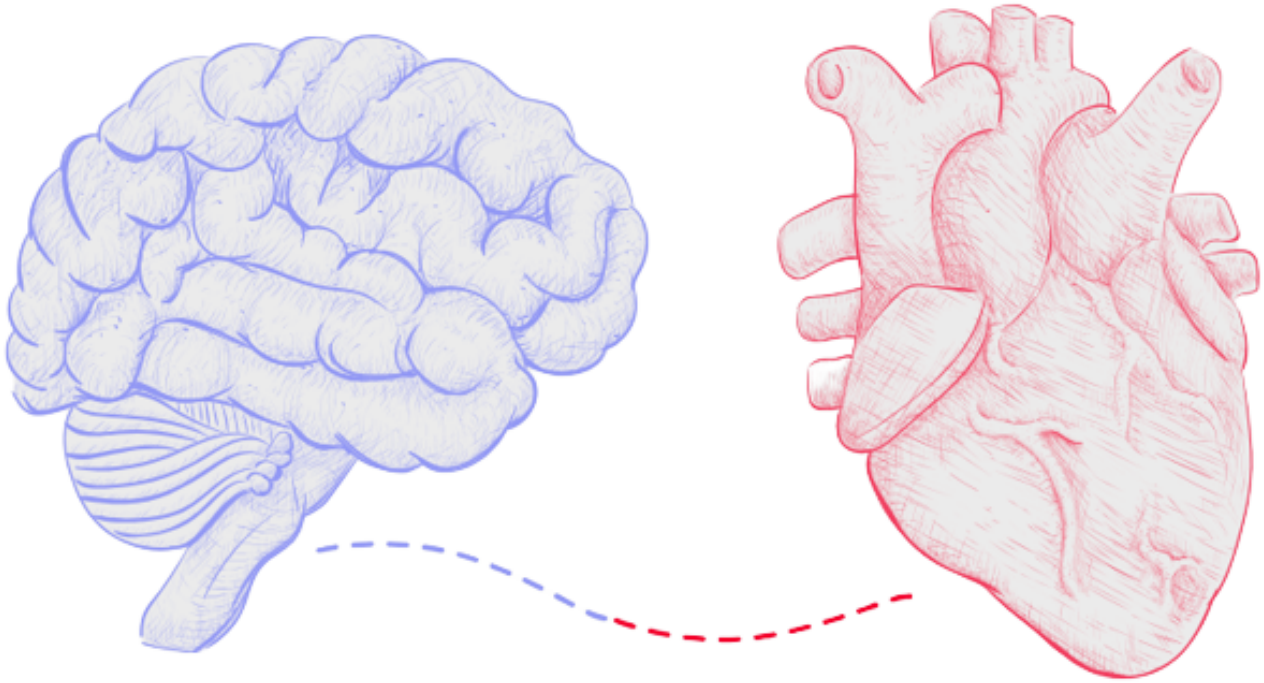
# Towards digital biomarkers to assess autonomic dysfunction in Parkinson's Disease:

## Feasibility of home-based photoplethysmographic signals for heart rate analysis

**K.I. Veldkamp**



UNIVERSITY OF TWENTE.

Radboudumc

*A thesis submitted for the degree of Master of Science*

---

# Towards digital biomarkers to assess autonomic dysfunction in Parkinson's Disease: Feasibility of home-based photoplethysmographic signals for heart rate analysis

---

**Kars Ingmar Veldkamp**
Nijmegen - January 24, 2023

**Technical Medicine**
Medical Sensing & Stimulation

Graduation committee:

| | |
|---|---|
| **Prof. dr. H.J. Zwart** | **Chairman & Technical supervisor** |
| *Professor Mathematics of Systems Theory, University of Twente* | |
| **Dr. M.A. Brouwer** | **Medical supervisor** |
| *Cardiologist, Radboudumc* | |
| **Dr. J. Thannhauser** | **Daily supervisor** |
| *Technical Physician & postdoctoral researcher, Radboudumc* | |
| **E.M. Walter, MSc** | **Process supervisor** |
| *Lecturer professional behavior, University of Twente* | |
| **M.P. Mulder, MSc** | **External Member** |
| *PhD candidate Cardiovascular and Respiratory Physiology, University of Twente* | |

Additional supervisor:

| | |
|---|---|
| **L.J.W. Evers** | **Daily supervisor** |
| *PhD candidate, Donders Institute for Brain, Cognition and Behaviour* | |

# Preface

Gedurende mijn gehele studie heb ik altijd een fascinatie gehouden voor zowel het hart als het brein. Tijdens mijn tweede M2 stage, juni 2020, kwam ik eigenlijk bij toeval terecht op de cardiologie in het Radboudumc in Nijmegen. Hoewel het reanimatieproject mijn eerste keuze was, werd ik eigenlijk geplaatst op de cardiologie in het UMC Utrecht. Echter zaten we midden in coronatijd en werden meerdere stageplekken geannuleerd waaronder de mijne. Het stagebureau heeft toen actief gepolst of het mogelijk was om mij als extra student toch op mijn eerste voorkeur te plaatsen. Toeval of niet, i.p.v. te werken aan het reanimatieproject, mocht ik als eerste student bezig gaan met een nieuw opgezette samenwerking tussen de cardiologie en neurologie met het doel om de cardiale kant van Parkinson beter te begrijpen. Ondanks alle restricties vanwege corona heb ik de toegankelijke en ontspannen sfeer als heel prettig ervaren. Deze ervaringen en het gave project zorgden ervoor dat ik hier heel graag mijn M3 wilde gaan doen. Nu aan het eind van mijn M3 kijk ik tevreden terug op een pittig, leerzaam en leuk jaar. Dit was echter niet mogelijk zonder de hulp van een aantal mensen.

Jos, als eerste wil ik jou bedanken. Je hebt wat met mij van doen gehad tijdens twee M2 stages (cardio en psychiatrie) en een M3. Wat ik altijd heel erg gewaardeerd heb is jouw enthousiasme en jouw gave om zaken te relativeren. Ik voelde mij altijd op mijn gemak tijdens onze gesprekken en heb ontzettend veel van jou kunnen leren. Ik vind het knap hoe je in je begeleidersrol mij heel erg een gelijkwaardig gevoel hebt gegeven maar tegelijkertijd op de juiste momenten ook de regie wist te pakken om mij tijdig bij te sturen als ik weer eens verdwaalde in een van mijn vele zijpaden. Heel veel dank prettige samenwerking en de intensieve begeleiding.

Marc, bedankt voor alle leerzame momenten en nieuwe inzichten. Ik kan mij onze eerste kennismaking, "de full experience met Marc in de gouden mini", onderweg naar het CWZ nog goed voor de geest halen. Tijdens de poli's op het hartcentrum, wanneer weer eens (heel irritant) al je voorspellingen uitkwamen, of tijdens de poli's op het Radboud, elke keer had je wel weer nieuwe manieren om mij aan het denken te zetten. Niet alleen op klinisch vlak heb ik veel van je geleerd, maar ook tijdens onze vele persoonlijke gesprekken waaronder die op het dak van het CWZ, buiten bij het Radboud en tijdens ons laatste etentje.

Luc, bedankt voor jouw "onofficiële" rol als dagelijks begeleider tijdens mijn afstuderen. Ik heb de manier waarop jij mij begeleid hebt als heel prettig en gelijkwaardig ervaren. Tijdens onze Parkinson-meetings kwam je altijd met goede inzichten en aanvullende ideeën. Elke keer maakte je veel tijd voor mij vrij, helemaal in de periode dat Jos op vakantie was. Dank dat ik je zowel in een formele als informele setting heb mogen leren kennen.

Hans, bedankt voor jouw rol als technisch begeleider en voorzitter van mijn commissie. Vanuit het vak BCS in de master kon ik je betrokken manier van begeleiden al als heel prettig ervaren wat zeker voortgezet werd tijdens mijn M3. Vooral onze gesprekken heb ik altijd als heel prettig ervaren, waarin je telkens uitgebreid de tijd nam en met jouw technische blik ook veel andere inzichten kon toevoegen.

Elyse, bedankt voor alle procesbegeleiding. Vanaf het eerste moment voelde ik mij prettig bij je om mijn "issues" te bespreken. Ik kon altijd bij je terecht en hebt echt bijgedragen om mij tot nieuwe inzichten te laten komen die mij verder zullen helpen om mij telkens te blijven ontwikkelen. Aan al mijn intervisiegenoten, dank voor alle ondersteuning en de leuke en leerzame gesprekken.

Marijn, bedankt voor het deelnemen in mijn afstudeercommissie als buitenlid.

Als laatste wil ik alle vrienden, familie en collega's bedanken. Zonder al jullie support was dit niet mogelijk geweest.

Nijmegen, 23 januari 2023
Kars Veldkamp

# Abstract

***Introduction:*** Recent evidence suggests autonomic dysfunction plays an important role in the pathogenesis of Parkinson's disease (PD). Only a few (subjective) clinical tests assess disease progression and are used to evaluate the entire spectrum of autonomic dysregulation in patients with PD. Long-term heart rate (HR) monitoring by wrist-worn photoplethysmography (PPG) sensors could enable patient-tailored follow-up in PD patients. We aimed to develop models that could assess data quality, analyzed whether several factors could influence the data quality and performed an exploratory HR analysis in relation to the symptoms of autonomic dysfunction.

***Methods:*** From an ongoing, prospective PD cohort study, we included all patients with a complete one-year follow up and analyzable photoplethysmography (PPG). We divided the PPG data into one-minute epochs and annotated epochs of a stratified subset of PD patients into different categories of signal quality. We extracted 18 time and frequency characteristics from the epochs and optimized (in terms of balanced accuracy) machine learning classifiers to assess data quality for two analysis of HR patterns: HR analysis and heart rate variability (HRV) analysis. We analyzed differences in the proportion of eligible data over time based on longitudinal assessment over three different weeks (Week 0 – Week 1 – Week 52). The influence of PD-specific motor symptoms was studied using tertiles based on the UPDRS symptom scores. Resting HR and maximum HR, obtained from eligible data in Week 0, were compared in relation to the gold standard in scoring non-motor symptoms, the UPDRS Part 1b.

***Findings:*** In total, PPG data of 20 PD patients were used to annotate, train, test and validate the two models for HR analysis and HRV analysis. We obtained balanced accuracies $> 93\%$ for HR analysis and $> 94\%$ for HRV analysis. We used PPG data of 431 subjects for longitudinal analysis and PPG data of 484 subjects for analyzing PD-specific factors. We did not find significant differences between time and PD-specific factors. Using 486 subjects, we showed an increase in resting HR during the day and night with increasing UPDRS Part 1b scores. The maximum HR decreased during the day whereas it increased during the night.

***Implications:*** This thesis showed that we could develop high-discriminative models to assess PPG data quality. In the exploratory HR analysis, using one of these models, we showed promising first results in linking HR parameters to the severity of autonomic dysfunction. This research should be further elaborated by focusing on longitudinal HR and HRV analysis during the night to enable patient-tailored follow-up.

# Contents

# List of abbreviations

| | |
|---|---|
| $^{18}$F-FDOPA | $^{18}$F-fluorodopamine |
| AF | atrial fibrillation |
| ANS | autonomic nervous system |
| AUC | area under curve |
| BPM | beats per minute |
| CNS | central nervous system |
| CV | cross-validation |
| ECG | electrocardiography |
| FN | false negatives |
| FP | false positives |
| GUI | graphical user interface |
| HR | heart rate |
| HRV | heart rate variability |
| IQR | interquartile ranges |
| LBFGS | limited Broyden–Fletcher–Goldfarb–Shanno algorithm |
| LR | logistic regression |
| ML | machine learning |
| MRMR | maximum relevance - minimum redundancy |
| NN | neural networks |

| | |
|---|---|
| PAC | premature atrial contractions |
| PAT | pulse arrival time |
| PD | Parkinson's disease |
| PPG | photoplethysmography |
| PPP | personalized parkinson project |
| PPV | positive predictive value |
| PVC | premature ventricular contractions |
| RBD | REM sleep behavior disorder |
| RBF | radial basis function |
| SGD | stochastic gradient descent |
| SpaRSA | Sparse Reconstruction by Seperable Approximation |
| SVM | support vector machines |
| TN | true negatives |
| TP | true positives |
| UPDRS | Unified Parkinson's Disease Rating Scale |

# List of Figures

# List of Tables

# 1 | Introduction

Parkinson's disease (PD) is the second most common degenerative brain disease, with a current prevalence of about 10 million patients worldwide [1]. The number of PD cases is expected to double in 2040 [2]. PD is commonly known for its classical 'motor' symptoms (e.g. tremor, bradykinesia, rigidity) which are the result of neurodegeneration in the central nervous system (CNS) [3, 4]. However, emerging evidence suggests that neurodegeneration also occurs in the autonomic nervous system (ANS) [5]. This so-called "autonomic dysfunction" can result in various non-motor symptoms such as constipation, urinary dysfunction and/or postural hypotension, and have a high impact on the quality of life [6].

PD symptoms have a high heterogeneity in disease onset and progression, posing major clinical challenges in terms of diagnosis and treatment. Hence, PD treatment importantly relies on adequate assessment of symptoms [7]. Currently, the gold standard for quantification of non-motor symptoms of PD is Unified Parkinson's Disease Rating Scale (UPDRS) [8]. However, the UPDRS has been regarded very time-dependent, subjective to symptom interpretation, by both patient and physician, and subjective to daily swings [9]. To enable patient-tailored long-term follow up and treatment of non-motor symptoms in PD, there is a need for new, objective measures to assess autonomic dysfunction.

Heart rate (HR) patterns, such as HR intervals, reflect autonomic functioning and could be a potential method to detect autonomic dysfunction [5, 10, 11]. Photoplethysmography (PPG) can detect HR intervals and is becoming increasingly popular in wrist-worn devices due to its technological and practical advantages [12, 13, 14]. A major challenge for using PPG in an ambulatory setting is the sensitivity of PPG-signals to motion artifacts [13, 14]. It is currently unkown whether accurate PPG-based HR estimation is feasible in PD patients with motor signs such as tremor. Although artifact suppression techniques are useful when the interference of such an artifact on the PPG signal is not fatal, alternative strategies are desired when the interference cause a serious-disturbed signal [15, 16, 17, 18]. Such an alternative strategy may be to develop models to distinguish signals eligible for heart rate analysis of interest from signals that are of insufficient quality. However, such models are lacking for PD populations. Moreover, it is unknown whether PPG is a reliable method to assess heart rates over a longer period in PD patients, and whether there are differences between patients with different levels of motor symptoms.

In the recent initiated Personalized Parkinson Project (PPP), a wrist-worn device collects physiological data from PD subjects continuously for a period of two years, including PPG. The PPP has been initiated to study established and novel biomarkers in relation to disease progression [19]. The amount of PPG data collected in the PPP allows for training and testing of complex algorithms, such as machine

learning (ML) models, to assess data quality in the quest towards detecting autonomic dysfunction in PD patients.

In this thesis, we aimed to develop supervised machine learning (ML) models to distinguish high-quality from low quality wrist-worn PPG data of PD patients in a home-based situation. Secondly, using these ML-models, we aimed to compare changes in the amount and the quality of the PPG data over the course of the PPP study period. Thirdly, we cross-sectionally assessed differences in the proportion of high-quality data between patients with and without specific PD motor symptoms. Lastly, we performed an exploratory heart rate analysis, in which we analyzed heart rate parameters (resting HR and maximum HR) in relation to the symptoms of autonomic dysfunction in PD patients.

## 1.1 Outline of thesis

In **Chapter 2** the thesis will continue with clinical background information on Parkinson's disease, the role of the autonomic nervous system in Parkinson's disease and the initiated PPP. **Chapter 3** consists of technical background information on PPG and the optimization of ML classifiers used in this study, namely logistic regression, support vector machine and feedforward neural networks. In **Chapter 4** we develop several supervised ML models to distinguish PPG-signals with high quality from signals with low quality. In **Chapter 5**, we evaluate the influence of time and clinical factors on the amount of eligible data for analysis in PD patients. As a last step in this chapter, we perform an exploratory HR analysis, in which we assess HR patterns in relation to the symptoms of autonomic dysfunction in PD patients. **Chapter 6** provides an overall discussion of the thesis.

# 2 | Clinical background

## 2.1 Parkinson's disease

### 2.1.1 Different phenotypes in Parkinson's disease

For decades, PD was thought to be a primary brain disorder characterized by loss of pigmented dopaminergic neurons residing in the substantia nigra [20, 21]. However, more recently, PD is used as an umbrella term for a group of underlying disorders which is highly heterogeneous and probably consists of several subtypes [6, 22]. In PD patients, neurodegeneration not only takes place in the CNS but also in the peripheral ANS, responsible for preserving homeostasis and serves as communication between the CNS and the peripheral tissue [5]. An emerging hypothesis poses that patients with PD could be globally classified into two groups according to disease progression [5, 23]. In the body-first phenotype, the ANS is damaged preceding measurable damage to higher Braak stage structures, including the substantia nigra. In the brain-first phenotype damage to the brain precedes measurable damage in the periphery [23]. Especially in the body-first phenotype there are multiple symptoms related to the ANS before the diagnosis of PD. A common cause of these symptoms is the deposition of of $\alpha$-synuclein aggregates in the ANS [22, 24, 25].

### 2.1.2 Cardiac autonomic dysfunction

The ANS is responsible for innervating every organ in the body, but deposition of $\alpha$-synuclein aggregates is particularly common in the cardiovascular system. Especially the heart, due to the high density of sympathetic innervation, is vulnerable to these depositions. Recently, researchers studied the effect of cardiac autonomic dysfunction in PD patients with $^{18}$F-fluorodopamine ($^{18}$F-FDOPA) positron emission tomography (PET) scanning (Figure 2.1 [26]. This study showed that over the duration of the disease, neurodegeneration reduces innervation of the heart. In this case, cardiac sympathetic denervation can be seen as a late finding.

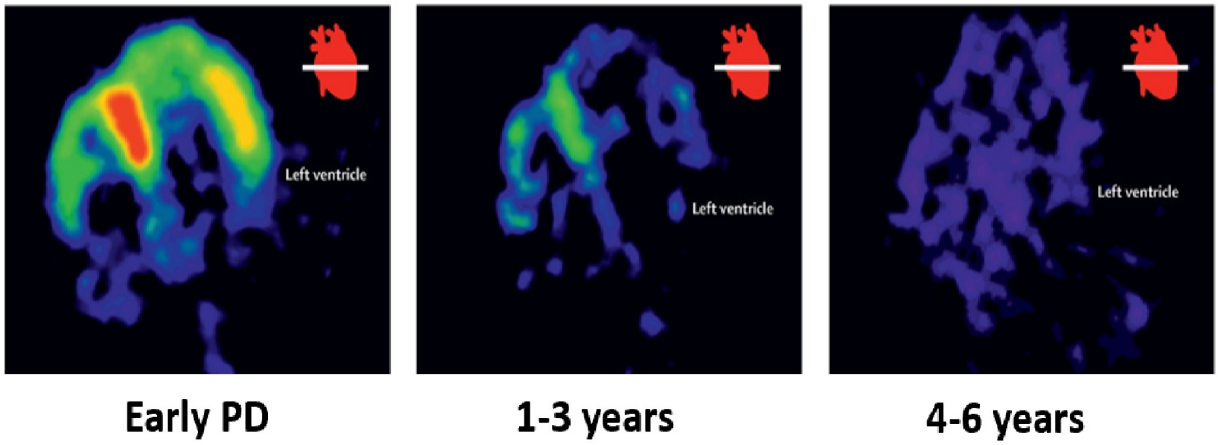



Figure 2.1: Cardiac sympathetic innervation in patients with Parkinson's disease using $^{18}$F-fluorodopamine ($^{18}$F-FDOPA) positron emission tomography scanning. Images are transverse sections of the cardiac ventricles of a patient with early PD (a), after 1-3 years (b), and after 4-6 years (c) after diagnosis. A more intense color represents more cardiac sympathetic innervation. Figure obtained from Goldstein and Sharabi [26].

Borghammer et al., studied cardiac innervation in three groups: healthy subjects, subjects with REM sleep behavior disorder (RBD; an important precursor of the PD body-first phenotype) and subjects with early PD and no RBD [23]. The first group was a healthy control group. They showed that in subjects with RBD (body-first), the dopamine storage in the substantia nigra is comparable to the healthy subject while these subjects show less cardiac innervation (Figure 2.2) These findings support the theory of the two main phenotypes. Furthermore, a case study of Goldstein et al. showed that cardiac sympathetic denervation can precede years before the diagnosis of Parkinson's disease [27].

This evidence suggests that the cardiovascular system could serve as an ideal window into the pathogenesis of PD. However, as stated in the introduction, the heterogeneous onset of the disease makes it difficult to evaluate disease progression based on infrequent measures such as a PET scan. Furthermore, scoring non-motor symptoms, using MDS UPDRS Part I: Non-Motor Aspects of Experiences of Daily Living, is subjective and has a high variability between days. Therefore, there is a urgent need to study the cardiovascular system in PD patients in daily-living conditions to obtain a reliable assessment of the disease progression.

Heart rate variability (HRV) is a key indicator of an individual's autonomic function and is generated by heart-brain interactions and processes of the ANS [13, 28]. Whereas heart rate refers to the quantification of heart beats per minute, HRV refers to the fluctuations between two successive heartbeats. Alterations in HRV may reflect pathological involvement of components of the ANS, including loss of sympathetic innervation[29]. Several studies showed associations between HRV fluctuations and having PD [30, 31]. These studies use the electrocardiogram (ECG). However to obtain insights in the long-term pathogenesis of the disease in an ambulatory setting, more unobtrusive methods are desired. To provide in this task, PPG could serve as an ideal tool [18, 12].

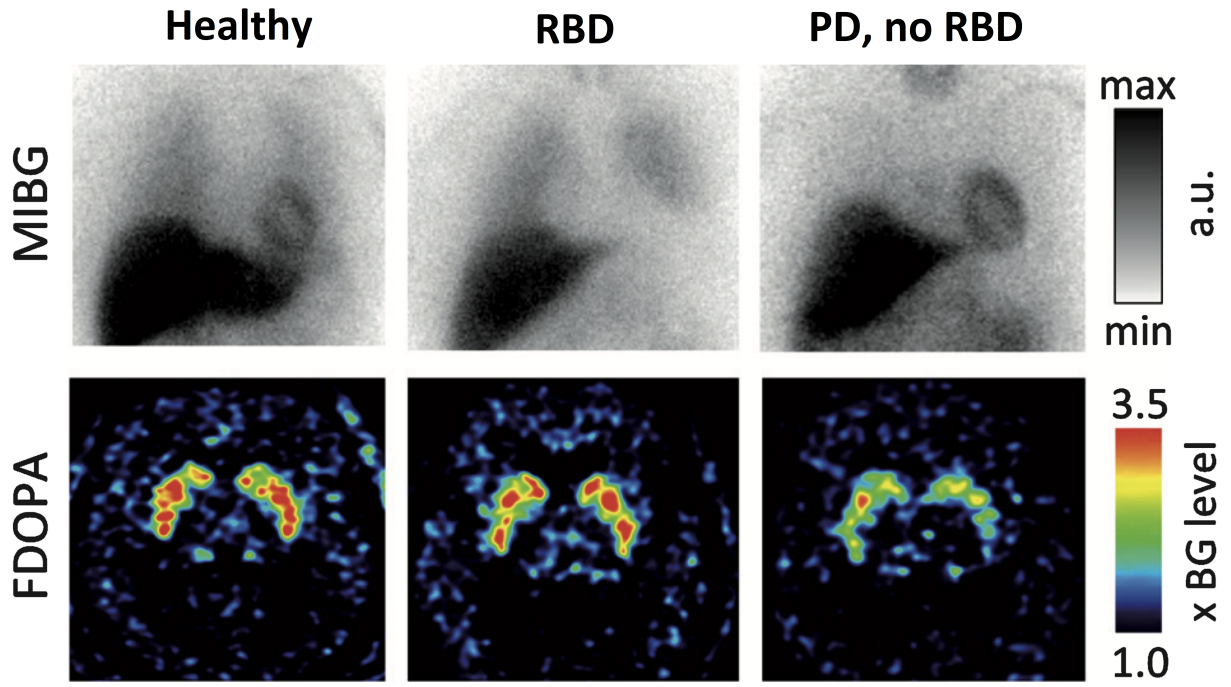



Figure 2.2: Cardiac sympathetic innervation, imaged with MIBG scintigraphy, and the dopamine storage in the substantia nigra for three different study groups. REM sleep behavior disorder (RBD) is a frequently seen precursor in PD patients in the body-first phenotype. The RBD patient shows reduced cardiac innervation but normal dopaminergic storage capacity. In contrast, in early PD patients without RBD, cardiac innervation is normal but there is a significant loss in dopaminergic storage capacity. Figure modified from Borghammer et al. [23].

## 2.2 Personalized Parkinson Project

As mentioned in Chapter 1, the PPP has been initiated to carry out an unbiased approach to biomarker development with multiple potential biomarkers measured longitudinally. The primary aims of the PPP are: (1) to perform analyses on the comprehensive data set, correlating established and novel biomarkers to the rate of disease progression and to treatment response; and (2) to create a widely accessible dataset for discovery of novel biomarkers and new targets for therapeutic interventions in PD.

In this project, a prospective, longitudinal cohort involving approximately 500 PD patients is continuously monitored by a multi-sensor investigational research device (Verily Study Watch) over the follow-up of a 2-year period. The Verily Study Watch enables continuous data collection of physiological parameters, such as photoplethysmography (PPG), single-lead ECG, acceleration/orientation, electrodermal activity and skin temperature. During the baseline visit, the assessor explains how the Study Watch works and emphasizes the importance of ambulatory monitoring. Data collection is intended to be as passive as possible, providing a minimum amount of

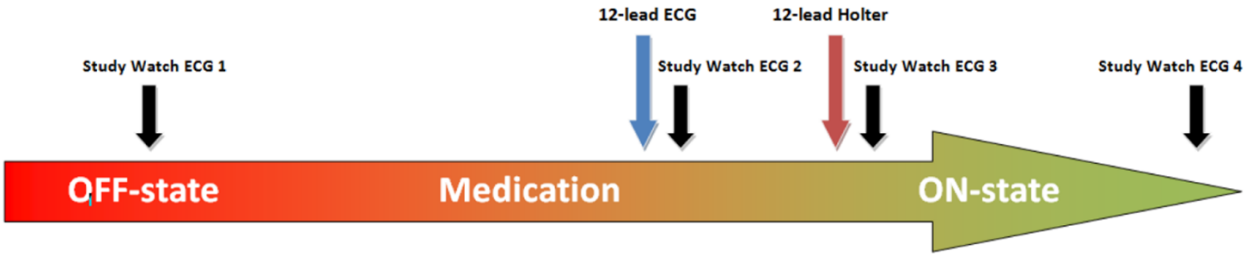



Figure 2.3: Overview of the different ECG measurements during an annual assessment. At 4 different time stamps, a single-lead study watch ECG is made. After medication, a standard 12-lead ECG is recorded and a 12 lead-holter ECG is attached.

information to the patient via the device. On a weekly basis, the encrypted data is securely sent to the Verily cloud, after which the data cannot be led back to patient-specific information.

During the follow-up period, participants visit the study site for 1) a baseline assessment, 2) one-year follow-up and 3) two-year follow up. Patients start these visits in the OFF state, which means that their last dopaminergic PD medication was taken at least 12h before the assessment. In the OFF state, the patients undergo several motor functioning, neuropsychological and other PD related tests. Directly after taking the regular first morning dose of medication(s), a standard 12-lead ECG is recorded and thereafter a 12-lead holter is attached. After subjectively reaching the typical ON-state, the above mentioned PD-tests are repeated.

During the day, four one-minute ECGs are recorded with the Verily Study Watch: one in the OFF-state at the start of the day, one simultaneously with the 12-lead ECG ($\pm$15 min after medication) and two simultaneously with the Holter ECG ($\pm$ 30 min after medication and one in the typical ON state). A schematic overview is given in Figure 2.3.

After each visit participants complete a set of validated questionnaires at home, including questionnaires about medication use, quality of life, lifestyle, neuropsychological symptoms, autonomic symptoms, sleep, and vision. These questionnaires are completed within 4 weeks after each visit, via an online survey module.

# 3 | Technical background

## 3.1 Photoplethysmography

Photoplethysmography (PPG) is a simple and low-cost alternative in performing HRV analysis. Like ECG, the PPG waveform provides information on the heart rate. However, the two measurement techniques are completely different. While ECG uses the electrical signal conduction of the heart, PPG indirectly measures the heart rate by measuring the absorption or reflection of infrared light by the pulsatile blood flow within the vessels [32]. Since light interacts with biological tissue it can be transmitted, reflected, refracted, scattered and/or absorbed. Tissue, arterial and venous blood are the primary absorbers of light with haemoglobin being one of the main components absorbing light passing through tissue. As the heart pumps blood to the periphery, the rapid increase in blood flow volume in the arteries attenuates the light source of the PPG device, allowing it to detect changes in blood volume during the cardiac cycle [33]. In Figure 3.1 the systolic and diastolic peaks of the PPG signal indicates the contraction and relaxation at the measurement site, while the dicrotic notch (only seen with high sample frequency) represents the aortic valve closure, which indicates the end of the systole of the heart. The Pulse Arrival Time (PAT) is the time taken for the pulse from the heart to travel to the PPG measurement site.

ECG is the gold standard for analyzing heart rate patterns, such as heart rate measurement and heart rate variability (HRV). However, PPG offers an expedient alternative that is suitable for ambulant monitoring of healthy subjects but also for various patients with particular cardiovascular diseases [34]. In Figure 3.2 it is seen that the PPG waveform comprises pulsatile (AC) and non-pulsatile (DC) components. The AC components provides information about the volumetric changes in blood vessels, which corresponds to the heart rate. The DC component measures light absorbed by the tissue, veins and blood at the measurement site. Slower changes in venous capacity, due to respiration are also captured by the DC component [35].

Various studies are performed using ECG and/or PPG in cardiovascular and PD research. However, there is some disconcordance in results of those studies. A couple of studies state that one should be careful while interpreting PPG signals in non-healthy patients for analyzing heart rate patterns [36, 13]. However, more recent studies showed promising results when using PPG as alternative for ECG in heart rate pattern analyses [33, 37].

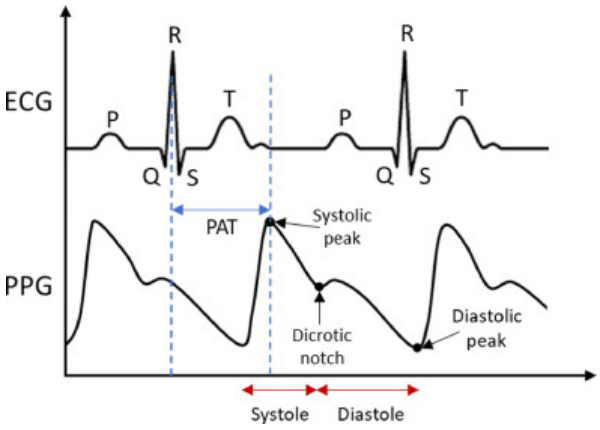



Figure 3.1: Schematic drawing of the PPG and ECG waveforms [33].

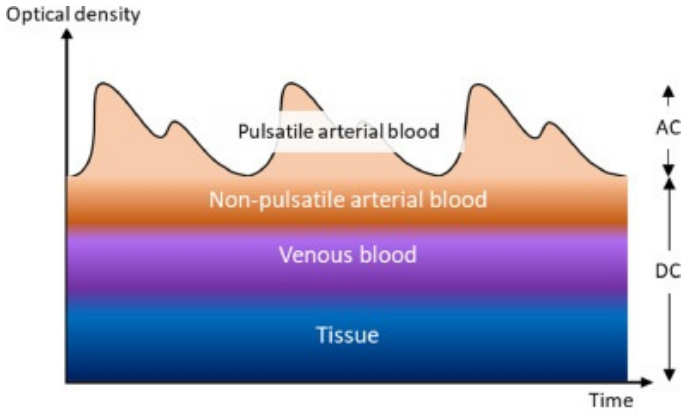



Figure 3.2: Schematic drawing of the AC and DC components of PPG signal [33].

## 3.2 Machine Learning

Significant progress in both statistical modeling techniques as well in computer power over the last decades have enabled the rapid rise of the field of data science, including artificial intelligence (AI) and machine learning (ML). Where AI is described as a much broader term with the goal to emulate human, "wide" intelligence with the ability to solve a range of different complex tasks with one algorithm, ML can be viewed as a form of "narrow" AI (focused on just one specific task). ML deals with learning problems by iteratively learning from experience (in the form of data). The algorithms used in ML can be viewed as mathematical functions that attempt to map an input feature vector $\mathbf{x}$ to a desired output $\mathbf{y}$, and experience can be stated as the iterative process to improve on the performance of a specific task by optimizing parameters within the mathematical function $f(\mathbf{x})$.

The two most widely used ML methods are either supervised or unsupervised learning methods. Furthermore, semi-supervised learning and reinforcement learning are known methods for ML. The method of learning depends on the available data and the type of the problem. In supervised learning, a set of input features $\mathbf{x}$ are used for training, e.g. meaningful variables such as clinical characteristics or signal characteristics to predict a known target variable $\mathbf{y}$. In unsupervised learning, only the input variables are available. Unsupervised learning refers to the analysis of a dataset without knowing a priori what should be learned. In the unsupervised method, the model aims to find structural coherence in the data, which can be leveraged in clustering.

### 3.2.1 Principles of ML

To reduce the dimensionality of the input data, various supervised ML learning approaches start with extracting features from the data. The goal of this process is to minimize the dimensionality while preserving all important aspects of the original data. Within supervised learning, a model can learn the pattern by linking the features to the target variable. This enables the model to use this pattern in the prediction of new data.

A multitude of ML algorithms exist, and choosing the "right" classifier for a specific task can be quite hard. Moreover, in ML there is no such a thing as a "free lunch", meaning that no algorithm can generally be considered superior for all given circumstances.

### 3.2.2 Optimization

The most important concept of nearly all ML and statistical modeling techniques is optimization. Optimization is the process of iteratively adjusting model parameters to improve performances of the model. Before discussing this topic, it is relevant to introduce the basic components of a ML classifier.

#### Components of a ML classifier

In classification problems, ML algorithms optimizes their model function $f(\mathbf{x})$ by learning, from an input observation $\mathbf{x}$, a vector of model parameters (weights $\mathbf{w}$) and a bias term. Each weight $w_i$ is a real number, and is associated with one of the input features $x_i$. Each weight represents how important that input feature is to the classification decision and can be positive (providing evidence that the observation being classified belongs to the positive class) or negative (providing evidence that the observation belongs to the negative class). To make a decision on a new observation, after learning the weights in training, the classifier multiplies $x_i$ by its weight $w_i$, sums up the weighted features and adds the bias term $b$:

$$z = \left(\sum_{i=1}^{f} w_i x_i\right) + b, \text{ with } [w_i,\ x_i] \in \mathbb{R}. \tag{3.1}$$

Here $f$ represents the length of the feature vector. An equivalent notation to the weighted sum notation is the dot notation:

$$z = \boldsymbol{w} \cdot \boldsymbol{x} + b. \tag{3.2}$$

To create a probability, $z$ is passed into a model function $f(\mathbf{x})$. For example, in a logistic regression (LR) classification model this model function can be defined as a sigmoidal function:

$$f_w(X) = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-(\boldsymbol{w}^T \boldsymbol{x} + b)}}. \tag{3.3}$$

Let's consider a binary classification problem (e.g. sick vs. not-sick) and consider a person being sick as the positive class. The return of the model function, in case of the sigmoidal function, is a number between 0 and 1. To make it a probability, all cases, $p$(sick=1) and $p$(sick=0) sum to 1. The probabilities of an observation $\mathbf{x}$ for a person belonging to each of the cases, parametrized by $\mathbf{w}$, are equal to:

$$\begin{aligned} P(sick = 1 \mid \boldsymbol{x}; \boldsymbol{w}) &= f_w(\boldsymbol{x}), \\ P(sick = 0 \mid \boldsymbol{x}; \boldsymbol{w}) &= 1 - f_w(\boldsymbol{x}). \end{aligned} \tag{3.4}$$

#### Processing multiple inputs

When training a ML algorithm, it's inefficient to optimize the function with one observation individually. Instead, we want to process an entire set with many examples. An efficient method to assign a class

to multiple inputs at the same time is to make use of a single matrix operation. The matrix $\mathbf{X}$ can be created by packing all input feature vectors where each row represents a single observation $\mathbf{x}$. Assuming each example has $f$ features and weights, and there are $n$ number of samples, the matrix $\mathbf{X}$ will be of size $[n \times f]$:

$$\boldsymbol{X} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & ... & x_f^{(1)} \\ x_1^{(2)} & x_2^{(2)} & ... & x_f^{(2)} \\ ... & ... & ... & ... \\ x_1^{(n)} & x_2^{(n)} & ... & x_f^{(n)} \end{bmatrix}. \tag{3.5}$$

We can introduce $\mathbf{b}$ as a repeated bias term column vector of length $n$, $\mathbf{b}$ = [b,b,b,..,b], and $\mathbf{y} = [y^{(1)}, y^{(2)}, \ldots, y^{(n)}]$ as the output vector (one scalar for each input vector). The weight vector $\mathbf{w}$ is equal for each input vector and is represented as a column vector. The matrix multiplication will be as follows:

$$\boldsymbol{y} = (\boldsymbol{X}\boldsymbol{w})^T + \boldsymbol{b}. \tag{3.6}$$

**Loss function**

The optimization process starts by randomly initializing $\mathbf{w}$. The values of the weights for a given classifier are then optimized by evaluating a so-called objective function. Commonly, these objective functions are called loss functions which describe the deviation of the predicted probabilities from the actual labels. In ML, the aim is to find the global minimum of the loss function.

There are various loss functions, but when using a probabilistic classifier in probabilistic (binary) classification algorithms the cross-entropy loss function is the most frequently used. Each output probability ($\hat{y}$=$f_w(\boldsymbol{x})$) is compared to sickness ($y$, which is 1 or 0) and a loss is calculated based on the distance between the probability and the true class label.

We can derive the cross-entropy loss function for a single observation $\mathbf{x}$. The goal is to learn weights that maximize the probability of the correct label $p(y|x)$. Since in this example, we have two discrete outcomes (1 or 0), we can express the probability for one observation as the following:

$$p(y \mid \boldsymbol{x}; \boldsymbol{w}) = \hat{y}^y (1 - \hat{y})^{1-y}, \text{ for } y \in \{0, 1\}. \tag{3.7}$$

Note that the equation simplifies to:

$$\begin{aligned} p(1|\boldsymbol{x}; \boldsymbol{w}) &= \hat{y}, & \text{if } y = 1 \\ p(0|\boldsymbol{x}; \boldsymbol{w}) &= 1 - \hat{y}, & \text{if } y = 0 \end{aligned} \tag{3.8}$$

The next step is to take the log of equation (3.7). The log has several advantages and whatever values maximize a probability will also maximize the log of the probability:

$$\begin{aligned} \ln(p(y \mid \boldsymbol{x}; \boldsymbol{w})) &= \ln \left[ \hat{y}^y (1 - \hat{y})^{1-y} \right] \\ &= y * \ln(\hat{y}) * (1 - y) * \ln(1 - \hat{y}). \end{aligned} \tag{3.9}$$

Equation (3.9) is a log likelihood equation that should be maximized. To obtain a loss function, with the purpose of minimizing the function, we can simply flip the sign. The resulting cross-entropy loss function $L_{CE}$ then becomes:

$$\begin{aligned} L_{CE}(\hat{y}, y) &= -\ln(p(y \mid \boldsymbol{x}; \boldsymbol{w})) \\ &= -[y * \ln(\hat{y}) * (1 - y) * \ln(1 - \hat{y})]. \end{aligned} \tag{3.10}$$

The penalty is logarithmic in nature yielding a larger score for larger differences in probability and actual label. For multiple inputs the $L_{CE}$ becomes:

$$L_{CE}(\hat{y}, y) = -\sum_{i}^{n} \left[ y_i * \ln\left(\widehat{y_i}\right) * (1 - y_i) * \ln\left(1 - \widehat{y_i}\right) \right]. \tag{3.11}$$

The cross-entropy function decreases when the predicted labels converge towards the actual labels. Therefore, the most optimal parameter values ($\mathbf{w}$) for classification are found in the global minimum of equation (3.11). To find this global minimum, $\mathbf{w}$ is optimized in an iterative manner, using the gradient of the loss function with respect to the parameters. If the error increases, the parameter values will be adjusted in opposite direction of the previous adjustment. If the error decreases, the parameter values are further modified in this specific direction. Examples of two iterative optimization methods used in this thesis can be found in Appendix A.

### 3.2.3 Generalization and regularization

In ML, the aim is to make accurate and generalizable predictions. One of the most common sources of trouble is overfitting. Overfitting occurs when a model tries to adjust too closely to the training set, and subsequently demonstrates poor performances on the test set. The model is impressive in finding a perfect fit to the training data, but it is unable to make accurate predictions on new observations. To understand what is going on in these situations it is important to introduce the concept of the bias-variance trade-off. Bias is defined as the error term introduced by approximating highly complicated real-life problems in a much simpler model. In other words, the model is underfitting the truly more complex nature of the data. On the other hand, variance is defined as the model learning random structures in the data irrelevant to the underlying true signal. Models with high variance can hallucinate patterns that do not truly represent the reality of the data. In Figure 3.3 this is illustrated in a classification problem. If a simple linear model is applied to data in which the frontier between classes is not linear, it is unable to find the true frontier between classes (A). The model is underfitting the data. If a highly complex or flexible model with high variance (C) is applied to the same data, it will learn random non-predictive structures which are unrelated to the true frontier between classes. In ML models, we strive towards the optimal fit, which has an appropriate level of model complexity without overfitting the data (B).

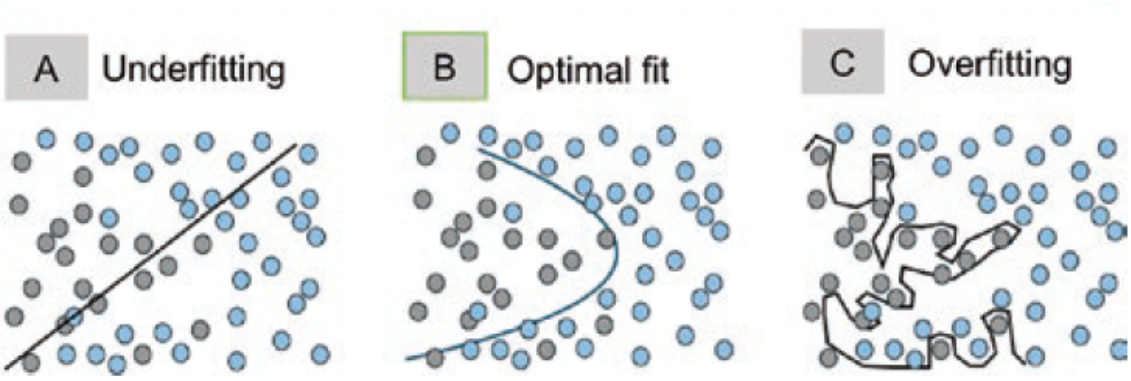



Figure 3.3: Visualization of the bias-variance trade-off. A predictive model which is underfitting has a high bias and low variance (A). The predictive model represents a optimal fit, using a hyperbolic function, and results in the lowest test error (B). In C, the model is overly flexible which leads to overfitting. It has a low bias but an extremely high variance. [38]

A possible solution for overfitting is called regularization. Model complexity can be modulated by adding a regularization term to the loss function. As doing so, the regularization term prevents the loss function to favor higher-complex model. In the case of overfitting, the weights of the model are extremely large. The regularization terms penalizes the weights based on its absolute values. Therefore, loss functions with lower weight values are favored. This reduces the variance in the model. In traditional supervised classifiers, two types of regularization terms are common: L1 penalty (also called Lasso regularization) and L2 penalty (Ridge regularization). The updated cross- entropy loss function for both regularization terms will be:

$$\begin{aligned} L_{CE}(\hat{y}, y) &= -\sum_{i}^{n} \left[ y_i * \ln\left(\widehat{y}_i\right) * \left(1 - y_i\right) * \ln\left(1 - \widehat{y}_i\right) \right] + \lambda \sum_{j}^{f} \left| w_j \right|, \text{ for Lasso regularization,} \\ L_{CE}(\hat{y}, y) &= -\sum_{i}^{n} \left[ y_i * \ln\left(\widehat{y}_i\right) * \left(1 - y_i\right) * \ln\left(1 - \widehat{y}_i\right) \right] + \lambda \sum_{j}^{f} w_j^2, \text{ for Ridge regularization.} \end{aligned} \tag{3.12}$$

Here lambda ($\lambda$) represents the regularization strength and $f$ the number of weights. Lasso regularization suffices in cases where only a small number of significant weights causes the overfitting. Ridge regularization works well if there are multiple number of large weights which causes the overfitting.

### 3.2.4 Classifiers

In this thesis, we will focus on three classification algorithms for classifying PPG data quality: Logistic Regression, Support Vector Machines and feedforward Neural Networks.

#### Logistic regression

Logistic regression (LR), despite its name, is a statistical probabilistic classification model rather than a regression model. A LR model approximates the relationship between a categorical variable and the

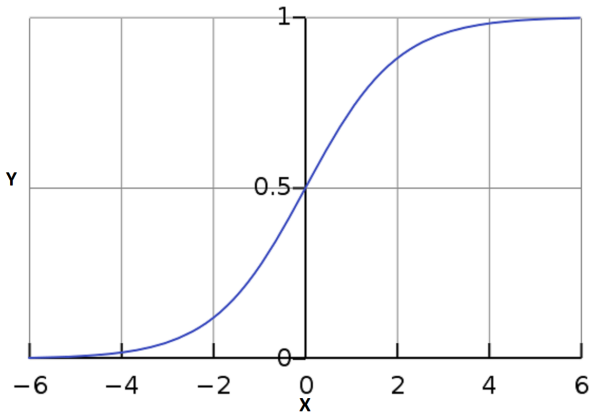



Figure 3.4: Sigmoidal function in logistic regression which shapes the predictions between 0 and 1. In this figure the threshold for a positive prediction is set at 0.5.

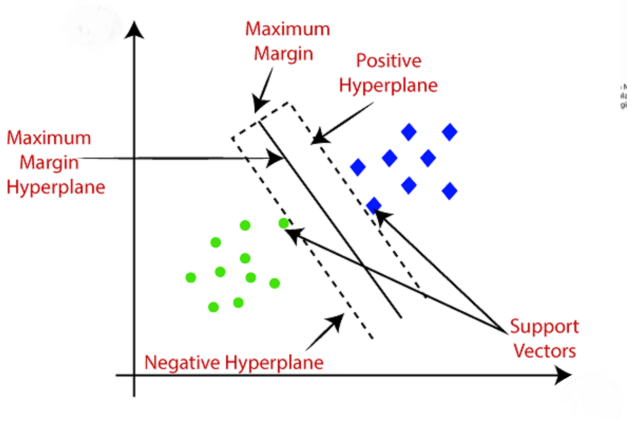



Figure 3.5: Defining the hyperplane in SVM. The figure shows a hypothetical decision boundary that maximizes the margins between the support vectors.

independent input variables $\mathbf{x}$. In this thesis we use the most simple form of logistic regression, namely binary logistic regression, in which the outcome variable $y$ has two possible outcomes $Y \in (0, 1)$. The model function for LR is the sigmoid function of input $\mathbf{x}$ and corresponding model weights $\mathbf{w}$, shown in equation (3.13) and Figure 3.4:

$$\hat{y} = p(y = 1 \mid \boldsymbol{x}; \boldsymbol{w}) = \frac{1}{1 + e^{-(\boldsymbol{w}^T \boldsymbol{x} + b)}} \tag{3.13}$$

One of the major advantages in LR is the mapping of real-valued numbers into the range (0,1). This is exactly what we need for a probability. The function is also differentiable which comes in handy for learning. But how can we make a decision for class assignment? Let's assume the binary classification problem sick vs. not sick. We can simply say, for a given $\mathbf{x}$, that a person is sick when the probability $p(y{=}sick|\boldsymbol{x}; \boldsymbol{w})$ is more than a certain threshold. The threshold, e.g. at $\hat{y} = 0.5$, for the LR can be described as:

$$\begin{aligned} \text{decision}(\boldsymbol{x}) &= 1, \text{if } \hat{y} \geq 0.5, \\ \text{decision}(\boldsymbol{x}) &= 0, \text{otherwise.} \end{aligned} \tag{3.14}$$

Suppose a LR model is trained on 3 normalized clinical features ($x_1$, $x_2$, $x_3$). F.e. these features could be the core temperature of the subject, the number of symptoms and/or the activity level during the day. Let's assume a new observation:

$$\boldsymbol{x} = [0.7, 0.1, 0.6], \tag{3.15}$$

and the LR model has the following weight vector and bias:

$$\begin{aligned} \boldsymbol{w} &= [2, -1, 3], \\ b &= 0.5 \end{aligned} \tag{3.16}$$

Then the resulting output $\hat{y}$ would be:

$$\begin{aligned}\hat{y} &= \frac{1}{1+e^{-(\boldsymbol{w}^T\boldsymbol{x}+b)}} \\ &= \frac{1}{1+e^{-(2*0.7-1*0.1+3*0.6+0.5)}} \\ &= 0.96\end{aligned} \tag{3.17}$$

The LR model will classify this new observation as sick.

**Support Vector Machine**

The Support Vector Machine (SVM) algorithm is well established as a non-probabilistic binary classifier. The SVM generates a model that represents the data as a point in space, and within this space a decision boundary or hyperplane is then determined. The principle of SVM is shown in Figure 3.5 . A maximum margin hyperplane is constructed using the support vectors: data points that have smaller distance to the other class and therefore more prone to misclassification. The negative hyperplane is constructed on the support vector of class 0 and the positive hyperplane on the support vector of class 1. The maximal margin hyperplane is constructed in a way that it is able to maximize the margin between data points of both classes. New test examples are then classified based on which side of the hyperplane they are positioned.

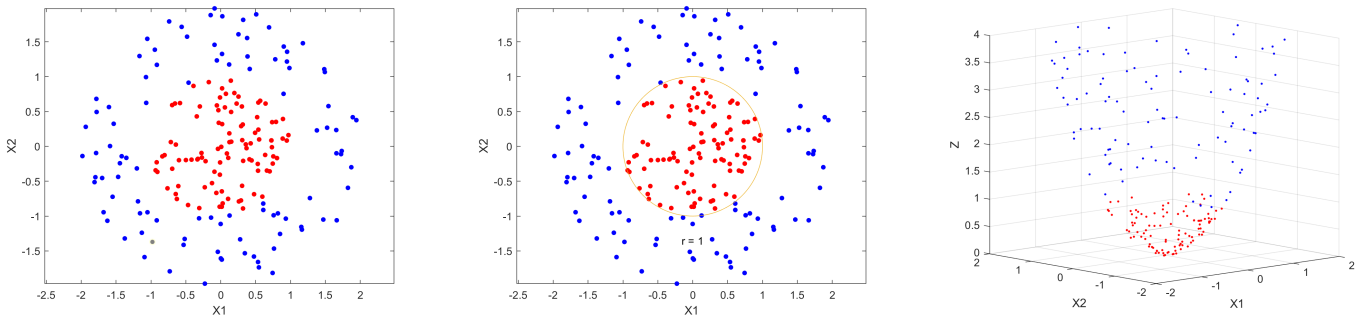
Since most data is not linearly separable, finding the correct hyperplane requires a preprocessing step. This is where the *kernel trick* comes in. A kernel is a function that takes the original non-linear classification task and transforms it into a linear classification task within a higher-dimensional space. We can take the non-linear classification problem in Figure 3.6a We cannot draw a line that would make a good estimation between the two classes. However, we can separate them by drawing a circle between the blue and red points (Figure 3.6b). We can add a third dimension $z$ to the problem by adding the squared input coordinates:

$$z = x_1^2 + x_2^2 \tag{3.18}$$

The result of this transformation is seen in Figure 3.6c. Now the classes can be separated by a hyperplane. In this case, we can use a simple equation to obtain linear separable classification task. However, in most of the cases usage of such a simple equation is not sufficient. One of the most widely used kernels in SVM is the Radial Basis Function (RBF) kernel. The RBF kernel function for two different input vectors computes the similarity to each other. The mathematical function of the kernel is as follows:

$$K(x_1, x_2) = \exp\left(-\frac{||\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)}||^2}{2\sigma^2}\right). \tag{3.19}$$

Here $\sigma$ represents a hyperparameter which determines the region of similarity, and $||\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)}||$ the Euclidean distance between the two observations $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$. This equation has a maximum value of 1 and occurs when two observations are the same. The more distance between the points, the lower the value of the RBF kernel. Since the kernel only stores the similarity between data points and not the entire data set itself, it is more efficient in training the classifier.




(a) (b) (c)

Figure 3.6: Performing the kernel trick for non-linear separable data (a). The two classes (red vs blue) can be seperated by drawing a circle between the classes (b) are transformed to a higher dimension, $z = x_1^2 + x_2^2$ (c), to make the classes linear separable.

### Neural Networks

The last classifier used in this thesis is a more sophisticated method for the classification problem. Neural networks (NN) are widely used within ML. The building block of a neural network is a single computational unit. As in most classification algorithms, the core concept consist of a neural unit that takes a weighted sum of its input, with an additional bias term ($z = \boldsymbol{w} \cdot \boldsymbol{x} + b$). Similar to LR, in most NN algorithms this weighted sum is applied in a non-linear function, known as the activation function. The most popular activation functions in NN are the sigmoidal function, the hyperbolic tangent function or the ReLU function. In Figure 3.7 a schematic picture of a neural unit is shown. The unit multiplies the input vector $\mathbf{x}$ (consisting of $x_1$, $x_2$ and $x_3$) is by its weight vector $\mathbf{w}$ ($w_1$, $w_2$ and $w_3$), computes the weighted sum and adds the bias term $b$. This output $z$ is passed into an activation function (e.g. the sigmoid function) to result in a number between 0 and 1.

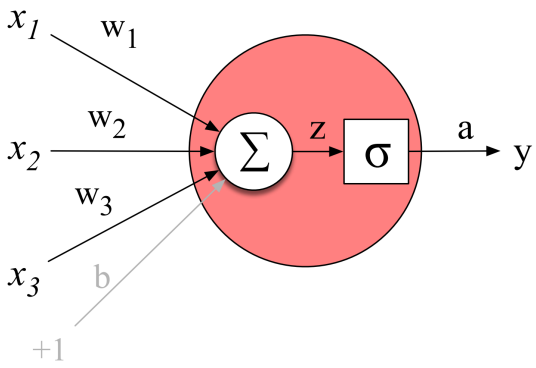



Figure 3.7: A neural unit, taking 3 weighted inputs $x_1$, $x_2$, and $x_3$ and a bias $b$. After calculating the weighted sum $z$, it produces an output $y$ by inserting the weighted sum to the activation function ($\sigma$). In this case, the output of the neural unit $a$ is the same as the total ouput $y$. In a network structure, $a$ represents the output/activation of an individual node and $y$ the final output of the entire network.

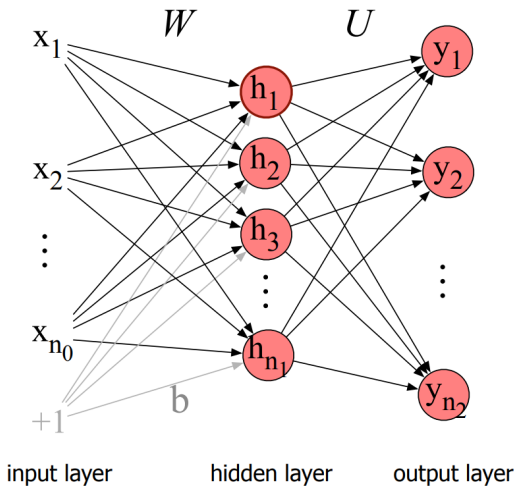



Figure 3.8: A simple feedforward neural network, with one hidden layer, one output layer, and one input layer. Each arrow connecting two units represents a weight. The weights between the input layer and the hidden layer are described in matrix **W**. The weights between the hidden layer and the output layer by matrix **U**.

A neural network consists of a variable number of layers. In each layer a number of neural units is present. The simplest kind of neural network is the feedforward network. These networks have three kinds of nodes: input units, hidden units and output units. In Figure 3.8 a representation of this structure is given. The core of the neural network is the hidden layer $h$. This hidden layer consists of the previous described neural units. In the standard architecture, every layer is fully-connected, meaning that each unit takes as input the outputs of all units in the previous layer. Every arrow can be represented as a weight connecting two units. For efficiency, **W** is a matrix where each row represents the weight vector corresponding to a hidden unit. For every unit the weighted sum is calculated by applying a given activation function.

For a binary task, the output layer consist of only one output node, and its value $y$ is the probability of the positive class. For multiclass problems, each output represent a probability of the observation belonging to a given class. For the output layer, there is also a weight matrix **U** which connects the input vector (**h**) to produce an output probability.

# 4 | Model development

## 4.1 Introduction

In the PPP study, we have continuous PPG data of subjects who are ambulatory monitored by a wrist-worn device over the follow-up of a 2-year period. During ambulatory monitoring, the PPG signal could be interfered by various types of artifacts could such as walking, tapping, hand movements and stretching [15, 17]. Although artifact suppression techniques are useful when the interference of such an artifact on the PPG signal is not fatal, alternative strategies are desired when the interference cause a serious-disturbed signal[15, 16, 17, 18]. Such an alternative strategy may be to distinguish signals that are eligible for analysis of heart rate patterns, from signals that are of insufficient quality.

In case of analysis of the heart rate (HR), high quality is desired, but one can often make a global estimate of the HR with a couple of artifacts in the data. On the other hand, when the aim is to assess information on the peak-to-peak intervals, like in heart rate variability (HRV) analysis, high quality is a requirement since the exact location of the pulse is crucial for reliable HRV parameter estimation [13, 37, 39].

Hence, the goal is to develop smart algorithms that can distinguish high-quality data from low-quality data. Traditional supervised ML models could be a tool to take this step. In previous studies, several supervised ML algorithms are used to assess data quality in PPG signals [17, 40].

In this chapter we aim to optimize different supervised ML models for two analysis of heart rate patterns: HR analysis and HRV analysis. We created a labeled data set using different categories of signal quality. Secondly, we extracted relevant features from the annotated PPG epochs which we used to train the ML classifiers and consequently evaluated model performances.

## 4.2 Methods

### 4.2.1 Data set

We used the patient cohort of the Personalized Parkinson Project (PPP) [9]. The PPP cohort consists of adult PD-patients (>18 years), included in the period 2017-2020, who were diagnosed with PD $\leq$ 5 years. From the PPP data set, we identified all patients who completed their baseline visit. We excluded patients without available PPG data.

### 4.2.2 Data acquisition and preprocessing

In the PPP study, patients undergo three extensive annual assessments and continuously wear a multi-sensor investigational research device (Verily Study Watch). Patients were instructed to wear the Verily Study Watch on the wrist of their own preference. All PPG sensor data obtained with the Verily Study Watch are stored as *protobuf* format first, and were subsequently converted to a new data format specifically designed by the PPP consortium (*tsdb*). The *tsdb* data format is designed to allow for queries on the metadata, clinical data and time-series data of the wearable sensor data. For data storage, transfer and analyses Snellius (SURF, Utrecht, The Netherlands) was used.

The PPG data was sampled at a sample frequency of 30 Hz over the course of the study period, with the exception of the period March 2021 – September 2021 in which a sample rate of 100Hz was used. For the current study, we only analyzed data segments that were sampled with 30 Hz.

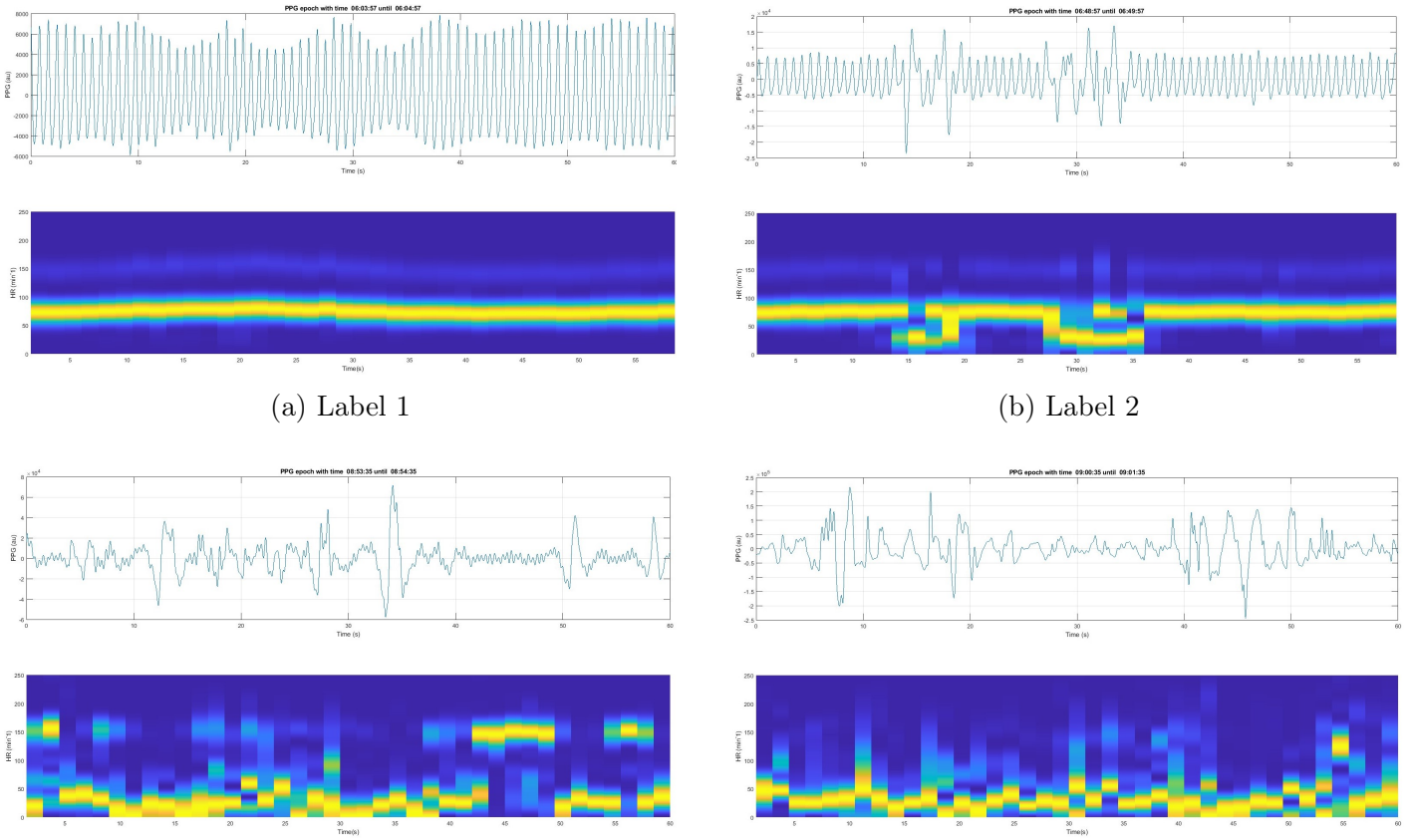
All preprocessing and analyses of the PPG data were performed using MATLAB® (Version 2022a, Natick, USA). We used a fifth order infinite impulse response filter (type band-pass Butterworth) with cut-off frequencies 0.4 - 3.5 Hz. After filtering, data was detrended using a first-order polynomial detrend function to remove the quasi DC component caused by changes in venous pressure from the time series signal.

### 4.2.3 Data quality

We divided all continuous PPG data recordings into non-overlapping 1-minute PPG-epochs. One-minute epochs were found the most optimal range for visual assessment of the data quality. We aimed to quantify all PPG-epochs based on their data quality, using categories based on the percentage of the 1-minute epoch that visually consists of high-quality PPG data. The following four categories were used (Table 4.1, Figure 4.1):

Table 4.1: Categorization of one-minute photoplethysmography (PPG) epochs based on the quality of the data into four types.

| Category | Description | Definition | To be used for |
|---|---|---|---|
| *Type 1* | High quality | 95% - 100% high-quality PPG data | - Heart rate variability<br>- Heart rate analysis |
| *Type 2* | Medium-high quality | 50% - 95% high-quality PPG data | Heart rate analysis |
| *Type 3* | Medium-low quality | 10% - 50% high-quality PPG data | Heart rata analysis |
| *Type 4* | Poor quality | 0% - 10% high-quality PPG data | N/A |



(a) Label 1

(b) Label 2

(c) Label 3

(d) Label 4

Figure 4.1: Overview of the four different categories, as described in Table 1, plotted as time-series data with corresponding spectrogram. Upper left panel: example of high quality PPG data (95-100%, label 1). Upper right panel: example of medium quality PPG (50-95%, label 2). Lower left panel: example of medium quality data (10-50%, label 3). Typically, this label contains a data segment with a higher heart rate (about 150bpm). Lower right panel: example of poor quality PPG data (0-10%, label 4). The spectrogram is normalized over 3 s windows.

### 4.2.4 Classifiers

The aim of the study was to develop supervised machine learning (ML) based classifiers, to distinguish between different epoch types (Table 4.2). The two binary models were optimized for both HR analysis (**model A**) and HRV analysis (**model B**). Furthermore, we optimized a multiclass model to evaluate the performances of all four classification types (**model C**). We trained the following classifier types: logistic regression (LR), support vector machine (SVM), and feedforward Neural Networks.

Table 4.2: Outline of the models developed in this thesis. For each model the corresponding classifiers are trained to obtain the optimal model. LR = Logistic Regression, SVM = Support Vector Machine, NN = feedforward Neural Networks.

| Name | Epoch types | Model | Classifiers | (Clinical) goal |
|---|---|---|---|---|
| *Model A* | Type [1, 2, 3] vs. Type 4 | Binary | LR, SVM, NN | Heart rate analysis |
| *Model B* | Type 1 vs. Type [2, 3, 4] | Binary | LR, SVM, NN | Beat-to-beat analysis |
| *Model C* | Type 1 vs. 2 vs. 3 vs. 4 | Multiclass | SVM, NN | Model performance |

### 4.2.5 Training protocol

#### Annotation data set

A representative sample of 20 subjects was selected from the data set to create the annotation set. Stratification was performed based on the following characteristics: gender, age, race and resting tremor scores (UPDRS III: Motor examination in OFF state). Manual annotation was performed on one-minute epochs of Week 1 and Week 52. As such, the annotation set incorporated representative data from both the start, as well as from a prolonged time during the study period. For each week, annotation was performed on the first full 24h of the PPG recording (1440 epochs). The total data set for manual annotation consisted of 20 * 2880 = 57,600 epochs with corresponding labels.

#### Annotation protocol

A graphical user interface (GUI) in Matlab was developed for data annotation (Appendix B). In the GUI, the filtered time-series signal of the PPG epoch is plotted above, together with its corresponding spectrogram (below). The PPG signal is plotted on a fixed y-axis range to keep the annotation process as constant as possible. The spectrogram is created based using the *pspectrum* function in MATLAB® and had a time-resolution of 3 s with an overlapping window of 50%. Subsequently, the spectrogram was normalized using a min-max normalization approach (i.e. leading to values between 0 and 1). Moreover, the exact time of the PPG-epoch was presented. Annotations were performed independently by two trained investigators. Before creating the final annotation set, these experts were trained on the definitions of the epoch types. Actual annotation was started after obtaining a Cohen's kappa of $> 0.9$ on three different PPG test sets, each consisting of a full 24h.

### 4.2.6 Feature extraction

From the one-minute PPG epochs, 18 "hand-crafted" features are constructed based on literature and expert insights while annotating the data. An overview, including formulas, is given in Appendix C. We distinguished three different feature categories:

I ***One minute epoch-based features***
The following features were calculated over the entire PPG epoch:

1 *Standard deviation of the PPG*
2 *Mean of the PPG*
3 *Median of the PPG*
4 *Skewness of the PPG*: which measures the assymmetry of the data distribution
5 *Kurtosis of the PPG*: which measures the "tailedness" of the classifier
6 *Dominant frequency of the PPG*: defined as the maximum peak in the power spectrum.
For calculation of the dominant frequency, Welch's periodogram is applied using a Hann window with an overlap of 50%
7 *Regularity index of the PPG*: defined as the ratio of the spectral power in the dominant frequency band (± 0.2 Hz) and the total spectral power.
For calculation of the regularity index the trapezoidal rule is used to calculate the spectral power of the Welch's periodogram.
8 *Organization index*: defined as the ratio of spectral power in the dominant frequency band (± 0.2 Hz) including the areas under harmonic frequencies and the total spectral power.
Similar to the regularity index, for calculation of the organization index the trapezoidal rule is used to calculate the spectral power.

II ***Mini-epoch based features***
The one minute PPG epoch was divided into sub epochs of 6 sec using an overlapping sliding window of 5 s to obtain 55 mini-epochs. Variation features were calculated over the mini-epochs:

9 *Variation in standard deviation of the PPG*
10 *Variation in mean of the PPG*
11 *Variation in median of the PPG*
12 *Variation in skewness of the PPG*
13 *Variation in kurtosis of the PPG*
14 *Variation in dominant frequency of the PPG*

An additional signal quality feature was calculated using an extra classification model based on six features of the mini-epochs: standard deviation, mean, median, skewness, kurtosis and dominant frequency of the PPG.

15 *Quality ratio*: defined as the ratio between the number of mini-epochs classified as high-quality data, using a LR mini-epoch classifier, and the total number of mini-epochs.
To train the classification model, mini-epochs were obtained from a random subset of the annotation set containing only epochs of high quality data (400 epochs with label 1) and low-quality data (400 epochs with label 4). The outcome of the classifier was 1 (high-quality mini-epoch) or 4 (low-quality mini-epoch).

III ***Spectrogram-based features***

A spectrogram was obtained using the Short-Time Fourier Transform with a Hann window with 50% overlap and a time resolution of 3 sec. From this spectrogram three features were calculated. In Figure 4.2 an illustration of the three different features is presented:

16 *Maximum sequence*: defined as the maximum of consecutive windows in the physiological HR range (40-180 bpm), which do not differ more than 15% in HR compared to the previous window .

17 *Sum of individual sequences*: defined as the sum of all independent sequences.
A sequence consist of at least a minimum of 2 conconsecutive windows in the physiological HR range (40-180 bpm), which do not differ more than 15% in HR compared to the previous window.

18 *Spectral standard deviation*: defined as the standard deviation of all dominant frequencies, in the physiological HR range, obtained in each separate window of the spectrogram.

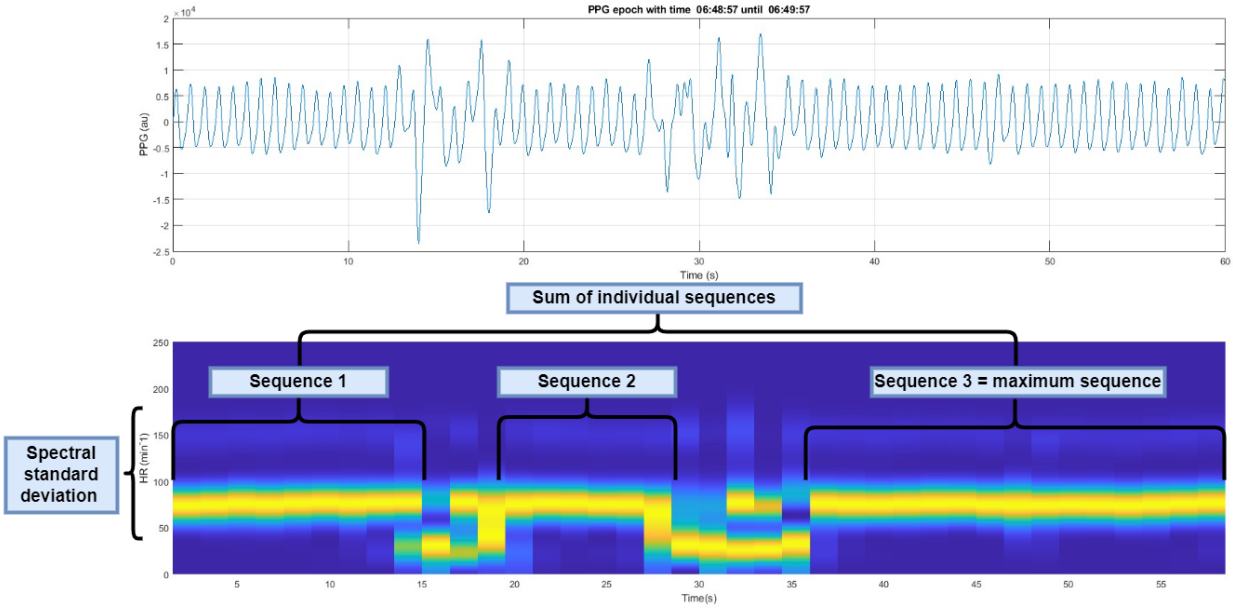



Figure 4.2: Visualization of the spectrogram-based features. A category 2 data quality is plotted. This epoch has three individual sequence from which sequence 3 is the longest (maximum sequence). The sum of indivdual sequences is defined as the sum of the three independent sequences. The spectral standard deviation is the standard deviation of all windows with the dominant frequency in the physiological HR range (40-180 bpm).

### 4.2.7 Feature selection

From the total of 18 calculated features, the most optimal feature set was determined using Maximum Relevance – Minimum Redundancy (MRMR). This algorithm ranks the features by relevance, based on

performance scores are calculated after each iteration. The optimal feature set is defined as the feature set with the highest accuracy. An overview of the MRMR algorithm can be found in Appendix D.

### 4.2.8 Training of machine learning classifiers

#### Hyperparameters

For the three models, described in Table 4.2, a grid search optimization method was chosen to optimize model to find optimal hyperparameter configuration. In Table 4.3, an overview of the hyperparameters used to optimize the different classifiers is presented.

For LR, the regularization strength ($\lambda$) and the solver used for the optimization problem are varied. $\lambda$ is set in a logarithmic range ($[\frac{10^{-5}}{N} \ \frac{10^{5}}{N}]$) with N equal to the number of observations. Three different solvers are used: stochastic gradient descent (SGD), limited BFGS (lbfgs) and Sparse Reconstruction by Separable Approximation (SpaRSA).

When using SVM, the kernel function and the regularization parameter C are optimized. Two different kernels are used: the Radial Basis Function (RBF) kernel and a linear kernel. The regularization parameter C is chosen in in a logarithmic range ($[10^{-2}\ 10^{2}]$). For NN, the regularization strength ($\lambda$) and the activation function are tuned. $\lambda$ is set in the same logarithmic range as in LR. Four different options for activation functions were used: the ReLu function, the hyperbolic tangent (tanh) function, a sigmoidal function and the option of no activation function.

Table 4.3: Overview of the adjusted hyperparameters using a specific grid search optimization procedure. [a] Grid names correspond to the input names in Matlab, with N equal to the number of observations. The grid values are log-scaled with correponding limits.

| Classifier | Hyperparameter | HP grid limits/names [a] |
|---|---|---|
| *LR* | $\lambda$ | $[\frac{10^{-5}}{N} \ \frac{10^{5}}{N}]$ |
| | Solver | ["sgd", "lbfgs", "sparsa"] |
| *SVM* | Kernel function | ["rbf", "linear"] |
| | C parameter | $[10^{-2}\ 10^{2}]$ |
| *NN* | $\lambda$ | $[\frac{10^{-5}}{N} \ \frac{10^{5}}{N}]$ |
| | Activation function | ["relu", "tanh", "sigmoid", "none"] |

#### Nested cross-validation

To obtain the optimal classifier for this classification problem, a nested cross-validation (CV)procedure, which nests CV and hyperparamter tuning, was performed for each classifier. Therefore, it can be used to evaluate the performance of an algorithm and also search for its most optimal combination of hyperparameter values. Nested CV consists of two optimization loops:

- The outer loop: for evaluating model performances

- The inner loop: for hyperparameter optimization

In Figure 4.3 an overview of the nested CV is given. In the nested CV procedure the data set is split into 10 folds and each fold is held out for testing (outer loop) while the remaining 9 folds are merged and split into 2 folds for training and validation (inner loop). Splits were made between subjects, so all epochs of one subject ended up in either the test or training set (outer loop) and in either the training or validation set (inner loop).

In the inner loop, the hyperparameters were optimized while performing grid search. The optimal hyperparameters were defined as the highest average performance over the two inner folds. Thereafter, in the outer loop, model performance was measured using the test fold on the optimal hyperparameter configuration. The performances were averaged over the 10 outer folds to obtain an average performance for each classifier. The average performances for each classifiers were compared and the best classifier is chosen.

As a last step, hyperparameters are then optimized using the same cross-validation procedure in the inner loop but then for the total data set ($N = 20$). After obtaining the final hyperparameters, the model can be generalized to a final model by retraining the classifier on all subjects with the corresponding hyperparameters. A detailed overview of the implementation in Matlab for the entire procedure is given in Appendix E.

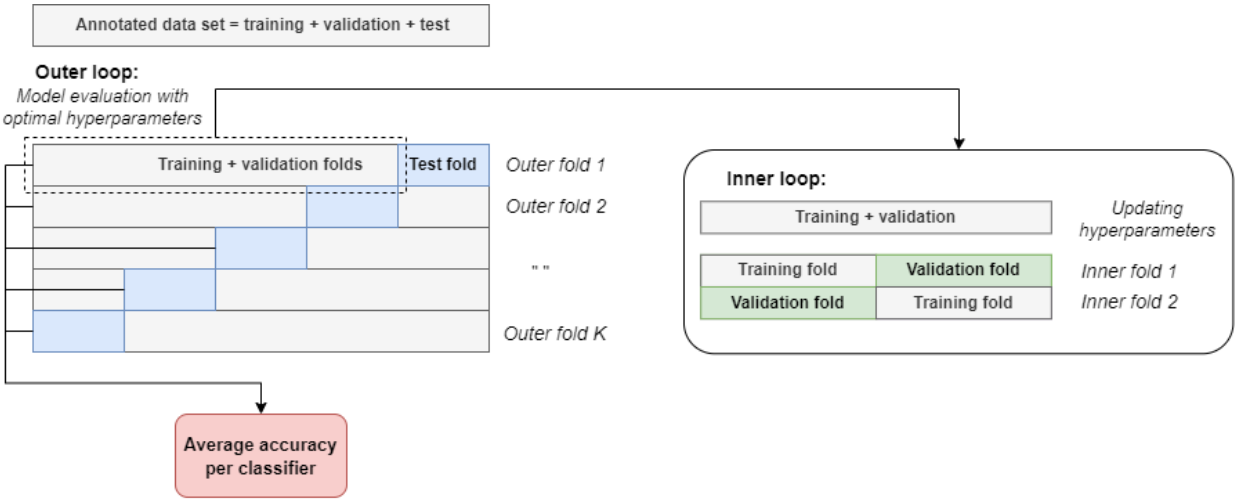



Figure 4.3: Overview of the nested CV procedure. The total data set is split into 10 outer folds (K) and 2 inner folds. In the outer loop the training and validation folds (K-1) are used for updating hyperparameters in the inner loop. These folds are merged and randomly divided into two inner folds. Grid search (not included in the figure) is performed to find the optimal hyperparameter configuration in the inner loop. In the outer loop, the optimal hyperparameter configuration is used to train the classifier and evaluate the classifier on the remaining test fold.

### Model performance metrics

The primary outcome measure to describe model performance was the balanced accuracy. Other performance scores were accuracy, sensitivity, specificity, positive predictive value (PPV), and F1-score. These performance measures were calculated using a probability of $\rho < 0.5$ for the negative class and a probability of $\rho \geq 0.5$ for the positive class. In Table 4.4, calculations and focus of all performance metrics are given for the different models.

Table 4.4: Performance metrics for binary and multiclass classification. For multi-class classification many classes $C_i$ can be used: $TP_i$ are true positives for $C_i$, $FP_i$ - false positives, $FN_i$ - false negatives and $TN_i$ - true negatives.

| **Measure** | **Formula (binary)** | **Formula (multiclass)** | **Focus** |
|---|---|---|---|
| *Accuracy* | $\frac{TP+TN}{TP+TN+FP+FN}$ | $\frac{\sum_{i=1}^{l} \frac{TP_i+FN_i}{TP_i+TN_i+FP_i+FN_i}}{l}$ | The average per-class effectiveness of a classifier |
| *Balanced accuracy* | $\frac{Sens+Spec}{2}$ | $\frac{Sens+Spec}{2}$ | Balanced effectiveness of a classifier |
| *Sensitivity* | $\frac{TP}{TP+FN}$ | $\frac{\sum_{i=1}^{l} \frac{TP_i}{TP_i+FN_i}}{l}$ | Quality of the classifier to recall TP |
| *Specificity* | $\frac{TN}{TN+FP}$ | $\frac{\sum_{i=1}^{l} \frac{TN_i}{TN_i+FP_i}}{l}$ | Quality of the classifier to recall TN |
| *Positive predictive value (PPV)* | $\frac{TP}{TP+FP}$ | $\frac{\sum_{i=1}^{l} \frac{TP_i}{TP_i+FP_i}}{l}$ | Quality of a positive predicition made by the model |
| *F1-score* | $\frac{2*TP}{2*TP+FN+FP}$ | $\frac{\sum_{i=1}^{l} \frac{2*TP_i}{2*TP_i+FN_i+FP_i}}{l}$ | Harmonic mean between PPV and sensitivity |

Table 4.5: Baseline characteristics of the annotation data set and the total PPP data set. Categorical variables are presented as n(%) and continuous variables as median (IQR).

| **Characteristics** | **Annotated data set (N = 20)** | **PPP data set (N = 484)** |
|---|---|---|
| **Demographic** | | |
| Age | 63 (54 - 73) | 62 (55 - 69) |
| Male | 12 (60) | 285 (59) |
| Caucasian race | 18 (90) | 470 (97) |
| **Clinical** | | |
| H & Y | 2 (2 - 2) | 2 (2 - 2) |
| Rest tremor score $\geq 1$ | 7 (35) | 170 (35) |
| Total dyskinesia $\geq 1$ | 3 (15) | 82 (17) |
| - Dyskinesia time $\geq 1$ | 3 (15) | 77 (17) |
| - Dyskinesia impact $\geq 1$ | 1 (5) | 46 (10) |

## 4.3 Results

### 4.3.1 Data set

In Table 4.5, the baseline clinical characteristics of the annotation data set and the total PPP data set are given. No numerical differences are found between age and sex. The percentage of Caucasian race was higher in the PPP data set. No numerical differences were found based on H & Y stage and rest tremor scores. Dyskinesia scores were higher in the PPP data set.

The distributions over the different categories is given in Figure 4.4. For model A, 65% of the PPG epochs is labelled as high quality and the remaining 35% as insufficient quality. For model B, 23% of the epochs is labelled as high quality PPG epochs while the remaining 77% of the epochs is of insufficient quality.

### 4.3.2 Optimal number of features

Figure 4.5 shows the performances of the model compared to the number of features ranked according to MRMR. As a plateau has already been reached for the performances of the HR model (Model A) around 7 features. This applies to LR (Figure 4.5a), SVM (Figure 4.5b) and NN (Figure 4.5c). For the HRV model (model B) it can be seen that 14 features are needed to reach an optimum for all classifiers (Figure 4.5d, 4.5e and 4.5f). For the multi-class model it can be seen in Figure 4.5g and Figure 4.5h that an optimum is reached after 9 features for both classifiers. The test accuracies of both the inner and the outer loop are almost the same for all classifiers.

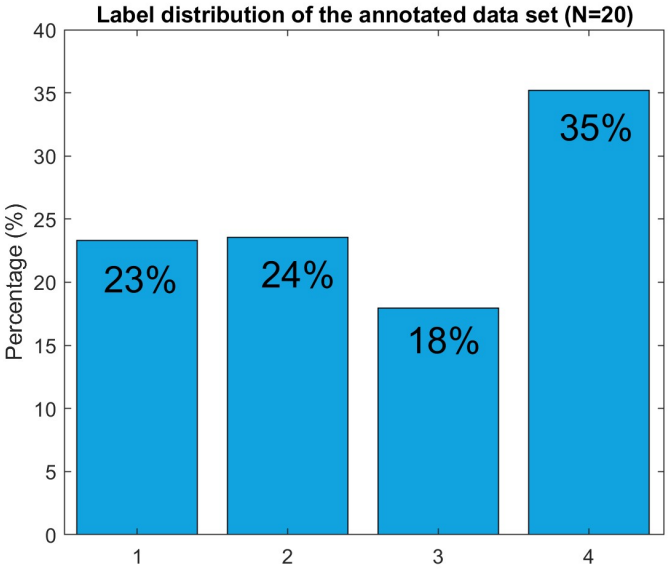



Figure 4.4: Distribution of the annotated data set. For the developement of model A, the distribution of the data set is 65% sufficient quality epochs and 35% poor quality PPG epochs. For model B, the distribution is 23% high-quality epochs vs 77% insufficient quality epochs

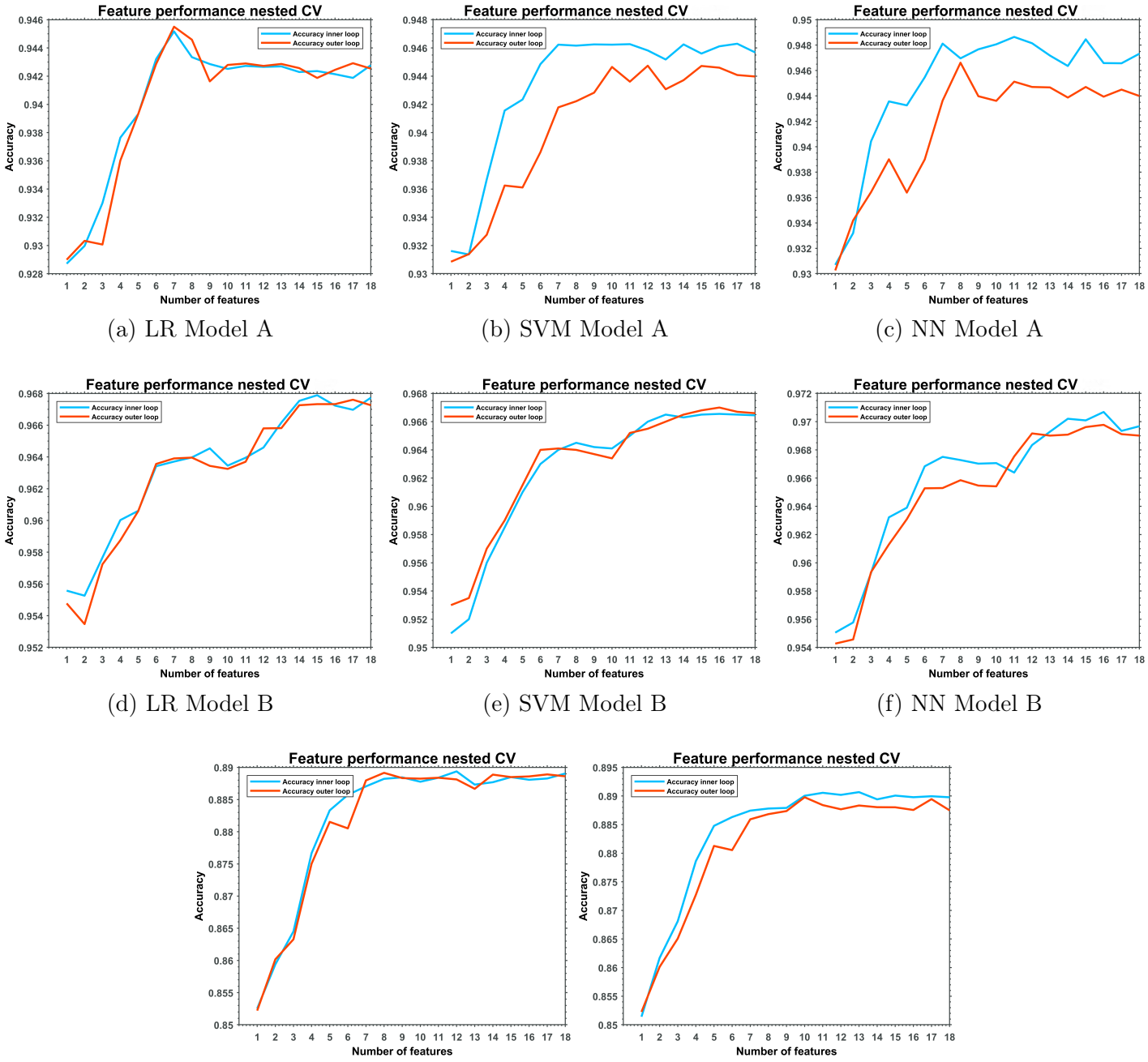
### 4.3.3 Model performances

The final model performance scores are calculated on the number of features needed to achieve the optimum for all three models. In Table 4.6 all performances of all classifiers are listed per type of model.

For model A it can be seen that the balanced accuracy is highest for LR (94.0%), followed by SVM (93.8%) and NN (93.7%). The sensitivity is lower for SVM (SVM: 91.8%; LR: 93.6%; NN: 94.0%). The specificity is highest for SVM (SVM: 95.8%; LR: 94.4%; NN: 93.3%). The PPV is highest for LR (LR: 97.0%; SVM: 96.6%; NN: 96.6%). F1 scores were highest for LR and NN (LR: 0.953; NN: 0.953; SVM: 0.941). In terms of time required to optimize the classifier, LR is the fastest (LR: 40s; SVM: 30 min; NN: 30 min).

For model B, the balanced accuracy is highest for LR (95.8%), followed by NN (95.6%) and SVM (94.3%) . Sensitivity is highest for LR (LR: 94.1%; NN:93.2%; SVM: 90.0%) while specificity is highest for SVM (SVM:98.7%: NN: 98.0%, SVM: 97.5%). PPV is highest for NN (NN: 93.5%; SVM: 92.9%, LR: 92.1%). F1 scores were highest for LR (LR: 0.933; NN: 0.931; SVM: 0.914). In terms of time required to optimize the classifier, LR is the fastest (LR: 40s; SVM: 30 min; NN: 30 min).

For the multi-class model C, the balanced accuracies and accuracies are identical for both the mean values as for the values per label. SVM has a slightly higher average accuracy (89.0%) vs. NN (88.7%). Furthermore, the balanced accuracies for label 1 and label 4 showed similar results to the results of the binary models of the corresponding classifiers: multiclass NN model (label 4: 93.3%; label 1: 95.7%) vs NN binary model A (93.7%) and model B (95.7%), and for the multiclass SVM model (label 4: 93.4%; label 1: 95.6%) vs SVM binary model A (93.9%) and model B (94.3%).




(a) LR Model A (b) SVM Model A (c) NN Model A

(d) LR Model B (e) SVM Model B (f) NN Model B

(g) SVM Model C (h) NN Model C

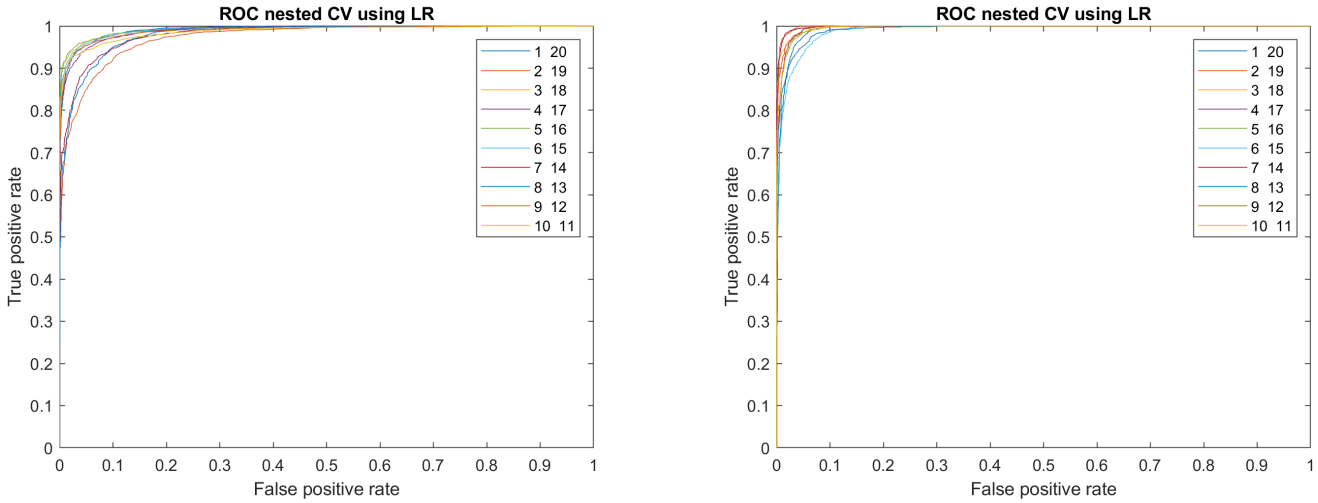
Figure 4.5: Feature performances of all models (A, B and C). The number of features, ranked by the minimum redundancy - maximum relevance algorithm, were compared based on accuracies of the inner (blue) en outer (red) loop of the nested CV procedure for model A (upper row), model B (middle row) and model C (lower row). An optimum was reached in all classifiers at around 7 features (model A), 14 features (model B) or 9 features (model C).

Table 4.6: Performance metrics of the classification models A, B and C. The last column represents the computational time to run the model. Model performances are based the optimal feature configuration: Model A = 7 features, Model B = 14 features, Model C = 9 features. For Model C, individual label performances are given including the average and the multiclass accuracy.

| Model | Classifier | Label | Balanced accuracy | Accuracy | Sensitivity | Specificity | PPV | F1-score | Time |
|---|---|---|---|---|---|---|---|---|---|
| *Binary* | | | | | | | | | |
| **Model A** | *LR* | 1,2,3 vs 4 | 94.0 | 94.5 | 93.6 | 94.4 | 97.0 | 0.953 | $\sim$ 40 s |
| | *SVM* | | 93.8 | 94.4 | 91.8 | 95.8 | 96.6 | 0.941 | $\sim$ 30 min |
| | *NN* | | 93.7 | 94.6 | 94.0 | 93.3 | 96.6 | 0.953 | $\sim$ 30 min |
| **Model B** | *LR* | 1 vs 2,3,4 | 95.8 | 96.8 | 94.1 | 97.5 | 92.1 | 0.931 | $\sim$ 40 s |
| | *SVM* | | 94.3 | 96.7 | 90.0 | 98.7 | 92.9 | 0.914 | $\sim$ 30 min |
| | *NN* | | 95.6 | 96.9 | 93.2 | 98.0 | 93.5 | 0.933 | $\sim$ 30 min |
| *Multiclass* | | | | | | | | | |
| **Model C** | *SVM* | 1 | 95.6 | 96.8 | 93.6 | 97.6 | 93.0 | 0.934 | $\sim$ 1 h |
| | | 2 | 91.1 | 92.7 | 87.5 | 94.9 | 87.4 | 0.875 | |
| | | 3 | 84.5 | 90.3 | 74.0 | 95.1 | 81.5 | 0.776 | |
| | | 4 | 93.4 | 93.3 | 94.4 | 92.4 | 90.7 | 0.925 | |
| | | *Average* | 91.2 | 93.3 | 87.3 | 95.0 | 88.2 | 0.878 | |
| | | *Multi* | | 89.0 | | | | | |
| | *NN* | 1 | 95.5 | 96.7 | 93.2 | 97.8 | 93.1 | 0.932 | $\sim$ 1 h |
| | | 2 | 91.0 | 92.6 | 87.5 | 94.7 | 87.0 | 0.873 | |
| | | 3 | 84.5 | 90.2 | 74.2 | 94.8 | 80.6 | 0.773 | |
| | | 4 | 93.3 | 93.2 | 94.0 | 92.5 | 90.7 | 0.924 | |
| | | *Average* | 91.1 | 93.2 | 87.2 | 94.9 | 87.8 | 0.875 | |
| | | *Multi* | | 88.7 | | | | | |

### ROC curves model A and B

In Figure 4.6, the example of the receiver operating (ROC) curve of both LR binary models is shown. Every line represents the performances (in terms of true positive and false positive rates) for a given threshold of the external test set in the nested CV procedure. In Figure 4.6a shows that three validation folds, consisting each two subjects, have a lower area under the curve (AUC), ranging from 0.976 - 0.980, than the remaining validation folds, ranging from 0.987 - 0.992. In Figure 4.6b the AUC of all validation folds ranged from 0.990 - 0.998.




(a) Model A (HR)

(b) Model B (HRV)

Figure 4.6: ROC of outer loop performance of the nested cross-validation procedure. Every line represents two subjects used for validation of the model.

## 4.4 Discussion

In this study, we aimed to develop supervised ML classification models to distinguish high quality from low quality PPG data segments using both time and frequency characteristics. We performed separate analyses for two purposes, HR analysis (model A) and HRV analysis (model B), which require different levels of data quality. Overall, high discriminative ability was seen for both model A and model B. We showed that for model A, we could develop models with similar results in balanced accuracy over the different classifiers (LR: 94.0%, SVM: 93.8% and NN: 93.7%). For model B, we obtained also similar results over the different classifiers (LR: 95.8%, SVM: 94.3% and NN: 95.6%). The LR classifiers seemed most optimal in terms of computational time.

During the annotation protocol, we annotated data from 20 subjects for the first 24 h in two different weeks. We obtained an imbalanced data set regarding the four levels of data quality (Figure 4.4). This distribution gave insights in the amount of PPG data we potentially could use. This implicates 65% of the annotated data is suited for HR analysis. Prior studies, regarding HR analysis from long-term PPG recordings, have shown that they could use lower percentages (55-60%) of the PPG data in their

study groups [41, 42]. For HRV analysis, 23% of the PPG data is suited during the 24h. Charlton et al. achieved up to 50 % high-quality signals in wrist-worn PPG data of healthy subjects [43]. In this paper, a signal quality index algorithm flowchart (containing several physiological thresholds), is used to assess epochs of 10 s [44]. However, the device used in this paper more advanced and larger in size, making it impractical for continuous PPG recordings over multiple years in an ambulatory setting.

While optimizing the models, the aim was to develop the most efficient classifier in terms of a minimal number of hand-crafted features, a high balanced accuracy and a low computational time. For model A, all classifiers reached an optimum in balanced accuracy after seven features, ranked by the MRMR algorithm. For model B, the optimum was reached after 14 features. In the feature selection process of both models, the most discriminative power was provided by the spectrogram features and the quality ratio feature.

In these situations the accuracy can be misleading and therefore a more careful interpretation of the model is required. The balanced accuracy takes into account the model's recall ability over both classes whilst the accuracy does not and is more simplistic. Still, interpretation-wise, the balanced accuracy requires more explanation since it is the mean of the sensitivity and the specificity of the model.

After annotating the data set and therefore obtaining the distributions of the labels, we concluded that for both binary models the data set was imbalanced. In these situations the accuracy can be misleading and therefore a more careful interpretation of the model is required. The balanced accuracy takes into account the model's recall ability over both classes. Therefore, the balanced accuracy was the primary outcome measure for both binary models. Still, interpretation-wise, the balanced accuracy requires more explanation since it includes both sensitivity and specificity. The sensitivity in model A ranged from 91.8% - 94.0%, while the specificity was in the range of 93.3% - 95.8%. In Table 4.7, the classifications of 100 PPG epochs for all three classifiers were given, regarding the imbalanced data set. For every 100 PPG epochs, 1.5-2.3 low quality epochs are classified as high quality and 3.9-5.4 high-quality epochs are classified as low quality epochs. In model B, the sensitivity ranged from 90.0 – 94.1%, while the specificity was in the range of 97.5 – 98.7%. Regarding the imbalanced data set in model B, this implicates that for every 100 PPG epochs, 1.0-1.9 low quality epochs were classified as high quality and 1.3-2.3 epochs falsely classified as low quality. The performance metrics were higher for model B, which is in line with the more imbalanced data set in model B.

The three classifiers obtained high performance measures for both classification tasks. Therefore, all classifiers could be used to assess data quality using PPG. However, the most efficient classifier, in terms of computational time, was LR. The obtained ROC curves, as showed in Figure 4.6, showed similar results over all validation folds, implicating a good generalizability of both models. This classifier was chosen to assess data quality for both HR analysis and HRV analysis.

Table 4.7: Overview of the expected classification of 100 PPG epochs for both model A and model B, based on the model performance metrics in Table 4.6. The data set is imbalanced for both model A (65-35) and model B (23-77). HQ = high-quality, LQ = low-quality

| | **Model A** | | **Model B** | |
|---|---|---|---|---|
| | *HQ epochs* | *LQ epochs* | *HQ epochs* | *LQ epochs* |
| *LR, +* | 60.8 | 2.0 | 21.7 | 1.9 |
| *LR, -* | 4.2 | 33.0 | 1.3 | 75.1 |
| *SVM, +* | 59.6 | 1.5 | 20.7 | 1.0 |
| *SVM, -* | 5.4 | 33.5 | 2.3 | 76.0 |
| *NN, +* | 61.1 | 2.3 | 21.4 | 1.5 |
| *NN, -* | 3.9 | 32.7 | 1.6 | 75.5 |
| **Expected total** | 65 | 35 | 23 | 77 |

### 4.4.1 Methodological considerations

In this study, we chose to perform supervised classification of one-minute PPG epochs into four categories. A requirement of supervised ML techniques is the existence of labels. To our knowledge, there was no gold standard for signal quality assessment of PPG data. Therefore, we developed an annotated data set consisting 57.600 epochs. Even with a high inter-rater agreement and a proper annotation protocol, it is likely that there are annotation errors included in the data set. An expert can be biased towards the labels of interest and/or can make a pre-emptive decision on the PPG epoch, based on prior knowledge of the previous annotated epochs [45].

One-minute epochs were chosen to be the most optimal range for visual assessment of the data quality. Moreover, spectral features obtained from a one-minute epoch had a high resolution which improved discriminative power of the models. The last advantage of choosing one-minute epochs is that for HRV analysis the minimum recording length for reliable parameter estimation is one minute [36].

We used relatively "simple" features in this supervised classification task. This caused a higher interpretability of the misclassifications. In the process of optimizing the models, we obtained insights which typical epochs caused the most trouble for the classifier. For example, periodic frequencies higher than the HR range, caused by e.g. tremor episodes (3.5 - 5.0 Hz), were classified as low-quality epochs after setting cut-off frequencies in the spectral features for solely HR range. Another advantage of composing simple, handcrafted features is the advantage of bringing in your own expertise [38]. By bringing in own expertise to distinguish quality in the four categories, the discriminative power of the models increased.

During optimization of the models, we used a nested CV procedure to perform this task. To make the model generalizable for PD, the validation in the outer loop consisted of a patient-wise 10-fold cross-validation. Most studies used k-fold CV with data both in test and training set. Such an approach overestimates the generalizability of a model since the PPG data within a subject is similar.

### 4.4.2 Future perspectives

In this study we showed promising results with regard to PPG quality discrimination. To improve further improve the reliability of our quality control tool, we could aim to diminish the influence of misclassifications. We could expand our tool by adding more physiological signals. An additional step could be to integrate the accelerometry signals, also measured by the wearable device, with the PPG signal to obtain a multi-modality quality control tool. Multiple studies used such an approach in PPG recordings for identification of motion artifacts [46, 47]. In our quality control tool, this would allow us to vary with the threshold for classification. We could allow more false positives, at the cost of a lower specificity, since this multi-modality tool would recognize noisy false positive epochs and reject them before performing further analysis.

However, the primary focus for further research should be to define strict criteria for the required eligible data size to perform HR(V) analyses during the day and during the night. Prior studies, use a wide range of signal durations on which HR(V) analyses are performed. This ranges from ultra-short recordings (15 s) towards ambulatory long-term (24 h) recordings. However, we should establish universal definitions for HR(V) analyses in long-term ambulatory PPG measures . In long-term HR(V) analyses in PD-patients, we need to determine which parameters could provide the most information to assess autonomic function. Based on these parameters, we could establish the length of consecutive high-quality PPG data needed for reliable parameter estimation and the interval at which these parameters should be calculated during the day and/or the night.

## 4.5 Conclusion

In this study, we developed several ML-based classifiers to assess data quality of wrist-worn PPG data in PD patients. These models demonstrated a high discriminative ability, for the two different types of heart rate analyses, with balanced accuracies $> 93\%$ for HR analysis and $> 94\%$ for HRV analysis. These results are promising for further research to monitor HR and HRV in PD patients in an ambulatory setting by a wearable device.

# 5 | Feasibility of PPG data in PD

## 5.1 Introduction

In the previous chapter, we developed algorithms to assess PPG quality for HR analysis (Model A) and HRV analysis (Model B). However, it is unknown whether these models can be used in the clinical setting for all patients.

In ambulatory monitoring of PD patients, there are several factors that possibly can influence the signal quality of the PPG signal. The first category is time factors. For example due to fluctuations in commitment to the clinical study and/or the deterioration of motor symptoms of the subjects over the follow-up period [48]. In addition, clinical factors like specific PD symptoms, such as tremor and involuntary movements (dyskinesia), and skin color could influence the data quality [49, 50]. These factors could give rise to selection bias.

To rule out whether these time and clinical factors influence the data quality, the objective is to compare the proportion of high quality PPG data in relation to the these factors. We aimed to longitudinally compare changes within subjects in the amount and the quality of the PPG data over the course of the study period. Secondly, we cross-sectionally analyzed differences in elgible data quality between subjects with or without the presence of the clinical factors. Thirdly, we performed an explorative HR analysis in which we cross-sectionally analyzed HR parameters (resting HR and maximum HR) in relation to symptoms of autonomic dysfunction in PD patients.

## 5.2 Methods

### 5.2.1 Data set

The PPP data set in section 4.2.1 was used as a starting point for feasibility analysis. We excluded all subjects with presence of cardiac arrythmias, such as atrial fibrillation, high premature atrial complex (PAC) burden and/or a high premature ventricular complex (PVC) burden, on the ECG at the baseline visit in week 0. A high burden was defined as two or more PACs and/or PVCs.

For the first aim (longitudinal analysis), subjects with insufficient wearing times (daily average PPG data $\leq$ 12 hrs in Week 0, Week 1 or Week 52) were excluded. For the second aim (clinical factor analysis), subjects with unavailable rest tremor scores (UPDRS Part III: Motor examination in OFF state), dyskinesia scores (UPDRS Part IV: Motor Examination in ON state) and/or demographic scores were excluded. Furthermore, subjects insufficient wearing times (daily average PPG data $\leq$ 12 hrs

in Week 0) were excluded. For the exploratory HR analysis, subjects with unavailable autonomic functioning scores (UPDRS Part I: Non-Motor Aspects of Experiences of Daily Living) were excluded.

### 5.2.2 Data preparation

All PPG data in a specific week is divided into non-overlapping 1-min segments whereafter all features were extracted. Given a specific model, the most optimal feature set (section 4.3.2) is selected. Subsequently, predictions are made based on a probability threshold of 0.5 ($\geq 0.5$ for the positive class). Furthermore, the segments were labeled with daytime (6:00 – 23:59) vs. nighttime (0:00 – 5:59) and the time of measurement (in hrs).

### 5.2.3 Study groups

For longitudinal analysis, we used the same study group for analyses in Week 0, Week 1 and Week 52.

For evaluating clinical factor, the included study population was divided into tertiles based on rest tremor score and dyskinesia. Two groups (caucasian vs non-caucasian) were created when evaluating the influence of race on the quality of the PPG signal.

### 5.2.4 Statistical analysis

Baseline variables and clinical characteristics were compared between the study groups. Categorical data were reported as frequencies(percentages) and analysed using the Chi square test. Continuous variables were analyzed for Gaussian distribution and reported as means $\pm$ standard deviations or medians (interquartile ranges (IQR)), whichever appropriate. Comparisons of dependent continuous variables were performed using a paired t-test or Wilcoxon rank sum test, whichever appropriate. Comparisons of continuous variables between the study groups were performed using the analysis of variance analysis (ANOVA) or with the Mann-Whitney U test, whichever appropriate.

### 5.2.5 Exploratory HR analysis

For the first HR analysis, we divided the selected 1-min epochs of all subjects in Week 0 into the same mini-epochs (as described in section 4.2.6).

Every mini-epoch of 6 sec was classified as high or low quality using the same LR classifier for the quality ratio feature (section 4.2.6. Per 1-min epoch, the HR was defined as the dominant frequency over all classified high-quality mini-epochs using Welch's periodogram using a Hann window with an overlap of 50%. In this analysis, the following two HR parameters were used:

- Resting HR, defined as the modal HR
- Maximum HR, defined as the 99th percentile

These HR parameters were calculated for each subject during the week, during the day and during the night and compared within study groups. In the HR analysis, these study groups (tertiles) are created based on the total score of the UPDRS Part I: Non-Motor Aspects of Experiences of Daily Living at the baseline visit of the study.

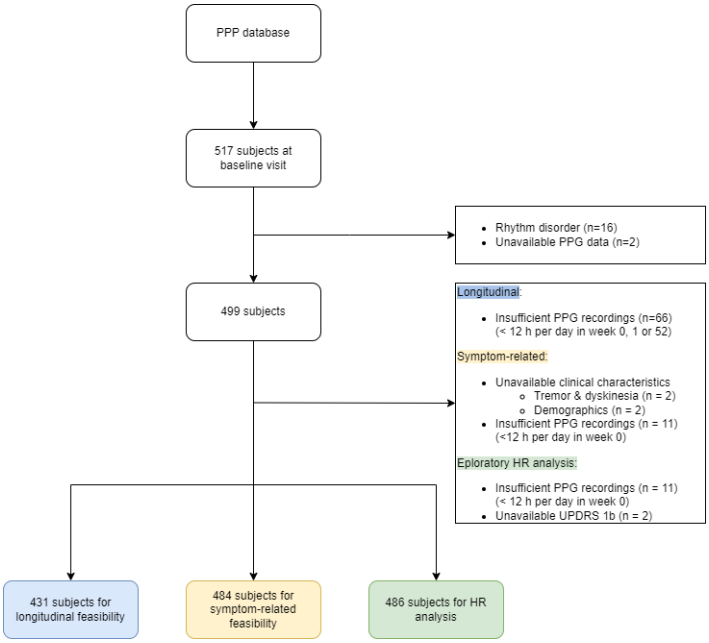



Figure 5.1: Flowchart of inclusion of the three different analysis

## 5.3 Results

### 5.3.1 Study population

From the PPP study population of 517 subjects, 16 patients with cardiac arrythmias and 2 with unavailable PPG data in Week 0 were excluded from the feasibility analysis. For evaluating the influence of time, we included 431 subjects. For evaluating clinical factors, we included 484 subjects. For HR analysis, we included 486 subjects. A schematic overview of the three different study populations and corresponding exclusion is given in Figure 5.1.

### 5.3.2 Longitudinal feasibility

Table 5.1 shows the data availability for the two models over the first year. The wearing time during the full 24 h is significantly lower in Week 1 (22.5 h) and Week 52 (21.6 h), compared to week 0 (22.5 h). The wearing time in week 52 is also significant lower compared to Week 1. The same decrease is seen during daytime (Week 0: 16.6 h; Week 1: 16.5 h; Week 52: 15.8 h) and nighttime (6.0 h; 6.0 h; 5.9 h).

For Model A, the total amount of high-quality data significantly decreased over the three different weeks (14.5 h; 14.5 h; 12.8 h), during daytime (9.0 h; 8.9 h; 7.5 h) and during nighttime (5.8 h; 5.7 h; 5.5 h). The proportion (in relation to the total wearing time) of eligible segments is significantly lower at Week 52 compared to both Week 0 and Week 1 during the week (65%; 65%; 60%), daytime (55%; 55%; 49%) and nighttime (97%; 97%; 95%).

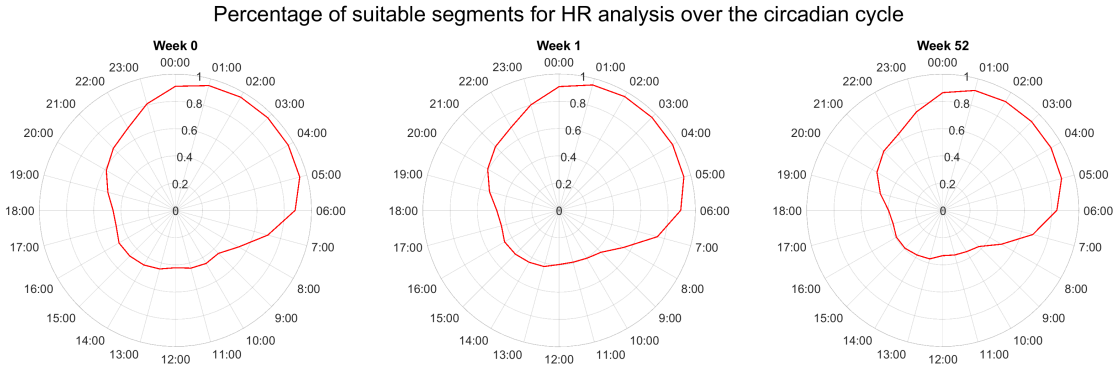
While using model B, the total hours of eligible data decreased from Week 0 and Week 1 to Week 52

Table 5.1: Data availability of PPG over time. Data are presented as medians (IQR) and percentages of the total wearing time; hrs = average number of hours per 24 hours over a week (00:00-23:59), daytime (06:00-23:59), or nighttime (00:00-05:59). * = Wilcoxon signed-rank test $p<0.05$ vs week 0, §= Wilcoxon signed-rank test $p<0.05$ vs week 1.

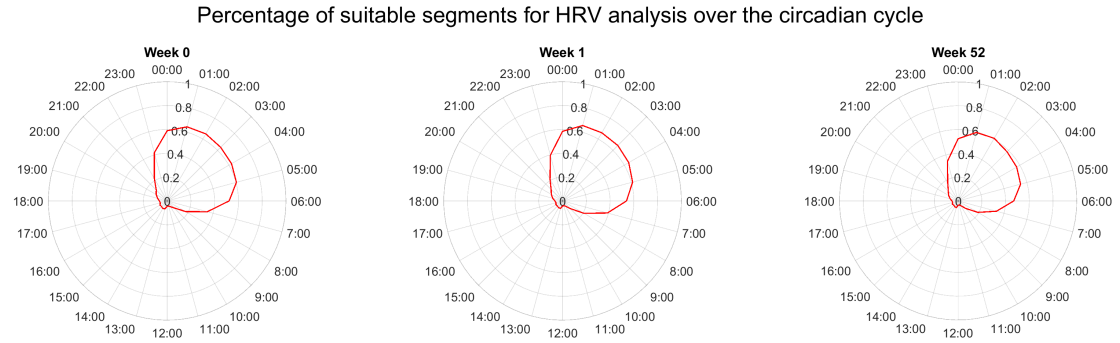
| | | **Week 0** | **Week 1** | **Week 52** |
|---|---|---|---|---|
| **Total wearing time** | *24h (hrs)* [b] | 22.5 (22.0 - 22.8) | 22.5 (21.8 - 22.8)* | 21.6 (20.3 - 22.3)*§ |
| | *Day (hrs)* | 16.6 (16.1 - 16.8) | 16.5 (16.0 - 16.8)* | 15.8 (15.0 - 16.5)*§ |
| | *Night (hrs)* | 6.0 (5.9 - 6.0) | 6.0 (5.9 - 6.0)* | 5.9 (5.1 - 6.0)*§ |
| **Model A** | *Eligible quality:* | | | |
| | *Week, hrs* | 14.5 (13.1 - 16.1) | 14.5 (12.7 - 16.1)* | 12.8 (10.9 - 14.6)*§ |
| | *Week, %* | 65 (60 - 72) | 65 (59 - 72) | 60 (53 - 68)*§ |
| | *Day, hrs* | 9.0 (7.5 - 10.5) | 8.9 (7.4 - 10.5)* | 7.5 (6.1 - 9.1)*§ |
| | *Day, %* | 55 (47 - 64) | 55 (46 - 64) | 49 (40 - 58)*§ |
| | *Night, hrs* | 5.8 (5.4 - 5.9) | 5.7 (5.3 - 5.9)* | 5.5 (4.6 - 5.8)*§ |
| | *Night, %* | 97 (94 - 98) | 97 (94 - 98) | 95 (93 - 97)*§ |
| **Model B** | *Eligible quality:* | | | |
| | *Week, hrs* | 6.1 (4.9 - 7.1) | 6.1 (5.0 - 7.1) | 5.1 (3.8 - 6.2)*§ |
| | *Week, %* | 27 (22 - 32) | 28 (23 - 32) | 24 (18 - 29)*§ |
| | *Day, hrs* | 2.2 (1.7 - 2.9) | 2.2 (1.7 - 2.9) | 1.8 (1.3 - 2.4)*§ |
| | *Day, %* | 14 (10 - 18) | 14 (10 - 18) | 11 (8 - 16)*§ |
| | *Night, hrs* | 3.8 (3.1 - 4.4) | 3.8 (3.1 - 4.4) | 3.2 (2.3 - 4.0)*§ |
| | *Night, %* | 65 (53 - 75) | 67 (56 - 76)* | 58 (45 - 70)*§ |

during 24 h (6.1 h; 6.1 h; 5.1 h), daytime (2.2 h; 2.2 h; 1.8 h) and nighttime (3.8 h; 3.8 h; 3.2 h). The percentage of eligible segments is also significantly lower in Week 52 during 24h (27%; 28%; 24%), daytime (14%; 14%; 11%) and nighttime (65%; 67%; 58%). Furthermore, the percentage in the night is significantly higher in Week 1 compared to Week 0. All statistical significant differences between Week 52 and Week 1 and/or Week 0 had p-values of $p<0.001$. All statistical significant differences between Week 1 and Week 0 had p-values of $p<0.05$.

In Figure 5.2 the average percentage of suitable PPG data for both binary models is plotted for Week 0, Week 1 and Week 52. The data is plotted per hour over the full circadian cycle. In model A, the proportion of eligible PPG data during the circadian cycle ranged from a minimum of 20% (during daytime) to 95% (during nighttime). The distribution of eligible PPG data is of a similar shape in Week 0 and Week 1. In Week 52 the percentage of eligible data is lower compared to Week 0 and Week 1.




(a) Model A




(b) Model B

Figure 5.2: Circadian cycle plots of the percentage of suitable data for model A and B. The available data is lower during the day than during the night.

For model B, The percentage of eligible PPG data during the circadian cycle ranged from around 10% (during daytime) to 60% (during nighttime). In multiple hours during the day, the mean percentage of available data for HRV analyses is less than 10% of the total recorded data. The pattern of the available data is similar for Week 0 and Week 1 whereas a reduced distribution of eligible data is seen in Week 52.

### 5.3.3 Clinical feasibility

Clinical feasibility was studied for the following clinical factors: tremor scores, dyskinesia scores and race.

#### Tremor

Table 5.2 showed the data availability for the different models comparing tertiles based on rest tremor scores. There were no significant differences between the three tertiles, no tremor vs. mild tremor vs. severe tremor, in total wearing time during 24h (22.5 h; 22.5 h; 22.4 h), day (16.6 h; 16.6 h; 16.4 h) and night (6.0 h; 6.0 h; 6.0 h)

Table 5.2: Data availability of PPG in week 0 comparing rest tremor scores (UPDRS III: Motor examination in OFF state). a Data are presented as medians (IQRs) and percentages of the total wearing time; hrs = average number of hours per 24 hours over a week (00:00-23:59), daytime (06:00-23:59), or nighttime (00:00-05:59). * = Mann-Whitney U test p<0.05 vs no tremor, § = Mann-Whitney U test test p<0.05 vs mild tremor.

| | | **No tremor (N = 314)** | **Mild tremor (N = 114)** | **Severe tremor (N = 56)** |
|---|---|---|---|---|
| **Total wearing time** | *24h (hrs)*[b] | 22.5 (21.9 - 22.8) | 22.5 (21.8 - 22.8) | 22.4 (21.9 - 22.7) |
| | *Day (hrs)* | 16.6 (16.0 - 16.8) | 16.6 (16.0 - 16.8) | 16.4 (15.0 - 16.5) |
| | *Night (hrs)* | 6.0 (5.9 - 6.0) | 6.0 (5.8 - 6.0) | 6.0 (5.9 - 6.0) |
| **Model A** | *Eligible quality:* | | | |
| | *Week, hrs* | 14.5 (13.0 - 15.9) | 14.5 (13.0 - 16.2) | 14.1 (12.0 - 15.8) |
| | *Week, %* | 66 (59 - 71) | 66 (60 - 72) | 65 (54 - 70) |
| | *Day, hrs* | 9.0 (7.5 - 10.4) | 9.0 (7.5 - 10.4) | 8.6 (6.4 - 9.9) |
| | *Day, %* | 54 (47 - 63) | 55 (47 - 64) | 54 (40 - 62) |
| | *Night, hrs* | 5.7 (5.4 - 5.9) | 5.7 (5.3 - 5.8) | 5.7 (5.3 - 5.9) |
| | *Night, %* | 97 (93 - 98) | 97 (94 - 98) | 97 (93 - 98) |
| **Model B** | *Eligible quality:* | | | |
| | *Week, hrs* | 5.9 (4.7 - 7.1) | 6.2 (4.9 - 6.9) | 6.2 (5.3 - 7.5) |
| | *Week, %* | 27 (22 - 32) | 27 (22 - 32) | 29 (24 - 33) |
| | *Day, hrs* | 2.1 (1.6 - 2.9) | 2.2 (1.6 - 2.9) | 2.4 (1.8 - 3.1) |
| | *Day, %* | 13 (10 - 17) | 14 (10 - 17) | 16 (11 - 19) |
| | *Night, hrs* | 3.8 (3.0 - 4.4) | 3.9 (2.9 - 4.4) | 3.7 (3.2 - 4.5) |
| | *Night, %* | 65 (52 - 74) | 67 (54 - 74) | 66 (54 - 78) |

For Model A, numeric differences in eligible data hours were seen comparing no tremor and mild tremor to severe tremor during 24 h (14.5 h; 14.5 h; 14.1 h) and daytime (9.0 h; 9.0 h; 8.6 h). Also percentage-wise, a smaller proportion of high-quality data is seen during 24 h and during daytime. In the night, the proportion of high quality data is the same for all study groups.

For Model B, the duration of eligible PPG data numerical increased with increasing tremor severity during 24 h (5.9 h; 6.2 h; 6.2 h) and daytime (2.1 h; 2.2 h; 2.4 h). This increase is also seen in the proportion of eligible PPG data. During the night, the proportion of high quality data is similar for all study groups.

Table 5.3: Data availability of PPG in week 0 comparing dyskinesia scores (UPDRS IV: Motor examination in ON state). Three groups are created based on the summation of the dyskinesia time and dyskinesia impact scores. a Data are presented as medians (IQRs) and percentages of the total wearing time; hrs = average number of hours per 24 hours over a week (00:00-23:59), daytime (06:00-23:59), or nighttime (00:00-05:59). * = Mann-Whitney U test p<0.05 no dyskinesia, § = Mann-Whitney U test p<0.05 vs mild dyskinesia.

| | | **No dyskinesia (N = 402)** | **Mild dyskinesia (N = 39)** | **Severe dyskinesia (N = 43)** |
|---|---|---|---|---|
| **Total wearing time** | *24h (hrs)*[b] | 22.5 (21.9 - 22.8) | 22.6 (22.1 - 22.8) | 22.2 (21.7 - 22.7) |
| | *Day (hrs)* | 16.6 (16.1 - 16.8) | 16.6 (16.3 - 16.8) | 16.3 (15.9 - 16.8) |
| | *Night (hrs)* | 6.0 (5.9 - 6.0) | 6.0 (5.9 - 6.0) | 6.0 (5.9 - 6.0) |
| **Model A** | *Eligible quality:* | | | |
| | *Week, hrs* | 14.5 (13.0 - 16.0) | 13.9 (12.9 - 14.5) | 14.1 (11.9 - 15.9) |
| | *Week, %* | 66 (60 - 72) | 64 (58 - 68) | 62 (55 - 71) |
| | *Day, hrs* | 9.0 (7.5 - 10.5) | 8.3 (7.5 - 9.7) | 8.3 (6.3 - 10.1) |
| | *Day, %* | 55 (47 - 64) | 53 (44 - 58) | 50 (42 - 61) |
| | *Night, hrs* | 5.7 (5.4 - 5.9) | 5.8 (5.4 - 5.9) | 5.7 (5.1 - 5.9) |
| | *Night, %* | 97 (93 - 98) | 97 (93 - 99) | 95 (93 - 99) |
| **Model B** | *Eligible quality:* | | | |
| | *Week, hrs* | 6.1 (4.8 - 7.1) | 5.9 (5.1 - 7.2) | 5.5 (4.2 - 6.6) |
| | *Week, %* | 27 (22 - 32) | 27 (22 - 32) | 25 (21 - 31) |
| | *Day, hrs* | 2.2 (1.7 - 2.9) | 2.3 (1.6 - 2.9) | 2.0 (1.3 - 2.7) |
| | *Day, %* | 14 (10 - 18) | 14 (10 - 17) | 12 (8 - 17) |
| | *Night, hrs* | 3.8 (3.0 - 4.4) | 3.6 (3.0 - 4.4) | 3.6 (2.6 - 4.2) |
| | *Night, %* | 66 (52 - 75) | 62 (54 - 73) | 60 (46 - 69) |

**Dyskinesia**

In Table 5.3 the data availability for the different models comparing tertiles based on total dyskinesia scores. The no dyskinesia group had more eligible PPG data, compared to the mild and severe dyskinesia group, during 24 h (14.5 h; 13.9 h; 14.1 h) and daytime (9.0 h; 8.3 h; 8.3 h) In the severe dyskinesia group, the proportion of eligible PPG data was slightly lower during 24 h, daytime and nighttime.

In Model B, the no dyskinesia group had more eligible data, compared to the other tertiles during the 24 h (6.1 h; 5.9 h; 5.5 h) and nightime (3.8 h; 3.6 h; 3.6 h). With increased dyskinesia, the proportion and the absolute amount of eligible data decreased.

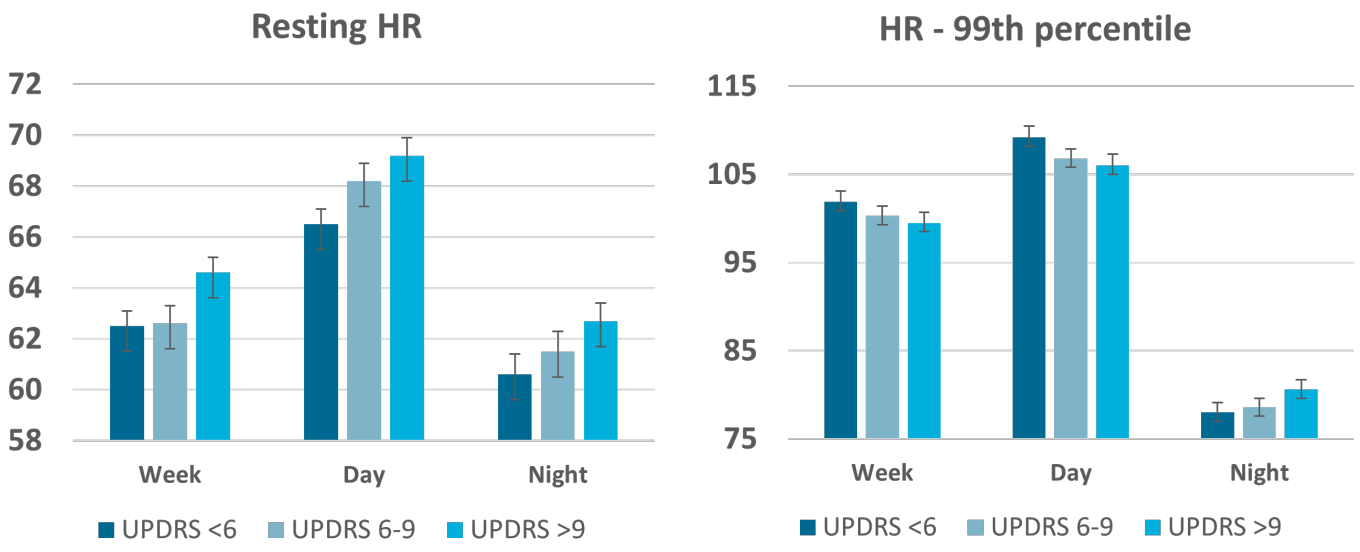
Table 5.4: Data availability of PPG in week 0 comparing the effect of race. Groups are divided into Caucasian race and other race. Data are presented as medians (IQRs) and percentages of the total wearing time; hrs = total number of hours per 24 hours (00:00-23:59), daytime (06:00-23:59), or nighttime (00:00-05:59). * = Mann-Whitney U test $p<0.05$ vs Caucasian.

| | | **Caucasian (N = 470)** | **Other (N = 14)** |
|---|---|---|---|
| **Total wearing time** | *24h (hrs)*[b] | 22.5 (21.9 - 22.8) | 22.5 (21.0 - 22.8) |
| | *Day (hrs)* | 16.6 (16.1 - 16.8) | 16.5 (15.4 - 16.8) |
| | *Night (hrs)* | 6.0 (5.9 - 6.0) | 6.0 (5.6 - 6.0) |
| **Model A** | *Eligible quality:* | | |
| | *Week, hrs* | 14.4 (13.0 - 15.9) | 13.5 (12.0 - 16.6) |
| | *Week, %* | 66 (59 - 71) | 62 (56 - 73) |
| | *Day, hrs* | 8.9 (7.5 - 10.3) | 7.9 (6.9 - 10.8) |
| | *Day, %* | 55 (47 - 63) | 50 (43 - 65) |
| | *Night, hrs* | 5.8 (5.4 - 5.9) | 5.8 (5.1 - 6.0) |
| | *Night, %* | 97 (93 - 98) | 97 (91 - 99) |
| **Model B** | *Eligible quality:* | | |
| | *Week, hrs* | 6.0 (4.8 - 7.1) | 5.0 (4.1 - 6.8) |
| | *Week, %* | 27 (22 - 32) | 23 (20 - 30) |
| | *Day, hrs* | 2.2 (1.7 - 2.9) | 1.9 (1.3 - 2.6) |
| | *Day, %* | 13 (10 - 18) | 11 (8 - 16) |
| | *Night, hrs* | 3.8 (3.0 - 4.4) | 3.6 (2.6 - 4.2) |
| | *Night, %* | 65 (52 - 74) | 64 (45 - 71) |

### Race

The last clinical characteristic which could influence was the race of the patient. In Table 5.4 the data availability was compared between caucasian race (N=470) and non-caucasian race (N=14).

No differences were found between wearing times. For Model A, the available data was lower, during 24 h and during the day, for the non-caucasian group. This difference was also seen in Model B. Furthermore, in Model B the availability was also less in the night for the non-caucasian group.




(a) Resting HR, defined as the modal HR

(b) Maximum HR, defined as the 99th percentile

Figure 5.3: The two different HR parameters compared within tertiles based on UPDRS Part I: Non-Motor Aspects of Experiences of Daily Living. HR parameters are plotted as means ± standard error during 24 h, during the day (6:00-23:59) and during the night (0:00 - 5:59)

### 5.3.4 Exploratory HR analysis

In the exploratory HR analysis, tertiles were created based on autonomic dysfunction scores of the 486 subjects. In Figure 5.3 the resting HR and maximum HR are shown. The resting HR (Figure 5.3a became higher, during 24h, day and night, with increasing autonomic dysfunction scores. The maximum HR (Figure 5.3b) showed an opposite trend during the week and day. The maximum HR during the night increased with more severe autonomic dysfunction scores.

## 5.4 Discussion

In this chapter, the first step was to evaluate whether the duration and proportion of high-quality PPG data varies according to subsequent time points during follow-up and differs in relation to clinical PD characteristics. For this analyses, we used the two binary LR models (model A: HR quality and model B: HRV quality) of the previous chapter. In both models, the total duration and the proportion of high quality PPG data, in terms of percentage of the total data, was significantly lower after one year during the 24 h average of the week, during the day and during the night. Using both models to classify data in the first week of the PPP study, we found no significant differences between tertiles with no, mild or severe PD-specific motor symptoms in the proportion of eligible data for analysis. In the last part of this chapter, we cross-sectionally performed an exploratory HR analysis in relation to the symptoms of autonomic dysfunction in PD patients. We showed that the resting HR increased in PD patients with more (severe) symptoms whereas the maximum HR showed a decreasing trend.

Similar to the distribution of eligible data in the annotation set (Figure 4.4), in the first two weeks 65% of the data is eligible for HR analysis. After one-year follow up, this percentage statistically significantly decreased to 60%. During the day, the percentages significantly decreased from 55% to 49% whereas in the night, the percentages significantly decreased from 97% to 95%. However, one could argue the clinical relevance of this decrease. In a previous study, regarding long-term PPG recordings, 67% of the data could be used in the night and 30.5% during the day [51]. However, in our study the percentage of eligible data is slightly overestimated since one-minute epochs containing less high-quality data are also calculated as a full minute of high quality data. For HRV analysis, slightly higher percentages of eligible data were seen in the first two weeks (Week 0: 27% – Week 1: 28%) compared to the 23% of the annotation set. After one-year follow up, this percentage significantly decreased to 24%. It seems that the percentages of eligible data quality remain stable over the course of the follow-up. However during daytime, the percentage decreased from only 14% to 11%, while during nighttime the percentages were decreasing from 65% to 58%. Charlton et al. obtained high-quality data, using a more advanced and larger device, in 90% during sleep whereas the percentage during daily activities ranged from 9% - 50% [43].

The statistically significant decrease over the one-year follow-up in data quality can be caused by multiple factors. Possible factors include fluctuations in commitment, increase in motor symptoms and/or a decline in the wearable device. However, finding an explanation is beyond the scope of this chapter and the question remains if this decline is clinical relevant. We concluded that over the course of one-year, we still have sufficient data for both HR analysis and HRV analysis.

After analyzing the clinical factors, we did not find statistical differences between study groups with no, mild or severe PD-specific motor symptoms. However using both models, a slight decrease in the amount and proportion of high-quality was seen with increasing tremor and dyskinesia scores. This numerical difference is the highest while comparing dyskinesia scores. This is in line with the hypothesis that longer and more severe involuntary movements during the day, causes less high-quality PPG-epochs. Furthermore, we saw no differences between the Caucasian race study group and the other race study group.

In our final HR analysis, using the eligible data classified by model A, we found that in PD patients with more severe symptoms of autonomic dysfunction the resting HR was higher, whereas the maximum HR was lower during the day. This is in accordance with the results from previous studies using ECG data to assess HR parameters. These studies showed an impairement in heart rate regulation with more severe PD symptoms [30, 31]. These first results are promising in our quest towards new digital biomarkers to assess autonomic dysfunction in PD patients.

### 5.4.1 Methodological considerations

We expected not to see differences in eligible data within the first weeks. Therefore, the reason why we analyzed Week 0 vs. Week 1 was to see the test-retest reliability. However, statistical differences in the total amount of eligible PPG data were found within the first two weeks. Since the proportion is not statistically different between the two weeks, this can be explained by the fact that the total wearing time also significantly differs in these weeks.

In this study, we analyzed the influence of two PD-specific motor symptoms on the amount of eligible data. For the tremor score, we created tertiles based on the rest tremor score at the arm of the device side. For analysis of dyskinesia as a possible influence, we created tertiles based on the summed score of the duration of the dyskinesia score and the functional impact of the dyskinesia score. This is, according to the official UPDRS questionnaire, the most reliable approach to describe the severity of the dyskinesia [8]. The tertiles were imbalanced in terms of number of subjects. This imbalance is also seen in the last factor we analyzed, the influence of skin color. Especially in the this factor, the results of this analysis should be interpreted with great caution.

In the exploratory HR analysis, there were no significant differences in age between the quartiles. This is important, since age has been considered an important confounder for cardiac autonomic dysfunction [52, 53]. Furthermore, we chose to built-in an extra quality check. Therefore, we divided every PPG epoch into mini-epochs of 6 s and classified these as high or low-quality. Thereafter, we extracted from the mini-epochs the dominant frequency. However, this approach required an extra classification before doing the HR analysis.

### 5.4.2 Future perspectives

Our first HR analyzes showed promising results in the quest towards new biomarkers to assess autonomic dysfunction in Parkinson's. However, there are several steps which needs to be taken for further research.

The first step could be to determine HR parameters over the follow-up of the study. This would allow us to gain insights into how reliable our HR parameters in relation to autonomic dysfunction over the different weeks. To perform this research, it is important to establish the length of consecutive high-quality PPG data needed for reliable parameter estimation and the interval at which these parameters should be calculated during the day and during the night.

The second step may be to increase the amount of HRV-suitable data throughout the day. One way to expand this is to see if we can add category 2 labels (50-95% high-quality data). By doing so, we can double our eligible data during the day. However, we could also initially focus on the nighttime. We have sufficient data for this analysis and it is precisely during the night that there are indications, according to various studies, that cardiac regulation is most disturbed in people with Parkinson's [10, 11].

## 5.5 Conclusion

In this study on potential factors influencing PPG data quality we found no impact of time and clinical factors one the amount of high-quality PPG data. In our first HR analysis, studying the resting HR and maximum HR, we showed promising associations between these HR parameters and autonomic dysfunction scores. These findings provide a stimulus for further research in the quest towards digital biomarkers.

# 6 | General discussion and future perspectives

## 6.1 General discussion

To our knowledge, this thesis is the first study to use long-term PPG data of PD patients, measured in an ambulatory setting, for heart rate analyses. The first step in this thesis was to develop supervised ML models to assess data quality using handcrafted features. Thereafter, we examined whether their were differences in data quality over the course of the PPP follow-up period and between patients with different severity of motor symptoms. Lastly, we used the ML models to perform an exploratory HR analysis.

In chapter 4 we showed that we can develop different models to determine data quality of the PPG using custom time and frequency characteristics. These models were developed in view of two heart rate analyses, HR and HRV, both of which require a different level of data quality. All classifiers performed well in this task, with balanced accuracies $> 93$ % for the HR model and balanced accuracies $> 94$ % for the HRV model. However, the most efficient classifier for this task was LR. These binary LR models also showed fairly similar results in the validation of the nested CV, indicating a high generalizability of the models to a larger population.

In chapter 5, we showed that the proportion of data quality was significantly lower after one year follow-up. However, we still obtained a sufficient amount of data quality. Futhermore, we concluded that there was no significant influence of motor symptoms on the amount of eligible data. In our first HR analysis, we showed promising first results in linking HR parameters in relation to the severity of autonomic dysfunction. These results supported the hypothesis that HR regulation is disturbed in PD patients with more symptoms of autonomic dysfunction. .

## 6.2 Future perspectives

The last 25 years the clinical-pathological concept of PD has been challenged. Despite remarkable progress in our understanding of many aspects of PD, the different aspects remain a topic of interest since no cures or treatments have been developed that are able to halt their relentless progression. Using the current diagnostic criteria, PD is clinically diagnosed when disease progression is already advanced. The latent phase of PD can vary from 5 to 20 years is called the prodromal phase of PD [54]. This phase represents an opportunity for earlier diagnosis, investigation in the pathophysiology and to prevent or

slow down the development of motor symptoms. The proposed methods to assess heart rate patterns in this study could assist in providing new insights in the pathogenesis of the autonomic nervous system. We could also use these insights in establishing the different phenotypes within our PD population.

The primary focus for further research should be to define universal definitions for segment length and the required intervals at which HR(V) parameters should be estimated. Thereafter, we should focus on performing large-scale HR and HRV analysis over a long-term period of years. This could provide reliable estimates of the autonomic function within a PD patient. The recommendation would be to start performing the analyses in the HR in the most controlled situation during the circadian cycle: the night. Even more, since emerging evidence suggests that especially during the night most signs of autonomic dysregulation, such as sleep disorders, could be present [11, 55]. However, one should be thoughtful when proposing a reliable method to analyze HR(V) parameters during the night since it is well established that sleep stages play a significant role in the parameter values [56, 57].

Using our proposed methods, during the night we could use around 65% of the data for HRV analysis and up to 97% for HR analysis. In this unique, prospective and longitudinal cohort study, this implicates we could use hours of PPG data per day for analysis over the course of a two-year follow-up in 500 subjects. This unique size of data availability enables us to average out the effect of the sleep stages.

Another relevant focus is to integrate more physiological signals, obtained by the study watch, into a multi-modality tool. Besides PPG, the study watch continuously measures acceleration and the electrodermal activity. These two signals could give insights into activity and stress levels of PD subjects at the time of the PPG recordings. Besides the proposed method to integrate acceleration data for quality control, this can give us more reliable estimates of the HR(V) parameters. Furthermore, such a multi-modality tool would enable to couple motor symptoms to autonomic functioning to obtain a much broader view towards patient-tailored long-term follow up. This may lead to novel insights into the exact pathogenesis in various phenotypes of PD.

My conviction is that the knowledge gap regarding to autonomic dysfunction in Parkinson's disease could be bridged in the coming years by further elaborating research on patient-tailored long-term follow up in an ambulatory setting. This will provide us one day such an extensive knowledge to develop treatments to halt and possibly cure the relentless disease progression in PD.

# Bibliography


[1] Alberto Ascherio and Michael A. Schwarzschild. The epidemiology of Parkinson's disease: risk factors and prevention, 11 2016.

[2] E. Ray Dorsey, Todd Sherer, Michael S. Okun, and Bastiaan R. Bloem. The emerging evidence of the Parkinson pandemic, 2018.

[3] Daniele Caligiore, Rick C Helmich, Mark Hallett, Ahmed A Moustafa, Lars Timmermann, Ivan Toni, and Gianluca Baldassarre. Parkinson's disease as a system-level disorder. 2016.

[4] Ronald B. Postuma, Dag Aarsland, Paolo Barone, David J. Burn, Christopher H. Hawkes, Wolfgang Oertel, and Tjalf Ziemssen. Identifying prodromal Parkinson's disease: Pre-Motor disorders in Parkinson's disease. Movement Disorders, 27(5):617–626, 4 2012.

[5] Yehonatan Sharabi, Gad D. Vatine, and Avraham Ashkenazi. Parkinson's disease outside the brain: targeting the autonomic nervous system, 2021.

[6] Connie Marras and K. Ray Chaudhuri. Nonmotor features of Parkinson's disease subtypes. Movement Disorders, 31(8), 2016.

[7] Julia C. Greenland, Caroline H. Williams-Gray, and Roger A. Barker. The clinical heterogeneity of Parkinson's disease and its therapeutic implications. The European journal of neuroscience, 49(3):328–338, 2 2019.

[8] Christopher G Goetz, Stanley Fahn, Pablo Martinez-Martin, Werner Poewe, Cristina Sampaio, Glenn T Stebbins, Matthew B Stern, Barbara C Tilley, Richard Dodel, Bruno Dubois, Robert Holloway, Joseph Jankovic, Jaime Kulisevsky, Anthony E Lang, Andrew Lees, Sue Leurgans, Peter A Lewitt, David Nyenhuis, Warren Olanow, Olivier Rascol, Anette Schrag, Jeanne A Teresi, Jacobus J Van Hilten, and Nancy Lapelle. MDS-UPDRS. Technical report, 2008.

[9] B R Bloem, W J Marks, A L Silva De Lima, M L Kuijf, T Van Laar, B P F Jacobs, M M Verbeek, R C Helmich, B P Van De Warrenburg, L J W Evers, J Inthout 10, T Van De Zande, T M Snyder, R Kapur, and M J Meinders. BMC Neurology.

[10] L. Ferini-Strambi, M. Franceschi, P. Pinto, M. Zucconi, and S. Smirne. Respiration and heart rate variability during sleep in untreated Parkinson patients. Gerontology, 38(1-2):92–98, 1992.

[11] V. Arnao, A. Cinturino, S. Mastrilli, C. Buttà, C. Maida, A. Tuttolomondo, P. Aridon, and M. D'Amelio. Impaired circadian heart rate variability in Parkinson's disease: A time-domain analysis in ambulatory setting. BMC Neurology, 20(1):1–5, 4 2020.

[12] N. Pinheiro, R. Couceiro, J. Henriques, J. Muehlsteff, I. Quintal, L. Goncalves, and P. Carvalho. Can PPG be used for HRV analysis? In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, volume 2016-October, pages 2945–2949. Institute of Electrical and Electronics Engineers Inc., 10 2016.

[13] Axel Schäfer and Jan Vagedes. How accurate is pulse rate variability as an estimate of heart rate variability? A review on studies comparing photoplethysmographic technology with an electrocardiogram. International journal of cardiology, 166(1):15–29, 6 2013.

[14] G. Lu, F. Yang, J. A. Taylor, and J. F. Stein. A comparison of photoplethysmography and ECG recording to analyse heart rate variability in healthy subjects. Journal of medical engineering & technology, 33(8):634–641, 11 2009.

[15] Yifan Zhang, Shuang Song, Rik Vullings, Dwaipayan Biswas, Neide Simões-Capela, Nick Van Helleputte, Chris Van Hoof, and Willemijn Groenendaal. Motion Artifact Reduction for Wrist-Worn Photoplethysmograph Sensors Based on Different Wavelengths. Sensors 2019, Vol. 19, Page 673, 19(3):673, 2 2019.

[16] Hyonyoung Han, Min Joon Kim, and Jung Kim. Development of real-time motion artifact reduction algorithm for a wearable photoplethysmography. Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference, 2007:1538–1541, 2007.

[17] Yue Zhang and Junjun Pan. Assessment of photoplethysmogram signal quality based on frequency domain and time series parameters. Proceedings - 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics, CISP-BMEI 2017, 2018-January:1–5, 2 2018.

[18] Hongxing Li and Shizhao Huang. A High-Efficiency and Real-Time Method for Quality Evaluation of PPG Signals. IOP Conference Series: Materials Science and Engineering, 711(1):012100, 1 2020.

[19] B. R. Bloem, W. J. Marks, A. L. Silva De Lima, M. L. Kuijf, T. Van Laar, B. P.F. Jacobs, M. M. Verbeek, R. C. Helmich, B. P. Van De Warrenburg, L. J.W. Evers, J. Inthout, T. Van De Zande, T. M. Snyder, R. Kapur, and M. J. Meinders. The Personalized Parkinson Project: Examining disease progression through broad biomarkers in early Parkinson's disease. BMC Neurology, 19(1):160, 7 2019.

[20] Per Borghammer. Is constipation in Parkinson's disease caused by gut or brain pathology?, 10 2018.

[21] Juan Segura-Aguilar, Irmgard Paris, Patricia Muñoz, Emanuele Ferrari, Luigi Zecca, and Fabio A. Zucca. Protective and toxic roles of dopamine in Parkinson's disease. Journal of Neurochemistry, 129(6):898–915, 2014.

[22] Bastiaan R. Bloem, Michael S. Okun, and Christine Klein. Parkinson's disease. The Lancet, 397(10291):2284–2303, 6 2021.

[23] Per Borghammer and Nathalie Van Den Berge. Brain-First versus Gut-First Parkinson's Disease: A Hypothesis, 2019.

[24] Conor Fearon, Anthony E Lang, Alberto J Espay, Edmond J Safra Program in Parkinson, and Gloria Shulman. The Logic and Pitfalls of Parkinson's Disease as "Brain-First" Versus "Body-First" Subtypes. Movement Disorders, 36(3):594–598, 3 2021.

[25] Horacio Kaufmann, Lucy Norcliffe-Kaufmann, Jose Alberto Palma, Italo Biaggioni, Phillip A. Low, Wolfgang Singer, David S. Goldstein, Amanda C. Peltier, Cyndia A. Shibao, Christopher H. Gibbons, Roy Freeman, and David Robertson. Natural history of pure autonomic failure: A United States prospective cohort. Annals of Neurology, 81(2):287–297, 2 2017.

[26] David S. Goldstein and Yehonatan Sharabi. The heart of PD: Lewy body diseases as neurocardiologic disorders. Brain Research, 1702:74–84, 1 2019.

[27] David S. Goldstein, Yehonatan Sharabi, Barbara I. Karp, Oladi Bentho, Ahmed Saleem, Karel Pacak, and Graeme Eisenhofer. Cardiac sympathetic denervation preceding motor signs in Parkinson disease. Clinical Autonomic Research, 17(2):118–121, 4 2007.

[28] Gari D Clifford. Signal Processing Methods for Heart Rate Variability. Technical report.

[29] Alvaro Alonso, Xuemei Huang, Thomas H. Mosley, Gerardo Heiss, and Honglei Chen. Heart rate variability and the risk of Parkinson disease: The Atherosclerosis Risk in Communities study. Annals of Neurology, 77(5):877–883, 5 2015.

[30] T. H. Haapaniemi, V. Pursiainen, J. T. Korpelainen, H. V. Huikuri, K. A. Sotaniemi, and V. V. Myllylä. Ambulatory ECG and analysis of heart rate variability in Parkinson's disease. Journal of Neurology Neurosurgery and Psychiatry, 70(3):305–310, 3 2001.

[31] Konstantin G. Heimrich, Thomas Lehmann, Peter Schlattmann, and Tino Prell. Heart Rate Variability Analyses in Parkinson's Disease: A Systematic Review and Meta-Analysis. Brain Sciences 2021, Vol. 11, Page 959, 11(8):959, 7 2021.

[32] Williebosseau Murray and Patrick Anthony Foster. The peripheral pulse wave: Information overlooked, 1996.

[33] Hui Wen Loh, Shuting Xu, Oliver Faust, Chui Ping Ooi, Prabal Datta Barua, Subrata Chakraborty, Ru-San Tan, Filippo Molinari, and U Rajendra Acharya. Application of photoplethysmography signals for healthcare systems: An in-depth review. Computer Methods and Programs in Biomedicine, 216:106677, 4 2022.

[34] Denisse Castaneda, Aibhlin Esparza, Mohammad Ghamari, Cinna Soltanpur, and Homer Nazeran. A review on wearable photoplethysmography sensors and their potential future applications in health care.

[35] Chungkeun Lee, Hang Sik Shin, and Myoungho Lee. Relations between ac-dc components and optical path length in photoplethysmography. Journal of Biomedical Optics, 16(7), 2011.

[36] Fred Shaffer and J. P. Ginsberg. An Overview of Heart Rate Variability Metrics and Norms. Frontiers in Public Health, 5:258, 9 2017.

[37] Ch Kiran kumar, M. Manaswini, K. N. Maruthy, A. V. Siva Kumar, and K. Mahesh kumar. Association of Heart rate variability measured by RR interval from ECG and pulse to pulse interval from Photoplethysmography. Clinical Epidemiology and Global Health, 10, 2021.

[38] Victor E. Staartjes, Luca Regli, and Carlo Serra. Machine Intelligence in Clinical Neuroscience: Taming the Unchained Prometheus. Acta neurochirurgica. Supplement, 134:1–4, 2022.

[39] Saime Akdemir Akar, Sadik Kara, Fatma Latifoğlu, and Vedat Bilgiç. Spectral analysis of photoplethysmographic signals: The importance of preprocessing. Biomedical Signal Processing and Control, 8(1):16–22, 1 2013.

[40] Mohamed Elgendi. Optimal Signal Quality Index for Photoplethysmogram Signals. Bioengineering, 3(4), 12 2016.

[41] E.-S Väliaho, P Kuoppa, J A Lipponen, T J Martikainen, H Jäntti, T T Rissanen, I Kolk, M Castrén, J Halonen, M P Tarvainen, and J E K Hartikainen. Wrist band photoplethysmography in detection of individual pulses in atrial fibrillation and algorithm-based detection of atrial fibrillation.

[42] Gaël Vila, Christelle Godin, Sylvie Charbonnier, and Aurélie Campagne. Real-Time Quality Index to Control Data Loss in Real-Life Cardiac Monitoring Applications. Sensors (Basel, Switzerland), 21(16), 8 2021.

[43] Peter H Charlton, Panicos Kyriacou, Jonathan Mant, and Jordi Alastruey. Acquiring Wearable Photoplethysmography Data in Daily Life: The PPG Diary Pilot Study †. 15, 2020.

[44] Christina Orphanidou, Timothy Bonnici, Peter Charlton, David Clifton, David Vallance, and Lionel Tarassenko. Signal-quality indices for the electrocardiogram and photoplethysmogram: derivation and applications to wireless monitoring. IEEE journal of biomedical and health informatics, 19(3):832–838, 5 2015.

[45] Rahul Pandey, Carlos Castillo, and Hemant Purohit. Modeling Human Annotation Errors to Design Bias-Aware Systems for Social Stream Processing.

[46] Nikhilesh Pradhan, Sreeraman Rajan, and Andy Adler. Evaluation of the signal quality of wrist-based photoplethysmography. Physiological Measurement, 40(6):065008, 7 2019.

[47] Serena Moscato, Stella Lo Giudice, Giulia Massaro, and Lorenzo Chiari. Wrist Photoplethysmography Signal Quality Assessment for Reliable Heart Rate Estimate and Morphological Analysis. Sensors (Basel, Switzerland), 22(15), 8 2022.

[48] Edward Joseph Caruana, Marius Roman, Jules Hernández-Sánchez, and Piergiorgio Solli. Longitudinal studies. Journal of Thoracic Disease, 7(11):E537, 2015.

[49] Peter J. Colvonen. Response To: Investigating sources of inaccuracy in wearable optical heart rate sensors. npj Digital Medicine 2021 4:1, 4(1):1–2, 2 2021.

[50] Michael W. Sjoding, Robert P. Dickson, Theodore J. Iwashyna, Steven E. Gay, and Thomas S. Valley. Racial Bias in Pulse Oximetry Measurement. New England Journal of Medicine, 383(25):2477–2478, 12 2020.

[51] Eemu Samuli Väliaho, Pekka Kuoppa, Jukka A. Lipponen, Juha E.K. Hartikainen, Helena Jäntti, Tuomas T. Rissanen, Indrek Kolk, Hanna Pohjantähti-Maaroos, Maaret Castrén, Jari Halonen, Mika P. Tarvainen, Onni E. Santala, and Tero J. Martikainen. Wrist Band Photoplethysmography Autocorrelation Analysis Enables Detection of Atrial Fibrillation Without Pulse Detection. Frontiers in Physiology, 12:576, 5 2021.

[52] David A. Gelber, Michael Pfeifer, Beth Dawson, and Mary Schumer. Cardiovascular autonomic nervous system tests: determination of normative values and effect of confounding variables. Journal of the autonomic nervous system, 62(1-2):40–44, 1997.

[53] Rachna Parashar, Mohammed Amir, Abhijit Pakhare, Preeti Rathi, and Lalita Chaudhary. Age Related Changes in Autonomic Functions. Journal of Clinical and Diagnostic Research : JCDR, 10(3):CC11, 3 2016.

[54] Jeremy Hunt, Elizabeth J. Coulson, Rajendram Rajnarayanan, Henrik Oster, Aleksandar Videnovic, and Oliver Rawashdeh. Sleep and circadian rhythms in Parkinson's disease and preclinical models. Molecular Neurodegeneration, 17(1):1–21, 12 2022.

[55] Eldbjørg Hustad and Jan O. Aasly. Clinical and Imaging Markers of Prodromal Parkinson's Disease. Frontiers in Neurology, 11:395, 5 2020.

[56] Emilio Vanoli, Philip B. Adamson, Ba-Lin, Gian D. Pinna, Ralph Lazzara, and William C. Orr. Heart Rate Variability During Specific Sleep Stages. Circulation, 91(7):1918–1922, 4 1995.

[57] David Herzig, Prisca Eser, Ximena Omlin, Robert Riener, Matthias Wilhelm, and Peter Achermann. Reproducibility of heart rate variability is parameter and sleep stage dependent. Frontiers in Physiology, 8(JAN):1100, 1 2018.

# A | Optimization Methods

The start of an optimization method is the initialization of the weights. Subsequently, the algorithm approximates the gradient of the loss function and makes a step in the direction of the steepest descent. After updating the weights, a new approximation of the gradient is calculated at this new point. In Section 3.2.2, we used $L_{CE}$ as our loss function. The general form of the loss function can be expressed as the function $E(w)$, where the loss function is parametrized by the weights of the model. We'll introduce two different optimization methods: Quasi-Newton methods, using an efficient form of the traditional Newton's methods, and stochastic gradient descent.

## A.1 Quasi-Newton methods

In Quasi-Newton methods, such as the limited BFGS method, the loss function $E(w)$, is approximated with a second order Taylor expansion:

$$E(w+\delta) \approx E(w) + \left(\frac{\partial E}{\partial w}\right)^T \delta + \frac{1}{2!}\delta^T \frac{\partial^2 E}{\partial w^2}\delta, \tag{A.1}$$

where $E(w)$ is the error function, $\mathbf{w}$ the vector containing the weights and $\delta$ a small step. Note that $\frac{\partial E}{\partial w}$ represents the gradient$\nabla$ and $\frac{\partial^2 E}{\partial w^2}$ the Hessian matrix. For every iteration, the algorithm approximates the first and second derivative of $E$ with respect to the weights, the approximated gradient $\hat{\nabla}$ and Hessian matrix $\mathbf{B}$, respectively. Using Equation (A.1) gives:

$$\hat{E}(\mathbf{w}+\delta) \approx E(\mathbf{w}) + \hat{\nabla}E(\mathbf{w})^T\delta + \frac{1}{2}\delta^T\mathbf{B}\delta. \tag{A.2}$$

To find a step $\delta$ that yields the minimum of this second order approximation of the error function, we set the derivative of Equation A.2, with respect to $\delta$, to zero:

$$\begin{aligned} &\hat{\nabla}E(\mathbf{w})^T + \mathbf{B}\delta = 0 \Rightarrow \\ &\delta = -\mathbf{B}^{-1}\hat{\nabla}E(\mathbf{w})^T. \end{aligned} \tag{A.3}$$

The minimum for updating the weights then becomes:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \mathbf{B}^{-1}\hat{\nabla}\mathrm{E}\left(\mathbf{w}_t\right)^T \tag{A.4}$$

The advantages of quasi Newton method is that it converges to the global minimum in relative few iterations. Also, since it approximates instead of calculating the Hessian, it is less computational expensive than in Newton's method.

## A.2 Stochastic gradient descent

A simpler form of an optimization method is the stochastic gradient descent (SGD) method. This first-order derivative method, is often used in training a ML algorithm. SGD is a method that finds a minimum of a function by calculating which direction the slope is rising the most steeply, and moving the opposite direction. In this thesis, only convex loss functions are used. These loss functions have at most one minimum; there are no local minima to get stuck in. The gradient descent algorithm finds the vector pointing in the direction of the increase in the function. The amount to move in SGD is the value of the slope ($\frac{d}{dw}E(\mathbf{w_t})$ weighted by a learning rate $\eta$. This results in the following function for updating the weights:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta * \frac{d}{dw}E(\mathbf{w_t}), \tag{A.5}$$

where $\frac{d}{dw}E(\mathbf{w_t})$ is the gradient of the loss function with respect to the weights $\mathbf{w}$ and $\eta$ the learning rate of the method. A main disadvantage of the SGD algorithm is that choosing the learning rate can be difficult. A high learning rate means that $\mathbf{w}$ is changed more on each step but it can overshoot the minimum of the convex loss function. A small learning rate means that the function converges more slowly to the global minimum.

# B | Annotation GUI

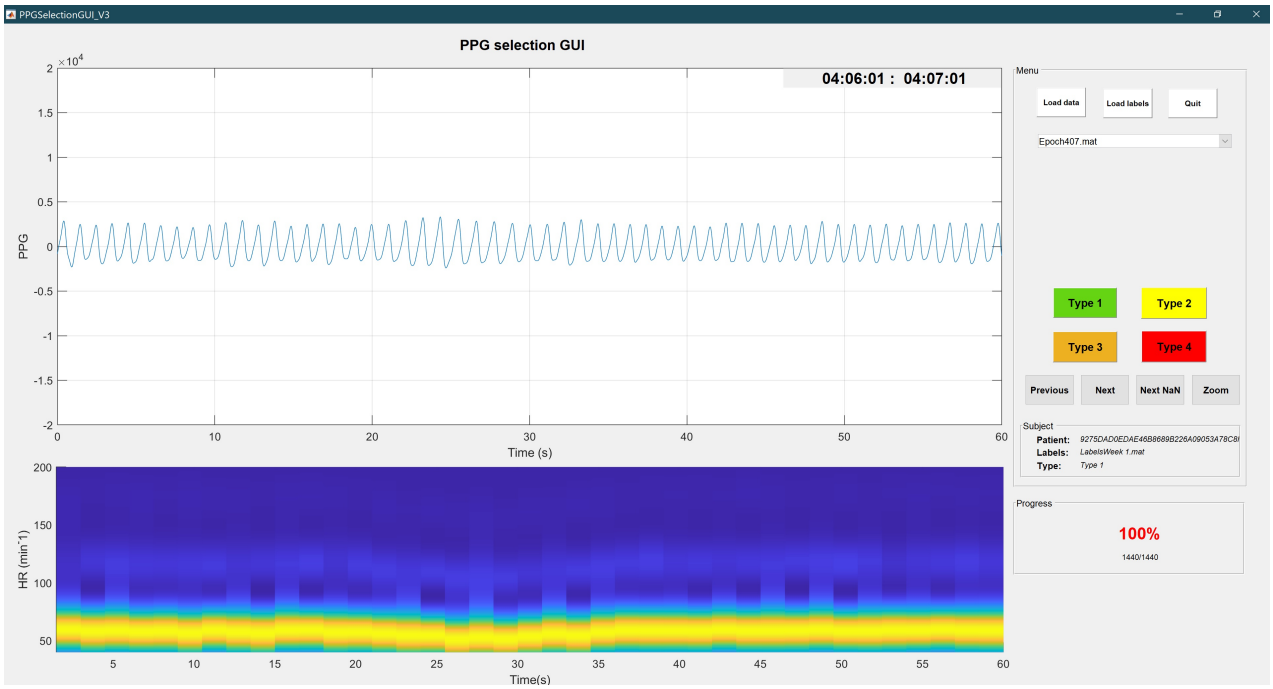



Figure B.1: Overview of the Graphical User Interface (GUI) in Matlab. This GUI was used for annotating the PPG epochs. Upper panel: time-series PPG signal. Lower panel: corresponding spectrogram.

# C | Feature definitions

Table C.1: Description of all extracted features from the PPG data. N is equal to the number of samples in the 1-min epoch, M is equal to the number of samples in the mini-epoch, A the area calculated using the trapezoidal rule and n the number of mini-epochs.

| **Feature category** | **Feature description** | **Formula (if applicable)** | **Number** |
|---|---|---|---|
| *One minute epoch-based features* | Standard deviation ($\sigma$) | $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\lvert PPG_i - \mu\rvert}$ | 1 |
| | Mean ($\mu$) | $\mu = \frac{1}{N}\sum_{i=1}^{N} PPG_i$ | 2 |
| | Median | $M = \frac{PPG\left(\frac{N}{2}\right)^{+PPG}\left(\frac{N}{2}+1\right)}{2}$, for even numbers | 3 |
| | Skewness (s) | $s = \frac{E(x-\mu)^3}{\sigma^3} = \frac{\frac{1}{N}\sum_{i=1}^{N}(PPG_i-\mu)^3}{\sigma^3}$ | 4 |
| | Kurtosis (k) | $k = \frac{E(x-\mu)^4}{\sigma^4} = \frac{\frac{1}{N}\sum_{i=1}^{N}(PPG_i-\mu)^4}{\sigma^4}$ | 5 |
| | Dominant frequency ($F_{dom}$) | Maximum power in the frequency domain | 6 |
| | Regularity index (RI) | $RI = \frac{A(\text{Fdom } \pm 0.2)}{A(\text{Ftotal })}$ | 7 |
| | Organization index | $OI = \frac{A(\text{Fdom}\pm 0.2)+\sum_{i=1}^{2} A((i+1)*Fdom\pm 0.2)}{A(\text{Ftotal})}$ | 8 |
| | *Mini-epoch mean function* | $\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i$ | |
| *Mini-epoch based features* | Variation of the standard deviation ($\sigma_{\sigma_{mini}}$) | $\sigma_{\text{mini}}(n) = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left\lvert PPG(n)_{\text{mini}(i)} - \mu_{PPG(n)_{\text{mini}}}\right\rvert}$ | 9 |
| | | $\sigma_{\sigma_{\text{mini}}} = \left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\lvert\sigma_{\text{mini}(i)} - \mu_{\sigma_{\text{mini}}}\right\rvert}\right)$ | |
| | Variation of the mean ($\sigma_{\mu_{mini}}$) | $\mu_{\text{mini}}(n) = \frac{1}{M}\sum_{i=1}^{M} PPG_{\text{mini}(i)}(n)$ | 10 |

| | | | |
|---|---|---|---|
| | | $\sigma_{\mu_{\text{mini}}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|\mu_{\text{mini}(i)} - \mu_{\mu_{\text{mini}}}\right\|)}$ | |
| | Variation of the median ($\sigma_{M_{mini}}$) | $M_{\text{mini}}(n) = \frac{1}{2}PPG(n)_{\left(\frac{M}{2}\right)} + PPG(n)_{\left(\frac{M}{2}+1\right)}$ | 11 |
| | | $\sigma_{M_{\text{mini}}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|M_{\text{mini}(i)} - \mu_{M_{\text{mini}}}\right\|)}$ | |
| | Variation of the skewness ($\sigma_{s_{mini}}$) | $S_{\text{mini}}(n) = \frac{E(x-\mu)^3}{\sigma^3_{\text{mini}}} = \frac{\frac{1}{M}\sum_{i=1}^{M}\left(PPG(n)_{\text{mini}(i)} - \mu\right)^3}{\sigma^3_{\text{mini}}}$ | 12 |
| | | $\sigma_{s_{\text{mini}}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|s_{\text{mini}(i)} - \mu_{s_{\text{mini}}}\right\|)}$ | |
| | Variation of the kurtosis ($\sigma_{k_{mini}}$) | $k_{\text{mini}}(n) = \frac{E(x-\mu)^4}{\sigma^4} = \frac{\frac{1}{M}\sum_{i=1}^{M}\left(PPG(n)_{\text{mini}(i)} - \mu\right)^3}{\sigma^4_{\text{mini}}}$ | 13 |
| | | $\sigma_{k_{\text{mini}}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|k_{\text{mini}(i)} - \mu_{k_{\text{mini}}}\right\|)}$ | |
| | Variation of the dominant frequency ($\sigma_{F_{dom}mini}$) | $\sigma_{F_{dom}mini} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|F_{\text{dom mini }(i)} - \mu_{F_{\text{dom mini}}}\right\|)}$ | 14 |
| | Quality index (QI) | $QI = \sum_{i=1}^{n}\text{Class}\left(PPG_{\text{mini }(i)} = 1\right)$ | 15 |
| *Spectrogram features* | Maximum sequence | N/A | 16 |
| | Sum of sequences | N/A | 17 |
| | Standard deviation | N/A | 18 |

# D | Feature selection using the MRMR algorithm

Feature selection, with the main purpose of accurately identify a subset of features among a larger set of features that are relevant for predicting an outcome, is one of the most important preprocessing steps in machine learning. Most feature selection algorithms can be broadly classified in two categories: minimal-optimal and all-relevant.

Minimal-optimal methods seek to have the maximum possible predictive power with one as small as possible feature set. The all-relevant algorithms seek to select all features that, individually, have any predictive power at all. Thus, if feature 1 and 2 are both relevant, but they are highly correlated (say they bring the same information), an all-relevant method will select them both, whereas a minimal-optimal method will select only one.

Maximum Relevance – Minimum Redundancy (MRMR) is a feature selection algorithm that uses such a minimal-optimal method. It selects features with a high correlation with the class/output and a low correlation between the previous selected features. For continuous features, as in this study, the F-statistic can be used to calculate correlation with the class (to maximize relevance) and the Pearson correlation coefficient is used to calculate correlation between features (to minimize redundancy).

MRMR works iteratively. At the first iteration, the feature with the highest correlation to the output variable is selected. At each following iteration, it defines the best features according to a formula and adds this to the basket of selected features. Once a feature goes into the bucket, it never comes out. At each iteration i, a score is calculated for each feature to be evaluated:

$$\text{Score}_i(f) = \frac{\text{relevance } (f \mid \text{ target })}{\text{redundancy } (f \mid \text{ features })} = \frac{F(f,\ \text{target })}{\sum_{s\in \text{ features selected until } i-1} |\operatorname{corr}(f,s)|/(i-1)} \tag{D.1}$$

The best feature at a particular iteration is the feature having the highest score. This means that the feature has overall the best ratio in correlation with the output variable and the correlation of the already selected features. One can simply iterate until the desired amount of features is reached.

# E | Matlab implementation of nested CV

The nested CV procedure (Figure E.1) was implemented in Matlab. The pseudo-code of the outer-loop implementation is given in Algorithm 1:

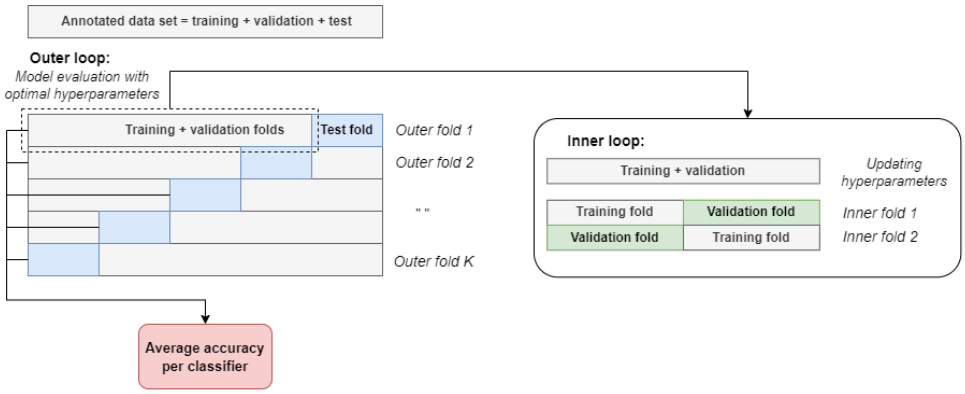



Figure E.1: Overview of the nested CV procedure.

**Algorithm 1** Outer loop of the nested CV

1: Initialization of the algorithm: load feature matrices with corresponding labels, adjust labels for specific binary classification, select most optimal feature set and specify the classifier.
2: **for** $i = 1 : K_{outer}$ **do**
3: Assign two subjects to the test fold
4: The remaining 18 subjects form the training + validation fold
5: Normalize the training + validation using Min-Max normalization
6: Normalize the test fold using the Min-Max values of the training + validation fold
7: Enter the inner loop **(Algorithm 2)**
8: Return the best hyperparameters
9: Train model on training + validation fold using the best hyperparameter values
10: Calculate performance measures on classification of the test fold
11: **end for**
12: Calculate mean performance metrics to compare results of the classifiers

The pseudo-code of the inner loop is given in Algorithm 2:

**Algorithm 2** Inner loop of the nested CV

```
Input: Train + validation folds
Create two folds (K_inner) of 9 subjects. One serves as the training fold while the other one is the validation fold (conversely in the second iteration)
Define hyperparameters HP_1, ..., HP_X
for i = 1 : length(HP_1) do
    for i = 1 : length(HP_X) do
        for i = 1 : K_inner do
            Train model on specific hyperparameter configuration using the training fold
            Calculate the loss on the validation fold
        end for
    end for
end for
Average the loss over the two folds
```