Beyond the Lab: Challenges to Detect Parkinson's Disease Symptoms in Remote Smartwatch Data

by

Inge Schut

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended on Wednesday January 31, 2024 at 14:00.

Master Program:	Mechanical Engineering - BioMechanical Design	
Student Number:	4549031	
Thesis Committee:	Prof. Dr. Ir. A.C. Schouten,	TU Delft, supervisor
	Dr. M. Kok	TU Delft

TU Delft

List of Abbreviations

Abbreviation	Definition
aDBS	Adaptive Deep Brain Stimulation
ADL	Activities of Daily Living
API	Application Programming Interface
CNN	Convolutional Neural Network
DBS	Deep Brain Stimulation
DL	Deep Learning
ECG	Electrocardiogram
FFT	Fast Fourier Transform
IMU	Inertial Measurement Unit
KL	Kullback-Leibler
LFP	Local Field Potential
ML	Machine Learning
MSE	Mean Squared Error
PD	Parkinson's Disease
PSD	Power Spectral Density
RAM	Random Access Memory
ReLU	Rectifier Linear Unit
SD	Standard Deviation
UMC	University Medical Center
UPDRS	Unified Parkinson Disease Rating Scale
VAE	Variational Autoencoder
VM	Vector Magnitude

Thesis Msc. Mechanical Engineering - Biomechanical Design
I. Schut, Delft, The Netherlands, compiled January 26, 2024

Beyond the Lab: Challenges to Detect Parkinson's Disease Symptoms in Remote Smartwatch Data

Inge Schut

***Abstract*—Parkinson's Disease (PD) is a neurodegenerative disorder with four cardinal motor symptoms: bradykinesia, tremor, rigidity, and postural instability. Adaptive Deep Brain Stimulation (aDBS) is a promising treatment for PD that provides stimulation based on the expression of PD symptoms, improving effectiveness and reducing side effects compared to continuous DBS. Smartwatches can facilitate aDBS by enabling continuous detection of tremor and bradykinesia. However, for bradykinesia detection, existing studies using smartwatch data from PD patients' natural environment were limited in study duration and sample size. To address this, the current study collected smartwatch data for up to seven months in PD patients' natural environments. From 22 PD patients, the smartwatch data was pre-processed, features were extracted and analysed, and a Variational Autoencoder (VAE) was trained to develop a bradykinesia detection model. However, the VAE could only learn from frequency-domain inputs and not from time-domain input data after current pre-processing methods. Limitations were identified in data quantity, distribution, and quality, including low-frequency artefacts and noise.**

Despite these limitations, feature analysis indicated that the data set contains valuable information about PD motor symptoms. The results of the feature analysis and VAE training on frequency-domain inputs suggest that, after addressing limitations in data quantity and quality, it could be possible to train the VAE using smartwatch data from PD patients' natural environments. In conclusion, although a bradykinesia model was not successfully developed, this study demonstrated the required steps for training the VAE and laid the groundwork for future studies to develop a bradykinesia detection model.

I. Introduction

Parkinson's Disease (PD) is one of the most prevalent neurodegenerative disorders, affecting over 10 million people worldwide [1, 2]. Over the last decade, the Netherlands has seen a 30 percent increase in Parkinson's Disease (PD) patients [3], mirroring a global trend of rising PD incidence [4].

PD is characterised by four cardinal motor symptoms: tremor (rhythmic body shaking), bradykinesia (decrease in speed or amplitude of movement), rigidity (stiffness of the body), and postural instability [5]. Although these common symptoms are typical for individuals with Parkinson's Disease (PD), the nature and degree of symptom expression varies strongly among patients. Moreover, in each patient, motor symptoms may exhibit fluctuations throughout the day or between days, making the disease highly heterogeneous [6]. The motor symptoms associated with PD seriously impact the quality of life of affected individuals. Since there is no cure for PD, there is a growing body of research focused on developing effective treatment methods to alleviate PD motor symptoms.

A standard treatment method for PD is dopamine replacement therapy, a pharmaceutical method with prescriptions of daily doses of dopamine agonists such as levodopa. Although effective in alleviating PD motor symptoms, due to the progressive nature of PD, long-term usage of levodopa often leads to disabling fluctuations and dyskinesias (involuntary movement of body parts) [5, 7]. In addition, specific symptoms may become resistant to these pharmacologic therapies [8]. In this case, patients become eligible candidates for another treatment option: deep-brain stimulation (DBS).

DBS is a surgical treatment method in which electrodes are implanted into deep regions of the brain [9]. The implanted electrodes provide continuous electrical stimulation to the brain and are connected to an internalised programmable stimulator [10]. Determination of the DBS stimulation parameter is typically achieved within 3 to 6 months through 4 to 5 programming sessions at the hospital [10]. Alongside DBS, treatment with medication is usually continued. DBS can effectively reduce the duration and severity of symptoms during the off-drug phase, reduce symptom fluctuations and minimise side effects such as dyskinesias [11, 12]. However, its widespread use is limited by side effects and partial efficacy. The current partial efficacy and risk of side effects may be due to the continuous application of DBS, where stimulation parameters are set at a fixed point in time. To address symptom fluctuations and mitigate associated side effects, adaptive treatment strategies, such as adaptive DBS (aDBS), could be more effective [8, 13, 14]. In aDBS, stimulation is only provided when necessary, improving therapeutic efficacy while limiting side effects and preserving battery life [15].

Real-time adjustments in aDBS based on the expression of PD symptoms require continuous symptom information. Utilising inertial sensors for measuring body motion provides an effective means to detect PD motor symptoms [16, 17]. Smartwatches with an inertial measurement unit (IMU) have emerged as a practical solution for continuous and remote monitoring of patients' wrist movement patterns over time [18–22]. Analysis of wrist movement patterns has been found useful in detecting PD symptoms such as tremor, dyskinesia and bradykinesia [23–35]. The widespread availability and cost-effectiveness of smartwatch devices make them an applicable option for continuous symptom monitoring, essential to ensure use in real-world settings [26, 36, 37].

In models for PD symptom detection, challenges arise in their performance for continuous application in real-world settings. The main challenges involve distinguishing symptoms from daily activities [38] and addressing the variability in patients' symptom manifestations. To overcome these challenges, it is essential to collect data from patients' natural environments when developing PD symptom detection algorithms. This approach enables capturing a broader range of symptom manifestations and daily activities, thereby enhancing the

accuracy and generalisability of the developed models [39].

Nevertheless, most studies focusing on developing PD symptom detection algorithms exclusively use data from controlled environments, such as hospitals or labs. This can be attributed to the complexity of the collection, storage, transfer, and interpretation of data from natural environments. Moreover, data labelling usually relies on symptom assessment from video recordings, which are unavailable in a patient's natural environment. Hence it is difficult to obtain accurate symptom event labels. Despite the challenges in utilising data from natural environments to develop PD detection algorithms, such research is essential since the performance of algorithms developed using laboratory data does not necessarily translate well to natural environments [29]. However, the few studies that have collected data from patients' natural environments [22, 23, 29, 36, 40–42] were limited in sample size[40, 42], study duration[41], or focused solely on data collection [36]. Only one study conducted by Powers et al. [22] involved a large sample size and a 6-month duration of smartwatch data collection and analysis. However, the study developed models for detecting tremor and dyskinesia but did not address one of the most characteristic symptoms of PD, which is bradykinesia. Given the previous studies' limitations, this study aimed to develop a bradykinesia detection model using smartwatch data collected in patients' natural environment over a period of seven months.

In previous research focused on developing algorithms to detect bradykinesia using wrist kinematic data from patients' natural environment, various techniques were employed. One of the first studies utilised empirically defined thresholds for signal classification [35]. However, more recent studies have utilised machine learning (ML) methods for bradykinesia detection, to enable the identification of more complex, nonlinear patterns. These methods include conventional ML models such as random forest and support vector machine [31, 32, 34, 43, 44], and deep learning (DL) methods such as convolutional neural network (CNN) [23, 25, 27, 30]. Among these, CNNs have shown superior performance for bradykinesia detection using data from natural environments. However, the limited interpretability of CNNs due to their "black box" approach often results in algorithms lacking transparency and explainability. This presents a challenge for their real-world application, especially in a medical context. A Variational Autoencoder (VAE) with convolutional layers offers an alternative method that enables improved interpretability of results by providing input reconstructions. This DL method is considered promising as it can offer better transparency and explainability compared to CNNs [45, 46].

In one previous study [42], a VAE was employed to successfully develop a bradykinesia detection model. However, the study's data collection was limited in duration and sample size, and the additional use of video data, which is impractical for home installation, limits the real-world application of the developed algorithm. Additionally, the lack of sharing details on the model development and the resulting algorithm limits the reproduction and applicability of the model [47].

This study aims to fill the research gap by exploring the use of a VAE for PD bradykinesia detection model development using a large dataset from a patient's natural environment. Smartwatch data was collected from each PD patient for up to seven months, and was used to train the VAE for developing a bradykinesia detection model. The study also performed a feature analysis to better understand the dataset. The publicly available and FDA-approved software used for data collection makes the process replicable. To optimise battery and memory usage on the smartwatch, the study limited pre-processing steps. By minimising these steps, the collected data was used to train the VAE to detect bradykinesia. Following these steps, the study aimed to assess the feasibility of using smartwatch data collected from PD patients' natural environment to train a VAE for developing a bradykinesia detection model.

To address the research objective, first, the working principle of a VAE is explained. This is followed by a description of the methodologies for data collection and data pre-processing, which includes segmentation, resampling and removal of data segments with small amounts of data and with detected non-wear instances of data. Next, the analysis of six characteristic features of PD is described, followed by the VAE model training procedures. Subsequently, the amount of collected and downloaded data are presented, followed by the results of the data processing, feature analysis, and VAE model training. To lay the groundwork for future research on this topic, limitations and corresponding recommendations for future research are extensively reported.

II. Variational Autoencoder

In the current study, a Variational AutoEncoder (VAE) is employed for the analysis of smartwatch data from PD patients recorded in their natural environment. The selection of a VAE was based on its general suitability for recognising complex patterns, like in bradykinesia, using large data sets. Moreover, through input reconstruction, results from a VAE have improved interpretability compared to other Deep Learning (DL) models [45, 46]. In this section, the working principle of a VAE is described.

VAEs are generative DL models typically consisting of two CNN components: an encoder and a decoder [48] (see fig. 1). The encoder network transforms high-dimensional input data into a lower-dimensional latent space with a normal distribution. Subsequently, the decoder network converts the latent vector back to higher dimensional space, to generate a reconstruction of the input. The loss function of a VAE consists of a reconstruction term and a regularisation term. The reconstruction term determines the Mean Squared Error (MSE) between the original input and the reconstructed signal. The regularisation term is expressed as the Kullback-Leibler (KL) divergence between the returned distribution and a standard Gaussian. This term serves as a regularisation metric to prevent overfitting and regulate the organisation of the latent space, maintaining important features. The regularity term in a VAE is introduced to encode an input as a distribution over the latent space, rather than a single point as in traditional Autoencoder models. This approach helps to avoid overfitting and creates a smooth and continuous latent space, which improves generalisation to unseen data.

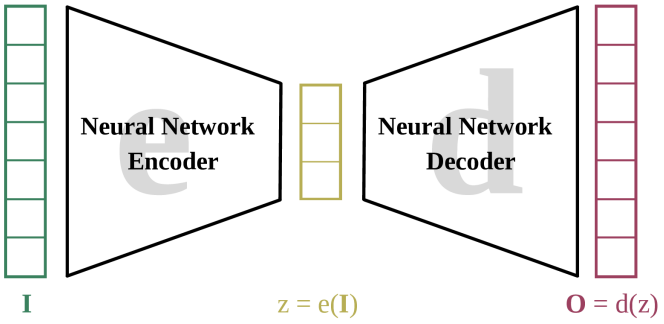

Fig. 1. Graphical representation of a Variational Autoencoder [48]. In green, the input data (**I**) is presented, yellow represents the latent space ($z = e(\mathbf{I})$), and the reconstructed data is given in red ($\mathbf{O} = d(z)$).

The process of training a VAE model involves a series of steps, in which the input data set, denoted as **I**, undergoes processing through the VAE model, resulting in the output **O**. In the initial step, the input data is passed through the encoder model, which can be identified in equation 1 as $e(\mathbf{I})$. In the encoder model, the input data is encoded as a distribution with some variance and is forced to be close to a standard normal distribution. Thus the parameters of the normal distribution, i.e. the mean μ and covariance σ, are outputted by the encoder model. The next step involves sampling a latent representation, represented by $\vec{z}$ in equation 1, from the distribution. Subsequently, the decoder model, represented by **O** in equation 1, is used to generate a reconstruction of the input data. In the final step, the reconstruction error is computed and backpropagated through the network.

$$\begin{aligned} e(\mathbf{I}) &= \mathcal{N}(\vec{\mu}, \vec{\sigma}) \\ \vec{z} &\sim \mathcal{N}(\vec{\mu}, \vec{\sigma}) \\ \mathbf{O} &= d(\vec{z}) \end{aligned} \tag{1}$$

The VAE is optimised by minimising both the reconstruction loss, found by calculation of the mean squared error and the Kullback-Leibler divergence between the distribution parameterised by the encoder outputs and a standard normal distribution. The total loss function is given in equation 2, where E_e denotes the expectation from the distribution of $e(\mathbf{I})$.

$$\begin{aligned} L &= \Sigma_i(||\mathbf{O} - \mathbf{I}||^2 + KL(\mathcal{N}(\vec{\mu}, \vec{\sigma}), \mathcal{N}(O, I))) \\ KL &= E_e[\log e(\mathbf{I}) - \log \mathcal{N}(O, I)] \end{aligned} \tag{2}$$

.

Using these two components in the total loss function, once the model has been trained, the loss function is plotted in a graph. The plot visualises the learning curve, illustrating how the loss function evolves over time. This plot comprises two key elements: the total loss and the validation loss. The total loss assesses how well the model fits the training data, whereas the validation loss measures the model's performance on unseen data. During training, the total loss and the validation loss should both decrease. However, if the model starts overfitting, a gap will emerge between the two curves in the loss graph [49]. The size of this gap signifies the extent to which the model is overfitting. During training, the model's performance is analysed based on the loss function, indicating how well the model is learning.

III. Method

Before implementing data to a VAE model, the current study performed collection, pre-processing, and feature analysis of smartwatch data. In this section the data collection method is elaborated first, including details about what data is collected throughout the collection period and the requirements for successful data collection. Subsequently, the data processing procedures are described, comprising data retrieval, data quality assessment, signal pre-processing and non-wear detection. Lastly, details on the data analysis steps are provided, including the computation and analysis of features and a description of the employed data classification model. The whole process from beginning to end is visualised in figure 3.

A. Data Collection

The current study was part of the first phase of an ongoing observational study at the Amsterdam University Medical Center (UMC), only including data collected between September 2022 and October 2023. The study focuses on patients with PD who received treatment with a Medtronic Percept$^{\text{TM}}$ PC STN DBS system [50]. A complete data collection period for a patient spans seven months. It involves six clinical testing visits that coincide with their routine care visits to minimise the required effort for participation. Various data types were gathered from patients at different stages of the data collection period. This is illustrated in figure 2, which represents the complete data collection period. The smartwatch data collection is conducted both in-clinic and remotely in the patient's natural environment. In the current study, every patient was included at a different point in time, with inclusion dates varying from several days up to a year. As a result, the duration of the data collection period varied for each patient. The following paragraphs outline the procedures used to continuously collect smartwatch data for this study.

1) The Data Collection Process: The data collection period starts at the first visit to the clinic, referred to as 'visit 0', during which the informed consent form is signed. At this visit, which takes place one month before the DBS surgery, the patient receives a smartwatch (Apple Watch Series 7 and 8) and a smartphone (Apple iPhone 6, 6s, 7, 8 and SE) with the StrivePD application [51] from Rune Labs installed. The smartwatch is placed on the wrist of the most affected upper limb, according to the clinical indication of the physician. The smartwatch records kinematic and heart rate data, which is temporarily stored on the device and transferred through the StrivePD application for long-term storage. The smartwatch data is pseudonymised and ingested into the Rune Labs platform, where researchers can access it.

The kinematic data recorded by the smartwatch consists of inertial measurements from the tri-axial accelerometer and gyroscope embedded in the smartwatch's IMU. The acceleration and angular velocity, referred to as rotation in this study, are registered with a sampling frequency of 50 Hz. The heart rate is recorded using an optical heart rate sensor, consisting

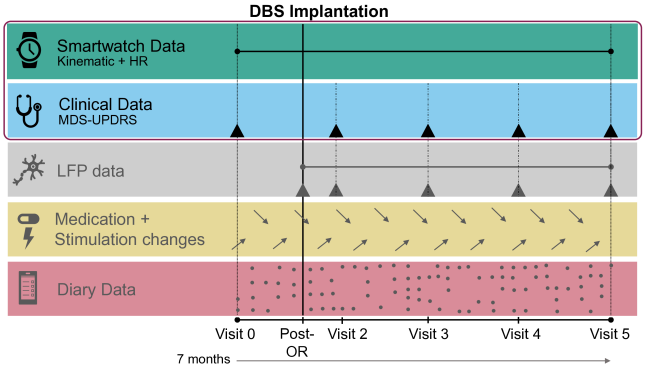

Fig. 2. Timeline of the data collection process. As illustrated, the patients undergo six clinical visits, before and after Deep Brain Stimulation (DBS) implantation. The 'post-OR' (post Operating Room) visit, takes place the day after DBS surgery. The smartwatch data, comprising kinematic data (acceleration and rotation) and heart rate (HR) data, is collected throughout the whole period. Visits encompass motor symptom evaluations using the MDS-UPDRS III (Appendix I) indicated by the triangles, with visit 5 incorporating evaluations under varied conditions: off medication and DBS switched off (OFFmed / OFFstim), off medication and DBS switched on (OFFmed / ONstim), ONmed / OFFstim, and ONmed / ONstim. Recording of the Local Field Potential (LFP) activity is measured by finding the differential in the LFP signal between two contact pairs [50]. The LFP data is measured before and after DBS activation, indicated by the triangles, and in one-minute readouts during diary entries. In addition, longitudinal LFP recordings indicated by the horizontal line, are passive out-of-clinic sensing of neuronal activity, which is continuously recorded after activation of the Medtronic Percept^{TM} PC DBS system, until the end of the data collection period. The diary data is obtained through an application belonging to the Medtronic Percept^{TM} DBS system, which patients use to fill out a clinical diary five times a day to assess their motor state. Stimulation and medication adjustments during the six-month recording period are documented in electronic case record forms (eCRFs). Data pseudonymisation, storage, and synchronisation are executed using the Rune Labs software platform, ensuring data privacy and access limited to the Amsterdam UMC researchers.

of an LED light that flashes with a frequency of 100 Hz [52, 53]. These measurements allow for calculating the heart rate in beats per minute (bpm), resulting in a heart rate value registration with a frequency of 0.2 Hz.

The second visit, referred to as the 'post-OR' (post Operating Room) visit, takes place the day after DBS surgery. During this visit, exclusively Local Field Potential (LFP) data is collected. The third clinical visit, referred to as 'visit 2', occurs two weeks after implantation, when the DBS system is activated. During this visit, standardised Movement Disorder Society Unified Parkinson Disease Rating Scale (MDS-UPDRS) Part III motor tasks are performed by the patients and videotaped. These video recordings are later used by a trained clinician to assess a patient's PD motor symptom severity. The PD motor symptom severity is rated by assigning MDS-UPDRS scores, making up the clinical data.

The two subsequent clinical testing visits, 'visit 3' and 'visit 4', occur two months and four months after surgery. Following the previously described procedure, clinical data is collected during both visits. Lastly, the final clinical testing visit, referred to as 'visit 5', is planned six months after implantation. During this last visit, the effect of DBS was evaluated using systematic MDS-UPDRS III scores in four different conditions.

From the smartwatch data types, the focus of the current study is on the use of kinematic data since heart rate measurements take up most part of a smartwatch's energy expenditure compared to other sensors [54, 55]. Therefore, developing a classification model that does not require heart rate data is desirable.

2) Monitoring Data Quality: The collected smartwatch data predominantly consists of recordings from patients' natural environment. Therefore the data collection could not be controlled directly, and data quality was monitored remotely through the Rune Labs platform throughout the complete data collection period. During the research process, Rune Labs stopped providing active support for the collection of raw kinematic data in 'high-fidelity mode', resulting in unresolved issues with the collection of this data.

The quality of the collected kinematic and heart rate data is influenced by a variety of factors, linked to the process of data transfer and storage. The data from the smartwatch is transferred online via the StrivePD application to the Rune Labs platform, directly or via the iPhone. The iPhone is required to provide an internet connection for data transfer. Therefore, a stable connection between the iPhone and the smartwatch, and to WiFi or mobile data is required. If the collection with the iPhone or the internet is lost, the raw kinematic data will not be transferred. If this connection loss persists for several days, also the transfer of data collected in the 'low-fidelity' mode, consisting of other data types (see Appendix G), will be disrupted.

Another factor that impacts the kinematic data storage and transfer is the activity of the StrivePD application. The StivePD application from Rune Labs is mainly developed for clinical application. It is designed for active patient use, offering tools such as medication reminders, insight into symptom progression and keeping a symptom diary. In the context of this research, employment of the StrivePD application was limited to the passive collection of smartwatch data. Patients were instructed to wear the smartwatch and carry the iPhone, without utilising the devices for any other purpose. However, lack of active engagement with the devices and the StrivePD application caused issues such as patients forgetting to carry the iPhone, disruptions in WiFi or Bluetooth connections that went unnoticed, or outdated software. In addition, periods of inactivity, such as non-wear periods, may cause the StrivePD application to close automatically, halting raw kinematic data collection. Therefore the patients were instructed to open the StrivePD application regularly to reduce instances of automatic closing of the StrivePD application. For the data collected in low-fidelity mode, this problem did not occur.

Another factor contributing to issues with data streaming relates to the device charging. The high-fidelity mode on the smartwatch contributed largely to the battery expenditure, which meant that patients sometimes had to charge the devices up to three times a day.

Thus, in order to obtain high-quality data, patients needed to be actively involved in the data collection process. This involved regularly charging their smartwatch and smartphone, consistently opening the StrivePD app, and adjusting settings in case of any streaming issues. The quality of the collected data was thus determined by the level of engagement from the

patients.

Due to the various challenges regarding data availability, and the minimal support from the platform to address these challenges, constant involvement of researchers was also required. When several days of missing data, or more extended periods of insufficient data quality were observed, patients were contacted by phone or email. This approach ensured that data streaming issues were often addressed promptly, minimising data loss and reduced data quality. More details on the data quality monitoring process are provided in Appendix A. In addition, an evaluation of using SIM cards to provide a stable internet connection is given in Appendix E.

B. Data Processing

Once the data collection phase was concluded, the next step involved retrieving and pre-processing the acquired data for subsequent analysis. The following paragraphs elaborate on the process of downloading and pre-processing of the data as preparation for input into the VAE.

1) Downloading Smartwatch Data: The Rune Labs platform was utilised to collect and store data from the smartwatch. To access the data, users can download it from the platform using an access application programming interface (API) in the Python programming language. The version 1 (V1) Access API for data downloading was found to be outdated and resulted in poor-quality data. Therefore, using the more advanced version 2 (V2) Access API is required to obtain data of sufficient quality. However, the Rune Labs platform was primarily designed to obtain data collected in the 'low-fidelity' mode, and it no longer actively supported the high-fidelity mode. Hence, initially, no code was available to download the raw kinematic data. The required raw kinematic data, could thus not be readily downloaded from the platform using their V2 access API. After multiple attempts to rewrite existing codes for downloading the kinematic data and discussing the results with the Rune Labs administrator, a code for parallelised data retrieval using the latest access API was shared by Rune Labs. This code was used to download the data from the platform. A more detailed description of the steps required for downloading kinematic data from the Rune Labs platform is provided in Appendix B.

After access was granted to the AMC cluster and to a storage disk, in the study's final month, the complete data set could be downloaded using the V2 Access API. However, interruptions occurred when running the data retrieval code, automatically freezing after a certain duration. Therefore, the current study did not manage to download the complete data set.

2) Data Quality Assessment: To prepare for the analysis and classification of kinematic data, it is important to have a clear understanding of the data quantity and distribution. This can help in interpreting results from subsequent analysis. Therefore, the first step in the data processing procedure was to visualise the data availability.

From the Rune Labs platform, it was found that there were instances of missing data for each patient, called data 'gaps'. These data gaps ranged from extended periods of several days to smaller gaps with data missing for just a few hours. Additionally, there were intermittent data points missing every few minutes or seconds, leading to a faltering data streaming pattern.

Assessment of the extent and the distribution of data gaps in the downloaded data was performed by visualising data availability across the entire data set and for each individual patient. This facilitated the determination of an approach to address these data gaps.

3) Data Pre-processing: Before pre-processing the data, the acceleration data was merged with the rotation data based on the timestamp for each data point, merging the data on the closest timestamp. The acceleration and rotation data were stored in separate folders. For the acceleration data, there was a folder with the acceleration data due to gravity and the acceleration data of the user, from which gravity was removed by Rune Labs. The acceleration data of the user was used in this study.

After merging of the kinematic data, the first step in data pre-processing was data segmentation. The length of each segment was determined in collaboration with the clinicians at the Amsterdam UMC. This decision was made, based on the duration of bradykinesia episodes, which vary from 10-15 minutes to one or two hours. Selecting a time segment approximately 10% of the average off-medication duration, aimed to avoid incorporating both normal and bradykinetic movements, as this could limit detection accuracy. Consequently, a five-minute segment length was recommended for bradykinesia detection. This decision was supported by studies such as Habets et al. [56], which found a five-minute segment length optimal for detecting bradykinesia.

The employed data segmentation method was based on a 5-minute duration determined by the timestamp data in the 'time' column, using a sliding window approach without any overlap [57]. The segmentation process started at the first timestamp in the 'time' column of the patient's data frame. This approach was implemented to ensure that, despite the gaps in data, the temporal information was retained.

Upon close observation of the data within the segments, an unusual pattern was observed in the data. As outlined in the Apple Watch documentation [52], the anticipated sampling frequency is 50 Hz, implying a time interval of 20 ms between consecutive data points. However, in addition to slight variations in the spacing length of 20 ms, very small spacing lengths were also observed, usually occurring after a data gap. This pattern was detected across all patient data. To compensate for this unevenly spaced data, the data was resampled using the mean of the data points within a 20 ms range, resulting in data points with an even spacing of 20 ms.

Due to periods of unevenly spaced and faltering data, missing data periods were contained within the segments, leading to empty segments or segments with data gaps. To address this, segments with data quantities below a certain threshold were removed, whereas those with some missing data points were preserved and imputed later to retain as much of the data as possible. In contrast to previous studies, the current study applied a segment-based threshold rather than daily-based thresholds, to ensure minimal data removal [58].

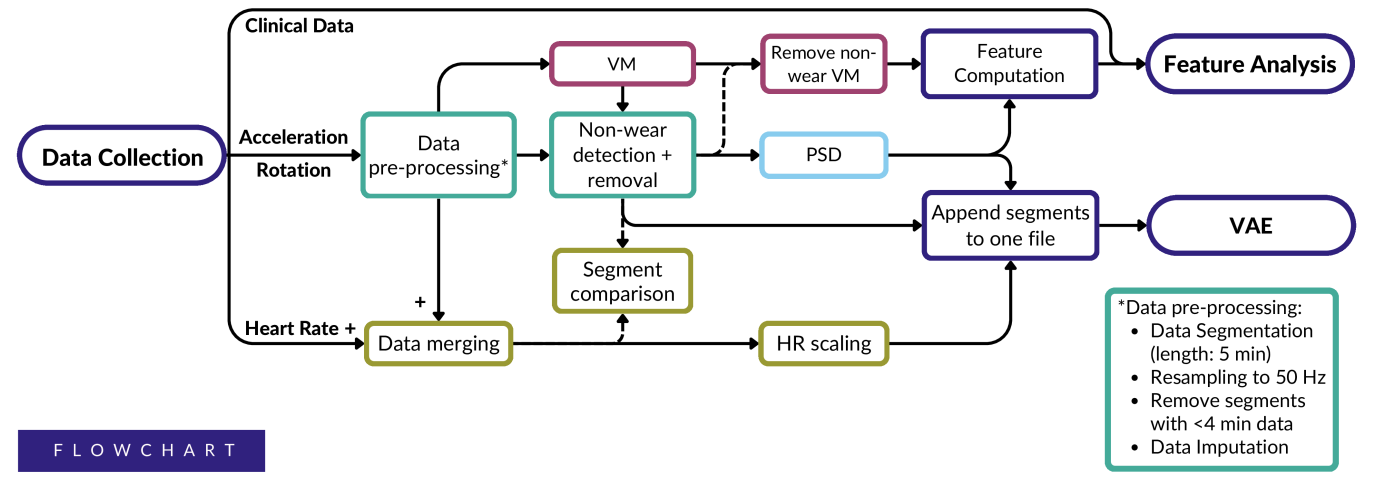

Fig. 3. Flowchart outlining the procedures from the smartwatch data collection to the smartwatch data analysis. The data pre-processing steps are defined in a separate block located at the bottom right. After data segmentation, all pre-processing steps are applied segment-wise, including the non-wear detection method. The heart rate data is merged with the kinematic data, and it is linearly interpolated to fill the segment with heart rate data. Furthermore, the heart rate data segments are compared with the non-wear data segments, and the heart rate is scaled to match the kinematic data values for use as input to the VAE. The vector magnitude (VM) is calculated for both acceleration and rotation signals at each time instance, within each segment. The power spectral density (PSD) is calculated using the Welch method. The PSD of both acceleration and rotation signals are computed for each axis and averaged over the three axes for feature computation. The feature computation includes the computation of six features characteristic of PD, including two time domain and four frequency domain features. These features are subsequently analysed to assess the value of the information contained in this data set to detect motor symptoms. The pre-processed data is used as input to the VAE in four different modalities. Two modalities consist of only kinematic data in the time domain, one modality consists of kinematic data in the frequency domain, and the last modality contains kinematic data and heart rate data in the time domain. The inputs to the VAE are first appended to a single file so that the data can be loaded into the VAE, after which the VAE is trained. The dashed lines indicate the use of data segments for comparison, not for further processing.

These prior studies used thresholds ranging from 25% to 50% as the maximal data missing rate over a day [59]. Using the highest threshold, the current study applied a minimum length of 4 minutes, meaning that data segments with 20% or more missing data points within 5-minute segments were removed. The number of removed segments and their corresponding segment numbers were tracked during this process.

The missing data in the remaining segments was addressed using data imputation. This was done to ensure all data segments contained the same number of data points. Since the data gaps in the remaining segments were limited, the methods selected for data imputation had limited influence and were chosen for their simplicity. To fill gaps in each data segment, linear interpolation, which is a simple and commonly applied method for imputing time-series data [60], was applied. Additionally, a first-order spline interpolation function was applied for comparison, as it was found to give good results for the imputation of kinematic data [61]. Both the linear and first-order spline interpolation methods can only impute data on the inner regions of the data segments. Therefore, zero-padding was used to impute data on the outer regions of the data segment. In this way, all data segments were ensured to have the same length of 5 minutes, which equals 15000 data points with a sampling frequency of 50 Hz ($5\text{min} \times 60\text{s} \times 50\text{Hz}$).

After resampling and interpolating the data segments, the heart rate data was merged with the kinematic data segments. The merging was based on the closest time, with a maximum variance of twice the sampling frequency of the heart rate data to be added as a value to the corresponding row with kinematic data. After merging the data, a lower missing rate threshold was set compared to the kinematic data due to the relatively lower sampling frequency of the watch for the heart rate data. The threshold was set to a maximum data points missing rate of 80%. For heart rate data imputation, linear interpolation was used to fill data gaps and forward fill as well as backward fill, with the last value present, were employed to ensure all segments were of equal length. For the input of segments with heart rate data into the VAE, the heart rate data was scaled to reduce bias caused by higher values of heart rate data in comparison to kinematic data [62]. The acceleration and rotation data values are mostly between -2 and 2 (G and rad/s), and heart rate values are mostly between 60 - 100 rpm [63]. Therefore, the heart rate data was scaled down by a factor of 50 to match the range of acceleration and rotation data values.

4) Non-wear Detection: After the data pre-processing phase, a cleaning step was introduced to eliminate non-wear instances from the data set. Non-wear instances, where the smartwatch continues recording despite not being worn, introduce erroneous information into the data set. This occurs because smartwatch noise prevents readings from being exactly zero. Consequently, segments that contain low acceleration and rotation values during non-wear periods seem to contain information about the patient's movement, but instead contain noise. Therefore, distinguishing non-wear segments from actual human movements during rest or sleep is challenging.

Conventional zero-value detection strategies described in

the literature [64] were not applicable, since the smartwatch readings during non-wear periods are not consistently zero. Instead, a review by Ahmadi et al. [65, 66] assessing five distinct non-wear detection methods for wrist-worn devices identified a superior approach. This approach utilises a threshold, based on the standard deviation (SD) of the signal's vector magnitude (VM) to detect non-wear instances using acceleration data. To determine the most appropriate threshold value for this data set, the distribution of the SD of the acceleration VM was analysed through a histogram. If the data could be divided into two distinct classes, the SD at the separation point would be used as the threshold value and otherwise, the threshold value of $0.013\mathrm{G}$ would be applied. Following this, the segments that were identified as 'non-wear' instances were removed from the data set.

To evaluate the non-wear detection method, the data segments identified as non-wear were compared to the heart rate data in these instances. The number of non-wear segments containing heart rate data can indicate the effectiveness of the non-wear detection method.

C. Data Analysis and Classification

After the data is pre-processed, feature analysis is performed to create a better understanding of the data. Subsequently, different modalities of the data are used to train the VAE. The methodologies for feature extraction and analysis are explained in the subsequent paragraphs, followed by an explanation of the VAE model architecture and training process.

1) Feature Analysis: Characteristic features of PD motor symptoms are extracted from the kinematic smartwatch data, and analysed through feature analysis. The aim of this analysis is to aid in the interpretation of the kinematic data, rather than to directly detect PD symptoms. Due to the complex nature of bradykinesia and the difficulty in isolating individual symptoms without additional filtering, it was decided not to focus solely on this particular symptom. Hence, features associated with multiple PD symptoms were selected.

The extracted features include characteristic features for bradykinesia and tremor, based on prior studies utilising wrist kinematic data for PD symptom detection [17, 35, 38, 56, 67]. Additionally, features representing the daily activity level for PD patients were extracted, to reflect the overall PD symptom severity [68, 69]. Features for both the time and frequency domain were extracted segment-wise across all data segments, after pre-processing and non-wear segment removal. Further details on the extraction procedures for time domain and frequency domain features, as well as their subsequent analysis, are discussed in the following paragraphs.

For the time domain features, among the key features for detecting bradykinesia are the mean [24, 35, 56, 70, 71] and the maximum [24, 56] of the vector magnitude (VM) of the acceleration and rotation signal. The first step to compute the time domain features was to calculate the VM [72], given by equation (3). The VM was calculated for each time instance i in the data segment, of the tri-axial (x_i, y_i and z_i) acceleration and rotation signal.

$$\mathrm{VM}_i = \sqrt{x_i^2 + y_i^2 + z_i^2} \tag{3}$$

Using the VM, the time-domain features were computed for both the acceleration and rotation signal. These include features:

1) Segment mean of the VM
2) Segment maximum of the VM

The values of feature 1 were stored in a data frame along with the start time of the segment. Using this data frame, the daily patterns of mean acceleration and rotation VM values were visualised by plotting them over the corresponding dates of the data segments. To capture the overall daily pattern for each patient, the mean of the acceleration and rotation VM were calculated for each segment and then averaged across the same time instances over all available days for each patient.

For the frequency domain features, the spectral power between certain frequency bands was analysed to characterise specific symptoms, while the total spectral power was analysed for a complete view of the patient's activity. The spectral power below 4 Hz, is one of the most dominant features for bradykinesia defined in literature [56]. This frequency range is typical for bradykinesia, because in episodes of bradykinesia, movements are typically reduced in amplitude, with frequencies mostly between 0.2 and 4 Hz [27, 41], and the bradykinesia dominant frequency between 0.32-1.40 Hz [23]. The spectral power between 4-8 Hz, which is the typical frequency band for Parkinsonian and Essential tremor [24, 35], reflects the presence of tremor symptoms. For all features, it should be taken into account in the analysis that dyskinesia (involuntary movements), which can also be measured from the wrist, is also typically observed within frequency ranges of 1-5 Hz [73, 74].

To compute the features in the frequency domain, first, the Power Spectral Density (PSD) [75] was generated for each segment of data. The PSD was calculated using the Welch Averaging method [76] using SciPy [77]. The Welch method involves dividing the signals into M windows of length L instead of taking the Fast Fourier Transform (FFT) over the entire signal, resulting in a smoother spectrum (see eq. (4)). Here $P_{\text{Welch}}(f)$ represents the Welch PSD, M the number of segments, and $|X_i(f)|^2$ is the magnitude squared of the Fourier transform of the i-th segment.

$$P_{\text{Welch}}(f) = \frac{1}{M} \sum_{i=1}^{M} |X_i(f)|^2 \tag{4}$$

The $X_i(f)$ term in the Welch method is calculated using Equation (5), where $x_i[n]$ represents the i-th segment of the input signal, $w[n]$ the window function, and FFT denotes the Fast Fourier Transform.

$$X_i(f) = \text{FFT}\{x_i[n]w[n]\} \tag{5}$$

This PSD estimation process is performed separately for acceleration and rotation data along each axis, resulting in three Welch PSDs (P_{xx}, P_{yy}, P_{zz}) per data type for each segment. The window length L was set to 10 seconds, and the overlap to 50%, determined based on a comparison of the

results of different settings, providing a good balance between resolution and a smooth, interpretable signal. The three PSDs, were combined by computing their mean, resulting in a single PSD for both the acceleration and rotation signals of each segment.

Using the combined PSDs, the frequency domain features [38, 68] could be computed. The selected frequency domain features are computed for each segment. These are features:

3) Total spectral power
4) Spectral power below 4 Hz
5) Spectral power between 4-8 Hz
6) Peak power between 4-8 Hz

Following Parseval's Theorem, the spectral power was found by calculating the area beneath the PSD [68, 75, 78]. For the total spectral power, reflecting the total activity of the patient, the total area under the PSD was calculated, with frequencies limited to the Nyquist frequency [79] of 25 Hz.

To analyse whether distinctions can be made between patients with more severe PD motor symptoms and those with milder symptom expressions, using the current data set, two distinct patient groups were composed based on the MDS-UPDRS scores assigned during the clinical visits. The decision to create only two groups, without distinguishing between different phenotypes like severe bradykinesia and rigidity or severe tremor group, was based on three reasons. Firstly, from the patients' MDS-UPDRS scores, it appeared that higher scores for tremor were often associated with higher scores for bradykinesia and vice versa. Moreover, distinguishing between different phenotypes through the current analysis is challenging since no frequency filtering was performed in the current study, making it difficult to isolate tremor expressions for bradykinesia analysis or isolate bradykinesia expressions for tremor analysis. Finally, the patient population is limited; thus, dividing the patients into multiple groups would further reduce group sizes, making it challenging to draw reliable conclusions.

The composition of the two PD severity groups was based on the overall MDS-UPDRS scores of the PD patients from which data was downloaded. The 'severe' patient group comprised the seven patients with the highest MDS-UPDRS scores, while the 'mild' patient group comprised the seven patients with the lowest MDS-UPDRS scores. Only patients from which at least 20 data segments are available, were selected for feature analysis. To investigate whether it was possible to distinguish between these patient groups using the current data set, both their time domain and frequency domain features were compared. Prior studies investigated the relationship between feature values and PD symptom severity using long-term wrist kinematic data [68, 69, 80]. In a three-year longitudinal study conducted by Nisha et al. [81], a positive correlation was found between changes in wrist kinematic data maxima and total MDS-UPDRS score. Based on these studies, it was hypothesised that patients in the 'severe' group would generally exhibit higher activity patterns and, hence, higher feature values with larger variations, than the 'mild' group.

Violin plots were employed to visually compare feature values between the two patient groups. If distinctions between the two severity groups are recognised, this would suggest that the data set contains relevant information for the detection of PD symptoms and, thus, also for further development of a bradykinesia detection model.

2) The Classification Model: The VAE model employed in this study is based on the model proposed by David et al. [82] developed for analysing the gait patterns of stroke patients. The VAE model leverages an unsupervised learning approach. In contrast to supervised learning, unsupervised learning methods do not rely on data labels. This is beneficial since accurate data labels are not available in the natural environment of patients. The accurate data labels for PD symptoms are obtained by rating the PD severity using the MDS-UPDRS guidelines from video recordings, which are not available from a patient's natural environment. Unsupervised learning focuses on pattern recognition without predefined targets and allows for identifying non-linear relationships without data labels [83]. Labels are exclusively required for performance evaluation, please refer to Appendix C for the data labelling method. The next paragraphs outline the VAE model architecture, followed by an elaboration on the training process.

3) Model Architecture: In this study, minor modifications have been made to the VAE model architecture to fit in with the context of the current problem. Adjustments made were to the kernel size, to fit the current data set, and to the number of latent features that span the latent space. The following paragraphs describe the VAE architecture from the input layer up to the output from the decoder.

In the first part of the VAE model, the input data shape is determined. Subsequently, the input data is provided to the encoder, which consists of three convolutional layers with 2, 10, and 20 filters, each having three 1D convolutions for the six signals. Each convolutional layer consists of the rectifier linear unit (ReLU) as an activation function [84]. ReLU is a widely used activation function for convolutional layers [84, 85], it returns zero for any negative input, and it will output the input directly for any positive value. The mathematical representation of the ReLU activation function involves the max function applied to the set of 0 and the input x, as shown in equation 6.

$$ReLU(x) = max(0, x) \tag{6}$$

The output of the last convolution is given to flatten and dense layers to predict the latent embedding of size 80. The decoder is symmetric to the encoder. The number of latent features was a hyperparameter that needed to be optimised using trial and error, for which 3 to 100 latent features were tested. The code was implemented in Python using Keras from TensorFlow [86].

4) Model Training: The VAE model was trained with multiple data input modalities to evaluate the most suitable input data type and representation for the VAE to be able to

learn. The following input data modalities were provided to the VAE:

1) Data without any non-wear segment filtering, from:
   a) Acceleration signal
   b) Rotation signal
   c) Acceleration and rotation signals
2) Data after non-wear segment filtering, from:
   a) Acceleration signal
   b) Rotation signal
   c) Acceleration and rotation signals
3) The PSD for all three axes after non-wear segment filtering, from:
   a) Acceleration signal
   b) Rotation signal
4) Scaled heart rate, acceleration and rotation data.

The data retrieved using the V2 access API were pre-processed to create the correct input shape, following the steps outlined in III-B3. This resulted in 15000 time instants for both acceleration and rotation signals in three spatial directions (x, y, z), within each segment of data. The input modalities 1 and 2 a-b were thus of size 15000 x 3, and the input modalies 1 and 2 c of size 15000 x 6.

The frequency content of the kinematic data was computed using PSDs for each axis of both acceleration and rotation signals of each segment. The PSDs were stored in a data frame per data type and used as input (modality 3 a-b) to the VAE. The frequency data included frequencies between 0 - 25 Hz, with a spacing of 0.1 Hz resulting in an input size of 251 x 3 for both acceleration and rotation frequency data.

The heart rate data was merged with the kinematic data and interpolated to match the length of the kinematic data segments, resulting in an input size of 15000 x 7.

Before training the VAE, first an external validation group was assigned, making a train-test split of 90% and 10% of the data segments, for later external validation purposes. For this split it was ensured that all segments within either the train or test group belonged to one patient, data segments of one patient were thus either contained in the train folder or in the test folder. For model validation during training, also in the model preparation code, a train-test split of 90% and 10% was made using the Scikit learn library function [87]. The number of epochs used for network training ranged from 10 to 50, with the final network training for 25 epochs using the Adam optimiser.

IV. Results

The results of the remote smartwatch data collection and corresponding analysis steps are presented in this section. Hereto, first, the results of the data collection are discussed, followed by the data pre-processing outcomes. Lastly, the results from the data analysis and performance of the VAE are presented.

A. Data Collection

Throughout the study period, kinematic smartwatch data was available for 43 patients, collected between September 2022 and October 2023. Patients were included at varying times, resulting in diverse data collection periods among patients. Ten patients completed the full seven-month data collection within the specified timeline. Notably, for each of the remaining patients, the duration of the data collection period was at least one month. Due to freezing issues encountered upon downloading the patient data, the data from only 22 patients was downloaded. The demographic details of the patients are provided in table I, differentiating between the patient group from which data is available and from which data is downloaded. A total number of 2738 days of movement data was collected from the smartwatch's triaxial accelerometer and a triaxial gyroscope.

The MDS-UPDRS scores for bradykinesia were assigned by a single trained clinician from the video data of the clinical visits. The average bradykinesia MDS-UPDRS scores for all patients for each hand task during the different visits are presented in table I and a visualisation of the mean and individual MDS-UPDRS scores is provided in Appendix F.1.

During data collection, several patients encountered notification issues with the smartwatch, experiencing sounds, vibrations, or light flashes. These technical challenges, together with the active involvement required from patients to maintain data quality, led to the exclusion of nine patients, which constitutes 17% of the initial, total number of patients participating in the smartwatch study. Notably, 78% of these exclusions occurred between visit 2 and visit 3, while the remaining patients were excluded between visit 0 and the post-OR visit. Thus from some of the excluded patients, valuable smartwatch data has been collected before exclusion from the study.

TABLE I
DEMOGRAPHIC DETAILS OF STUDY PARTICIPANTS
LEFT: DETAILS OF PATIENTS AND FROM WHICH DATA IS AVAILABLE, RIGHT: DETAILS OF PATIENTS AND FROM WHICH DATA IS DOWNLOADED
TOTAL SCORES ARE FOR: BRADYKINESIA, TREMOR AND RIGIDITY.

	Available	**Downloaded**
Number of Subjects	43 PD Patients	22 PD Patients
Gender	11 females, 33 males	6 females, 16 males
Age	65.2 $\pm$ 7.1 years	65.9 $\pm$ 6.2 years
PD Dominant Side	17 Right, 12 Left, 12 Unknown	6 Right, 10 Left, 6 Unknown
Mean of Total MDS-UPDRS scores (visit 2-4)	33.5 $\pm$ 14.3 (max. 80)	34.4 $\pm$ 16.4 (max. 80)
Mean of Hand Bradykinesia MDS-UPDRS scores (visit 2-4)	16.1 $\pm$ 7.7 (max. 40)	17.2 $\pm$ 5.5 (max. 40)
Quantity of Data	8802 days	2738 days

For the categorisation of patients into two PD severity groups, the MDS-UPDRS scores for wrist-related symptoms were compared, including bradykinesia, rigidity, and tremor. Table II provides an overview of the MDS-UPDRS scores used for the group composition. Patients were classified as 'severe' or 'mild' based on their total MDS-UPDRS scores for these

TABLE II
MEAN MDS-UPDRS SCORES FOR ALL SYMPTOM TASKS OVER VISITS
RED: 'SEVERE' PATIENT GROUP, BLUE: 'MILD' PATIENT GROUP, GREY: INSUFFICIENT NUMBER OF DATA SEGMENTS, WHITE: NOT ASSIGNED

	Bradykinesia	Rigidity	Tremor	**Total**
Sub003	17.0	8.7	12.3	29.3
Sub005	12.0	4.5	12.0	24.0
Sub006	15.0	4.0	11.0	26.0
Sub010	11.0	1.0	2.0	13.0
Sub014	8.0	3.5	3.7	11.7
Sub018	15.3	5.7	2.3	17.7
Sub020	21.3	4.0	5.7	27.0
Sub021	5.5	2.0	1.0	6.5
Sub025	13.3	7.7	6.3	19.7
Sub031	14.5	0.0	0.0	14.5
Sub033	3.5	1.0	0.0	3.5
Sub038	10.0	7.5	14.5	24.5
Sub039	27.0	5.0	3.0	30.0
Sub041	6.5	2.0	9.0	15.5
Sub042	20.0	8.5	0.0	20.0
Sub043	8.5	3.5	4.0	12.5
Sub044	5.5	2.0	0.0	5.5
Sub046	17.0	2.0	2.0	19.0
Sub047	20.5	6.5	1.0	21.5
Sub048	9.0	6.0	0.0	9.0
Sub050	9.5	2.0	1.0	10.5
Sub052	14.0	6.5	5.0	19.0

three symptoms. The 'severe' patient group includes the seven patients with the highest total scores, highlighted in red. The 'mild' patient group includes the seven patients with the lowest total scores, highlighted in blue. Three patients are marked in grey in the table as their total number of data segments was below 20, which was considered insufficient for feature analysis. The remaining patients with moderate MDS-UPDRS scores, are not highlighted and are marked in white. The final composition of the two PD severity groups is thus visualised by the red and blue colours in table II.

B. Data Processing

The following paragraphs provide details on the quantity and quality of the employed dataset, as well as the methodological choices employed and the resulting outcomes. First, the amount of downloaded data is elaborated on, followed by a discussion on the quality of the current data set. Subsequently, the results from data resampling and imputation are presented. Finally, the non-wear detection results are explained.

1) Data Downloading: For downloading the data from the Rune Labs platform using the V2 access API on the cluster of the AMC, the freezing of the data downloading process resulted in only 20% of the available data being successfully downloaded. The amount of data available and downloaded from the complete data collection period of one patient is visualised in figure 4. The period spans from the first to the last day of data availability for each patient. This figure visualises the distinct differences between the available data and the downloaded data. The variations in dates of downloaded data are not limited to any particular time period. Instead, the downloaded data occurrences are distributed throughout the patient's entire data collection period.

Furthermore, a global visualisation was generated by plotting the available and downloaded data for all patients across the entire study period, as illustrated in figure 5. This graph represents each downloaded file as a count of one for the respective date, aggregated to make up the collective daily patterns for all patients. Upon analysing these graphs, it was also found that no pattern could be identified, and the total downloaded data set is distributed over the entire study period.

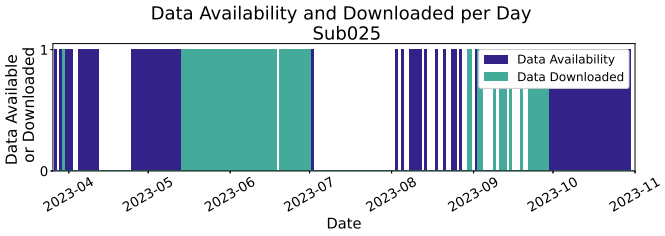

Fig. 4. Bar diagram with the available and downloaded data of a patient, throughout the data collection period. Each bar in the diagram represents a single day, with the blue bar indicating the availability of data and the green bar representing that data has been downloaded. This allows for a patient-wise visualisation of the distribution of the downloaded data over time.

2) Data Quality Assessment: Despite the close monitoring of data quality, periods of missing data and non-wear instances are present for all patients, varying from hours to complete days or weeks. To assess the data quality, various plots were generated to illustrate the number of data points per data segment for each patient, highlighted in the following paragraphs.

The bar diagram from figure 6 visualises the total number of resampled data segments for each patient. From the diagram, it is observed that most patients have a limited number of data segments, ranging from 50 to 250 segments per patient. This distribution demonstrates the generally low quantity of data segments included for analysis and the large differences in segment quantities per patient.

In addition, a bar diagram was created to display the hourly count of data segments for all patients across all days. This diagram offers insight into the distribution of data segments throughout the day (see figure 7). The diagram for each individual patient is provided in Appendix F.2. The majority of data segments in the dataset contain daytime data.

Furthermore, the number of data points per segment was plotted in a histogram, representing the quality of the included data segments (see Appendix F.3). The evident peak in the outermost bin shows that most data segment lengths approach 5 minutes.

Finally, the histogram constructed to visualise the data point spacing (figure 8) presented the largest bin for a data point spacing of 20 ms, which is consistent with the smartwatch's sampling frequency. However, the histogram also revealed a variety in data point spacing and a considerable percentage of data points with a very small spacing of only 0-5 ms. Upon verification with Rune Labs, it was found that this is how the

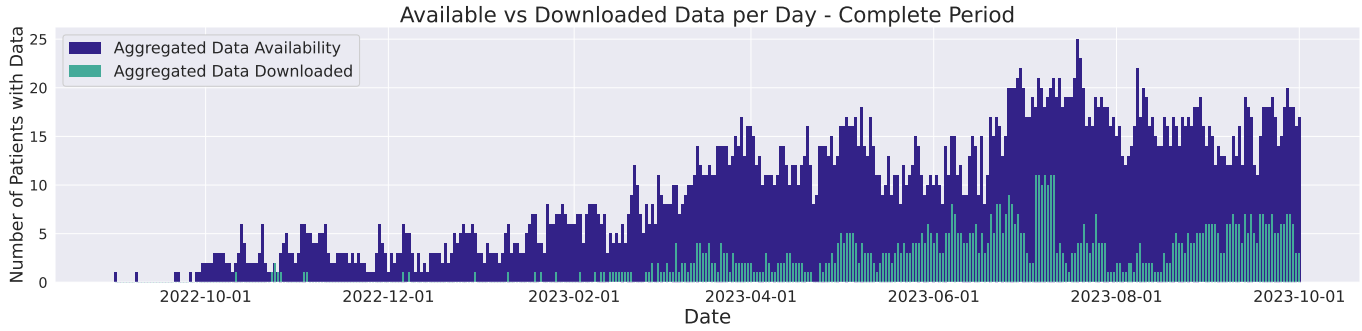

Fig. 5. Bar diagram displaying the total amount of data available (blue) and downloaded (green) throughout the entire study period, aggregated for all patients. The diagram consists of bars representing each day, with a count of one added for each patient from which data is downloaded on that day. The counts are then aggregated for all patients, showing the distribution of the total available and downloaded data.

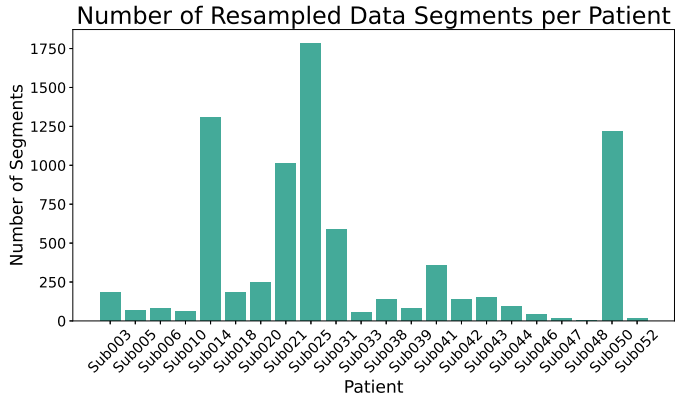

Fig. 6. Bar diagram for the total number of included data segments per patient. Each bar corresponds to a single patient, with the height of the bar corresponding to the total number of data segments included for analysis, resampling of the data segments.

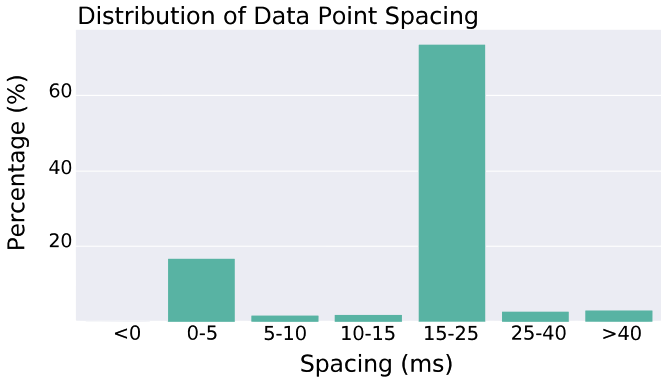

Fig. 8. Histogram of spacing between consecutive data points, over the complete data set. The highest percentage falls within the 15-25 ms bin for data point spacing, which is in line with the sampling frequency of 50 Hz and, thus, a data spacing of 20 ms. Remarkable, is the high percentage of occurrences of a data point spacing of 0-5 ms.

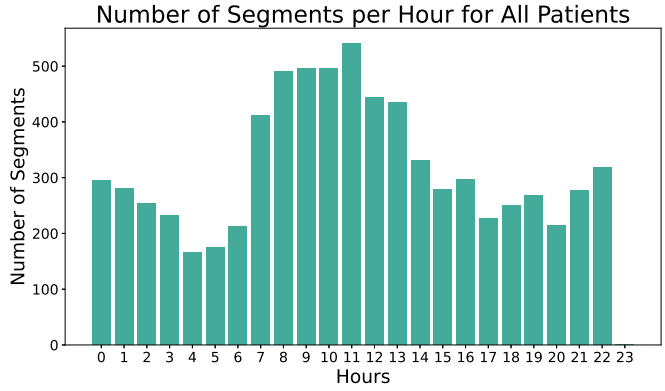

Fig. 7. Bar diagram presenting the number of data segments for all patients over a complete day. The x-axis represents the hour in time for which a data segment is present, with the time ranging from 00:00 to 23:00. This diagram shows the distribution of data segments for each patient over a complete day, with the majority of data obtained from daytime hours.

data from the Apple Watch is obtained, and the reason for this small data point spacing is unclear.

3) Data Resampling and Imputation: To address unevenly spaced data, the data segments were resampled to obtain a data spacing of 20 ms. To evaluate the plausibility of this method, the signals before and after resampling are plotted in one figure (see Appendix F.4). Due to averaging, resampled values were sometimes slightly lower than the original values. However, these differences are minor.

After data resampling, the gaps in data segments were filled using linear interpolation and zero padding, selected for their simplicity and effectiveness in imputing time-series data. To verify the correct working of these methods, a visualisation was made of the signal before and after imputation.

The resulting data segments of merged heart rate and kinematic data, after heart rate data interpolation and scaling, were also visualised. This visualisation served as a means to verify the correct functioning of the method (see Appendix F.5). The number of segments with heart rate data is 477 segments, which is much less than the number of kinematic data segments.

4) Non-Wear Detection: To determine the threshold for non-wear detection, a histogram was plotted (figure 9) using the standard deviation of the VM for the acceleration signal to represent the distribution. From this histogram, it was not possible to make a distinction between two classes for defining a threshold based on the current data set. Therefore, the threshold value of $0.013\mathrm{G}$ [66] was applied, indicated by the

red line in the histogram.

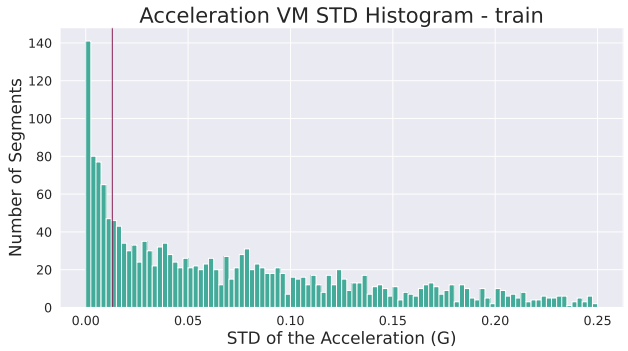

Fig. 9. Histogram for identification of threshold for maximal data missing rate, using the SD of the VM from the acceleration signal. The red line indicates the threshold value of 0.013G [66]. No clear distinction can be found between two classes, for determining a non-wear detection threshold based on this data.

The non-wear detection method resulted in the removal of a total of 857 data segments. Upon visual inspection, most segments showed signals with values around zero for both the acceleration and rotation signals or with a noise signal with zero mean. In Appendix F.5, examples of segments can be found identified as non-wear instances.

The comparison between the non-wear segments and the heart-rate data revealed that most of the non-wear segments did not contain heart-rate data. In 1.5% of the non-wear segments, heart rate data was present, indicating that these segments are not actual non-wear instances. Notably, the heart rate values of these non-wear segments ranged between 40 - 50 rpm, typical of human heart rates during sleep. This observation highlights the challenge of distinguishing between non-wear and sleep or rest instances.

5) Signal Artefacts and Noise: Upon visual examination of data segments, distinct low-frequency artefacts resembling non-linear baseline drift were observed in specific acceleration signal segments. Examples of such segments are visualised in Appendix F.8. These artefacts were more frequently present in segments categorised as 'non-wear' than those classified as 'wear'. No prior studies were found with identical signal artefacts in acceleration signals. However, these artefacts may be a combination of multiple sensor errors, such as scale factor errors, constant bias, drift and random errors [88–90]. Since errors in accelerometer and gyroscope data are common, a variety of studies are focused on developing error reduction methods [88–92]. However, in electrocardiogram (ECG) signal processing, a similar phenomenon, known as 'baseline wandering,' has been previously identified [93–95].

During the analysis of the rotation signal, it was observed that certain inaccuracies were present in the form of noise signals. Examples of such segments are visualised in Appendix F.8. These noise signals were not always associated with instances of artefacts in the acceleration signal.

C. Data Analysis

After pre-processing the data, characteristic features for PD were computed, which were then used to train a VAE. The computed features are presented in the following paragraphs, followed by a discussion on the training performance of the VAE.

1) Feature Analysis: For data interpretation, characteristic features for PD were computed in the time domain and in the frequency domain, from which a visualisation of the results is presented in the following paragraphs.

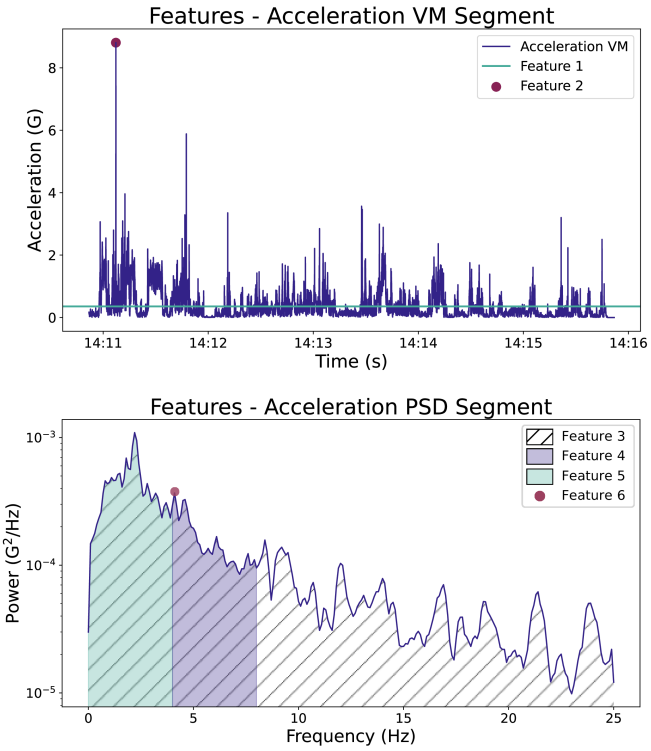

Fig. 10. Vector magnitude (VM) plot and power spectral density (PSD) plot examples of the acceleration signal with the features indicated in the figure. In the top figure, the two time domain features are indicated: Feature 1) the mean acceleration and rotation of the VM per segment, and feature 2) the maximum acceleration and rotation of the VM per segment. The spectral features are indicated in the bottom figure: Feature 3) The area under the total PSD (total spectral power), feature 2) the area under the PSD below 4 Hz, feature 3) the area under the PSD in the tremor frequency band (4 - 8 Hz) and 4) the peak power in the tremor frequency band (4 - 8 Hz).

In the time domain analysis, two features (feature 1: segment mean of the VM, feature 2: segment maximum of the VM) were computed from both acceleration and rotation signals from each data segment (see figure 10). The mean VM values were then averaged across all available days for each patient, enabling the observation of average daily activity patterns. Figure 11 illustrates the average day for the patient from which the most data is available. Observations from the average daily activity patterns suggest higher mean VM values for both signals between 6 am and 9 pm, aligning with expected daily activity patterns [96, 97], suggesting the plausibility of the data in this data set.

To investigate whether the 'mild' patient group can be distinguished from the 'severe' patient group, a violin plot was generated for each feature (see figure 12). The observations from the violin plots in the time domain show that generally, the 'severe' patient group shows higher mean values for

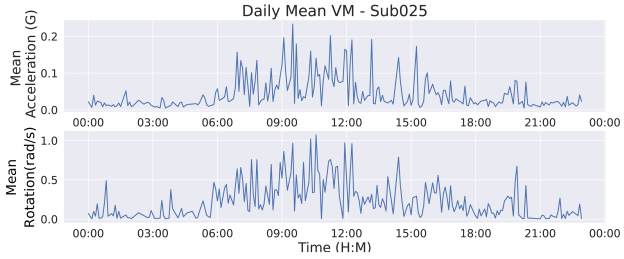

Fig. 11. Plot of the acceleration and rotation VM values, showing the average values over all data segments from one patient during a day. This plot provides a visualisation of the daily pattern for the acceleration and rotation magnitude values of Sub025.

both features 1 and 2, a larger variance and a larger data distribution than the 'mild' patient group. Especially a large difference is visible in the feature value distribution between the two groups. For the acceleration signal specifically, a clear difference is seen for both features between the distribution and the mean of the two patient groups. The difference in variance of the groups is smaller, but also slightly higher for the severe patient group. The rotation signal features showed an even greater difference in distribution, especially for feature 2. The mean of the severe patient group was also higher for both features, compared to the mild patient group. The variance of the severe patient group, however, is lower for feature 2 compared to the mild group.

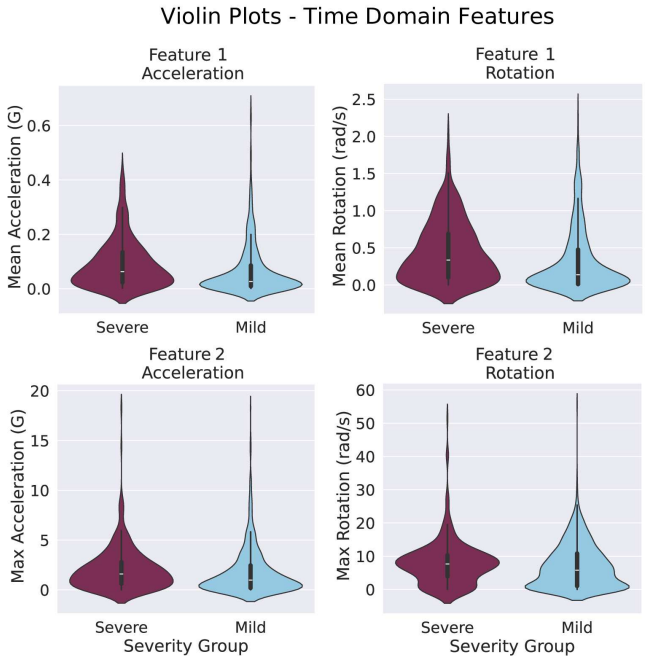

Fig. 12. Violin plot of the time domain features of the 'severe' (red) and 'mild' (blue) patient groups. Feature 1) is the mean acceleration and rotation of the VM per segment, and feature 2) is the maximum acceleration and rotation of the VM per segment. On the left are violin plots of acceleration signal features, and on the right are violin plots of rotation signal features.

For each of the frequency domain features (features 3-6), a violin plot was also computed for both signals to distinguish between the 'mild' and 'severe' patient groups (see figure 13). For the acceleration signal, relatively small differences were

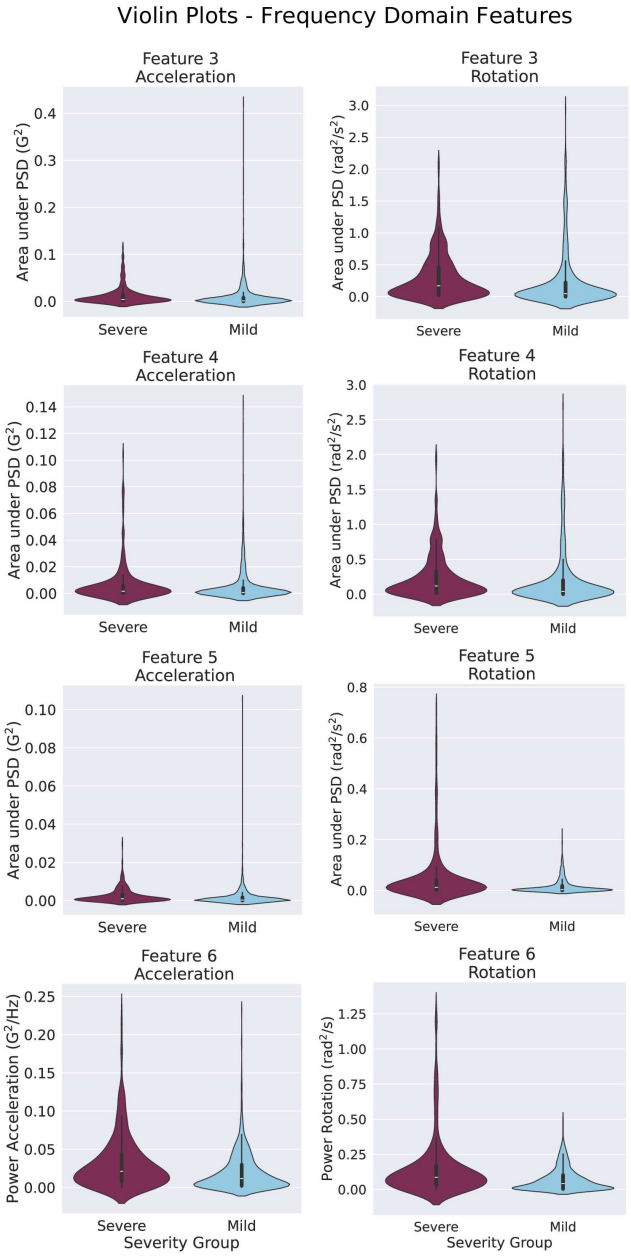

Fig. 13. Violin plot of the frequency domain features of the acceleration and rotation signals of the 'severe' (red) and 'mild' (blue) patient groups. From top to bottom, the spectral features are presented: 3. Total area under PSD 4. Power under PSD for bradykinesia frequency band (below 4 Hz) 5. Power under PSD for tremor frequency band (4-8 Hz) 6. Peak within the tremor frequency band (4-8 Hz). On the left are violin plots of acceleration signal features, and on the right are violin plots of rotation signal features.

observed between the violin plots of the features, compared to the rotation signal, with a large outlier across all features in the mild patient group. For feature 5 (spectral power between 4-8 Hz), the differences between the two patient groups are the smallest. For feature 3-5, the 'severe' group shows slightly higher mean and variance compared to the 'mild' group and slightly larger distributions. Notably, larger differences between patient groups are visible for feature 6 (peak power between 4-8 Hz). With a larger distribution, a higher mean, and a slightly higher variance of the severe patient group.

For the spectral features of the rotation signal, more distinct variations are observed between the two patient groups, than for the acceleration signal. For feature 3 (total spectral power), it is clear that the distribution is larger, and the mean and variance are higher for the severe patient group. For feature 4 (spectral power below 4 Hz), the severe patient groups also show larger distribution, mean, and variance but with smaller differences for the latter two compared to feature 3. For feature 5 (spectral power between 4-8 Hz), it is notable that the mild patient group has a very low distribution of feature values, approaching zero, with clearly a larger feature value distribution for the severe group. For this feature, small differences are observed between the mean and variance, with slightly higher values for the severe patient group. Finally, for feature 6 (peak power between 4-8 Hz), the largest difference is observed in distribution with a slightly higher mean and variance for the severe patient group.

2) Data Classification Model: The VAE model was trained using different input modalities as described in section III-C4. Modalities 1a-c, with non-wear segments not removed, contained 2319 segments as input. After the removal of non-wear segments from input modalities 2a-c, 1653 segments remained for use as input. The PSD of the segments was computed after non-wear segment removal, and thus input modalities 3a-b contained 1653 segments. Due to the limited availability of heart rate data, input modality 4 contained only 436 segments for use as input to the VAE. The model architecture changes were limited, with adjustments made only to the number of latent features, varying between 3 - 100 latent features.

For training the VAE model using the time-series input data, despite using various input modalities (modalities 1 and 2 a-c), the reconstruction and validation losses for both models remained constant. The input modality 4, using heart rate data, resulted in even higher loss values. These results indicate that the model did not learn from the input data throughout the training epochs. Adjusting the number of latent features or adjusting layers and filter sizes, did not yield any differences in the loss values, supporting the observation that the VAE was not capable of learning from the current versions of the time-series input data.

In contrast, the VAE model was able to learn from input modality 3a, the frequency-domain input of the rotation data. For this, applying three latent features resulted in optimal learning and model reconstruction curves. In figure 14, the learning curve of one training instance is displayed, where the blue line shows the total loss as defined in Section II, and the orange line represents the validation loss. Both the training and validation loss curves displayed a decreasing trend throughout the learning process, suggesting learning progress. However, the observed gap for the current results between the total and validation loss curves suggests the model may be slightly overfitting [49]. The number of epochs, displayed on the y-axis, is limited to 18 due to the implementation of an early stopping rule in the model, stopping the model training before reaching 25 epochs. This prevents the model from continuing to learn if no significant improvement is observed in several epochs, saving computational costs and preventing overfitting.

The VAE model was also capable of learning using input modality 3b, the frequency domain acceleration data. The initial learning phase was indicated by a rapid decrease in the learning curves. However, after a few epochs, a plateau in the learning curves suggests the learning did not improve further. In addition, the input reconstructions were noticeably different from the original signal, which suggests that the model has limitations in its ability to reproduce input data accurately. The VAE model thus showed superior performance in learning from the frequency-domain rotation data.

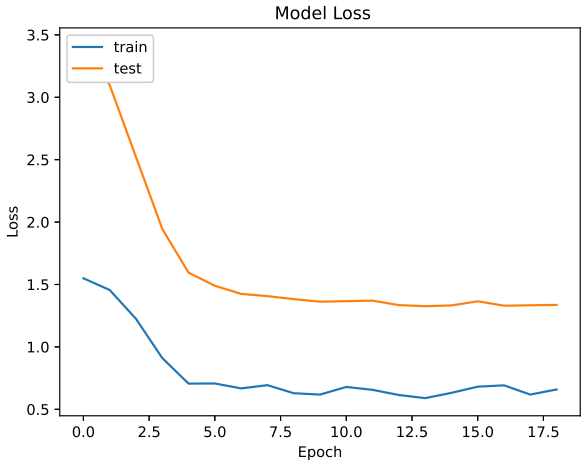

Fig. 14. The loss function for input modality 3a, which is the Power Spectral Density (PSD) of the rotation signal, from training the Variational Autoencoder (VAE) model with three latent features. In the graph, the blue line represents the total loss (eq. 2), while the orange line represents the validation loss. Both lines show a decreasing trend across epochs, indicating that the model is learning.

V. Discussion

According to the findings of this study, the current data pre-processing techniques utilised on this particular smartwatch data set present some limitations in training a VAE for the development of a bradykinesia detection model. Nevertheless, the VAE model's capability to train on the frequency domain input data modality and the identification of differences between two PD severity groups in feature analysis, demonstrate the potential of using smartwatch data recorded in a patient's natural environment and training a VAE. Therefore, by identifying limitations and outlining the necessary steps for future studies, this study has laid the foundation for further research into the development of a bradykinesia detection model.

In this section, a detailed description of the results of VAE training and feature analysis is provided, to elaborate on this study's findings. Following this, limitations of the current study in data quantity and quality, and current VAE training procedures are discussed, along with the required steps for future studies to be able to train a VAE using this data. Finally, further limitations of the current methods for data collection, pre-processing, and analysis are highlighted, along with recommendations for improving these methods in future research.

A. Interpretation of Results

In training the proposed VAE model on the kinematic smartwatch data, it encountered challenges in learning from any of the input modalities with the time domain version of the data, even with additional heart rate data. However, it was found that the VAE model effectively learned from the frequency domain version of the data.

The VAE's inability to learn from time domain input modalities can be caused by various factors. One of these factors is the presence of low-frequency artefacts in the acceleration data and noise in the rotation data. These low-frequency artefacts could be an important cause of the model's inability to identify patterns from the time domain version of the kinematic data. The frequency domain representation, on the other hand, breaks the signals down into their frequency components, making identification of any patterns by a VAE less susceptible to noise and artefacts. However, while the frequency domain input (modalities 3 a-b) allowed the VAE to learn, if the input data contains inaccuracies, training the VAE on that data can result in erroneous pattern recognition. Additionally, the frequency domain version of the data disregards time information and, given the current data segment length, is thus not appropriate for real-time symptom detection.

Moreover, the impact of the data pre-processing method has not been thoroughly investigated in the current study. These methods might have possibly altered the data in such a way that it posed challenges for the VAE in extracting meaningful features. In addition, data imbalances due to varying data availability among patients and freezing of the data downloading process might also have contributed to the limitations for the VAE to learn from the time domain modalities. These findings emphasise the importance of guaranteeing sufficient quality and quantity of the data set, before further training of the VAE model.

Although the possibility of training the VAE was limited, the results from the feature analysis indicate that relevant information is captured in the data about PD symptoms. This is supported by the differences in feature values between the 'mild' and 'severe' patient groups. Specifically, the 'severe' patient group showed a larger distribution of feature values, as well as higher mean and variance values for the features.

These observations correspond with previous studies that found a positive correlation between the MDS-UPDRS scores and wrist activity [81, 98, 99]. The greater variance and feature value distribution can be explained by fluctuations associated with PD symptoms [98]. Moreover, the larger distribution of the feature values for the severe patient group indicates that variations in feature values are consistent and substantial. Higher mean feature values indicate higher wrist activity and are likely higher due to tremor and dyskinesia in the wrist [99].

This assumption is supported by the observed presence of tremor in the spectral features. In the rotation signal, features 5 and 6 (spectral power and peak power within the tremor frequency band), show noticeable differences between the two patient groups. From the acceleration signal, for feature 6, the 'severe' and 'mild' patient groups can also clearly be distinguished. These findings provide clear evidence that tremor is present in the data.

On the other hand, identifying bradykinesia through the current feature analysis is challenging. This is because the analysis of feature 4 (spectral power below 4 Hz), which relates to bradykinesia, shows minor differences in mean, variance, and distribution for both rotation and acceleration. These values are comparatively higher in the 'severe' patient group, which suggests the presence of bradykinesia. However, the differences are so small that they do not provide strong evidence for the presence of bradykinesia. Moreover, other factors, such as dyskinesia, which share similar frequencies, are not considered during the analysis. Therefore, a more detailed analysis is required to correctly identify bradykinesia.

Another observation based on the feature plots is that all features, in both time and frequency domains, showed greater differences between the 'severe' and 'mild' patient groups for the rotation signals compared to acceleration signals. These findings are in accordance with the results from the VAE, which showed a better learning performance using the rotation input data compared to the acceleration input data. This suggests the relevance of analysing rotation signals in analysing PD symptoms. However, the rotation signal is important for detecting tremor, which should be taken into consideration since it could largely influence the results.

Finally, the daily activity patterns of the VM of the acceleration and rotation signals, align with expected times of increased and decreased activity. This also indicates the relevance of the current data set and shows that, despite the data quality issues, global patterns can be identified in the data.

Overall, the differences identified from the feature plots between the two severity groups, consistent with prior research, suggest that the data collected in this study contains information indicative of PD symptom severity. Despite the limitations in training the VAE from the time domain modalities of the current input data, the presence of these distinctive data patterns suggests that, after obtaining a more evenly distributed data set and performing additional data pre-processing, it is possible to develop classification models for this data in the future.

B. Limitations and Future Research

The current study has identified several limitations related to data collection, retrieval, and data quality. In the following paragraphs, firstly, the limitations that need to be considered in future studies working with the same dataset and the proposed VAE model are highlighted, along with suggestions to address these limitations to develop a bradykinesia detection model. This is followed by additional limitations and corresponding recommendations that could enhance the data collection, pre-processing and analysis methods and result validation.

1) Further Model Development: In the current study, the key limitation that was identified relates to data anomalies, which include low-frequency artefacts in the acceleration data, and noise in the rotation data. Such anomalies can compromise the reliability of the data and negatively affect the model's

ability to identify patterns from the kinematic data in the time-domain version. Since the data is collected in an uncontrolled environment and no previous studies were found that highlight and address similar artefacts in acceleration signals, it is hard to determine the root cause of the acceleration artefacts and identify potential resolution methods.

Therefore, to address this issue, it is recommended to determine the root cause of the acceleration artefacts that facilitate further development of a bradykinesia detection model using a VAE trained on this study's data set. One approach is to start with a small-scale study in a controlled environment focused on hardware inspection, to ensure the reliability of the data and understand the implications of the artefacts in the current data set. Employing experiments on a smaller scale within a controlled setting enables assessment of the feasibility of individual research components [100].

Examining signal responses during specific movements and conditions could help identify the origins of acceleration artefacts. This way, it can be determined whether these artefacts arise from specific movements, sensor hardware-related factors such as mechanical wear, hysteresis, or the sensor's sensitivity to environmental variables like humidity and temperature [88, 101]. Additionally, it can be assessed what the noise patterns in the rotation data stem from and whether the rotation data is reliable during these acceleration artefacts.

After identifying the causes of signal artefacts, the next step is to develop strategies for detecting and addressing these artefacts. No prior studies with identical signal artefacts were found. In addition, since these artefacts could be a combination of several sensor errors, it is difficult to distinguish and address these errors. There has been extensive research into accelerometer and gyroscope error reduction methods, including static and dynamic device calibration [90, 91] and filtering [90] and ML methods [92] to remove deterministic and stochastic errors. These methods can be applied if sensor errors are found to be the cause of the acceleration artefacts and rotation noise.

Additionally, the low-frequent acceleration artefacts resemble a nonlinear baseline drift pattern known as 'baseline wandering' in ECG signals. These have been addressed using filtering techniques [93] or baseline cancellation using wavelets like the discrete wavelet transform [93, 94]. Advanced approaches, leveraging DL methods [95], have shown even greater efficacy in eliminating baseline wandering but at the cost of higher computational requirements. These methods could be applied in case the acceleration artefacts are found to be identical to ECG baseline wandering.

An alternative consideration is to explore the possibility of entirely removing segments with signal artefacts or noise from the data set. However, doing so can pose a risk of discarding important data and introducing bias into the analysis. Therefore, deciding whether to remove data segments with artefacts or noise in the kinematic data or remove signal errors depends on the root causes of the low-frequency artefacts and should be carefully considered.

Secondly, to work with this data set in future studies, the limitations posed by the freeze in data downloading need to be addressed, to be able to acquire the complete data set. The problem with the freeze in the downloading process for the complete data set was not resolved in the current study. This is due to time limitations, as this problem emerged during the final month of the study while attempting to download the complete data set using the V2 access API code. From communication with the Rune Labs platform regarding this data downloading issue, it was identified that the disruption in the data downloading process might be due to random access memory (RAM) overload. This problem is likely related to how parallelisation is implemented in the code for data retrieval since the code was developed for the Rune Labs cluster. The code's batch processing of API calls might conflict with settings in the AMC cluster, causing the freezing of code execution [102]. To solve this issue, it is advised to contact the AMC cluster administrator.

Thirdly, to improve VAE model training on the current data set, it is advisable to also train the model with smaller segment lengths. Smaller segment lengths minimise the likelihood of capturing both normal and bradykinetic movements and enable the model to identify patterns more easily. However, if the segments are too small, there is an increased chance of capturing only noise, hence requiring a careful trade-off. A viable option is a segment length of 30 seconds, for which the segmentation process and the initial pre-processing steps were already conducted in this study (see Appendix D). However, it takes longer to process these smaller data segments. Therefore, due to time limitations and the presence of the low-frequency acceleration artefacts, no further steps were performed using the 30-second data segments. However, the data set segmented into 30-second segments can be utilised in future studies.

Fourthly, the VAE model showed successful learning with frequency domain input, but this version of the 5-minute data segments is not optimal for real-time application, as it lacks time information. Therefore, future studies should explore wavelets as an alternative input approach, which contains both time and frequency content [33, 103, 104]. This will ensure that the input data is more representative of the time domain in real-time application, while allowing improved pattern recognition from the frequency domain.

Fifthly, to address limitations regarding data quantity and distribution, it is advised to train the VAE on the complete data set in future studies. To train the VAE effectively, it is advisable to start with a small well-distributed data set and progress to larger ones once successful learning is achieved. However, the current study's data processing codes have not been optimised yet to run efficiently on the entire dataset. As future studies move to larger versions of the data set, it is advised to optimise the code and improve parallel processing to process large amounts of data efficiently. To avoid RAM memory overload issues when loading the complete dataset to the VAE, a possible solution would be to incorporate a data loader for batch input of the complete dataset [105]. This approach would prevent loading all the data into the system memory simultaneously.

Finally, for the VAE model to learn from time domain input data, it is advised to first address the input data limitations using the previously outlined recommendations. After this, it becomes valuable to evaluate the VAE model and

experiment with different architectures and hyperparameters to improve model performance. Additionally, alternative ML models could also be applied in future for comparative analysis and evaluating the suitability of the VAE for the current purpose. This comparative analysis not only facilitates model assessment but can also serve as a validation step after successful VAE training.

2) Additional Improvements: The recommendations for future research for additional improvements are related to the pre-processing and feature analysis of the data and to improving the data collection process, described in the following paragraphs.

The data pre-processing methods used in the current study were chosen for simplicity and evaluated through visual inspection. However, a more thorough evaluation of these techniques is recommended to better understand their effects on VAE training. Future studies should compare the results of current pre-processing methods for feature analysis or VAE training to those obtained after applying alternative, more advanced methods. The following paragraphs elaborate on possible alternative pre-processing methods.

An alternative method to address unevenly spaced data is the Lomb-Scargle Periodogram [106]. This method allows the identification of periodicity in the signal and can aid in accounting for uneven sampling. The current method of resampling, to resolve uneven spacing of data points, shows slight deviations in resampled values at certain time instances. Therefore, this alternative method could be used to more effectively address unevenly spaced data.

For data imputation, alternative methods range from statistical methods to ML methods [58, 61, 107–109]. In selecting an alternative data pre-processing method, a trade-off should be made between the complexity of the method and the computational resources required. More advanced methods may require more computational power, so the benefits of using a more advanced method should be weighed against its computational costs.

Another approach for investigating the effect of data imputation is to remove parts of complete data segments, replace the missing data using different data imputation methods, and calculate features for the resulting segments [60]. By comparing the results from different data imputation methods with those obtained using the complete signal, it is possible to evaluate the effect of different data imputation methods and determine the most suitable one.

An alternative non-wear detection method can be used to evaluate the effectiveness of the current method. Alternative methods can also be combined with the current method to improve non-wear detection. An alternative approach to identify non-wear, involves utilising the ratio of PSD between frequencies associated with activities of daily living (ADL) and remaining frequencies. During non-wear, an even power distribution across frequencies is expected, mainly determined by noise, and thus the PSD ratio is expected to be small. In contrast, during wear periods, more power in frequencies is expected, linked to daily activities, resulting in a larger PSD ratio. Therefore, setting a minimal threshold for the power ratio can serve as a non-wear detection method.

A second alternative non-wear detection method involves using heart rate data to identify 'wear' segments if they contain heart rate data. However, this method requires more heart rate data than what is currently available in the data set to minimise the risk of disposing of relevant kinematic data segments. Additionally, since recording the heart rate data increases battery consumption [54, 55], a trade-off between battery life and data reliability must be considered.

The final recommendations for future research relate to the feature analysis method applied in this study. The study's analysis was based on a restricted set of features without additional filtering to attenuate certain symptoms, providing an analysis of patients' overall symptom severity. Further analysis will help determine if bradykinesia manifestations are adequately represented in this data set. To achieve this, a larger and more evenly distributed data set is required, after which an extended feature analysis process can be implemented. For the feature analysis, patients can be categorised into three distinct groups based on their individual MDS-UPDRS symptom scores. These groups can be categorised as follows: patients who predominantly experience bradykinesia and rigidity, patients with more pronounced tremors and dyskinesias, and patients with mild overall symptom expressions. This categorisation improves the assessment of the data set's suitability for detecting bradykinesia by separating the effects of different symptoms. In addition, to test if there are significant differences between different patient groups, more factors need to be considered, including their medical history, age and medication intake, and the date and time from which the data is available. By conducting a more comprehensive feature analysis, it becomes feasible to gain a better understanding of the data, which can help enhance the interpretation of VAE training results.

VI. Conclusion

In this study, a VAE model was trained using kinematic smartwatch data obtained from PD patients in their natural environments. The model demonstrated efficacy in learning from the frequency content of the smartwatch data. The study also analysed PD features derived from kinematic data collected in natural environments. It revealed differences between patient groups with 'mild' and 'severe' PD symptoms, demonstrating that this data contains valuable information about PD motor symptoms. Based on the study's findings, it can be concluded that with the current data pre-processing steps and the current data set, it is not possible to develop a bradykinesia detection model using a VAE. However, the relevant information in the data, as demonstrated by the feature analysis and the learning capability of the VAE with frequency input data, indicates that after additional data cleaning and pre-processing and a more balanced data set, the VAE can further be trained. While the study's approach did not result in bradykinesia model development, the necessary steps for utilising this data for further analysis and as input for training the proposed VAE model have been demonstrated, and thus, the groundwork has been laid for the development of a bradykinesia detection model using this data set.

.

Acknowledgements

First of all, I would like to express my gratitude to my supervisor, Alfred Schouten, for his guidance during my research project. He encouraged me to think critically about my methodology by zooming out to the general overview and purpose of the project. I would also like to thank Sina David for sharing her VAE model with me. She was always available to answer my questions, and her insights were helpful in finding ways to address inaccuracies in the data. Additionally, I thank Mariëlle Stam and Martijn de Neeling for providing valuable clinical insights. I collaborated with them for monitoring data quality, and thanks to them, I got to see the hospital from the inside, which was a special experience. Finally, I would like to thank Ro'ee Gilron from Rune Labs for answering my questions regarding data downloading and quality. His code enabled me to download the data in the end.

References

[1] B. R. Bloem, M. S. Okun, and C. Klein, "Parkinson's disease," *The Lancet*, vol. 397, no. 10291, pp. 2284–2303, Jun. 2021.

[2] C. Marras and A. Lang, "Parkinson's disease subtypes: Lost in translation?" *Journal of Neurology, Neurosurgery and Psychiatry*, vol. 84, no. 4, pp. 409–415, 2013.

[3] B. L. Van Der Gaag, D. H. Hepp, J. I. Hoff, J. Van Hilten, S. K. Darweesh, and B. R. Bloem En Wilma Dj Van De Berg, "Risicofactoren voor de ziekte van Parkinson Mogelijkheden voor preventie en interventie," *Nederlands Tijdschrift voor Geneeskunde*, vol. 67, no. 38, Sep. 2023. [Online]. Available: https://proparkinson.nl

[4] E. Ray Dorsey, A. Elbaz, E. Nichols, G. Logroscino, D. M. Pereira, and M. J. Postma, "Global, regional, and national burden of Parkinson's disease, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016," *The Lancet Neurology*, vol. 17, no. 11, pp. 939–953, Nov. 2018.

[5] R. Xia and Z. H. Mao, "Progression of motor symptoms in Parkinson's disease," *Neuroscience Bulletin*, vol. 28, no. 1, pp. 39–48, Feb. 2012.

[6] J. Jankovic, "Motor fluctuations and dyskinesias in Parkinson's disease: Clinical manifestations," *Movement Disorders*, vol. 20, no. Suppl. 11, 2005.

[7] B. Thanvi, N. Lo, and T. Robinson, "Levodopa-induced dyskinesia in Parkinson's disease: clinical features, pathogenesis, prevention and treatment," *Postgraduate Medical Journal*, vol. 83, no. 980, pp. 384–388, 6 2007.

[8] T. M. Fraczek, B. I. Ferleger, T. E. Brown, M. C. Thompson, A. J. Haddock, B. C. Houston, and J. G. Ojemann, "Closing the Loop With Cortical Sensing: The Development of Adaptive Deep Brain Stimulation for Essential Tremor Using the Activa PC+S," Dec. 2021.

[9] E. Rojas, S. L. Schmidt, A. Chowdhury, M. Pajic, D. A. Turner, and D. S. Won, "A comparison of an implanted accelerometer with a wearable accelerometer for closed-loop DBS," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2022-July. Institute of Electrical and Electronics Engineers Inc., 2022, pp. 3439–3442.

[10] A. L. Benabid, "Deep brain stimulation for Parkinson's disease," *Current Opinion in Neurobiology*, vol. 13, no. 6, pp. 696–706, Dec. 2003.

[11] A. L. Silva de Lima, T. Hahn, N. M. de Vries, E. Cohen, L. Bataille, M. A. Little, H. Baldus, B. R. Bloem, and M. J. Faber, "Large-Scale Wearable Sensor Deployment in Parkinson's Patients: The Parkinson@Home Study Protocol," *JMIR Research Protocols*, vol. 5, no. 3, p. e172, Aug. 2016.

[12] O. Rascol, C. Goetz, W. Koller, W. Poewe, and C. Sampaio, "Treatment interventions for Parkinson's disease: an evidence based assessment," *The Lancet*, vol. 359, no. 9317, pp. 1589–1598, May 2002.

[13] C. E. Clarke, "Parkinson's disease," *British Medical Journal*, vol. 335, no. 7617, pp. 441–445, 9 2007.

[14] J. M. Bronstein, M. Tagliati, R. L. Alterman, A. M. Lozano, J. Volkmann, A. Stefani, F. B. Horak, M. S. Okun, K. D. Foote, P. Krack, R. Pahwa, J. M. Henderson, M. I. Hariz, R. A. Bakay, A. Rezai, W. J. Marks, E. Moro, J. L. Vitek, F. M. Weaver, R. E. Gross, and M. R. DeLong, "Deep brain stimulation for Parkinson disease an expert consensus and review of key issues," *Archives of Neurology*, vol. 68, no. 2, pp. 165–171, Feb. 2011.

[15] S. Little, A. Pogosyan, S. Neal, B. Zavala, L. Zrinzo, M. Hariz, T. Foltynie, P. Limousin, K. Ashkan, J. FitzGerald, A. L. Green, T. Z. Aziz, and P. Brown, "Adaptive deep brain stimulation in advanced Parkinson disease." *Annals of neurology*, vol. 74, no. 3, pp. 449–457, 2013.

[16] H. B. Kim, W. W. Lee, A. Kim, H. J. Lee, H. Y. Park, H. S. Jeon, S. K. Kim, B. Jeon, and K. S. Park, "Wrist sensor-based tremor severity quantification in Parkinson's disease using convolutional neural network," *Computers in Biology and Medicine*, vol. 95, pp. 140–146, Apr. 2018.

[17] E. Rastegari, H. Ali, and V. Marmelat, "Detection of Parkinson's Disease Using Wrist Accelerometer Data and Passive Monitoring," *Sensors*, vol. 22, no. 23, Dec. 2022.

[18] A. J. Hadley, D. E. Riley, and D. A. Heldman, "Real-World Evidence for a Smartwatch-Based Parkinson's Motor Assessment App for Patients Undergoing Therapy Changes," *Digital Biomarkers*, vol. 5, no. 3, pp. 206–215, Sep. 2021.

[19] A. S. Chandrabhatla, I. J. Pomeraniec, and A. Ksendzovsky, "Co-evolution of machine learning and digital technologies to improve monitoring of Parkinson's disease motor symptoms," Dec. 2022.

[20] C. Morgan, M. Rolinski, R. McNaney, B. Jones, L. Rochester, W. Maetzler, I. Craddock, and A. L. Whone, "Systematic Review Looking at the Use of

Technology to Measure Free-Living Symptom and Activity Outcomes in Parkinson's Disease in the Home or a Home-like Environment," *Journal of Parkinson's Disease*, vol. 10, no. 2, pp. 429–454, 2020.

[21] J. Fujikawa, R. Morigaki, N. Yamamoto, H. Nakanishi, T. Oda, Y. Izumi, and Y. Takagi, "Diagnosis and Treatment of Tremor in Parkinson's Disease Using Mechanical Devices," *Life*, vol. 13, no. 1, p. 78, Dec. 2022.

[22] R. Powers, M. Etezadi-Amoli, E. M. Arnold, S. Kianian, I. Mance, M. Gibiansky, D. Trietsch, A. Singh Alvarado, J. D. Kretlow, T. M. Herrington, S. Brillman, N. Huang, P. T. Lin, H. A. Pham, and A. V. Ullal, "Smartwatch inertial sensors continuously monitor real-world motor fluctuations in Parkinson's disease," *Sci. Transl. Med*, vol. 13, p. 7865, 2021. [Online]. Available: http://stm.sciencemag.org/

[23] A. Channa, G. Ruggeri, N. Mammone, R. C. Ifrim, A. Iera, and N. Popescu, "Parkinson's Disease Severity Estimation using Deep Learning and Cloud Technology," in *2022 IEEE International Conference on Omni-Layer Intelligent Systems, COINS 2022*. Institute of Electrical and Electronics Engineers Inc., 2022.

[24] A. Channa, R. C. Ifrim, D. Popescu, and N. Popescu, "A-wear bracelet for detection of hand tremor and bradykinesia in parkinson's patients," *Sensors (Switzerland)*, vol. 21, no. 3, pp. 1–23, Feb. 2021.

[25] S. H. Roy, B. T. Cole, L. D. Gilmore, C. J. De Luca, C. A. Thomas, M. M. Saint-Hilaire, and S. H. Nawab, "High-resolution tracking of motor disorders in Parkinson's disease during unconstrained activity," *Movement Disorders*, vol. 28, no. 8, pp. 1080–1087, Jul. 2013.

[26] L. Sigcha, I. Pavón, N. Costa, S. Costa, M. Gago, P. Arezes, J. M. López, and G. D. Arcas, "Automatic resting tremor assessment in parkinson's disease using smartwatches and multitask convolutional neural networks," *Sensors (Switzerland)*, vol. 21, no. 1, pp. 1–29, Jan. 2021.

[27] L. Sigcha, B. Domínguez, L. Borzì, N. Costa, S. Costa, P. Arezes, J. M. López, G. De Arcas, and I. Pavón, "Bradykinesia Detection in Parkinson's Disease Using Smartwatches' Inertial Sensors and Deep Learning Methods," *Electronics (Switzerland)*, vol. 11, no. 23, Dec. 2022.

[28] L. Tong, J. He, and L. Peng, "CNN-Based PD Hand Tremor Detection Using Inertial Sensors," *IEEE Sensors Letters*, vol. 5, no. 7, Jul. 2021.

[29] R. San-Segundo, A. Zhang, A. Cebulla, S. Panev, G. Tabor, K. Stebbins, R. E. Massa, A. Whitford, F. de la Torre, and J. Hodgins, "Parkinson's disease tremor detection in the wild using wearable accelerometers," *Sensors (Switzerland)*, vol. 20, no. 20, pp. 1–23, Oct. 2020.

[30] B. M. Eskofier, S. I. Lee, J. F. Daneault, F. N. Golabchi, G. Ferreira-Carvalho, G. Vergara-Diaz, S. Sapienza, G. Costante, J. Klucken, T. Kautz, and P. Bonato, "Recent machine learning advancements in sensor-based mobility analysis: Deep learning for Parkinson's disease assessment," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2016-October. Institute of Electrical and Electronics Engineers Inc., Oct. 2016, pp. 655–658.

[31] J. Watts, A. Khojandi, R. Vasudevan, F. B. Nahab, and R. A. Ramdhani, "Improving medication regimen recommendation for parkinson's disease using sensor technology," *Sensors*, vol. 21, no. 10, May 2021.

[32] N. Shawen, M. K. O'Brien, S. Venkatesan, L. Lonini, T. Simuni, J. L. Hamilton, R. Ghaffari, J. A. Rogers, and A. Jayaraman, "Role of data measurement characteristics in the accurate detection of Parkinson's disease symptoms using wearable sensors," *Journal of Neuro-Engineering and Rehabilitation*, vol. 17, no. 1, Apr. 2020.

[33] M. Lang, F. M. J. Pfister, J. Fröhner, K. Abedinpour, D. Pichler, U. Fietzek, T. T. Um, D. Kulić, S. Endo, and S. Hirche, "A Multi-layer Gaussian Process for Motor Symptom Estimation in People with Parkinson's Disease," *IEEE Transactions on Biomedical Engineering*, Aug. 2018. [Online]. Available: http://arxiv.org/abs/1808.10663

[34] N. Mahadevan, C. Demanuele, H. Zhang, D. Volfson, B. Ho, M. K. Erb, and S. Patel, "Development of digital biomarkers for resting tremor and bradykinesia using a wrist-worn wearable device," *npj Digital Medicine*, vol. 3, no. 1, Dec. 2020.

[35] A. Salarian, H. Russmann, C. Wider, P. R. Burkhard, F. J. Vingerhoets, and K. Aminian, "Quantification of tremor and bradykinesia in Parkinson's disease using a novel ambulatory monitoring system," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 2, pp. 313–322, Feb. 2007.

[36] A. L. S. De Lima, T. Hahn, L. J. Evers, N. M. De Vries, E. Cohen, M. Afek, L. Bataille, M. Daeschler, K. Claes, B. Boroojerdi, D. Terricabras, M. A. Little, H. Baldus, B. R. Bloem, and M. J. Faber, "Feasibility of large-scale deployment of multiple wearable sensors in Parkinson's disease," *PLoS ONE*, vol. 12, no. 12, Dec. 2017.

[37] K. M. Giannakopoulou, I. Roussaki, and K. Demestichas, "Internet of Things Technologies and Machine Learning Methods for Parkinson's Disease Diagnosis, Monitoring and Management: A Systematic Review," *Sensors*, vol. 22, no. 5, Mar. 2022.

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

[39] M. Sica, S. Tedesco, C. Crowe, L. Kenny, K. Moore, S. Timmons, J. Barton, B. O'Flynn, and D. S. Komaris, "Continuous home monitoring of Parkinson's disease using inertial sensors: A systematic review," *PLoS ONE*, vol. 16, no. 2 February, Feb. 2021.

[40] R. Atri, K. Urban, B. Marebwa, T. Simuni, C. Tanner, A. Siderowf, M. Frasier, M. Haas, and L. Lancashire,

"Deep Learning for Daily Monitoring of Parkinson's Disease Outside the Clinic Using Wearable Sensors," *Sensors*, vol. 22, no. 18, Sep. 2022.

[41] R. I. Griffiths, K. Kotschet, S. Arfon, Z. M. Xu, W. Johnson, J. Drago, A. Evans, P. Kempster, S. Raghav, and M. K. Horne, "Automated assessment of bradykinesia and dyskinesia in Parkinson's disease," *Journal of Parkinson's Disease*, vol. 2, no. 1, pp. 47–55, 2012.

[42] F. Heidarivincheh, R. McConville, C. Morgan, R. McNaney, A. Masullo, M. Mirmehdi, A. L. Whone, and I. Craddock, "Multimodal classification of parkinson's disease in home environments with resiliency to missing modalities," *Sensors*, vol. 21, no. 12, Jun. 2021.

[43] B. Ghoraani, M. D. Hssayeni, M. M. Bruack, and J. Jimenez-Shahed, "Multilevel Features for Sensor-Based Assessment of Motor Fluctuation in Parkinson's Disease Subjects," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 5, pp. 1284–1295, May 2020.

[44] S. Patel, K. Lorincz, R. Hughes, N. Huggins, J. Growdon, D. Standaert, M. Akay, J. Dy, M. Welsh, and P. Bonato, "Monitoring motor fluctuations in patients with parkinsons disease using wearable sensors," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 6, pp. 864–873, Nov. 2009.

[45] N. Siddharth, B. Paige, A. Desmaison, J.-W. Van de Meent, F. Wood, N. D. Goodman, P. Kohli, and P. H. S. Torr, "Inducing Interpretable Representations with Variational Autoencoders," Nov. 2016. [Online]. Available: http://arxiv.org/abs/1611.07492

[46] N. Takeishi and A. Kalousis, "Physics-Integrated Variational Autoencoders for Robust and Interpretable Generative Modeling," in *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*., 2021.

[47] H. V. Fineberg, *Reproducibility and Replicability in Science*. Washington, D.C.: National Academies Press, Sep. 2019.

[48] J. Rocca and B. Rocca, "Understanding Variational Autoencoders (VAEs)," Sep. 2019, accessed on Nov. 6, 2023. [Online]. Available: https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

[49] M. Cogswell, F. Ahmed, R. Girshick, L. Zitnick, and D. Batra, "Reducing Overfitting in Deep Networks by Decorrelating Representations," *CoRR*, Nov. 2015. [Online]. Available: http://arxiv.org/abs/1511.06068

[50] J. Jimenez-Shahed, "Device profile of the percept PC deep brain stimulation system for the treatment of Parkinson's disease and related disorders," *Expert Review of Medical Devices*, vol. 18, no. 4, pp. 319–332, 2021.

[51] "StrivePD Application," 2023, accessed on Dec. 18, 2023. [Online]. Available: https://www.strive.group/app

[52] Apple Inc., "Monitor your heart rate with Apple Watch," Sep. 2023, accessed on Dec. 18, 2023. [Online]. Available: https://support.apple.com/en-us/HT204666#:~:text=By%20flashing%20its%20LED%20lights,30%E2%80%93210%20beats%20per%20minute.

[53] H. G. Espinosa, D. V. Thiel, M. Sorell, and D. Rowlands, "Can We Trust Inertial and Heart Rate Sensor Data from an APPLE Watch Device?" in *The 13th Conference of the International Sports Engineering Association*. MDPI AG, Jun. 2020, p. 128.

[54] A. Vyas and S. Pal, "Power Saving Approach of a Smart Watch for Monitoring the Heart Rate of a Runner," *IEEE Transactions on Consumer Electronics*, vol. 69, no. 3, pp. 490–498, Aug. 2023.

[55] S. Aras, T. Johnson, K. Cabulong, and Gniady Chris, "GreenMonitor: Extending battery life for continuous heart rate monitoring in smartwatches," in *17th International Conference on E-health Networking, Application & Services (HealthCom) : 14-17 Oct. 2015.*, 2015, pp. 317–322.

[56] J. G. Habets, C. Herff, P. L. Kubben, M. L. Kuijf, Y. Temel, L. J. Evers, B. R. Bloem, P. A. Starr, R. Gilron, and S. Little, "Rapid dynamic naturalistic monitoring of bradykinesia in parkinson's disease using a wrist-worn accelerometer," *Sensors*, vol. 21, no. 23, Dec. 2021.

[57] M. Jaén-Vargas, K. M. R. Leiva, F. Fernandes, S. B. Goncalves, M. T. Silva, D. S. Lopes, and J. J. S. Olmedo, "Effects of sliding window variation in the performance of acceleration-based human activity recognition using deep learning models," *PeerJ Computer Science*, vol. 8, 2022.

[58] J. Ae Lee and J. Gill, "Missing value imputation for physical activity data measured by accelerometer," *Statistical Methods in Medical Research*, vol. 27, no. 2, pp. 490–506, Feb. 2018.

[59] A. Sano, W. Chen, D. Lopez-Martinez, S. Taylor, and R. W. Picard, "Multimodal Ambulatory Sleep Detection Using LSTM Recurrent Neural Networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 4, pp. 1607–1617, Jul. 2019.

[60] L. Weed, R. Lok, D. Chawra, and J. Zeitzer, "The Impact of Missing Data and Imputation Methods on the Analysis of 24-Hour Activity Patterns," *Clocks and Sleep*, vol. 4, no. 4, pp. 497–507, Dec. 2022.

[61] A. Tlija, K. Węgrzyn-Wolska, and D. Istrate, "Missing-data imputation using wearable sensors in heart rate variability," *Bulletin of the Polish Academy of Sciences: Technical Sciences*, vol. 68, no. 2, pp. 255–261, 2020.

[62] T. Jayalakshmi and A. Santhakumaran, "Statistical Normalization and Back Propagationfor Classification," *International Journal of Computer Theory and Engineering*, pp. 89–93, 2011.

[63] Y. Ostchega, K. S. Porter, J. Hughes, M. P. H. . Charles, F. Dillon, and T. Nwankwo, "Resting Pulse Rate Reference Data for Children, Adolescents, and Adults: United States, 1999-2008," Tech. Rep., 2011.

[64] L. Choi, Z. Liu, C. E. Matthews, and M. S. Buchowski, "Validation of accelerometer wear and nonwear time classification algorithm," *Medicine and Science in Sports and Exercise*, vol. 43, no. 2, pp. 357–364, Feb. 2011.

[65] M. N. Ahmadi, N. Nathan, R. Sutherland, L. Wolfenden, and S. G. Trost, "Non-wear or sleep? Evaluation of five non-wear detection algorithms for raw accelerometer data," *Journal of Sports Sciences*, vol. 38, no. 4, pp. 399–404, Feb. 2020.

[66] V. T. van Hees, S. Sabia, S. E. Jones, A. R. Wood, K. N. Anderson, M. Kivimäki, T. M. Frayling, A. I. Pack, M. Bucan, M. I. Trenell, D. R. Mazzotti, P. R. Gehrman, B. A. Singh-Manoux, and M. N. Weedon, "Estimating sleep parameters using an accelerometer without sleep diary," *Scientific Reports*, vol. 8, no. 1, Dec. 2018.

[67] M. C. P. M. Vos, "Quantification of parkinsonian tremor, bradykinesia and rigid-ity using the Power-Glove in combination with a force sensor," 2015.

[68] L. di Biase, S. Summa, J. Tosi, F. Taffoni, M. Marano, A. C. Rizzo, F. Vecchio, D. Formica, V. Di Lazzaro, G. Di Pino, and M. Tombini, "Quantitative analysis of bradykinesia and rigidity in Parkinson's disease," *Frontiers in Neurology*, vol. 9, no. MAR, Mar. 2018.

[69] F. M. Skidmore, C. A. Mackman, B. Pav, L. M. Shulman, C. Garvan, R. F. Macko, and K. M. Heilman, "Daily ambulatory activity levels in idiopathic Parkinson disease," *Journal of Rehabilitation Research and Development*, vol. 45, no. 9, pp. 1343–1348, 2008.

[70] M. Kheirkhahan, A. Chakraborty, A. A. Wanigatunga, D. B. Corbett, T. M. Manini, and S. Ranka, "Wrist accelerometer shape feature derivation methods for assessing activities of daily living," *BMC Medical Informatics and Decision Making*, vol. 18, Dec. 2018.

[71] R. J. W. Dunnewold, C. E. Jacobi, and J. J. Van Hilten, "Quantitative assessment of bradykinesia in patients with parkinson's disease," *Journal of Neuroscience Methods*, vol. 74, pp. 107–112, Sep. 1997.

[72] Y. Athavale and S. Krishnan, "A device-independent efficient actigraphy signal-encoding system for applications in monitoring daily human activities and health," *Sensors (Switzerland)*, vol. 18, no. 9, Sep. 2018.

[73] M. G. Tsipouras, A. T. Tzallas, G. Rigas, P. Bougia, D. I. Fotiadis, and S. Konitsiotis, "Automated Levodopa-induced dyskinesia assessment," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC'10*, 2010, pp. 2411–2414.

[74] N. L. W. Keijsers, M. Horstink, and S. Gielen, "Online Monitoring of Dyskinesia in Patients with Parkinson's Disease," *IEEE Engineering in Medicine and Biology Magazine*, pp. 96–103, Jun. 2003.

[75] R. N. Youngworth, B. B. Gallagher, and B. L. Stamper, "An overview of power spectral density (PSD) calculations," in *Optical Manufacturing and Testing VI*, vol. 5869. SPIE, Aug. 2005, p. 58690U.

[76] *PSD computations using Welch's method.*, 1991. [Online]. Available: https://api.semanticscholar.org/CorpusID:61307541

[77] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, I. Henriksen, H. Audren, and T. J. Pingel, "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, no. 3, pp. 261–272, Mar. 2020.

[78] R. A. Ramdhani, A. Khojandi, O. Shylo, and B. H. Kopell, "Optimizing clinical assessments in Parkinson's disease through the use of wearable sensors and data driven modeling," *Frontiers in Computational Neuroscience*, vol. 12, Sep. 2018.

[79] F. Mignard, "About the nyquist frequency," 2005. [Online]. Available: https://api.semanticscholar.org/CorpusID:160008640

[80] J. L. Hale, G. Knell, M. D. Swartz, E. J. Shiroma, T. Ellis, I. M. Lee, and K. P. Gabriel, "The prospective association of accelerometer-measured physical activity and sedentary behavior and time to Parkinson's disease diagnosis in older women: The Women's Health Study (WHS)," *Preventive Medicine Reports*, vol. 35, Oct. 2023.

[81] B. Nisha, J. Rohini, P. Shinde, and R. R. Deshpande, "Actigraphy Monitoring of Symptoms in Patients with Parkinson's disease," *International Journal of Innovative Research in Technology*, 2020. [Online]. Available: http://dx.doi.org/!

[82] S. . David, S. . Georgievska, C. . Geng, Y. . Liu, and Punt, "Using Deep learning to personalize stroke rehabilitation," in *International Society of Biomechanics - Fukuoka, Japan*. Fukuoka: JSB, Aug. 2023, pp. 218–219. [Online]. Available: https://www.isb-jsb2023.com/program-proceedings/

[83] M. W. B. Azlinah, M. Bee, and W. Yap, *Unsupervised and Semi-Supervised Learning*, M. Emre Celebi, Ed. Springer, 2020. [Online]. Available: http://www.springer.com/series/15892

[84] S. Sharma, S. Sharma, and A. Athaiya, "Activation Functions In Neural Networks," *International Journal of Engineering Applied Sciences and Technology*, vol. 4, pp. 310–316, 2020. [Online]. Available: http://www.ijeast.com

[85] T. Szandałaâ, "Review and Comparison of Commonly Used Activation Functions for Deep Neural Networks," in *Bio-inspired Neurocomputing*. Springer, Singapore, 2021, vol. 903, pp. 203–224.

[86] F. Chollet and others, "Keras," *Github*, 2015. [Online]. Available: https://github.com/fchollet/keras

[87] J. Vanderplas, P. Prettenhofer, l. Mans, G. Louppe, O. Grisel, and T. Fan, "Scikit learn - train_test_split," 2011, accessed on Dec. 15, 2023. [Online]. Available: https://github.com/scikit-learn/scikit-learn/blob/3f89022fa/sklearn/model_selection/_split.py#L2532

[88] A. Ibrahim, A. Eltawil, A. Eltawil, Y. Na, and S. El-Tawil, "Accuracy limits of embedded smart device accelerometer sensors," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 8, pp. 5488–5496, Aug. 2020.

[89] M. Šipoš, P. Pačes, J. Roháč, and P. Nováček, "Analyses of triaxial accelerometer calibration algorithms," *IEEE Sensors Journal*, vol. 12, no. 5, pp. 1157–1165, 2012.

[90] S. Lambrecht, S. L. Nogueira, M. Bortole, A. A. Siqueira, M. H. Terra, E. Rocon, and J. L. Pons, "Inertial sensor error reduction through calibration and sensor

fusion," *Sensors (Switzerland)*, vol. 16, no. 2, Feb. 2016.

[91] A. Nez, L. Fradet, P. Laguillaumie, T. Monnet, and P. Lacouture, "Comparison of calibration methods for accelerometers used in human motion analysis," *Medical Engineering and Physics*, vol. 38, no. 11, pp. 1289–1299, Nov. 2016.

[92] H. Chen, P. Aggarwal, T. M. Taha, and V. P. Chodavarapu, "Improving Inertial Sensor by Reducing Errors using
}Deep Learning Methodology," in *NAECON 2018 - IEEE National Aerospace and Electronics Conference.*, 2018, pp. 197–202.

[93] G. Lenis, N. Pilia, A. Loewe, W. H. Schulze, and O. Dössel, "Comparison of Baseline Wander Removal Techniques considering the Preservation of ST Changes in the Ischemic ECG: A Simulation Study," *Computational and Mathematical Methods in Medicine*, vol. 2017, 2017.

[94] H. Li and P. Boulanger, "An automatic method to reduce baseline wander and motion artifacts on ambulatory electrocardiogram signals," *Sensors*, vol. 21, no. 24, Dec. 2021.

[95] F. P. Romero, D. C. Piñol, and C. R. V. Seisdedos, "DeepFilter: an ECG baseline wander removal filter using deep learning techniques," *Biomedical Signal Processing and Control*, vol. 70, Jan. 2021. [Online]. Available: http://arxiv.org/abs/2101.03423

[96] J. Fridolfsson, M. Börjesson, and D. Arvidsson, "A biomechanical re-examination of physical activity measurement with accelerometers," *Sensors (Switzerland)*, vol. 18, no. 10, Oct. 2018.

[97] Q. Ni, Z. Fan, L. Zhang, B. Zhang, X. Zheng, and Y. Zhang, "Daily Activity Recognition and Tremor Quantification from Accelerometer Data for Patients with Essential Tremor Using Stacked Denoising Autoencoders," *International Journal of Computational Intelligence Systems*, vol. 15, no. 1, Dec. 2022.

[98] W. Pan, S. Kwak, F. Li, C. Wu, Y. Chen, Y. Yamamoto, and D. Cai, "Actigraphy monitoring of symptoms in patients with Parkinson's disease," *Physiology and Behavior*, vol. 119, pp. 156–160, Jul. 2013.

[99] D. W. Kim, L. M. Hassett, V. Nguy, and N. E. Allen, "A Comparison of Activity Monitor Data from Devices Worn on the Wrist and the Waist in People with Parkinson's Disease," *Movement Disorders Clinical Practice*, vol. 6, no. 8, pp. 693–699, Nov. 2019.

[100] A. J. Silman, G. J. Macfarlane, and T. Macfarlane, "Feasibility and pilot studies," in *Epidemiological Studies: A Practical Guide*. Oxford University Press, Oct. 2018, pp. 136–142.

[101] S. Łuczak, "Experimental studies of hysteresis in MEMS accelerometers: A commentary," *IEEE Sensors Journal*, vol. 15, no. 6, pp. 3492–3499, Jun. 2015.

[102] J. Oleszkiewicz, L. Xiao, and Y. Liu, "Parallel Network RAM: Effectively Utilizing Global Cluster Memory for Large Data-Intensive Parallel Programs," in *Proceedings of the 2004 International Conference on Parallel Processing (ICPP'04)*, 2004.

[103] M. D. Hssayeni, J. Jimenez-Shahed, M. A. Burack, and B. Ghoraani, "Wearable sensors for estimation of parkinsonian tremor severity during free body movements," *Sensors (Switzerland)*, vol. 19, no. 19, Oct. 2019.

[104] T. Gutowski and M. Chmielewski, "An Algorithmic Approach for Quantitative Evaluation of Parkinson's Disease Symptoms and Medical Treatment Utilizing Wearables and Multi-Criteria Symptoms Assessment," *IEEE Access*, vol. 9, pp. 24 133–24 144, 2021.

[105] P. Prakash, "Torch Dataset and Dataloader â Early Loading of Data," Sep. 2017.

[106] J. T. VanderPlas, "Understanding the Lomb-Scargle Periodogram," *The Astrophysical Journal Supplement Series*, Mar. 2017. [Online]. Available: http://arxiv.org/abs/1703.09824http://dx.doi.org/10.3847/1538-4365/aab766

[107] S. Yue Xu, S. Nelson, J. Kerr, S. Godbole, R. Patterson, G. Merchant, I. Abramson, J. Staudenmayer, and L. Natarajan, "Statistical approaches to account for missing values in accelerometer data: Applications to modeling physical activity," *Statistical Methods in Medical Research*, vol. 27, no. 4, pp. 1168–1186, Apr. 2018.

[108] N. M. Butera, S. Li, K. R. Evenson, C. Di, D. M. Buchner, M. J. LaMonte, A. Z. LaCroix, and A. Herring, "Hot Deck Multiple Imputation for Handling Missing Accelerometer Data," *Statistics in Biosciences*, vol. 11, no. 2, pp. 422–448, Jul. 2019.

[109] S. Chakrabarti, N. Biswas, K. Karnani, V. Padul, L. D. Jones, S. Kesari, and S. Ashili, "Binned Data Provide Better Imputation of Missing Time Series Data from Wearables," *Sensors*, vol. 23, no. 3, Feb. 2023.