman, J.H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer Series in Statistics; Springer: New York, NY, USA, 2009.
41. Molnar, C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*; Christoph Molnar: Munich, Germany, 2019. Available online: https://christophm.github.io/interpretable-ml-book/ (accessed on 4 April 2021).
42. Zhong, Y.; He, J.; Chalise, P. Nested and Repeated Cross Validation for Classification Model With High-Dimensional Data. *Rev. Colomb. Estad.* **2020**, *43*, 103–125. [CrossRef]
43. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
44. Li, S.; Feng, L.; Ge, Y.; Zhu, L.; Zhao, L. An Ensemble Learning Method for Robot Electronic Nose with Active Perception. *Sensors* **2021**, *21*, 3941. [CrossRef] [PubMed]
45. Head, T.; Kumar, M.; Nahrstaedt, H.; Louppe, G.; Shcherbatyi, I. Scikit-optimize/scikit-optimize. *Zenodo* **2020**, 4014775. [CrossRef]
46. Lemaître, G.; Nogueira, F.; Aridas, C.K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *J. Mach. Learn. Res.* **2017**, *18*, 1–5.
47. Cawley, G.C.; Talbot, N.L.C. On Over-Fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *J. Mach. Learn. Res.* **2010**, *11*, 2079–2107.
48. Brownlee, J. Nested Cross-Validation for Machine Learning with Python. 2020. Available online: https://machinelearningmastery.com/nested-cross-validation-for-machine-learning-with-python/ (accessed on 4 April 2021).
49. Raschka, S. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. *arXiv* **2020**, arXiv:1811.12808.
50. Brodersen, K.H.; Ong, C.S.; Stephan, K.E.; Buhmann, J.M. The Balanced Accuracy and Its Posterior Distribution. In Proceedings of the 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; IEEE: Istanbul, Turkey, 2010; pp. 3121–3124. [CrossRef]
51. Keijsers, N.L.W.; Horstink, M.W.I.M.; Gielen, S.C.A.M. Ambulatory Motor Assessment in Parkinson's Disease. *Mov. Disord.* **2006**, *21*, 34–44. [CrossRef]
52. Hssayeni, M.D.; Burack, M.A.; Jimenez-Shahed, J.; Ghoraani, B. Assessment of Response to Medication in Individuals with Parkinson's Disease. *Med. Eng. Phys.* **2019**, *67*, 33–43. [CrossRef]