A Machine Learning Approach to Objective Measurement of Tremor Severity in Parkinson's Disease: Clinical and User Perspectives on Wearable Devices

Ghayth Mohammad Maddallah ALMahadin

School of Science and Technology

A thesis submitted in partial fulfilment of the requirements of

Nottingham Trent University for the degree of

Doctor of Philosophy

September 2021

This thesis is dedicated to my father, mother, wife and and family, my son Adam and my daughter Tala with great gratitude. I know you will be proud of this milestone accomplished. Undoubtedly, without their prayers and support this thesis would have been impossible.

Acknowledgements

The work described in this thesis was carried out at Nottingham Trent University between January 2018 and September 2021, while I was working as a full-time doctoral research student.

All pray and glory to Allah Almighty for making this a reality.

This thesis would not have been possible without the inspiration and support of a number of wonderful individuals. My thanks and appreciation to all of them for being part of this journey and making this thesis possible. I owe my utmost sincere gratitude to my honourable director of studies, Prof. Philip Breedon, and my second supervisor Prof. Ahmad Lotfi for their guidance over the last four years. Your consistent motivation and kind words will never be forgotten. Many thanks for making your office accessible despite your busy schedules. Also, your relentless effort and extensive suggestions around my work were indispensable to the accomplishment of this research. Also, I would like to extend a huge thank to the team of co-supervisors I had at different stages of my research: Dr. Eva Zysk and Mrs. Marie Mc Carthy, I am grateful for the creative ideas you always had to offer and for the many corrections you made to the publications leading to this thesis. I would like also to thank Prof. Amin Al-Habaibeh for his guidance and support. Also, I would like to thank Mr. Paul Watts for his help in data collection and Dr. Francesco Luke Siena for his help in data analysis and correction in early stage of the research.

My sincere appreciation to my family for their unconditional support, sympathy, love, encouragement and patience throughout the journey of this study through the course of my research. To my mother and my father words fail me to describe what a blessing it is to be your son. Your love, constant prayers, words of wisdom and discipline have contributed in my journey to achieving this feat. To my beloved wife and our lovely kids, Adam and Tala. Your joy and love have always put me up during the well-being and sadness of the last four years. To my brothers Anaa, Layth and Malek, and to my lovely sister Duha, thank you very much. Also, I would like thank my Father and mother in law for their love and encouragement.

Finally, I would like to appreciate my friends and colleagues at NTU for the time shared during the past few years of my research. Specifically, my fellow PhD researchers in the Medical Engineering Design Research Group and the Computational Intelligence and Applications research group. To everyone who in one way or the other contributed to the success of my research, I am most grateful.

This research project was supported by Nottingham Trent University and ICON PLC, for the duration 2018 - 2021.

Ghayth Mohammad Maddallah AlMahadin
September 2021

Abstract

Tremor is the most typical common and simply recognised symptom of Parkinson's disease (PD) and presents in 70% - 90% of PD patients. In addition, tremor severity often indicates PD progress and severity and it can be used to evaluate treatment efficiency. Currently, the severity of Parkinson's tremor is scored based on the Movement Disorders Society's Unified Parkinson's Disease Rating Scale (MDS-UPDRS), however, the MDS-UPDRS is subjective and can be a lengthy process. Advances in wearable technologies combined with Machine Learning (ML) techniques have enabled the development of new approaches for the objective assessment of PD motor symptoms. A limited number of commercial systems are available with limited adoption and implementation due to the apparent lack of clinicians' and patients' perspectives. The goal of this research is to develop and validate a comprehensive solution to measure and quantify PD tremor severity objectively that incorporates the analysis of the perspective of the patients and healthcare professionals and provide an appropriate technology based solution. A holistic approach was adopted comprising of qualitative and quantitative methods divided into three stages. Firstly, a qualitative method using semi-structured interviews identified the perspectives of both healthcare professionals and patients linked to current assessment methods and their requirements for wearable devices. The results showed that a well-known assessment process such as MDS-UPDRS was not used routinely in clinics as it is time consuming, subjective, inaccurate and dependent on patients' memories. Participants suggested that objective assessment methods are needed to increase the chance of effective treatment. The participants' perspectives were positive toward using wearable devices. Healthcare professionals stated a need for an economical solution that provides concise information and is easy to use and interpret and should mimic the current scale. Secondly, a novel framework is proposed to enhance the tremor severity classification. The proposed approach is a combination of signal processing and resampling techniques integrated with well-known classifiers. The results show that over-sampling techniques performed better than other resampling techniques. The proposed approach has solved the imbalanced data problem and it has improved tremor severity detection significantly without neglecting minority classes and achieved 95.04% overall accuracy, 96% G-mean, 93% IBA and 99% AUC with Artificial Neural Network based on Multi-Layer Perceptron (ANN-MLP) with Borderline SMOTE. Finally a recommended system was identified to measure tremor severity. The system comprises of recommended tasks, classifier, classifier hyper-parameters and resampling technique. In this stage, a novel comprehensive method is developed to discriminate tasks' effect on tremor severity detection by developing an efficient and unique metric rule-based algorithm to identify recommended and non recommended tasks to be performed for tremor data collection. This establishes a novel quantitative framework that is based on an exhaustive sequential filtering algorithm that takes into consideration various combinations based on different advanced metrics instead of depending on a single metric. Results showed that ADL tasks that involve direct wrist movements are not suitable for tremor severity assessment. The findings of this research suggest that the recommended system is the SVM classifier combined with Borderline SMOTE over-sampling technique and the tasks are sitting, stairs up and down, walking straight, walking while counting, and standing and achieved 98% accuracy, 98% F1-score, 97% IBA, 98% G-mean and 99% AUC. The novel system solutions and the results presented in this thesis demonstrate a significant contribution towards the objective measurement of tremors in Parkinson's disease. New data is also presented for policy-makers and healthcare professionals which provides new perspectives in relation to the objective assessment of PD in current clinical practice.

Publications

As a result of the research presented in this thesis, the following publications have been published:

Journal Papers:

AlMahadin, G., Lotfi, A., Zysk, E., Siena, F.L., Carthy, M.M. and Breedon, P., 2020. Parkinson's disease: Current assessment methods and wearable devices for evaluation of movement disorder motor symptoms-A patient and healthcare professional perspective. BMC neurology, 20(1), pp.1-13.

AlMahadin, G., Lotfi, A., Carthy, M.M. and Breedon, P., 2021. Task-Oriented Intelligent Solution to Measure Parkinson's Disease Tremor Severity. Journal of Healthcare Engineering, 2021.

AlMahadin, G., Lotfi, A., Carthy, M.M. and Breedon, P., 2022. Enhanced Parkinson's Disease Tremor Severity Classification by Combining Signal Processing with Resampling Techniques. SN Computer Science, 3(1), pp.1-21.

Conference Proceedings:

AlMahadin, G., Lotfi, A., Carthy, M.M. and Breedon, P., 2021, December. Parkinson's Disease Tremor Severity Classification-A Comparison Between ON and OFF Medication State. In International Conference on Innovative Techniques and Applications of Artificial Intelligence (pp. 364-370). Springer, Cham.

AlMahadin, G., Lotfi, A., Carthy, M. M., and Breedon, P. 2019. Objective measurement of tremor symptom in Parkinson's disease remotely and within a clinic. [Poster and Presentation]. Smart Industry Workshop: Recent Advances in Industrial Digitalisation, Robotics and Automation, 9-11 January, Nottingham Trent University.

Contents

Nomenclature

Acronyms

ADADELTA	Adaptive Delta
ADAGRAD	Adaptive Gradients
ADAM	Adaptive Moment Estimation
ADMAX	Adaptive Moment Estimation Max
ADASYN	Adaptive Synthetic Sampling Approach
ADL	Activities of Daily Living
AI	Artificial intelligence
ANN	Artificial Neural Network
AUC	Area Under the Curve
CG	Conjugate Gradient
CID	Complexity-Invariant Distance
CNN	Condensed Nearest Neighbour
CURS	Columbia University Rating Scale
DBS	Deep Brain Stimulation
DI	Drug-Induced
DT	Dystonic Tremor
DT	Decision Tree
DWT	Discrete Wavelet Transform
EDS	Extensive Disability Scale
EDS	Extensive Disability Scale
EEG	Electroencephalogram
ELU	Exponential Linear Units
EMG	Electromyogram

ENN	Edited Nearest Neighbours
ET	Essential Tremor
FFT	Fast Fourier Transform
FT	Finger Tapping
GD	Gradient Descent
G-mean	Geometric Mean
GP	General Practitioner
H&Y	Hoehn and Yahr
HFE	Human Factors Engineering
HMM	Hidden Markov Models
IBA	Index of Balanced Accuracy
IHT	Instance Hardness Threshold
IMU	Inertial Measurement Unit
ISAPD	Intermediate Scale for Assessment of Parkinson's Disease
JICEC	Joint Inter-College Ethics Committee
KNN	K-Nearest Neighbours
KT	Kinetic Tremor
KW	Kruskal–Wallis
LASSO	Least Absolute Shrinkage and Selection Operator
L-BFGS	Limited-memory Broyden–Fletcher–Goldfarb–Shanno
LOOCV	Leave One Out Cross-Validation
LR	Logistic Regression
MDS	Movement Disorders Society
MEG	Magnetoencephalography
MEMS	Microelectromechanical Systems
MJFF	Michael J. Fox Foundation for Parkinson's Research
ML	Machine Learning
MLP	Multi-Layer Perceptron
MMG	Mechanomyogram
Nadam	Nesterov-accelerated Adaptive Moment Estimation
NAG	Nesterov Accelerated Gradient
NICE	National Institute for Health and Care Excellence
NUDS	North-western University Disability Scale

NYU	New York University Disease Evaluation
OvO	One-vs-One
OvR	One-vs-Rest
PCS	Principal Component Analysis
PD	Parkinson's Disease
PDIS	Parkinson's Disease Impairment Scale
PI	Postural Instability
PKG	Parkinson's Kinetigraph
PSD	Power Spectral Density
PT	Postural Tremor
RBF	Radial Basis Function
ReLU	Rectified Linear Unit
REN	Elman Neural Network
RF	Random Forest
RMS	Root Mean Square
RMSE	Root Mean Square Error
RMSprop	Root Mean Squared Propagation
RT	Rest Tremor
SAD	Sum of Absolute Differences
SAG	Stochastic Average Gradient
SAGA	Stochastic Average Gradient
SCA	Spectral Centroid Amplitude
SELU	Scaled Exponential Linear Unit
SFS	Sequential Forward Selection
SGD	Stochastic Gradient Descent
SMOTE	Synthetic Minority Over-sampling Technique
SMOTEENN	SMOTE with Edited Nearest Neighbour
SMOTETomek	SMOTE with Tomek link
SPES	Short Parkinson's Evaluation Scale
SVM	Support Vector Machine
UCD	User-Centered Design
UCLA	University of California Los Angeles Scale
UKPDSBB	UK Parkinson's Disease Society Brain Bank
UPDRS	Unified Parkinson's Disease Rating Scale

Greek Symbols

β_2	Kurtosis
γ_1	Skewness
γ	is used to set the spread of the kernel
σ	Scaling parameter of the input samples
$\sigma(z)$	Sigmoid function
α	Weight parameter of IBA

Other Symbols

W^+	Window subset contains elements above the mean
W^-	Window subset contains elements below the mean
W_l	Window length (number of samples)
a_t	The acceleration at time t
l	The lag
$\overline{a}_w$	Window's samples mean
s_w	Window's samples standard deviation
$A^m\,(r)$	The probability that two vectors of m points would match
$A^{m+1}\,(r)$	The probability that two vectors of $m+1$ points would match
$W_l^{(\mathcal{O})}$	Window length is odd
$W_l^{(\mathcal{E})}$	Window length is even
i	An element position (index) in the window $\{W\}$
n	Number of neighnours
$a_{(n+m+k)}$	The acceleration at a time $(n+m+k)$
W	The selected window
$e^{\frac{-j2\pi}{W_l}}$	The primitive Nth root of unity
f_{dis}	The dispersion frequency in the selected window
f_{med}	Median frequency
f	Frequency bin
f_l	The lowest frequency in the selected window
f_h	The highest frequency in the selected window
f_{step}	The range between the f_{med} and the f_{dis}
PSD_{fund}	The PSD at fundamental frequency.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

Parkinson's Disease (PD) is the second common long-term chronic, progressive, neurodegenerative disease. It mainly affects the motor system, and the cardinal motor symptoms are tremor (rhythmic shaking movement), bradykinesia or akinesia (slowness of movement), rigidity (muscle stiffness), and postural instability (impaired balance) [1, 2].

PD prevalence varieties from 41 people per $100,000$ in the fourth decade of life to over 1900 people per $100,000$ in people over 80 years of age [3]. According to the Parkinson's Foundation [4], more than 10 million people worldwide are living with PD, and about one million people in the United States (US) alone, and the estimated direct and indirect cost of Parkinson's in the US is $52 billion per year. In the United Kingdom, two people are diagnosed every hour, and the estimated number of people diagnosed with PD in 2018 was around $145,000$ [5] with an overall direct and indirect cost of over $\pounds 20,000$ per patient. PD is linked with morbidity, mortality, high economic burden, and a decreased quality of life. However, studies show that positive results can be achieved in the management of motor symptoms in the early stages. The consequences of late or incorrect diagnosis have a negative impact on individuals' patients and health service system [6].

Extensive research has shown that tremor is the most typical common and simply recognised symptom of PD and usually affecting upper limb and unilateral

[1]. Tremor severity often indicates PD progress and severity. It can also be used to evaluate treatment efficiency.

Currently, Parkinson's tremor severity is scored based on the Movement Disorders Society's Unified Parkinson's Disease Rating Scale (MDS-UPDRS) from 0 to 4 with 0: normal, 1: slight, 2: mild, 3: moderate, and 4: severe [7]. However, The MDS-UPDRS is a subjective and lengthy process [8, 9]. Furthermore, many aspects of the MDS-UPDRS scale depend on the patient's memory, which is inaccurate and hindered by reporting bias [8]. Therefore, an early objective and reliable assessment method could improve treatment and reduce direct and indirect healthcare services cost [2].

Advances in wearable technologies combined with Artificial Intelligence (AI), specifically Machine Learning (ML) techniques, have enabled the development of new approaches for objective assessment of PD motor symptoms [10], and they have shown promising results in research and clinical trials to objectively measure and monitor symptoms, both on-site and remotely [11, 12]. However, a limited number of commercial systems are available for such purposes [13] such as SENSE-PARK system[1], Kinesia system[2], Parkinson's Kinetigraph (PKG)[3] and Physilog[4]. Where such systems are in operation, they show limited adoption and implementation [14] despite the fact that these devices show high reliability, validity and responsiveness [13, 15], and have been used in the evaluation of PD symptoms and signs by individuals apart from the development team and reported successful clinical trials [15, 16]. Data from several studies suggest that these inconsistencies in adoption and implementation may be due to the lack of users' perspectives in devices design and development [17].

The research to date has tended to focus on proposing a tremor severity classification approach by combining signal processing with ML without discrimination tasks and medication state effect on classification and tremor severity detection [10]. Furthermore, A well-presented challenge to implement ML algorithms in real-world classification problems, including biomedical engineering is the issue of imbalanced classes representation or the lack of density

[1] http://www.sense-park.eu/sense-park-system.php

[2] http://glneurotech.com/kinesia/products/kinesia-360/

[3] https://globalkineticscorporation.com/the-pkg-system/

[4] Physilog:https://gaitup.com/physilog-sensor/

of one or more classes in the data, which cause false-negative predictions [18]. This miss-classification can lead to erroneous diagnosis or evaluation, consequently affect treatment efficiency.

The limitations mentioned above strongly suggest that there is a research need for exploring the perspectives of healthcare professionals and patients linked to current diagnosis and assessment methods and to identify their preferences and requirements of wearable devices. In addition, there is a need to automate tremor severity assessment using an accurate, reliable and objective method and forms the focus of this thesis.

1.2 Overview of the Research

This research aims to develop and validate a recommended solution to measure and quantify PD tremor severity objectively, and to identify the perspectives of healthcare professionals and patients to current assessment methods and to identify their preferences, and requirements of wearable devices. A schematic representation of the proposed framework in this thesis is illustrated in Figure 1.1. The aim is to find a recommended system to measure tremor severity objectively.

The proposed framework comprises qualitative and quantitative methods. The qualitative method thematically analysed semi-structured interviews and focus group discussions to define the perspectives of healthcare professionals and patients in the context of existing diagnosis and assessment methods (i.e. MDS-UPDRS). It also enabled to identify their preferences and requirements of wearable devices to ultimately evaluate and monitor PD tremor. Secondly, their expectations and outlooks on potential solutions.

The quantitative method utilises different datasets collected through wearable devices including a tri-axial accelerometer. The patients have performed a set of scripted activities (predefined motor tasks) to find the best tasks, best-optimised classifiers and best resampling techniques. Also, this approach is used to identify medication state effects on tremor severity estimation. To begin this process, datasets were split into sub-datasets according to tasks performed during data collection (i.e. walking, standing, drinking, sitting, etc.), then raw sensors

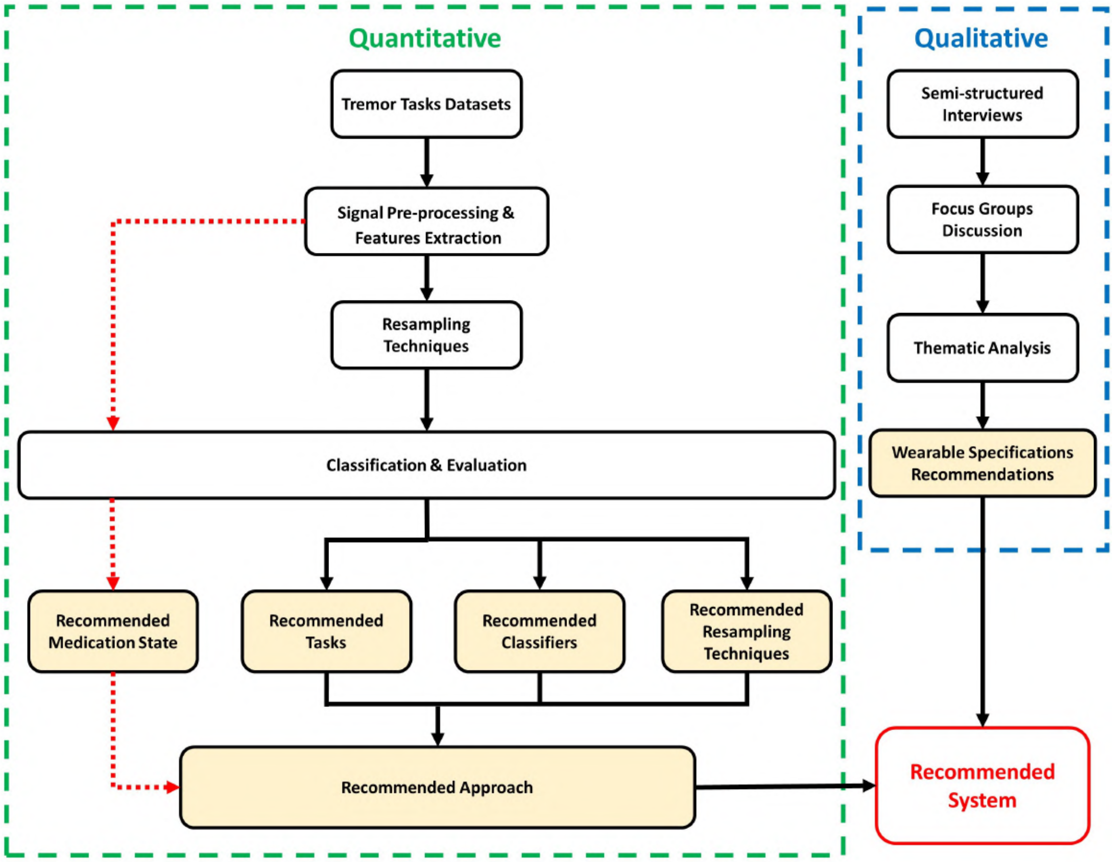

Quantitative
Tremor Tasks Datasets
Signal Pre-processing & Features Extraction
Resampling Techniques
Classification & Evaluation
Recommended Medication State
Recommended Tasks
Recommended Classifiers
Recommended Resampling Techniques
Recommended Approach
Qualitative
Semi-structured Interviews
Focus Groups Discussion
Thematic Analysis
Wearable Specifications Recommendations
Recommended System

Figure 1.1: Schematic representation of the proposed framework to identify a recommended system to detect PD tremor severity.

signals of each sub-dataset were filtered to eliminate sensor non-tremor data and artefacts. The results were segmented into windows to extract meaningful features that representing relevant information about input patterns or classes. After that, imbalanced data was resampled through different resampling techniques and passed to different supervised classifiers separately. The results were evaluated through various advanced metrics and analysed by a developed framework to identify potential recommended approaches (recommended tasks, classifiers, resampling techniques, medication state). The potential recommended systems were evaluated independently to identify the final recommended approach.

1.3 Research Questions

Following the research overview, this thesis primarily explores the usefulness of wearable devices in the prediction of PD tremor severity and to identify a recommended system that is significantly correlated with MDS-UPDRS and with high levels of patients' and healthcare professionals' acceptance. In particular, this thesis attempts to answer the following questions:

- What do PD patients and healthcare professionals want from wearable medical devices? How can they benefit from wearable technologies?
- What is the role of wearable devices aesthetics and design for acceptance and adoption?
- Can we perform automatic PD tremor severity classification from different wearable devices contains tri-axial accelerometer of multiple tasks using machine learning techniques?
- On which types of tasks must we focus while performing tremor severity assessment? In other words, Which tasks or types of tasks maximise inter-class separations?
- How much improvement for imbalanced data classification can be achieved by employing resampling techniques?
- Can we utilise advanced classification metrics to identify a cross-platform recommended system that comprises recommended tasks, the recommended classifier(s) and recommended resampling technique(s) to quantify tremor severity?
- What is the effect of medication state on tremor severity classification?

To address the above questions, the following section outlines the aim and objectives of the performed research reported in this thesis.

1.4 Research Aims and Objectives

The aim of this research is to develop a reliable, responsive and cross-platform sensors solution that quantify tremor severity objectively for people affected with Parkinson's disease that could be used by a patient, with or without the help of a caregiver. In addition, to identify the perspectives of healthcare professionals and patients about current assessment methods and to identify their preferences, and requirements of wearable devices.

To achieve the above research aim, the following objectives were identified:

1. To conduct an extensive review of the literature on the PD tremor evaluation approaches and to identify their limitations.
2. To understand the key requirements and needs of patients and healthcare professionals of objective monitoring solutions and wearable devices.
3. Extract meaningful features from raw accelerometer signals that can distinguish between PD tremor severity.
4. Develop and evaluate ML methods that automatically predict PD tremor severity score aligned with MDS-UPDRS scale.
5. Investigate the effectiveness of various resampling techniques on tremor severity classification performance.
6. Explore the impact of tasks performed during data collection on tremor severity classification performance.
7. Develop and design an algorithm to identify a recommended system to measure tremor severity based on advanced metrics evaluation.
8. Investigate medication state effects on tremor severity classification.

1.5 Major Contributions of the Thesis

The major contributions of the work presented in this thesis are summarised as follows:

- An extensive literature review in relation to state-of-the-art tremor assessment.

- This research presents one of the first comprehensive qualitative studies to explore both patient and healthcare professionals' perspectives. This is specifically linked to current diagnosis and assessment methods, wearable device design and materials, and the requirements and specifications of a combined PD monitoring solution. These findings will be of interest in the development of a wearable device that meets both clinicians and patients needs and requirements.

- This thesis contributes to the existing knowledge on objective tremor severity quantification by solving the imbalanced data problem by applying different resampling techniques with different classifiers. Also, it has improved tremor severity detection significantly without neglecting minority classes. In addition, it offers important insights into advanced metrics and how standard metrics can mislead classification results.

- This work provides one of the first attempts to discriminate tasks' effect on tremor severity detection by developing an efficient and unique metric rule-based algorithm to identify recommended and non recommended tasks to be performed for tremor data collection. The findings will be of interest for future research that investigates the measurement of tremor severity objectively by laying on a data collection protocol instead of the traditional trial and error approach.

- This thesis establishes a novel quantitative framework to identify a recommended system to measure tremor severity. The framework is based on an exhaustive sequential filtering algorithm. The output of the proposed framework produces a comprehensive recommended system that takes into consideration various combinations based on different advanced metrics instead of depending on a single metric. Also, the recommended system comprises of three main components, namely: tasks for data collection, resampling technique(s) and classifier(s).

- This is the first study that predicts PD tremor severity from two different wearable devices while patients perform different scripted tasks, including ADLs and MDS-UPDRS motor tasks. Significantly, the proposed approach can quantify tremor severity regardless of the wearable devices that have been used to collect tremor data as long as the wearable device contains an accelerometer.

- This work contributes to the existing knowledge of PD tremor measurement by investigating medication state effects linked to classification accuracy. This new understanding should help to improve PD tremor prediction by collecting sensor data that represents tremor severity more accurately.

  The outlined contributions of the thesis are addressed in different chapters of this thesis. A summary of these chapters is presented in the following section.

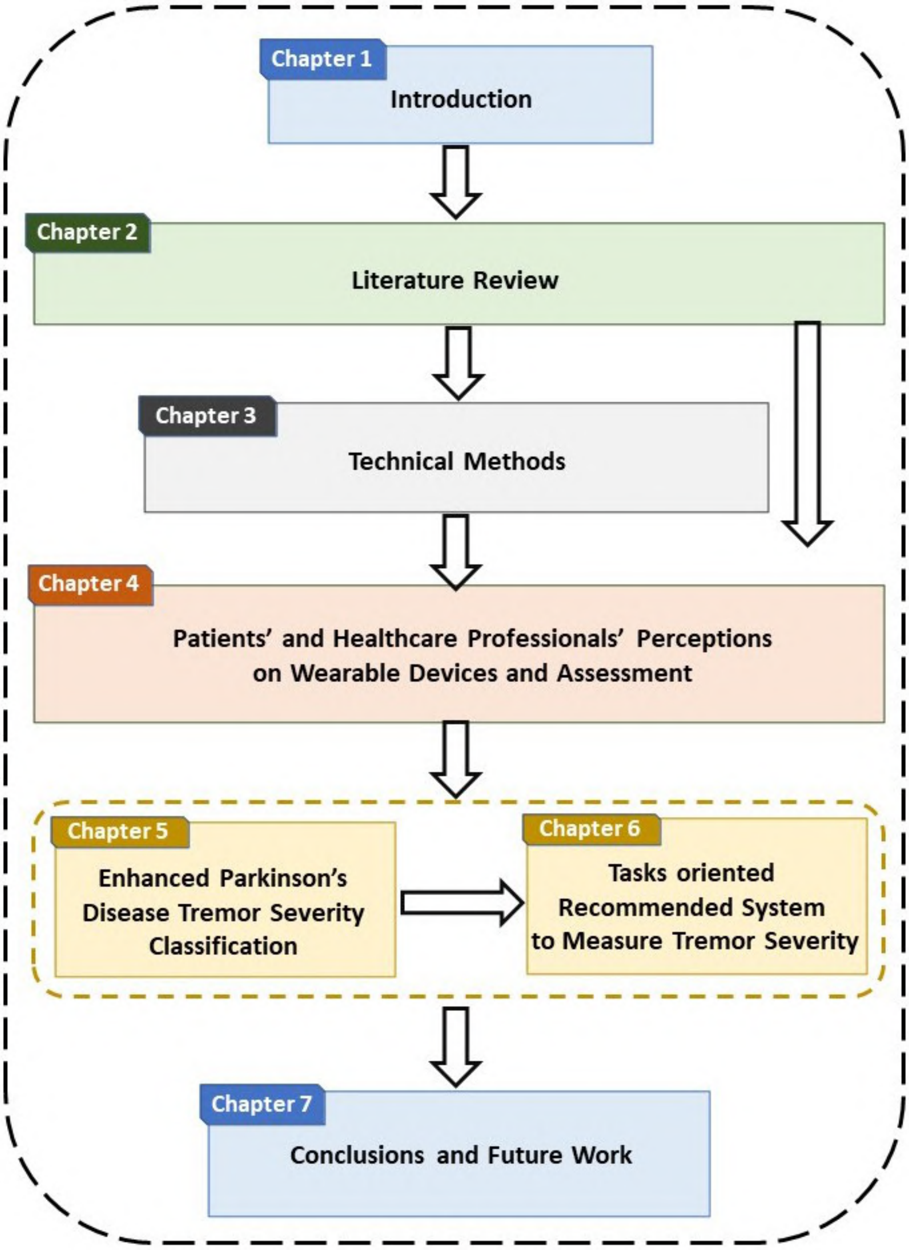

Chapter 1
Introduction
Chapter 2
Literature Review
Chapter 3
Technical Methods
Chapter 4
Patients' and Healthcare Professionals' Perceptions on Wearable Devices and Assessment
Chapter 5
Enhanced Parkinson's Disease Tremor Severity Classification
Chapter 6
Tasks oriented Recommended System to Measure Tremor Severity
Chapter 7
Conclusions and Future Work

Figure 1.2: Thesis structure showing the organisation of the chapters and their respective dependencies.

1.6 Thesis Outline

This thesis consists of seven chapters. Figure 1.2 shows the structure of the thesis with an indication of how the chapters are linked. This gives readers an overview of the organisation of the thesis and a direction on how the chapters are grouped. The summary of contents of this thesis are presented as follows:

Chapter 2: Literature Review - This chapter presents a comprehensive review of the relevant literature under three sections: PD symptoms, clinical assessment methods and objective assessment methods. The first section gives a brief description of PD history and symptoms. The second section discusses the current clinical assessments and monitoring methods and highlights their limitations. The third section present and discuss the work that has been carried out into the use of technology in PD tremor assessment and can be categorised into three main areas based on the application area: PD diagnosis, response to treatment and PD progression particularly automatic scoring of PD tremor.

Chapter 3: Technical Overivew - This chapter provides an explanation about common used in tremor severity estimation. this includes the description of datasets, signal processing, features extraction, resampling techniques, machine learning classification techniques and evaluation metrics.

Chapter 4: Patients' and Healthcare Professionals' Perceptions on Wearable Devices and Assessment - This chapter presents a comprehensive qualitative work that explores the perspectives of healthcare professionals and patients linked to current assessment methods and to identify preferences, and requirements of wearable devices.

Chapter 5: Enhanced Parkinson's Disease Tremor Severity Classification - This chapter presents a solution to enhance the tremor severity classification. The proposed approach is a combination of signal processing and resampling techniques; over-sampling, under-sampling and a hybrid combination. Resampling techniques are integrated with well-known classifiers, such as Artificial Neural Network based on Multi-Layer Perceptron (ANN-MLP) and Random Forest (RF). The experiment is conducted using the datasets which are explained in Chapter 3. The performance of the proposed approach is evaluated through advanced metrics that have been described in Chapter 3, which achieved a significant result compared to earlier works.

Chapter 6: Tasks oriented Recommended System To Measure Tremor Severity - This chapter presents a novel above-average rule-based approach to identify a recommended system that is used to measure tremor severity, including the influence of tasks performed during data collection on classification performance. The recommended system includes recommended tasks, classifier, hyper-parameters and resampling technique. The proposed approach is based on various advanced metrics results of datasets which are described in Chapter 3. In addition, this chapter investigates medication state effects on tremor severity classification.

Chapter 7: Conclusions and Future Works - This chapter presents a summary of the findings of the research conducted in this thesis. The key findings from this thesis are discussed in light of the research questions identified in Chapter 1. The chapter also formulates future work directions for applications of the work in this thesis. As well, some possible areas for enhancing the work has been done in this thesis in future.

Chapter 2

Literature Review

2.1 Introduction

PD was originally described by James Parkinson in 1817 and defined as the "shaking palsy" [19], later refined by Jean-Martin Charcot [20] and separated from other disorders characterised by the tremor. PD is caused by degeneration of nerve cells in the substantia nigra that produce dopamine, causing the movement impairments that characterize the disease [21]. PD cannot be cured at present and it is unpredictable; also the progress can vary between patients [22].

Accurate diagnosis and evaluation of PD tremor has a pivotal role in PD management and reduces disability over time using the appropriate interventions. Consequently, it improves PD patients' quality of life and reduces direct and indirect healthcare costs. Therefore, it is important to understand PD symptoms, clinical diagnosis and assessment methods and their limitations. In this regard, several objective methods have been proposed for measuring and quantifying PD tremor, hence it is important to review the current state-of-the-art related to PD tremor diagnosis and evaluation to justify the intent of the work in this thesis.

This chapter is structured as follows: Section 2.2 provides an overview of PD motor symptoms. Section 2.3 discusses the current clinical diagnosis, clinical assessment and clinical rating scale methods. Section 2.4 reviews the existing literature studies on objective tremor assessment under two headings as follows: PD diagnosis and PD progression, in particular automatic tremor scoring, which forms the focus of this thesis. Section 2.5 presents the conclusions and the research opportunity derived from the literature.

2.2 Parkinson's Disease Motor Symptoms

PD clinical symptoms first appear when about 50% - 70% of dopaminergic cells degenerated [23]. The four cardinal motor symptoms or features of PD are often abbreviated as **TRAP** are **T**remor, **R**igidity, **A**kinesia (or bradykinesia), and **P**ostural instability [24]. The motor symptoms are often accompanied by non-motor symptoms such as dementia, fatigue, depression, insomnia and other neuropsychiatric and autonomic complications [25]. The symptoms of Parkinson's disease are progressive over time, and it can vary from one individual to another both in terms of the intensity and the way they progress [1]. The main four cardinal motor symptoms are described in more detail below:

Tremor is the most common initial and easily recognised symptom of PD and presents in 70% - 90% of PD patients [2]. PD patients show different types of tremors: rest and action tremors [24]. Rest tremor (RT) describes unilateral involuntary, rhythmic and alternating movements in relaxed and supported limbs, mostly hands, and typically disappears with action. RT might also appear in lips, chin, jaw and legs but rarely involve the neck, head or voice as essential tremor (ET) [1]. RT occurs at a frequency between 4 - 6 Hz, which is different from ET that occurs between 5-10 Hz [2]. The action tremor has two types: Kinetic tremor (KT) and postural tremor (PT) [26]. KT is a type of tremor present during voluntary hand movements such as touching the tip of the nose, typing or writing. PT occurs when a person maintains a position against gravity, such as stretching arms. PT tremor occurs at a frequency between 6-9 Hz, while KT occurs at a frequency between 9 - 12 Hz [26].

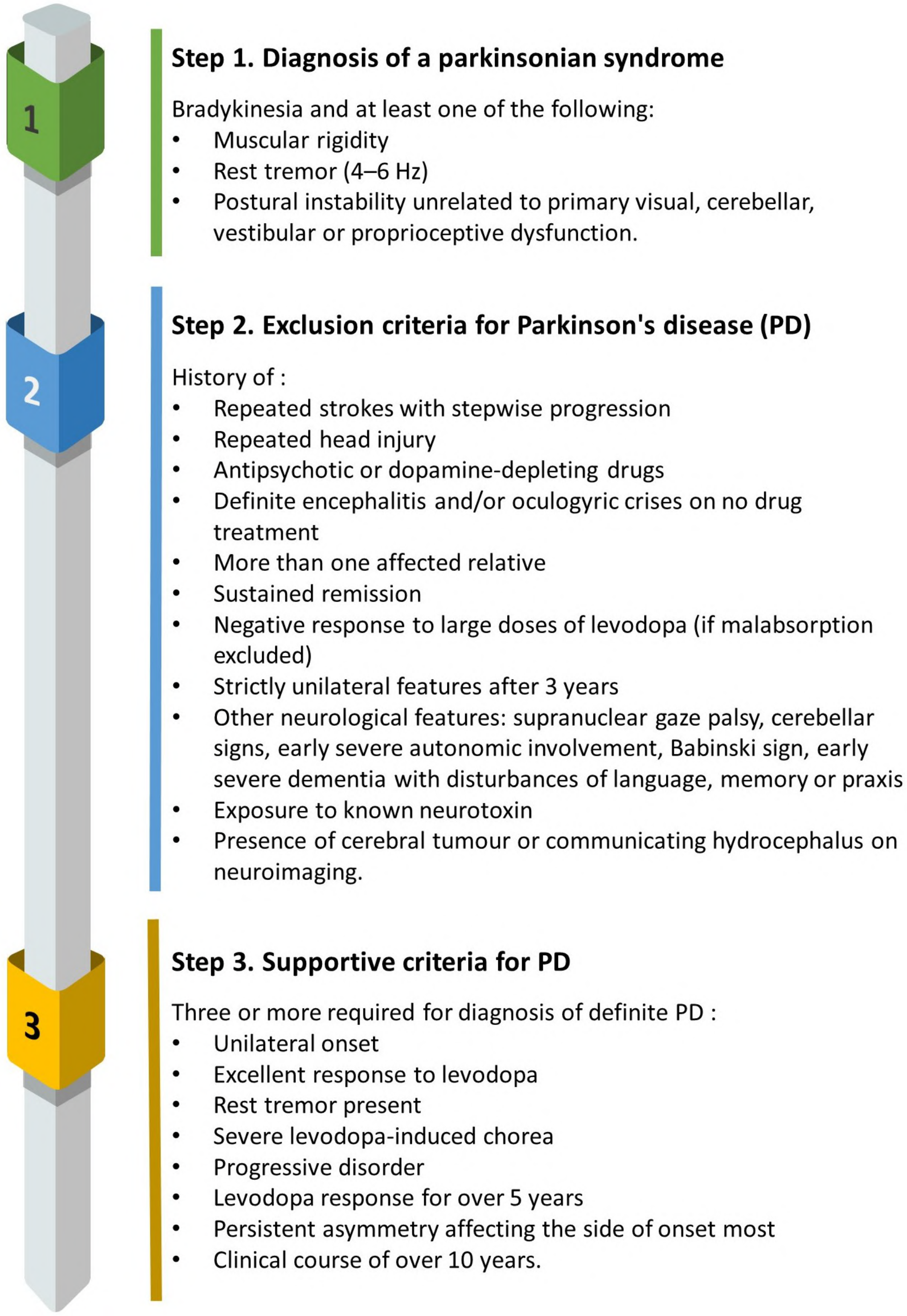
Rigidity presents in more than 90% of PD patients, which is increased resistance to the passive movement throughout the length of movement, such as physical examination, due to high muscle tone or tension (tightness and stiffness). Rigidity is commonly labelled as "lead-pipe" rigidity as the affected limb does not respond to forced movement as the lead-pipe resists movement [2]. Rigidity often increases with reinforcing manoeuvres, and it could involve pain, mostly shoulder pain; hence, it is commonly misdiagnosed as other medical conditions such as arthritis or bursitis [1].

Akinesia or **Bradykinesia** is one of the cardinal manifestations of PD and presents in 80% - 90%of PD patients. It refers to the movement slowness and a reduced ability to move the body. Patients experiencing bradykinesia may be unable to initiate and perform movement require several successive steps or fine motor control such as getting up from a chair, walking, buttoning a shirt, talking or showing facial expressions (hypomimia). Also, they may not control the size and speed of the movement. Other bradykinesia indicators include loss of involuntary movements and facial expressions, reduced blinking, and decreased arm swing while walking [1].

Postural instability (PI) or **balance impairment** is the last, and it is one of the most disabling motor symptoms in the advanced stages of the disease. Patients who experience PI cannot keep their bodies in a stable or balanced position due to the loss of postural reflexes. PI is more remarkable during position change, such as standing up or turning or pivoting while walking [1, 2]. People who are experiencing postural instability may fall easily since they could not maintain an upright position. Previous research has established that falls in people with Parkinson's are very common, as 60.5% of people with PD reported at least one fall, with 39% reporting recurrent falls with average 20.8 falls per person per year [27].

2.3 Clinical Diagnosis and Assessment Methods

Currently, The pathology of PD is a clinically based process. As the diagnosis, progress monitoring and treatment efficacy are subject to history taking and neurological examination of motor and non-motor symptoms. The following three sections summarise these methods under three heading: clinical diagnosis, clinical assessment and clinical rating scales.



Step 1. Diagnosis of a parkinsonian syndrome

Bradykinesia and at least one of the following:

- Muscular rigidity
- Rest tremor (4–6 Hz)
- Postural instability unrelated to primary visual, cerebellar, vestibular or proprioceptive dysfunction.

Step 2. Exclusion criteria for Parkinson's disease (PD)

History of :

- Repeated strokes with stepwise progression
- Repeated head injury
- Antipsychotic or dopamine-depleting drugs
- Definite encephalitis and/or oculogyric crises on no drug treatment
- More than one affected relative
- Sustained remission
- Negative response to large doses of levodopa (if malabsorption excluded)
- Strictly unilateral features after 3 years
- Other neurological features: supranuclear gaze palsy, cerebellar signs, early severe autonomic involvement, Babinski sign, early severe dementia with disturbances of language, memory or praxis
- Exposure to known neurotoxin
- Presence of cerebral tumour or communicating hydrocephalus on neuroimaging.

Step 3. Supportive criteria for PD

Three or more required for diagnosis of definite PD :

- Unilateral onset
- Excellent response to levodopa
- Rest tremor present
- Severe levodopa-induced chorea
- Progressive disorder
- Levodopa response for over 5 years
- Persistent asymmetry affecting the side of onset most
- Clinical course of over 10 years.

Figure 2.1: UK Parkinson's Disease Society Brain Bank diagnosis criteria.

2.3.1 Clinical Diagnosis

Nowadays, there is no definite diagnostic before postmortem confirmation. The clinical diagnosis for probable PD used based on medical history, clinical observation and evaluation of the symptoms [21].

Diagnosis of PD is based on clinical interpretation of symptoms evoked through history taking, visual observation and motor examination. The formal diagnosis criteria by UK Parkinson's Disease Society Brain Bank (UKPDSBB) [28] were proposed by Gibb and Lees in 1988 [29] are described in Figure 2.1. These criteria comprise of three steps: The first step involves the inclusion criteria where the subject must have bradykinesia and at least one of other cardinal symptoms(4 - 6 Hz rest tremor, rigidity, postural instability) to be diagnosed with Parkinsonian syndrome. The second step involves the exclusion criteria to differentiate PD from other causes of parkinsonism, where PD diagnosis is excluded if any of these criteria is identified. The third step refers to supportive diagnosis criteria, where definite PD diagnosis is confirmed if three or more of these criteria are identified along with previous criteria.

Even though UKPDSBB applies strict diagnostic criteria and become the gold standard for clinical diagnosis of PD, misdiagnosis is common in PD diagnosis and it can be affected by subjectivity of the examiner and often made in patients with essential tremor or other parkinsonian syndromes[2]. Several studies have shown that the clinical diagnosis using UKPDSBB does not achieve 100% accurate diagnostic. A recent systematic review and meta-analysis concluded that the clinical diagnosis did not improve in the last two decades, for example the accuracy of UKPDSBB diagnostic was 82.7% [30].

There is much scope for Type I (False Positive) and Type II (False Negative) errors in diagnosis [31]. A false-negative diagnosis may increase disease progression due to late or lack of treatment or inappropriate interventions if diagnosed as a different disease. On the other hand, a false-positive diagnosis may increase the risks of medications side effects that have been given inappropriately. To minimise misdiagnosis rates and risks, the National Institute for Health and Care Excellence (NICE) recommend that people in the UK with suspected PD should be referred quickly and untreated to an expert in movement disorders diagnosis [28].

2.3.2 Clinical Assessment

PD is heterogeneous in terms of symptoms, progression rates and response to treatments, so it is impossible to come up with an optimal treatment for all patients [32]. Therefore, a disease management plan should be in place by regularly reviewing symptoms development, new symptoms and response to medications for individual patients [33]. Furthermore, disease management can help to determine or quantify PD severity in terms of impairment and disability measurements and to evaluate advantages and disadvantages of new drugs or surgical procedures in clinical trials [34].

The primary method for PD progress monitoring is clinical based assessment in outpatient clinics. the NICE recommends PD patients to be reviewed by a movement disorder specialist (neurologist or PD nurse specialist) at least every 6-12 months [33]. The frequency of the assessment depends on patient conditions and symptoms' severity. The Duration and the frequency of assessment is not fixed and depends on patient conditions and symptoms' severity, but typically in 15 to 20 minutes appointment [35]. The assessment includes physical examination, taking history and gathering information about symptoms impact, fluctuation, severity and medication response. Sometimes, the assessment could lead to reconsider the diagnosis if new symptoms changed or new symptoms developed. The clinical assessment comprises of two parts, history taken and physical or motor examination.

The history is taken by listening to the patient, sometimes accompanied by a family member or a caregiver, about symptoms, response to medication, and difficulties experienced in daily living activities. It also includes a symptoms diary completed by patients to capture motor fluctuations, the diary provides clinicians with great information about symptoms, but it places a burden on patients [36]. These information can support or doubt the diagnosis and treatment efficacy [37].

The second part of the clinical assessment is the physical examination, which is the patient's movement observation and analysis during the review appointment [36, 38]. The examiner starts watching and analysing patient's movements from the moment they walk into the clinic, looking for abnormal movements such as slowed movement, rest tremor, mask face and impaired posture, so most of the examination can be done alongside history taking. Other examination techniques are used to evaluate some symptoms by asking the patient to perform some tasks such as finger tapping, nose touching, draw an Archimedes spiral or getting up from a chair and walk. Also, there is a passive examination to assess rigidity, where the examiner moves limbs passively looking for resistance or stiffness [38].

Based on the observations from history taking and physical examination, the clinician makes a judgement on diagnosis if it needs reconsideration or to prescribe medications or to refer the patient to alternative treatments such as supportive therapies or surgery for some people.

Clinical assessment techniques, including history taking and physical examination, carry with them well-known limitations [36]. History taking mainly depends on patient recall which is unreliable due to several factors. Firstly, PD can cause cognitive dysfunction that affects patients recall capabilities even in the early stages of the disease [25]. Secondly, many patients are unaware of some symptoms combined with dyskinesis, and sometimes they do not associate some symptoms with PD [39], and thus these symptoms are not reported to clinicians.

Data from several studies suggest that patients' diaries are subjective and there is a weak correlation between diaries and clinical assessment, where patients' assessment tended to be more severe than clinicians' assessment [40, 41]. This inconsistency may be due to symptoms fluctuations from time to time [2], beside the fact that many patients find that filling diaries is exhausting, particularly for extended periods [42]. In addition, 80% of PD patients affected by dementia and 25% of PD patients prone to cognitive impairment [43].

In terms of physical examination, previous research has showed significant variances in motor symptoms measurements between the clinic and home assessment [44]. As the home environment and activities patterns are different from the clinic environment, and they cannot be replicated in the clinic. Moreover, the clinic's assessment does not reflect day-to-day symptoms, which might vary during the day or the last few weeks or months, and can only capture the patient's symptoms at that moment [36]. Furthermore, the examination in the clinic and the presence of the examiner can be stressful for patients and may lead to wrong symptoms interpretation or alert non existing symptoms or increase symptom severity such as tremor [36]. Another limitation is that the clinical assessment is subjective and depending on the clinician's expertise, skills and knowledge and it is different from one expert to another, and this may lead to inconsistent assessment [8, 45].

2.3.3 Clinical Rating Scales

Clinical rating scales are essential to quantify neurological disorders symptoms, impairment and disability [46]. Clinical rating scales enable researchers and clinicians to evaluate disease symptoms, severity, progression, treatment efficacy, response and side effects [47, 48].

Several PD clinical rating scales have been developed since the 1960*s*. Table 2.1 presents the most common rating scales. The scales can be categorised into impairment scales, disability scales and multi-modular scales (impairment scales and disability scales) [48]. Below a summary of these scales:

Table 2.1: Clinical rating scales of Parkinson's disease

Impairment	**Disability**	**Multi-modular**
Columbia University Rating Scale (CURS)	Hoehn and Yahr (H&Y)	Unified Parkinson's Disease Rating Scale (UPDRS)
Webster scale	North-western University Disability Scale (NUDS)	Movement Disorders Society Sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS)
PD Impairment Scale (PDIS)	Intermediate Scale for Assessment of PD (ISAPD)	New York University PD evaluation (NYU)
	Schwab and England Activities of Daily Living (ADL)	University of California Los Angeles scale (UCLA)
	Extensive Disability Scale (EDS)	Short Parkinson's Evaluation Scale (SPES)
	Extensive Disability Scale (EDS)	

Columbia University Rating Scale (CURS) was created in 1969 by

Columbia University to study PD symptoms and demonstrated moderate to good validity and reliability to evaluate the symptoms degree from 0 to 4, with 0 representing normal and 4 representing severe including bradykinesia, gait, posture, rest tremor and dyskinesias. The main drawback of CURS is that it does not evaluate the activities of daily living [47, 48].

Webster scale was developed in 1968, and it includes nine impairment elements (bradykinesia of hands, rigidity, posture, arm swing, gait, tremor, facies, seborrhea and speech) and one disability element (self-care). Each of these elements scored from 0 to 3. The total score from 0-30 and the disease severity increases when the score increases. The disability is divided into three categories based on total score: 1-10 early illness, 11-20 moderate and 21-30 severe. Webster scale showed poor to moderate reliability because it has only one disability element and nine impairment elements [47, 48].

PD Impairment Scale (PDIS) was created in 1987, and it includes ten items (balance, posture, extra steps, overflow movements, masked faces, slowness, arm swing, short steps, resting tremor and postural tremor). Each of these items scored from 0 to 3, with 0 normal and 3 is severe [47].

North-western University Disability Scale (NUDS) was created in 1980 by the North-western University to evaluate patients with PD disability to perform six tasks (walk, dress, eat, activities for food, hygiene, language), and each task scored from 0-10, where 0 is normal, and 10 is the highest disability. NUDS scales demonstrated a high correlation with CURS and Webster scales, but the reliability was found moderate by others while it was found excellent by the developers [47, 48].

Intermediate Scale for Assessment of PD (ISAPD) scale was developed in 1987 to be used in clinical trials and daily practice assessments. The scale includes 13 activities of daily life items (self-care, turning in bed, getting out of the bed, hygiene, bathing, dressing, speech, eating, feeding, walking, stairs, arising from chair and gait) with a score from 0-3 for each element with a total score from 0-39, the highest score the highest disability [49] ISAPD showed moderate to good correlation with the H&Y scale [50], but it showed good validity and the inter-rater reliability evaluation only by the developer, which need independent evaluation [47].

Schwab and England ADL scale was created in 1969 to measure patients of PD abilities to perform life daily activities [51]. This scale rates patient abilities on a percentage scale from 0% to 100%, where 0% represents the patient is bedridden, and 100% represents patients is independent. The Schwab and England ADL is done by interviewing the patient by an examiner or a caregiver [36]. The scale has been used in many studies to investigate other scales' characteristics and shows moderate to substantial validity and decent reliability. This scale's main drawback that there is no instruction or guidance for users [2].

Extensive Disability Scale (EDS) was developed in 1991 to evaluate PD patients' disability. EDS consists of two sections, the first part concerned about physical disabilities, and composed of 15 elements (stair climbing, walking, moving to bed or toilet or chair, bowel function, bladder function, bathing, dressing, grooming, feeding, sleeping, speaking, medical problems, mood, and sexual functions). The second part regarding environmental conditions and composed of 6 elements (working, financial status, personal assistance, conveyance, community, services and social activities). All elements scored from 0 to 4. The EDS scale is interview-based, and it requires about 15-20 minutes to be completed [52]. However, The scale validity and inter-rater reliability only reported by developers, which was moderate to good [47].

University of California Los Angeles (UCLA) scale was developed in 1981, and it is rarely used in clinical trials, and showed moderate to good inter-rater reliability by one study [47]. There is no access to other information about this scale.

Short Parkinson's Evaluation Scale (SPES) was developed in 1997 to address some limitation in UPDRS, such as low inter-rater reliability in some items and redundancy in others [53]. SPES comprised of three parts: mental axis, ADL and motor examination. The mental axis includes three elements: memory, thought disorders, and depression. ADL includes eight elements: speech, eating, feeding, dressing, hygiene, handwriting, walking, and turning in bed. The motor examination includes eight elements: speech, tremor, rest and posture, rigidity, finger tapping, arising from a chair, gait, and postural stability. Each element from the three parts scored from 0 to 3, where 0 is normal, and 3 is severe. The advantages of SPES is short and easy to do the examination in 7 to 10 minutes. Regarding the validity and inter-rater reliability is good but it was only stated by the scale developers [47].

Hoehn and Yahr (H&Y) scale was designed originally in 1967 to estimate the disease severity combining deficiency and disability based on bilateral motor involvement and compromised balance and gait [50]. It is a simple scale describes the stage of PD from 1 to 5 based on motor impairment severity and disability.

A modified version of H&Y scale developed in the 1990s and included two additional stages 1.5 (Unilateral and axial involvement) and 2.5 (Mild bilateral disease with recovery on pull test). However, the clinimetrics characteristics of this modified version have not been investigated, so it is recommended to use the original version [54].

The H&Y scale has been used widely and has universal acceptance as a scale to describe PD stages due to it is simplicity and ability to group PD patients based on motor and functionality severity and progress [36]. H&Y defined five stages, as shown in Table 2.2

The main advantage of H&Y scale is well known and easy to use, but it is a classification scale and not a rank order [55], i.e., the stage does not reflect the disability as someone in stage 2 may be more impaired for ADL than someone in stage 3.

The **UPDRS** was published in 1987 and become the most widely used rating scale [47]. However, the UPDRS scale has some ambiguities and limitations. The MDS has identified some of the limitations, including ambiguities in questions, poor instructions and the absence of important aspects of non-motor symptoms. Their findings have led to revised version MDS-UPDRS to resolve identified

Table 2.2: Hoehn and Yahr Scale.

Stage	Description
1	Unilateral involvement only, usually with minimal or no functional disability.
2	Bilateral or midline involvement without impairment of balance.
3	Mild to moderate bilateral disease; some postural instability; physically independent.
4	Severely disabling disease; still able to walk or stand unassisted.
5	Confinement to bed or wheelchair unless aided.

problems in the UPDRS and to enable better detection of small changes and mild disabilities [7]. MDS-UPDRS consists of 65 elements requires 30 minutes administration time distributed among four parts; I) Non-motor experiences of daily living (13 elements), II) Motor experiences of daily living (13 elements), III) Motor examination (33 elements), IV) Motor complications (6 elements) [7]. Each element scored from 0 to 4 where (0: normal, 1: slight, 2: mild, 3: moderate, and 4: severe), some elements are self-administrated by patients without any help, and some elements are completed with or without help from a caregiver, but independently from the examiner, and some elements are rated by the examiner based on observation and physical examination. The MDS-UPDRS scale designed to avoid medical terms to be easier for PD patients, and it applies to PD patients with different levels of disabilities [46].

Part one assesses the non-motor symptoms impact ADL, such as cognitive impairment, hallucinations and psychosis, depression, anxiety, apathy, sleep disorders, pain, urinary constipation problems, dizziness and fatigue. Part two assesses motor symptoms that affect ADL, such as speech problems, saliva and drooling, chewing and swallowing problems, eating tasks and dressing difficulties, movement difficulties and freezing.

Part three assesses motor symptoms seen by the examiner during the session, including speech, facial expression, rigidity, finger and toe-tapping, hand movements, pronation-supination of hands movements, leg agility, arising from a chair, gait impairment and freezing, posture, postural stability and tremor. Part four assess two motor complications, dyskinesias and motor fluctuations. The assessment is based on historical information from the patient and caregiver as well as information from the examination.

Even though MDS-UPDRS assessment is internationally accepted rating scale to assess PD, and enhanced the quality of clinical trials outcomes, and has undergone strict clinimetrics validation, but it is clinically based scale, that the clinician assigns numerical scores based on qualitative observations of the patient in various postures and are often insensitive and subjective, so, the assessment depends on the examiners' skills and knowledge, and it is various from one examiner to another, so examiners' disagreement on assessment and scores [8, 45]. There is evidence showed that MDS-UPDRS has high inter and intra-rater variability between nurses' and neurologists' assessments [56]. Thus, a patient's tremor may be assigned MDS-UPDRS score by one examiner and in the next appointment assessed by a different examiner and assigned a higher score. In this situation, it is difficult to interpret these two different scores, whether symptoms worsen or due to subjectivity.

The MDS-UPDRS scale is time-consuming and requires lengthy administration time, approximately 30 minutes, besides it requires specialised official training to improve the coherence of data acquisition and interpretation, these make it unhandy for routine clinical practice [7, 9]. Another time burden that many elements in MDS-UPDRS need to be completed by patients, so additional time is required besides the time required to review these elements by the examiner. This time burden limits the use of the MDS-UPDRS in routine clinical practice. Therefore, MDS-UPDRS scale is mainly used in clinical research.

Similar to clinical assessments, The MDS-UPDRS assessment performed in a clinical environment does not mirror day to day symptoms, which might vary during the day or the last few weeks or months and can only capture a snapshot of patient's symptoms at that moment [36]. Moreover, many elements of the MDS-UPDRS scale do not apply to each patient, so these questions may raise anxiety in patients, consequently influence their assessment, besides incompetent usage of examiner's time.

Additionally, many elements in MDS-UPDRS scale depends on patients memory, which is unreliable and limited by recall bias [8, 45] due to that fact that most patients are elderly, they exhibit cognition and dementia issues. PD can also cause cognitive dysfunction even in early stages which affect patients recall capabilities [25]. Also, long time between appointments makes it difficult to remember symptoms during this period, besides it is an inconvenience for patients to travel to the clinic due to the weather conditions, transportation, distance, and their conditions, particularly in advanced stages of the disease [33]. Therefore, an early objective and more detailed, accurate, and reliable diagnosis and assessment could be a contributing factor to help increase the chance of effective treatment. This could decrease disability over time, thus cutting down on direct and indirect healthcare costs [2].

2.4 Objective Diagnosis and Assessment Methods

As discussed in Section 2.3 the clinical monitoring and assessment methods of PD are subjective, infrequent, time-consuming, depending on patients recall and prone to inter and intra-rater reliability issues. Consequently, they are unhandy in routine clinical practice. Fortunately, the advances in sensing technologies have enabled the development of new approaches for objective assessment of PD motor symptoms. A large and growing body of literature have explored different objective methods to monitor and evaluate PD tremor, particularly wearable devices combined with soft computing techniques and statistical analysis.

Wearable devices with built-in accelerometers, force sensors, gyroscopes and magnetometers have been the most popular approach for objective assessment [57] because they are inexpensive, portable, easy to use and low power consumption. The objective assessment methods can be categorised into two main groups based on application area: PD diagnosis and PD progression are presented in Sections 2.4.1 and 2.4.2 respectively.

2.4.1 Objective Diagnosis

To date, several studies have proposed approaches to differentiate the people affected with PD tremor from healthy people or people those affected with other movement disorders using motion sensors, such as accelerometers, gyroscopes and magnetometers [11, 57]. In this section, a literature review is presented on the usage of these sensors for PD tremor diagnosis, and can can be classified into two three categories based on classification approach which are frequency analysis, statistical analysis and machine learning approaches.

Authors in [58, 59] proposed a methods to identify PD patients based on frequency of the tremor using an an accelerometer attached to the index finger. In [58], the data were collected from 15 participants (4 PD patients, 4 ET patients, and 7 healthy controlled participants) , while in [59], the data were collected from 12 patients having different movement disorders, PD, ET and drug-induced (DI). The authors in [58] showed that the mean frequency of PD tremor 5.92 Hz for the x-axis, 6.42 Hz for the y-axis, 6.46 Hz for the z-axis, while in [59], the diagnosis was based on the calculated frequency (PD 3-7 Hz resting tremor; ET 4-12 postural tremor, DI 3-12, kinetic tremor) and achieved 95% correlation neurologist's diagnosis. These results confirmed what has been already established regrading PD tremor frequency.

A number of research studies have been carried out on the diagnosis of PD tremor utilising statistical techniques. For example, Braybrook et al. [60] used commercially available device PKG[1] to detect PD tremor. The system contains a tri-axial accelerometer, and it is wearable on the wrist. In this study, 85 people with PD and 28 control healthy subjects. The data were collected from the accelerometer and segmented into one second windows. Each window converted to the frequency domain through FFT. The tremor identified when at least 10 consecutive steps spectral peak have power more than 6 dB, and a frequency between 2.8 Hz and 10 Hz and the frequency should not differ more than 0.4 Hz from the two adjacent steps. The results were promising and achieved 0.92 Area Under the Curve (AUC), 92.55% sensitivity and 92.9% specificity.

Bove et al. [61] conducted a study to distinguish between three tremor disorders; PD, ET and dystonic tremor (DT). The study involved 60 patients (20 patients each group), where a tri-axial accelerometer was placed on hands metacarpals, the data was recorded during rest, posture. Different features were extracted after converting the signals to the frequency domain, including

[1] https://globalkineticscorporation.com/the-pkg-system/

frequency, peak dispersion, spectral coherence, the amplitude of tremor (action, resting) and unilateral tremor. The criteria to identify each tremor are described in Table 2.3. The result was promising, and the performance metrics calculated according to these diagnostic criteria were significant: DT (85% sensitivity; 87.5% specificity), ET (95% sensitivity; 90% specificity), and PD (100% sensitivity; 93% specificity).

A number of authors have considered ML techniques for PD tremor diagnosis, for example, Molparia et al. [62] tried to employee genetic information along with accelerometer data to improve ET, PD diagnosis. In this study, data collected from 33 PD patients and 24 ET patients worn a smartwatch (contains tri-axial accelerometer) on their tremor dominant hand for two weeks. Raw data from the accelerometer was sampled every 20 milliseconds, and the overall magnitude (the root of the sum of squares) of each data point calculated resulting a vector of acceleration processed through a 3-7 Hz Butterworth band-pass filter. Accelerometer features were calculated from the filtered data included total energy, average amplitude, and maximum amplitude. Linear model utilised to differentiate PD from ET from accelerometer features and achieved 76% sensitivity, 65% specificity and 75% AUC. PD polygenic risk score was calculated and added as a genetic feature to the model, but the results did not improve and achieved 80% sensitivity, 65% specificity and 73% AUC.

Surangsrirat et al. [63] used Support Vector Machine (SVM) classifier to discriminate PD from ET patients. In this study, an inertial measurement unit (IMU) contains an accelerometer and a gyroscope attached to the middle finger of the tremor dominant hand of 32 PD and 20 ET patients. Data were collected

Table 2.3: Tremor identification criteria of PD, ET and DT.

	Dystonic tremor (DT)	Essential Tremor (ET)	PD Tremor
Frequency (Hz)	4-10	5 - 15	4 - 7
Peak dispersion (Hz)	≥ 3	≤ 2.5	2.5 - 35.5
Spectral coherence	$\leq 60\%$	$\geq 80\%$	$\geq 70\%$
Tremor Amplitude	Action > Rest	Action > Rest	Rest > Action
Unilateral	No	No	Yes

from the accelerometer and the gyroscope in the IMU while subjects were performing two tasks, kinetic and resting for 10 seconds. Raw data were filtered using a Butterworth ban-pass filter between 3-10 Hz. The temporal fluctuations of filtered tremor signal during kinetic and resting tasks were used to train the SVM model. Features were extracted from the gyroscope (angular velocity) only used in this study. The classification results were highly significant for data extracted from X-axis, with 100% accuracy, 100% sensitivity and 100% specificity.

The previous studies used sensors such as accelerometer, gyroscope, and smartwatches as wearable devices to distinguish PD from ET or healthy control subjects. Researchers in the last few years have explored the ability to utilise built-in smartphone sensors, such as accelerometer and gyroscope, to quantify and diagnosis PD tremor from other movement disorders or healthy people. For example, a study by Woods et al. [64] used a smartphone's tri-axial accelerometer to differentiate 14 PD and 18 ET patients. The Data were recorded while patients held the smartphone in their hands and performed different tasks, including open eyes, closed eyes, attending to laser target at one and two meters and counting backwards by three. Data were processed through discrete wavelet transform (DWT) to produce frequency bins, pins' energy were used as features for the SVM model with RBF kernel for classification. The result was significant and achieved 96.4% accuracy.

Likewise, in [65] a smartphone built-in tri-axial accelerometer was used to distinguish 21 PD patients from 21 healthy subjects. Four features were extracted from collected data, including energy, mean value, variance and entropy. For classification, an ANN model was built and trained based on the mean square error to adjust model weight and thresholds values. 10-fold cross-validation technique used to evaluate the model and achieved highly significant results, 95% accuracy, 95% sensitivity and 95% specificity.

The previous studies mainly focused on accelerometers and gyroscopes actuators/sensors or combining these sensors. Researchers also explored biopotential sensors that detect small electrical signals generated within the body, such as muscle or heart. Sensors such as electroencephalogram (EEG), electromyogram (EMG) and mechanomyogram (MMG), independently or combined with other actuators to diagnose PD from a healthy subject or other movement disorders. Arvind et al. [66] conducted a study to detect rest tremor using the EMG signal Power Spectral Density (PSD). The PSD was extracted by two methods, namely autoregressive Burgs and Welch and feed independently to Elman Neural Network (REN) for classification. Results of both methods were evaluated in terms of accuracy. Burgs method performance was better in terms of accuracy and achieved 95.66%, while Welch method obtained 90.41%.

Nanda et al. [67] used wavelet analysis on accelerometer and EMG signals to differentiate PD tremor from ET. The study involved one PD patient and one ET patient. The data were collected by attaching the accelerometer to the patient's thumb and the EMG to the patient's extensor digitorum muscle. Data were processed through low and high pass filters. Eight wavelet features were calculated, including mean absolute value, variance, RMS value, log detector, average amplitude change, difference absolute standard deviation value, standard deviation and maximum fractal length. Four features only were selected (variance, RMS, standard deviation and maximum fractal length value) based on the highest percentage difference between PD tremor and ET. ANN model was used to classify the first 45 seconds of the acquired signal windowed into 20 seconds length with 50% overlap. The results were significant and higher with accelerometer data to achieve 100% accuracy, while it was slightly lower with EMG signal and obtain 91.66% accuracy.

Another study by Ghassemi et al. [68] also combined accelerometer and EMG to differentiate ET from PD using a machine learning algorithm. In this study, data were collected from 13 PD and 11 ET patients by fixing two accelerometers to the dorsal side of both hands, and two electrodes of EMG fixed on the extensor and flexor muscles of both arms. Signals were recorded and processed through low and high pass filters with cut off frequencies 70 Hz and 20 Hz, respectively. Filtered signals were decomposed using DWT into 10 levels corresponding to different frequency bands to extract 524 features. The principal component analysis (PCA) technique was used to reduce features dimensionality. The results used to train a binary SVM classifier. In this study, two SVM kernels (linear and radial basis function (RBF)) were examined using the grid-search technique to tune classifier parameters. The highest accuracy was 83% with three PCA components and RBF kernel.

2.4.2 Progression and Automatic Scoring of PD Tremor

PD is a chronic and progressive neurodegenerative disorder which means that the symptoms can become worse overtime [2]. The progression of PD is highly heterogeneous and variable across the patients in terms of the presence and severity of symptoms [69]. In addition, motor fluctuations and the confounding effects of treatment such ON-OFF periods when the medication suddenly and unpredictably starts or stops to take effect [70]. The progression is subject to numerous factors including family history of PD, genetics, gender and age of onset [71]. The ability to objectively quantify disease progression would improve the understanding of symptom fluctuations. Therefore, many studies have investigated the use of different sensors and technologies to monitor and measure PD progression remotely and within a clinic.

Researchers in [72] used tri-axial accelerometers and tri-axial gyroscopes embedded in IMU. The study involved six PD patients treated with DBS and seven healthy subjects. Tremor data collected from IMU placed on the trunk and the wrist, the thigh, and the foot of the most affected side of patients and the dominant side of the healthy group, data were collected under three DBS settings (optimal settings, 80% of optimal settings and off), while the subjects were performing UPDRS motor tasks (rest, postural and kinetic). All tasks were videotaped and scored by a physician. The data were filtered between 3.5-9 Hz for rest tremor and between 3.5-12 Hz for Kinetic tremor. Filtered data were segmented into three seconds windows. For rest tremor, the frequency spectrum and RMS were calculated for each window. For Kinetic tremor, the variance of frequency was calculated. Tremors were classified using a decision tree (DT) classifier and evaluated using accuracy performance measure through the LOOCV technique. The classifies achieved an accuracy of 94.1%, 81.7% for rest and kinetic respectively.

Even though the previous studies achieved good results of monitoring PD objectively, most clinicians continue to use clinical assessment. According to Martinez-Manzanera et al. [73], a possible explanation for this might be that the clinicians are more familiar with clinical assessment, besides the gap between clinical assessment methods and the objective assessment approaches. In addition to other factors such as lack of users' perspectives in devices design and development which will be discussed later in Section 2.4.3. Therefore, researchers tried to fill this gap by developing an objective measurement as close as possible to clinical rating scales. The underlying concept is to develop an objective monitoring system that automatically scores PD symptoms similar or as close as possible to golden standard rating scales such as UPDRS, and this forms the core of this thesis. Several studies have explored different sensors with different methods to estimate tremor severity, including soft computing techniques and statistical analysis. A few of these studies are reviewed below.

Giuffrida et al. [74] used Kinesia™ system, which is a sensor integrates accelerometer and gyroscope, for PD tremor severity score assessment. In this study, the data were collected from 60 PD patients, while the sensor was placed on the middle finger of the most affected hand. The data were collected while the subject was performing three UPDRS motor tasks, including rest, postural and kinetic tremor. Collected data were passed through a 3-10 Hz band-pass filter, set of features were calculated from all six axes (three accelerometer axes and three gyroscope axes). This study utilised a multiple linear regression algorithm that correlated calculated features with the UPDRS score. The coefficient of determination, $r2$, with LOOCV used to evaluate the model. The results showed that the natural logarithm of peak power obtained the strongest correlations in rest and postural tremors for both accelerometer and gyroscope, $r2 = 0.89$ and $r2 = 0.90$ respectively, while the RMS of all signals (accelerometer and gyroscope) had the highest correlation with kinetic tremor with $r2 = 0.69$. However, this study is limited in that patients can not wear the device on the finger for a long time. Moreover, collected data tremor severities were not balanced and biased toward less severe tremor and no severity 4 data recorded, which may affect the classification accuracy.

Niazmand et al. [75] have used an accelerometer to estimate tremor severity utilising data collected from 10 PD patients and 2 healthy control subjects. The Data were collected from integrated pullover tri-axial accelerometers while subject performing rest and posture UPDRS motor tasks. The total acceleration was calculated from the accelerometer axes, then normalised by subtracting earth gravity. Normalised data were filtered and used to calculate relative acceleration by calculating the difference between absolute acceleration and filtered acceleration. After that, the movement frequency was calculated and tremor assessment based on frequency bands, as shown in Table 2.4. The correlation between the measurements from accelerometers and UPDRS scores was calculated and achieved 71% sensitivity of detecting rest tremor and 89% sensitivity of detecting posture tremor. However, the study suffered from the fact that the data collected came from pullover fits exactly to the patients achieved good results. At the same time, it is lower for the loose-fitting pullover, and this limitation can be a barrier from using this pullover for PD assessment in routine and continuous assessment since it might be not comfortable for patients and it is difficult to design a pullover to fit all patients. Moreover, using a fit pullover might increase the tense of muscles, particularly in posture tremor, which can change accelerometer position depending on executed movements. Also, this study is limited by the lack of information about patients' UPDRS severities and might be biased towards some severities.

In a follow up study [76], the authors used a glove that contains a set of sensors to quantify motor symptoms. For tremor evaluation, three tri-axial accelerometers were placed on the middle finger and the wrist within the glove. Data were collected from four PD patients and one healthy subject while they performing rest and postural tasks, each task for for 15 seconds. Similar to [75] a frequency ranges used to quantify tremor. The results showed that the correlation between the calculated frequencies and the offline UPDRS score achieved 100% sensitivity and 83% specificity. However, the authors also does

Table 2.4: Tremor severity vs frequency ranges.

Frequency range (Hz)	Tremor severity
0 - 0.50	0
0.5 - 0.9	1
0.9 - 1.8	2
1.8 - 3.4	3
> 3.4	4

not offer information about patients' UPDRS scores. In addition the research does not take into account the difficulty of wearing gloves during ADLs.

Rigas et al. [77] conducted a study to estimate tremor severity using a set of wearable accelerometers were placed on arms and rest of the subjects. This study involved ten PD patients with tremor range from 0 to 3 according to the UPDRS score, 8 PD patients without tremor and 5 healthy subjects. Data were collected from the subjects while performing ADL tasks. Recorded data were processed through low and band-pass filters with cut-off frequencies 3 Hz and 3-12 Hz respectively. The filtered signals were segmented into 3 seconds windows with 1.5 seconds overlap. Set of features were extracted from each window including dominant frequency, the energy of dominant frequency, high and low frequencies energy, spectrum entropy and mechanical. A HMM with LOOCV employed to estimate tremor severity. They have achieved 87% overall accuracy with 91% sensitivity and 94% specificity for tremor 0, 87% sensitivity 82% specificity for tremor 1, 69% sensitivity and 79% specificity for tremor 2, 91% sensitivity and 83% specificity for tremor 3. However, this study suffered from a lack of severe severity (tremor severity 4) in the collected data; thus, it cannot be generalized and used to assess all patients, particularly PD patients with severity 4 tremor, besides the relatively low sensitivity of 69% and specificity of 79% for classification severity 2 tremor.

Wagner et al. [78] collected tri-axial accelerometer data from 19 PD patients using a smartwatch while they are performing five motor tasks including sitting quietly, folding towels, drawing, hand rotation and walking. In this study, data were segmented into 10 seconds windows with 50% overlap. A wavelet features extraction technique was used to process the acquired signal and extract relative energy and mean relative energy for each tri-axial accelerometer axis. Extracted features were used to predict tremor severity into three tremor levels 0, 1 and 2 where 2 represents tremor severities 2, 3 and 4 using SVM classifier, the prediction made by summing all axis prediction since the tremor is often in only one axis. The model was evaluated using LOOCV and achieved 78.91% overall accuracy, 67% average precision ,79% average recall and 75% AUC. However, severity 2 prediction precision was 28%, and it is very low in comparisons with severity 1 and 3 as they achieved 98% and 75% respectively. In addition, a major problem with this experiment was that severities 2,3 and 4 combined into one score 2, which does not reflect patients' actual severity and might not be helpful for neurologists to assess the tremor and does not identify tremor development, especially in advanced stages.

Jeon et al. [79] used tri-axial accelerometer and tri-axial gyroscope attached to the index finger to estimate tremor severity. Resting tremor signals were acquired from 85 PD patients when they were seated, and their forearms rested on the arms of the chair for one minute. Signals were passed through a band-pass filter between 1 and 16 Hz. Features were extracted from the time and the frequency domains including mean amplitude, average regularity, the standard deviation of the regularity, peak frequency, mean frequency, peak power, mean power. Two features reductions techniques were used independently; a pairwise selection strategy based on correlation coefficients. Several machine learning algorithms were evaluated; DT, SVM, discriminant analysis, RF, and k-nearest-neighbour (kNN). The highest accuracy achieved was 85.5% with the DT and the pairwise correlation method for features reduction. The average sensitivity and precisions were 62.3%, and 74.6% respectively. Despite that this study achieved a relatively high accuracy of tremor severity assessment, the sensitivity was low. A possible explanation for this might be that collected data tremor severities were not balanced, and bias toward less severe tremor. This bias can cause significant changes in classification output as the DT algorithm is sensitive to the perturbations in training data [80]. Besides, it is unlikely that patients can attach the sensors to finger for a long time. In a follow-up study [81], the authors tried to improve performance by adding new features to the earlier calculated ones. In this study, the newly added features were logarithms and relative powers of calculated features in their previous study. The highest accuracy was achieved with SVM and obtained 92.3% accuracy, 95.48% average sensitivity and 96.68% average specificity. These results demonstrated the impact of selecting the best feature for tremor severity classification, but both studies suffer from that collected is biased towards low sever tremors. Also, the latter study did not include the UPDRS 0 tremor score.

Pierleoni et al. [26] used tri-axial accelerometer attached to the wrist to estimate resting, postural and kinetic tremor severity. In this study, data were collected from 30 PD patients. Raw signals band-pass filtered from 3 to 12 Hz. Filtered signals were segmented into 4 seconds windows with 25% overlap. The PSD distributions were used to estimate tremor severity including F50 (median frequency) which is the frequency band below that 50% of PSD is present, and SF50 (frequency dispersion) which is the frequency band that contains 68% of the PSD that is centred on F50. The results were significant and correlated with neurologists scores and achieved 100% sensitivity and 100% specificity. However, this study ignored the non-tremor windows. Hence, the proposed approach only can classify only tremor data and need data cleaning before classification. Moreover, no other performance metrics reported such as error rate, sensitivity, specificity, precision and F-score which are very important to evaluate classification models, especially in medicine classification problems, since accuracy neglects the difference between types of errors [82].

Angeles et al. [83] used signals were acquired from a tri-axial accelerometer and MMG sensors placed on the upper arm and the forearm of the most severely affected arm of seven PD patients. Acquired signals were passed through a band-pass filter range from 2-15 Hz for tri-axial accelerometer and 20-30 Hz for MMG sensor. Features were extracted from the filtered signals including mean and standard deviation of acceleration and mean of MMG. Several classification algorithms were evaluated, and the highest accuracy achieved was 87.7% with KNN. However, a significant problem with this experiment was the unbalanced distribution of the UPDRS ratings, as patients' tremor UPDRS scores were between 0 and 3 and did not include UPDRS score 4. This may have affected the performance of the classifier used to predict the UPDRS scores.

Butt et al. [84] used IMUs contain a tri-axial accelerometer and tri-axial gyroscope attached to thumb, index and middle fingers of 59 PD patients. The data were collected while participants were performing MDS-UPDRS motor tasks. PD patients were divided into two groups, group 0 belongs to slight and mild tremor, and group 1 belongs to moderate and severe tremor. Acquired signals, except rest's task, were passed through low- and high-pass filters with cut off frequencies 5 Hz and 0.5 Hz. Set of features were extracted from the filtered signals such as angular velocity, energy, amplitude, average and peak power. Least Absolute Shrinkage and Selection Operator (LASSO) and

Kruskal–Wallis (KW) were used for features selection. For classification, three models were evaluated; SVM, ANN, and logistic regression (LR). SVM achieved the highest accuracy of 79.66%, sensitivity of 92.10%, specificity of 57.4% and AUC of 87.09%. This study is limited in that it only used two classes that do not represent PD progress precisely. In addition, the classifier did not achieve high accuracy either specificity.

Similarly, Dai et al. [85] used IMUs contain a tri-axial accelerometer and tri-axial gyroscope attached to index finger of 45 PD patients and 30 healthy subjects. The data were collected while participants were performing RT, PT and finger tapping (FT) tasks from MDS-UPDRS. Tasks were videotaped for tremor severity scoring by seven neurologists offline. The collected signals were passed through 3-10 Hz band-pass filter. Features were extracted from filtered signals including logarithm of squared magnitudes of linear accelerations and angular velocities, logarithm of squared summation of the differences between two adjacent acceleration, logarithm of peak amplitude from angular velocities, tapping frequency, amplitude, average tap angle, and standard deviation of angles. Three classifier were tested to estimate tremor severity including SVM, RF and KNN, The highest results were achieved using SVM classifier for all tremor with 97.33% for PT, 96% for RT, and 96% for FT. However, similar to previously discussed studies [74, 79, 81, 84, 86], this study is limited in that sensor position on fingers might not be practical for continuous monitoring. Moreover, no information have been provided about severity distribution and without evaluation with advanced metrics, the classifier overall accuracy might be biased to the majority classes.

Zajki-Zechmeister et al. [87] tried to measure rest tremor of 14 PD patients while they were on-medication and off-medication using and accelerometer was attached to index finger and while it was held by patient's hand. The collected signals transformed to the frequency domain and filtered between 3 Hz and 20 Hz. Three features were extracted from each accelerometer's axis including power of main peak, peak frequency and energy. The extracted features analysed statistically to identify the correlation with MDS-UPDRS. The results demonstrated that tremor measurement correlation while patients were on-medication (0.779) is higher than the correlation while patients were off-medication (0.638). However, this study also did not take into accounts tremor severity distribution.

Table 2.5 summarised the studies that tried to quantify PD tremor. It is apparent from this table that two main approaches have been employed to measure tremor: statistical analysis and machine learning. Also, the most commonly used devices were wearable accelerometers and gyroscopes. What stands out in the table is that only few studies included data from all level of tremor severity and some studies combined severities together. The most interesting aspect of this table is that the most commonly used metrics are the traditional metrics such as accuracy, precision, sensitivity, and specificity which are illusory and insufficient to evaluate an imbalanced data , since they are sensitive to data distribution [88], which can cause issues, especially in the medical diagnosis field where misclassified true negative can lead to unnecessary and expensive treatment, only few studies have evaluated the results through advanced metrics such as F1-score [89] and AUC [90]. Moreover, none of these studies used more advanced metrics such as geometric mean (G-mean) [91] and index of balanced accuracy (IBA) [92]. Closer inspection of the table shows that the studies explored different types of tremor and some of these studies measured all types of tremor and some of them measured only one type of tremor.

In terms of machine learning approaches, the highest accuracy achieved was 98% in [93], however this study only measured kinetic tremor. On the other hand, the highest overall accuracy achieved to measure all types of tremor was 97% in [94] and in [85] where the achieved accuracy were 96% for rest, 97.33% for postural and 96% for kinetic. In addition, none of these studies measured all tremor severities from 0 to 4.

Table 2.5: Literature of automatic scoring of PD

Article	Tremor	Patients(n)	Sensor	Approach	Performance Metrics		Tremor Severity
Giuffrida et al. [74]	Rest Postural Kinetic	60	Accelerometer Gyroscope	LR	Coefficient of Determination	Rest 0.89 Postural 0.90 Kinetic 0.69	0, 1, 2, 3, 4
					RMSE	Rest 0.32 Postural 0.35 Kinetic 0.45	
Rigas et al. [95]	Rest	0	Accelerometer	KNN	Accuracy Specificity Sensitivity Precision	92% 95.8% 91.8% 91.5%	Simulated 0, 1, 2, 3, 4
Niazmand et al. [76]	Rest Postural	4	Accelerometer	Statistical	Specificity Sensitivity	83% 100%	Unknown
Niazmand et al. [75]	Rest Postural	10	Accelerometer	Statistical	Sensitivity	Rest 71% Postural 89%	Unknown
Darnall et al. [96]	Rest Postural Kinetic	10	Gyroscope	DT	Accuracy	82%	0,1
Rigas et al. [77]	Rest Postural Kinetic	18	Accelerometer	HMM	Accuracy Specificity Sensitivity Precision	87% 84.5% 84.5% 66.3%	0, 1, 2, 3
Cole et al. [97]	Rest Postural Kinetic	8	Accelerometer Electromyographic	Bayesian maximum likelihood	Specificity Sensitivity	96.43% 98%	*mild* *moderate* *severe*
Tzallas et al. [98]	Rest Postural	24	Accelerometer	HMM	Accuracy Sensitivity	87% 84.8%	0, 1, 2 & (3,4) together

Table 2.5 continued from previous page

Article	Tremor	Patients(n)	Sensor	Approach	Performance Metrics		Tremor Severity
Pierleoni et al. [26]	Rest Postural Kinetic	30	Accelerometer	Statistical	Specificity Sensitivity	100% 100%	0, 1, 2, 3, 4
Pan et al. [99]	Rest	40	Accelerometer	Lasso Regression	Specificity Sensitivity	82% 77%	1, 2, 3, 4
Bazgir et al. [100]	Rest Postural Kinetic	52	Accelerometer	ANN	Accuracy Specificity Sensitivity Relative error	91% 90.64% 89.64% 2.5	0, 1, 2, 3, 4
Kostikis et al. [101]	Rest Postural	25	Accelerometer Gyroscope	Bagged Tree (BgT)	Specificity Sensitivity AUC	90% 82% 94%	0, 1, 2, 4
Dai et al. [102]	Rest Postural Kinetic	7	Accelerometer Gyroscope	Statistical	Correlation Coefficient	0.98	1, 2, 3
Wagner, Fixler and Resheff [78]	Rest	19	Accelerometer	SVM	Accuracy Specificity Sensitivity Precision AUC	78.9% 81.5% 66.93% 79.2% 75%	0, 1, 2
Yohanandan et al. [103]	Postural Kinetic	12	Electromagnetic Motion tracker	RF	kappa		
Coefficient	0.81	0 to 10 (BTRS)					

Table 2.5 continued from previous page

Article	Tremor	Patients(n)	Sensor	Approach	Performance Metrics		Tremor Severity
Alam et al. [86]	Rest Postural	0	Accelerometer Gyroscope	SVM	Accuracy	Rest 88.9% Postural 81.8%	Simulated 1, 2, 3
Lugo et al. [104]	Rest Postural	21	Motion tracker	Statistical	Correlation coefficient	Postural 0.45 Rest 0.43	0, 1, 2, 3, 4
Jeon et al. [79]	Rest	85	Accelerometer Gyroscope	DT	Accuracy Specificity Sensitivity Precision	85.5% 95.6% 62.3% 74.61%	0, 1, 2, 3, 4
Jeon et al. [81]	Rest Postural Kinetic	85	Accelerometer Gyroscope	SVM	Accuracy Specificity Sensitivity Precision RMSE	92.31% 97.3% 95.5% 95.6% 0.039	0, 1, 2, 3, 4
Angeles et al. [83]	Rest Postural Kinetic	7	Accelerometer Mechanomyography	KNN	Accuracy	87.3%	0, 1, 3
Butt et al. [84]	Rest Postural Kinetic	59	Accelerometer Gyroscope	SVM	Accuracy Specificity Sensitivity AUC	79.66% 57.14% 92.1% 87.09%	*slight* & *mild* = 0 *moderate* & *severe* = 1
Bazgir et al. [94]	Rest Postural Kinetic	52	Accelerometer	NB	Accuracy	97%	0, 1, 2, 3, 4
Soltaninejad et al. [105]	Postural	8	Camera Infrared	RF	Accuracy Sensitivity Precision F1-score	81.07% 73.76% 83.55% 71.21%	0,1,2

Table 2.5 continued from previous page

Article	Tremor	Patients(n)	Sensor	Approach	Performance Metrics		Tremor Severity
Cai et al. [106]	Rest Postural	34	Accelerometer	Statistical	Coefficient of Determination	Rest 0.95 Postural 0.93	1, 2, 3, 4
Kim et al. [107]	Rest	92	Accelerometer Gyroscope	CNN	Accuracy Specificity Sensitivity Precision kappa Correlation Coefficient RMSE	85% 94.2% 79.4% 81.3% 0.85 0.93 0.35	0, 1, 2 & (3,4) together
Vivar et al. [93]	Kinetic	21	Motion tracker	BgT	Accuracy Specificity Sensitivity Precision	98% 97.98% 97.62% 98%	0 = *normal* 1 = *slight* 2 = *mild*
López-Blanco et al. [108]	Rest	22	Gyroscope	Statistical	Spearman's Correlation	0.81	0, 1, 2, 3
Sigcha et al. [109]	Rest	18	Accelerometer	AdaBoost	Specificity Sensitivity AUC	86.1% 86.1% 93.6%	0, 1, 2
Zajki-Zechmeister et al. [87]	Rest	14	Accelerometer	Statistical	Correlation Coefficient	Off-med 0.638 On-med 0.779	Unknown

Table 2.5 continued from previous page

Article	Tremor	Patients(n)	Sensor	Approach	Performance Metrics		Tremor Severity
Dai et al. [85]	Rest Postural Kinetic	45	Accelerometer Gyroscope Magnetometer	SVM	Accuracy	Rest 96% Postural 97.33% Kinetic 96%	0, 1, 2, 3, 4
					Specificity	Rest 96.67% Postural 96.67% Kinetic 95%	
					Sensitivity	Rest 100% Postural 97.78% Kinetic 96.36%	
Kostikis et al. [110]	Rest	23	Accelerometer Gyroscope	Statistical	Correlation coefficient	Right-hand 0.7706 Left-hand 0.8793	Unknown
de Oliveira Andrade et al. [111]	Rest Postural	11	Accelerometer Gyroscope Magnetometer	Statistical	Correlation Coefficient	0.75	0, 1, 2, 3, 4
Hoff, Wagemans and Hiltten [112]	Kinetic	7	Accelerometer	LDA	Spearman's correlation	Standing 0.70 Sitting 0.75	0, 1, 2, 3

Statistical analysis approaches that explored all types of tremors achieved the highest correlation of 0.98 in [102]. However, even though this study measured all types of tremor, but it only measured three level of tremor severity (1, 2, 3). On the other hand, the authors in [26] achieved 100% specificity and 100% sensitivity for measuring all types of tremors as well for all severities.

Only few studies have explored different aspects of tremor measurement. For example, in [87], the tremor severity were quantified under two conditions, while patients were on-medication and off-medication. While in [112], the authors explored two tasks (standing, sitting) influence on tremor measurement. In [110], authors reported tremor measurement of the left and the right hands. This indicates a need to explore different aspects of tremor measurement that might improve the objective evaluation PD tremor.

2.4.3 The Patients' and Healthcare Professionals' Perspectives

As discussed in the previous sections, various systems have been developed to monitor and evaluate PD tremor objectively, and they have shown promising results in research and clinical trials. However, a few number of commercial systems are available that can be used [13], and where they do exist, they show limited uptake and adoption in clinical environments [14] despite of the fact that these systems shows high reliability, validity and responsiveness [13, 15], and have been used in the evaluation of PD symptoms and signs by individuals apart from development team and reported successful clinical trials [15, 16]. Data from several studies suggest that these inconsistencies in adoption and implementation may be due to the lack of users' perspectives in devices design and development [17]. Moreover, the adoption and acceptance of the new systems influenced by their intended usage, efficacy and their suitability in terms of clinical context as well as patients' and clinician's' needs [113].

Even though users' preferences need to be considered if wearable devices are to gain acceptance at home or within a clinic [114], a recent systematic review shows that the research to date has tended to focus on wearable devices development from quantitative perspectives such as type of sensors, data extraction, and classification methods [11]. Therein lies a problem, there is little attention paid to users' preferences with few studies reporting patients' or clinicians' experiences, preferences, and expectations of using wearable devices.

High-quality medical devices with high levels of patients' and healthcare professionals' acceptance requires engineering methods such as user-centered design (UCD) philosophy [17] that places an emphasis on patients' and healthcare professionals' individual needs as well the environment where the devices will be used. It is suggested that innovations that are often driven by technology evolution may increase the 'risks' that researchers develop products that only a few people need and are willing to use [115]. Product development must also take into consideration users' involvement in the early stages of device design and development where challenges and possibilities, ideas, and concepts are presented and discussed, thus minimising costly device modifications and reducing recalls. This could lead to more robust usable devices that are better suited to users' needs and ultimately lead to patient benefit [115].

An extensive search of the literature revealed few studies have explored patients' preferences or healthcare professionals' preferences or both [116–120]. However, such studies remain narrow in focus dealing only with device acceptance rather than users' needs, and in addition, most of these studies have investigated either patients' perspectives or healthcare professionals' perspectives and have not included both in the same study. For example, some studies [116] utilised questionnaires to get feedback only from patients linked to existing developed wearable devices and focusing on user satisfaction, comfort and wearing the device publicly. Also, these studies did not involve or explore the design requirements from patients' point of view or from healthcare professionals. Conversely, some studies explored healthcare professionals' perspectives, for example, Santiago et al. [119] conducted a survey to evaluate the impact of using a commercially available wearable device (KinetiGraph™ or PKG) in routine clinical appointments of PD patients from physicians' point of view. The study has explored how a developed device could add value to current assessment, but it did not take into consideration the design aspects and what are the requirements since it is limited by feedback about gathered data.

Few studies explored the needs and the requirements of patients and healthcare professionals. For example, Bergmann et al. [114] used an online questionnaire to identify the preferences of medical wearable devices for people affected with arthritis, so these preferences can be used in devices design and developments. However, this study did not take into consideration clinicians' or physiotherapists' preferences, which ultimately might limit device adoption. In addition, arthritis conditions such as joint pain might influence arthritis patients' preferences; for example, the preferable part of the body to place the wearable device, which might be different from other diseases such as PD. Similarly, Bruno et al. [121] used an online survey to identify users' perspectives toward digital technology and wearable devices, but in this study, people with epilepsy, caregivers, and healthcare professionals were included. However, most of these studies have utilised questionnaires, and approaches of this kind carry with them various well-known limitations [122]. For example, questions understanding and interpretations, the difficulty of conveying feelings and emotions, and they are not flexible and do not allow probing to get in-depth information [123]. In contrast, focus group discussions and interviews can explore a range of views, perceptions, thoughts, and sentiments and can provide insight into complicated subjects [124].

Few studies utilised focused group discussions to gain an in-depth insight into patients' perspectives, and these can help to design and develop wearable devices with high acceptance and adoption. For example, authors in [113] used focus group discussions to identify design requirements and mode of use of wearable technology for patients with osteoarthritis. Similarly, Thilo et al. [120] have used the same approach to identify elderly people perspectives toward wearable technologies, but, of these only one study has explored PD patients' and healthcare professionals' preferences in-depth, utilising focus group discussions [118].

2.5 Discussion and Research Opportunity

Based on an extensive review of the literature, it can be concluded that the current diagnosis and assessment methods of PD have many limitations. Clinical assessment and clinical rating scales are subjective that mainly relies on visual observations and on the clinicians' skills and experience and involves advanced official training to improve the coherence of data acquisition and interpretation. In addition, these methods are infrequent, time-consuming, depending on patients recall and prone to inter and intra-rater reliability issues. These limitations mentioned above might lead to poor management of PD and wasteful use of resources besides that they are unwieldy in routine clinical practice.

The evolution of sensing technologies has enabled the development of new approaches for the objective diagnosis and assessment of PD motor symptoms. A considerable amount of literature has been published on PD diagnosis utilising sensors such as accelerometers, gyroscopes and magnetometers. These approaches tried to differentiate PD patients from healthy people or other movements disorders such as ET. Different features have been extracted from sensors' signals in the time and the frequency domains then analysed and classified statistically or by machine learning techniques. However, these approaches are only limited to diagnosis. Therefore, various studies tried to monitor PD Progression and the efficacy of treatment objectively. Response to treatment can help neurologists to confirm PD diagnosis. In addition, it can help to understand motor fluctuations. These approaches followed the same objective diagnosis methods in terms of used sensors and analysis.

While the previous research has objectively obtained successful outcomes of PD monitoring, most clinicians continue to use clinical evaluation to the disparity between clinical and objective evaluation methods. A likely explanation is that the clinicians are more familiar with clinical evaluation. In addition, the results were not quantified and can not be interpreted easily. Therefore, researchers tried to develop systems that are mimicking clinical rating scales and scores of PD tremors.

Several objective methods have been proposed for measuring and quantifying PD tremors from data collected during performing scripted and unscripted tasks using ML algorithms combined signal processing techniques. These approaches comprise three main steps: (1) Signal processing which includes the removal of non-tremor data or artefacts and the segmentation of filtered signals into windows that represent tremor patterns, (2) Features extraction which is the process of transforming raw signals into a set of features that represent relevant information about tremor severity, (3) Classification or statistical analysis which is the process of predicting tremor severity or correlating extracted features with clinicians' scoring of tremor.

A common limitation in most of the reviewed studies was that the authors did not take into consideration all tremor levels and imbalanced classes distribution among collected data. Also, some of these studies only used data collected while subjects performing specific tasks which do not necessarily include ADLs. Most of these studies did not report advanced performance metrics such as F-score, AUC and IBA, which are very important to evaluate classification models. In addition, the research to date has tended to focus on proposing a tremor severity classification approach without discrimination tasks effect on classification and tremor severity detection, even though motor examination of PD is a key aspect of tremor assessment.

Another crucial aspect that has been highlighted in the literature is the limited adoption of objective assessment in clinics due to the lack of patients' and healthcare professionals' perspectives in terms of wearable devices design and implementation.

To address the identified gaps, this research employed qualitative and quantitative methods to identify a recommended solution that is coherent in healthcare professionals' and patients' points of view to avoid the off-the-shelf solution. In addition, the efforts of this thesis will be focused on enhancing tremor severity by combining signal processing with ML techniques. Besides, exploring the influence of tasks performed during data collection on tremor classification. In the following chapter, a description of the proposed methodology framework is presented.

Chapter 3

Technical Methods

3.1 Introduction

This chapter presents a technical overview of the methods utilised to estimate the severity of tremors including signal processing, features extraction, classifiers, and performance metrics. Methods relevant to each experiment will be described in the respective experimental chapters. Besides, the dataset utilised in this thesis is also explained in this chapter.

3.1.1 MEMS Accelerometers

Recently, the use of Microelectromechanical systems (MEMS) accelerometer-based systems for quantification and characterisation of human movements has recently increased considerably. Consequently, this would introduce new possibilities for continuous and objective monitoring of movement disorders remotely and within a clinic [125]. An accelerometer is an electromechanical device that measures the static or dynamic force of acceleration based on Newton's second low ($Force = Mass \times Acceleration$). There are uniaxial, biaxial and triaxial accelerometers that can measure acceleration in one, two or three dimensions, respectively.

Accelerometers are relatively cheap, portable, small and can measure 3D acceleration directly [126]. In addition, accelerometers can record data for long periods with small memory and low processing power and low energy consumption [127]. Therefore, accelerometers have been utilized in a wide range of applications, such as automotive, biology, industry, medical applications, navigation, orientation sensing, image stabilization, gravimetry, volcanology, transportation, building and structural monitoring [128]. Moreover, Accelerometers has been used in medical applications for diagnosis, assessment and physical activity analysis such as fall detection [129], and evaluation of walking impairment and chronic and neurodegenerative disease [127, 129].

3.1.2 Dataset

Tremor dataset[1] was taken from Levodopa response trial wearable data from the Michael J. Fox Foundation for Parkinson's research (MJFF) [130]. The data were collected from 30 PD patients from two clinical sites over four days from wearable sensors in both laboratory and home environments using different devices that contains triaxial accelerometer; a Pebble Smartwatch [2] on the wrist of the least affected limb, GENEActiv accelerometer [3] on the wrist of the most affected limb and a Samsung Galaxy Mini smartphone in a fanny pack worn in front at the waist. The data were collected at 50 Hz sample rate.

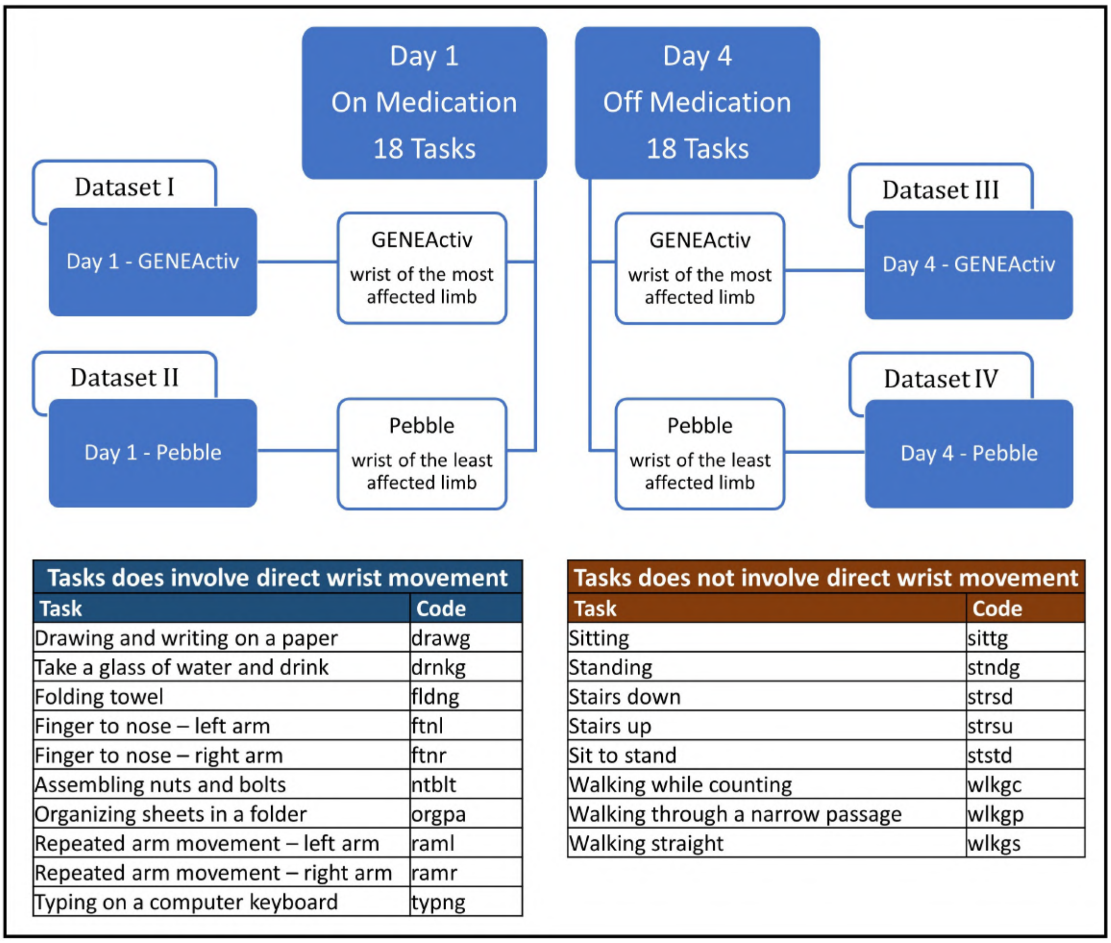
On the first day of data collection, participants came to the laboratory on their regular medication regimen (On Medication) and performed set ADL tasks and tasks of motor examination of the MDS-UPDRS [7] which is used to assess motor symptoms. The list of tasks performed includes; standing, walking straight, walking while counting, walking upstairs, walking downstairs, walking through a narrow passage, finger to the nose (left and right hands), repeated arm movement (left and right hands), sit to stand, drawing on a paper, writing on a paper, typing on a computer keyboard, assembling nuts and bolts, taking a glass of water and drinking, organising sheets in a folder, folding a towel, and sitting. The tasks lasted approximately 20 minutes, excluding walking up and downstairs, and was repeated 6-8 times at 30 minute intervals.

On the second and third days, accelerometers data were collected while participants were at home and performing their usual activities. On the fourth day, the same procedures that were performed on the first day were performed

[1] It is available at https://www.michaeljfox.org/news/levodopa-response-study

[2] https://www.fitbit.com/pebble

[3] https://www.activinsights.com/products/geneactiv/



Tasks does involve direct wrist movement	
Task	**Code**
Drawing and writing on a paper	drawg
Take a glass of water and drink	drnkg
Folding towel	fldng
Finger to nose – left arm	ftnl
Finger to nose – right arm	ftnr
Assembling nuts and bolts	ntblt
Organizing sheets in a folder	orgpa
Repeated arm movement – left arm	raml
Repeated arm movement – right arm	ramr
Typing on a computer keyboard	typng

Tasks does not involve direct wrist movement	
Task	**Code**
Sitting	sittg
Standing	stndg
Stairs down	strsd
Stairs up	strsu
Sit to stand	ststd
Walking while counting	wlkgc
Walking through a narrow passage	wlkgp
Walking straight	wlkgs

Figure 3.1: Tremor dataset.

once again, but the participants were off medication for twelve hours. For each task, on the first day and the fourth day symptom severity scores (rated 0-4) were provided by a clinician.

The list of tasks performed can be categorised into two groups as shown in Figure 3.1 : The first group includes tasks that involve direct wrist movement such as drawing on a paper or writing on a paper. The second group includes tasks that do not involve direct wrist movement such as sitting or standing. In this work, only labelled data was used, which is the data was collected on day one and day four from the GENEActiv accelerometer and Pebble Smartwatch.

Table 3.1 shows classes (severities) distribution of $103,080$ instances (windows), segmented from collected data. It is clear how data distribution being

Table 3.1: Imbalanced classes (severities) distribution.

Tremor Severity	GENEActiv		Pebble		Total
(Class)	Day 1	Day 4	Day 1	Day 4	($n = 103080$)
0	18843	16860	19389	17215	72307
1	5845	6534	4491	4421	21291
2	2185	2921	1357	1112	7575
3	845	676	117	103	1741
4	43	53	11	59	166

skewed towards less severe tremors, and this bias can cause significant changes in classification output, in this situation the classifier is more sensitive to identifying the majority classes but less sensitive to identifying the minority classes. Thus, different resampling techniques are described later in Section 3.1.5 were utilised to eliminate the imbalanced class distribution effect.

3.1.3 Signal Processing

The accelerometer signal is mainly composed of gravity acceleration and human movement body acceleration, in addition to the intrinsic noise of the electronic system and measurement circumstances [131]. Therefore, some preprocessing techniques are performed to eliminate non-tremor data or artefacts in order to extract meaningful features from accelerometer data. For example, to avoid dependency on sensor orientation and to avoid processing signal in three dimensions, the vector magnitude of three orthogonal acceleration is calculated.

Afterwards, digital band-pass Butterworth filters are employed to remove low and high-frequency bands and retain tremors bands from the data. The filter's low and high cut off frequencies are chosen based on tremor frequency which is between 3-12 Hz [26].

The next step is to split the filtered signals into consecutive windows or segments using the overlap sliding window technique, which has been shown in the literature to be effective in activity recognition [132]. This step helps to identify the most relevant characteristics of the signal and to identify the

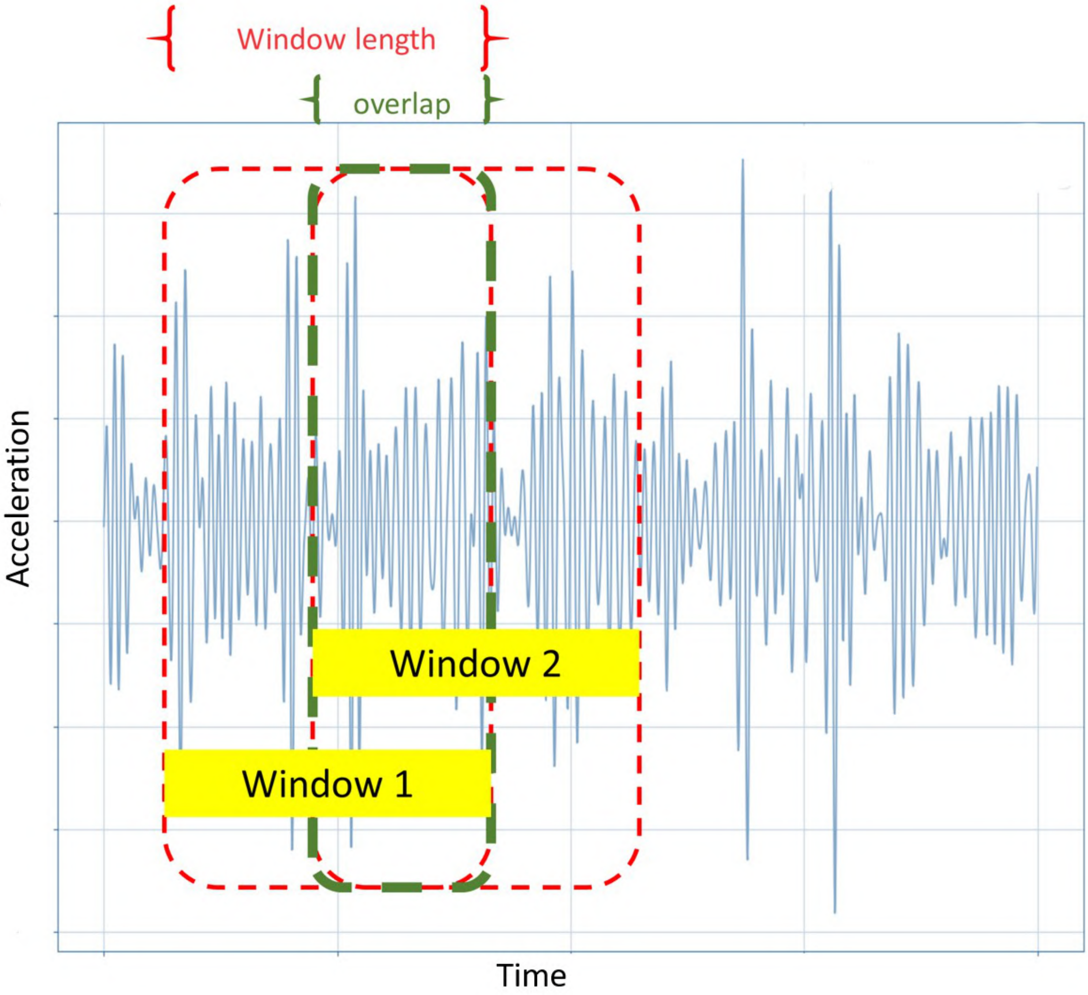


Figure 3.2: Sliding window segmentation technique

boundaries of tremors that are present in the signal. The underlying idea is to split a continuous signal into labelled chunks with a defined length and overlap between windows that can be used to extract features and as inputs to a ML algorithm. Figure 3.2 shows the working principle of the sliding window technique.

3.1.4 Features Extraction

Features extraction is the process of transforming raw data into a set of features that representing relevant information about input patterns or classes [133]. Thus, features extraction is a crucial phase in the development of any classifier. Features can be extracted in the original space or in transformed space, such as transforming the time domain to the frequency domain. These features capture signal behaviour over time or frequency, such as intensity, variation, range of motion and signal complexity. Consequently, that may contribute to identifying tremor severity. Therefore, an optimal feature shows a considerable variation between classes and a small variation between input data belongs to the same class.

Frequency domain features were calculated after transforming the raw signal from the time domain to the frequency domain using Fast Fourier Transform (FFT) as given below:

$$F(k) = \sum_{t=0}^{W_l-1} a_t e^{\left(\frac{-j2\pi kt}{W_l}\right)} \qquad \text{For} \quad k = 0 \ldots W_l - 1 \tag{3.1}$$

where $F(k)$ complex sequence that has the same dimensions as the input sequence $(\mathrm{a_t})_{t=0}^{w_l}$ and $e^{\frac{-j2\pi}{W_l}}$ is a primitive N^{th} root of unity.

Based on the different works reviewed in Section 2.4, various features in time and frequency domains are presented below:

- Above mean:

$$\left|W^+\right| \ : \ W^+ = \left\{ a_t \in W \ : \ a_t > \left(\frac{1}{W_l} \sum_{t=0}^{W_l} a_t \right) \right\} \tag{3.2}$$

  where W^+ is the window subset contains elements above the mean, a_t is the acceleration at time t, and W_l is the corresponding window length (number of samples).

- Below mean:

$$\left|W^-\right| \ : \ W^- = \left\{ a_t \in W \ : \ a_t < \left(\frac{1}{W_l} \sum_{t=0}^{W_l} a_t \right) \right\} \tag{3.3}$$

  where W^- is the window subset contains elements below the mean, a_t is the acceleration at time t, and W_l is the corresponding window length (number of samples).

- Autocorrelation:

$$\hat{R}(l) = \frac{1}{(W_l - l)\, s_w^2} \sum_{t=0}^{W_l - l} (a_t - \overline{a})\,(a_{t+l} - \overline{a}_w) \tag{3.4}$$

where l is the lag, a_t is the acceleration at time t, W_l, $\overline{a}_w$, and s_w are the window length (number of samples), the mean acceleration and the sample standard deviation of the corresponding window respectively

- Complexity-invariant distance:

$$CID = \sqrt{\sum_{t=1}^{W_l - 1} (a_t - a_{t+1})^2} \tag{3.5}$$

where W_l is the corresponding window length (number of samples), a_t is the acceleration at time t.

- Sample entropy:

$$SampEn(m, r) = \log_e \left(\frac{A^{m+1}(r)}{A^m(r)} \right) \tag{3.6}$$

where $A^m(r)$ is the probability that two vectors of (m) points within one window would match, while $A^{(m+1)}(r)$ is the probability that two vectors of $m + 1$ points within one window would match. m is the length of sequences to be compared , r is the tolerance value for accepting matches.

- Kurtosis:

$$\beta_2 = \frac{1}{W_l} \frac{\sum_{t=0}^{W_l} (a_t - \overline{a}_w)^4}{s_w^4} \tag{3.7}$$

where a_t is the acceleration at time t, W_l, $\overline{a}_w$, and s_w are the window length (number of samples), the mean acceleration and the sample standard deviation of the corresponding window respectively.

- Skewness:

$$\gamma_1 = \frac{1}{W_l}\frac{\sum\limits_{t=0}^{W_l}\left(a_t - \overline{a}_w\right)^3}{s_w^3} \tag{3.8}$$

where a_t is the acceleration at time t, W_l, $\overline{a}_w$, and s_w are the window length (number of samples), the mean acceleration and the sample standard deviation of the corresponding window respectively.

- Standard deviation:

$$s_w = \sqrt{\frac{\sum\limits_{t=0}^{W_l}\left(a_t - \overline{a}_w\right)^2}{W_l - 1}} \tag{3.9}$$

where a_t is the acceleration at time t, W_l, $\overline{a}_w$, and s_w are the window length (number of samples) and the mean acceleration of the corresponding window respectively.

- Maximum acceleration:

$$\max(a) = \max_{t=0}^{W_l} a_t \tag{3.10}$$

where a_t is the acceleration at time t, and W_l is the window length (number of samples) of the corresponding window.

- Mean:

$$\frac{1}{W_l}\sum_{t=0}^{W_l} a_t \tag{3.11}$$

where a_t is the acceleration at time t, and W_l is the window length (number of samples) of the corresponding window.

- Median:

$$\begin{cases} a_t^{(i)} & : i = \frac{W_l^{(\mathcal{O})}+1}{2} \\ \dfrac{a_t^{(i)} + a_t^{(i+1)}}{2} & : i = \frac{W_l^{(\mathcal{E})}}{2} \end{cases} \tag{3.12}$$

where $W_l^{(\mathcal{O})}$ window length is odd, $W_l^{(\mathcal{E})}$ window length is even, and i is the element position (index) in the window $\{W\}$

- Sum of absolute differences:

$$SAD = \sum_{t=0}^{W_l-1} \left| a_{(t+1)} - a_t \right| \tag{3.13}$$

where a_t is the acceleration at time t, and W_l is the window length (number of samples) of the corresponding window.

- Energy:

$$E = \sum_{t=0}^{W_l} a_t^2 \tag{3.14}$$

where a_t is the acceleration at time t, and W_l is the window length (number of samples) of the corresponding window.

- Peaks:

$$|P| \ : \ P = \left\{ max \left\{ a_{(n+m+k)} \right\}_{k=-n}^{n} \right\}_{m=0}^{W_l-(2n-1)} \tag{3.15}$$

where n is the number of neighbours, $a_{(n+m+k)}$ is the acceleration at time $(n+m+k)$, and W_l is the window length (number of samples) of the corresponding window.

- Amplitude of peak PSD:

$$A_{(PSD_P)} = \max_{W} \left(\sqrt{PSD} \right) = \max_{a_t \in W} \left(\sqrt{ \frac{1}{W_l} \left| \sum_{t=0}^{W_l-1} a_t e^{\left(\frac{-j2\pi kt}{W_l} \right)} \right|^2 } \right) \tag{3.16}$$

where a_t is the acceleration at time t, W is the window number, W_l is the window length (number of samples) of the corresponding window, and $(e^{\frac{-j2\pi}{W_l}})$ is the primitive Nth root of unity.

- Median frequency:

$$f_{med} \ : \ \left(\sum_{f=f_l}^{f_{med}} PSD \right) = \left(\sum_{f=f_{med}}^{f_h} PSD \right) = \left(\frac{1}{2} \sum_{f=f_l}^{f_h} PSD \right) \tag{3.17}$$

where f is the frequency bin, PSD, f_l and f_h are the power spectral density, lowest and the highest frequency in the corresponding window respectively.

- Frequency dispersion:

$$f_{disp} = 2f_{step} \ : \ \left(\sum_{f_{med}-f_{step}}^{f_{med}+f_{step}} PSD \ = \ \frac{68}{100} \sum_{f=f_l}^{f_h} PSD \right) \tag{3.18}$$

where f is the frequency bin, PSD, f_l and f_h are the power spectral density, lowest and the highest frequency in the corresponding window respectively. f_{step} is the range between the median frequency and the lower bound of dispersion frequency, which is equal to the range between median frequency and the higher bound of dispersion frequency. i.e. $2fstep$ is the range between lower and higher bound of of dispersion frequency in the corresponding window.

- Fundamental frequency:

$$f_{fund} \ : \ PSD_{fund} = \max_{f_l}^{f_h} \ \{PSD\} \tag{3.19}$$

where f is the frequency bin, PSD, f_l and f_h are the power spectral density, lowest and the highest frequency in the corresponding window respectively. PSD_{fund} is the power spectral density at the fundamental frequency in the corresponding window.

- Frequency difference:

$$f_{\Delta} = f_{med} - f_{fund} \tag{3.20}$$

where f_{med} and f_{fund} are the median and the fundamental frequencies in the corresponding window respectively.

- Spectral Centroid Amplitude:

$$SCA = \frac{\sum\limits_{f=f_l}^{f_h} (f)(PSD)}{\sum_{f=f_l}^{f_h} (f)} \tag{3.21}$$

where f is the frequency bin, PSD, f_l and f_h are the power spectral density, lowest and the highest frequency in the corresponding window respectively.

- Maximum weighted PSD:

$$PSD(w)_{max} = \max_{f_l}^{f_h} \{(f)(PSD)\} \tag{3.22}$$

where f is the frequency bin, PSD, f_l and f_h are the power spectral density, lowest and the highest frequency in the corresponding window respectively.

3.1.5 Resampling Techniques

In many real-world datasets, the imbalanced data problem emerges when the classes data distributions are excessively skewed [134]. Consequently, the majority classes dominate the minority classes; hence, the ML classifiers are much more inclined to the majority classes and their results are not reliable since most machine learning algorithms perform better when the number of samples of each class is approximately equal. However, in many applications detecting the minority classes is very important such as disease diagnosis in the medical field, or fault detection in the industrial field. Therefore, various approaches have been developed that can handle the imbalanced data problem.

The solutions approaches for imbalanced classes can be divided into two categories: algorithm level approaches and data level approaches (resampling) [135]. The algorithm level approaches modify existing classification algorithms to increase their learning ability with regard to the minority classes and can be categorised into cost-sensitive approaches and ensemble approaches. The cost-sensitive approaches assign different weights to the types of misclassification errors that can be made, while ensemble approaches are designed to combine the predictions from several ML models in order to take misclassification cost into account, hence obtain improved predictive performance when compared to the use of a single model. The data level approaches try to balance classes distribution or remove samples that are difficult to classify in the training data before building the classifier. However, the algorithm level approaches require a strong understanding of the modified classifier and the application domain in which it will be implemented. Therefore, the data level approaches are most common and preferred in many applications [135].

Resampling methods can be categorised into three groups: over-sampling, under-sampling, and hybrid (Combination of over- and under-sampling). These approaches are described in more detail in the following sections.

3.1.5.1 Over-sampling Techniques

Over-sampling techniques are used to increase the minority classes instances or samples by creating new instances or duplicating some instances. Considering a sample c_i, and c_ji is one of the nearest neighbours selected according to its k neareast neighbour, then the new sample denoted by c_i' will be generated as follows:

$$c_i' = c_i + r \times (c_ji - c_i) \tag{3.23}$$

where r is random number between $(0, 1)$, this will generate a new sample on the line between c_i and c_ji, as shown in Figure 3.3.

There are various techniques used for over-sampling, the most commonly used techniques are listed below:

a) *Synthetic Minority Over-sampling Technique (SMOTE)* [136] synthetically creates samples in the minority class instead of replacing original samples, which lead to an over-fitting issue. The SMOTE create samples based on similarities in feature space along the line segments joining the minority instance and its 'k' minority class nearest neighbours in feature space.

b) *Adaptive Synthetic Sampling Approach (ADASYN)* [137] generate samples in the minority class according to their weighted distributions using KNN. The ADASYN assign higher weights for instances that are difficult to classify using the KNN classifier, where more instances are generated for higher weights classes.

c) *Borderline SMOTE* [138] identifies decision boundary (borderline) minority samples and then SMOTE algorithm is applied to generate synthetic samples along decision boundary. It works by classifying each instance to one of

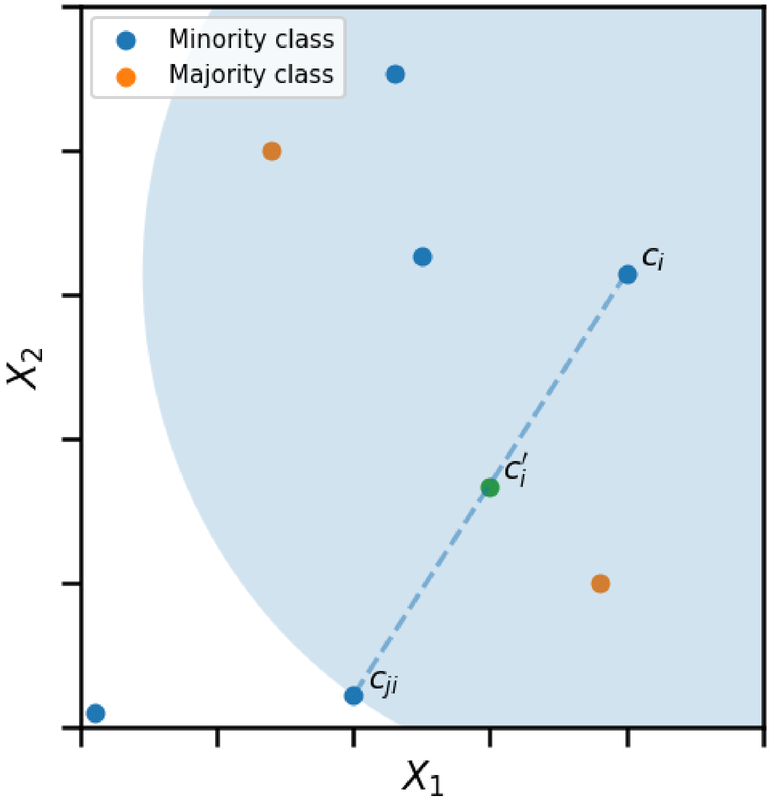


Figure 3.3: Creating new instance using over-sampling technique

three categories based on nearest neighbours: (1) "noise" when all nearest-neighbours are from a different class, (2) "danger" when at least half of the nearest neighbours are from a different class, (3) "safe" when all nearest neighbours are from the same class. Then it will use the instances that belong to "danger" to synthesis new instances either by creating an instance that belong to "danger" which is called Borderline-1, or by creating instances from instances belong to the "danger" class and all its neighbours, which called Borderline-2.

3.1.5.2 Under-sampling Techniques

Under-sampling techniques work by removing samples from the majority classes. Under-sampling techniques can be categorised into two groups: prototype generation algorithms and prototype selection algorithms. Prototype generation algorithms reduce the number of instances in the majority classes by creating new instances that replace the original instances in the majority classes as below:

$$d' \ : \ |d'| < |d| \ \wedge \ d' \not\subset d \tag{3.24}$$

where d' is the new instances set, d is the original majority instances.

On the other hand, the prototype selection select instances from the majority classes and does not create new instances as given by the following expression:

$$d' \ : \ |d'| < |d| \ \wedge \ d' \in d \tag{3.25}$$

where d' is the new instances set, d is the original majority instances.

There are various techniques used for under-sampling, the most commonly used techniques are the prototype selection methods are listed below:

a) *Condensed Nearest Neighbour (CNN)* [139] was originally designed to reduce the memory used by K-nearest neighbours algorithm. It works by iterating over majority classes and selecting subset samples that are correctly classified by 1-nearest neighbour algorithm, thus including only relevant samples and eliminating insignificant samples from majority classes.

b) *Tomek–links* [140] is an enhancement of the CNN technique, as the CNN initially chose samples randomly, but the Tomek-links firstly finds Tomek link samples, which are pairs samples that belong to different classes and are each other's 1-nearest neighbours. Then removes Tomek's link samples belong to the majority classes or alternatively both. In this work, the only majority of Tomek's link classes are removed to retain minority classes and increase distances between classes by removing majority classes near the decision boundary.

c) *AllKNN* [141] is an under-sampling technique based on Edited Nearest Neighbours (ENN), [142], which is an under-sampling technique that applies KNN classifier on majority class and removes samples that are misclassified i.e. removes samples whose class differs from a majority of its k-nearest neighbours. So, the AllKNN technique removes all samples that are adjacent to the minority class, in order to make classes more separable. It works by removing samples from the majority class that has at least 1-nearest neighbour in the minority class.

d) *Instance Hardness Threshold (IHT)* [143] is a technique based on removing samples with high hardness value. The hardness value indicates the probability of being misclassified. This approach removes majority class samples that are classified with low probabilities (overlap the minority classes samples).

e) *NearMiss* [144] technique based on the average distance of majority classes samples to minority classes samples. There are 3 versions of NearMiss technique: NearMiss-1 selects the majority class samples with the smallest average distance to three closest minority class samples, NearMiss-2 selects majority class samples with the smallest average distance to three farthest minority samples, NearMiss-3 selects majority class samples with the smallest distance to each minority class sample.

3.1.5.3 Hybrid Resampling (Combination of Over- and Under-sampling)

The last group that has been examined is the hybrid approach that combines over- and under-sampling techniques. This approach basically cleans the noise that has been generated from over-sampling techniques by removing majority classes samples that overlaps minority classes samples.

a) Synthetic Minority Over-sampling technique combined with Tomek link (SMOTETomek) [145] works by increasing the number of minority class samples by generating synthetic samples using SMOTE as discussed in Section 3.1.5.1, and, subsequently, Tomek link under-sampling technique is applied to the original and new synthetic samples as discussed in Section 3.1.5.2.

b) Synthetic Minority Over-sampling Technique combined with Edited Nearest Neighbour (SMOTEENN) [146] is the second hybrid approach SMOTEENN starts with SMOTE which has been discussed in Section 3.1.5.1, followed by ENN which has bee discussed in 3.1.5.2.

3.1.6 Classification Algorithms

Classification is the process of predicting the class of given instances on the basis of training data. In classification algorithm, a discrete output $f(y)$ is mapped to input variable (x). In this section, some of the machine learning algorithms used for tremor severity classification are reviewed.

3.1.6.1 Decision Trees

Decision Trees (DT) is a supervised ML algorithm that predicts the target variable based on decision rules derived from the data features [147]. DT is a tree-structured algorithm where nodes represent the features of the data, branches represent the decision rules and leaves represent the target variables. The classification process starts from the root node and splits the training data into possible branches based on the decision rules, the branches then lead either to other branches or end in leaf nodes which is the target variable. There are two approaches for identifying the root node: Information Gain and Gini Index. Information Gain is the measurement of changes in entropy after splitting the data based on a decision rule where the feature with the highest information gain should be selected for splitting. The Information Gain is given by:

$$I_G(S,A) = Ent(S) \quad - \sum_{i \in Values(A)} \frac{S_i}{S} \cdot Ent(S_i) \tag{3.26}$$

where S is the dataset, $Ent(S)$ is the entropy of the dataset S before any split, A is an attribute in S, $Values(A)$ is the possible value of attribute A, S_i is the subset of S for which attribute A has a value i, $Ent(S_i)$ is the entropy of the subset dataset S_i.

Gini Index is a measure of impurity where features with the low Gini Index should be selected for splitting. The Gini Index is given by:

$$G_I = 1 - \sum_{i=1}^{N} (P_i)^2 \tag{3.27}$$

where P_i is the probability of an instance being classified to a particular class, and N is the total number of classes in the dataset.

3.1.6.2 Random Forest

The Random Forest (RF) classifier is an ensemble learning algorithm [148] composed from a set of decision trees to overcome the over-fitting problem of decision trees. The decision trees were randomly selected from the original training dataset using the bootstrap method [149]. The remaining training data is used to estimate the error and features importance to decrease the correlation between constructed trees in the forest, hence, decrease the final model variance. The final classification is based on a majority vote of the decision trees in the forest.

3.1.6.3 Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm that works by constructing decision boundaries (hyperplanes) in the feature space maximising the margin between samples which belong to different classes [150]. The feature space dimension depends upon the number of features in the dataset. The position and direction of the hyperplanes are influenced by support vectors, which are the closest samples to the hyperplanes. SVM can be of two types: Linear SVM and Non-linear SVM. Linear SVM is used for linearly separable datasets, in this case, the hyperplane is a single straight line. Non-linear SVM is used for non-linearly separated datasets. The hyperplane can be defined as follows in the context of a binary classification problem:

$$W^T X + b = 0 \tag{3.28}$$

where W is a weight vector, X is the input feature vector and b is the bias.

Considering m training inputs $X = x_1, x_2, x_3, \ldots, x_m$ with classes $y = \{-1, 1\}$, The class y is determined as follows for new samples:

$$y_i = \begin{cases} -1 & : W^T X + b \leq -1 \\ 1 & : W^T X + b \geq 1 \end{cases} \tag{3.29}$$

In SVM, kernel function are used to transform or map original training dataset into a new higher dimensional space where the margin between the samples of the different classes can be maximised. There are four basic kernel functions; linear, polynomial, radial basis function (RBF) and sigmoid [151], are given below for inputs x_i and x_j:

- Linear kernel:

$$K(x_i, x_j) = x_i^T x_j \tag{3.30}$$

- Polynomial kernel:

$$K(x_i, x_j) = \left(1 + x_i^T x_j\right)^d \tag{3.31}$$

  where d is degree of kernel function.

- Radial basis function (RBF) kernel:

$$K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2} \tag{3.32}$$

  where γ is used to set the spread of the kernel.

- Sigmoid kernel:

$$K(x_i, x_j) = \tanh(\sigma X_i^T X_j + r) \tag{3.33}$$

  where σ is a scaling parameter of the input samples, r is the shifting parameter.

Other important hyper-parameters in SVM are the regularisation parameters "C", and "*Gamma*". C is used to control error (misclassification). Whereas, a large value of C, a smaller margin will be chosen to classify all training instances correctly. On the other hand, small values of C encourage a larger margin at the cost of training accuracy. Simply, it is a trade-off between margin and training accuracy. *Gamma* is used only with RBF kernel and it controls the curvature of the decision boundary. Whereas, a high value of *Gamma* increases the curvature, and low values of *Gamma* decreases the curvature.

3.1.6.4 Logistic Regression

Logistic Regression (LR) logit model is a supervised machine learning algorithm used to predict the probability of target variable (class) based on a set of independent variables (features) based on the maximum likelihood estimation [152]. The name logistic regression derives from the employment of logistic or sigmoid function, which transforms the predicted probabilities into binary values 0 and 1. The term sigmoid means S-shaped which is the shape of logistic function, and is given by:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{3.34}$$

where $\sigma(z)$ is the output between 0 and 1, e is base of natural log and z is the input to the function, and is given by:

$$z = w_b + w_1 x_1 + w_2 x_2 + w_3 x_3 + \ldots + w_n x_n \tag{3.35}$$

where w values are the model's learned weight, x values are the feature values, and w_b is the bias.

LR, by default, is limited to the binary classification problem. However, it can be used for a multiclass problem using One-vs-One (OvO) and One-vs-Rest (OvR) methods [153]. Both methods split the multiclass task into multiple binary classification tasks. The OvO is a pairwise classification method in which all datasets that belong to a particular class are paired with datasets belong to another class. The number of binary models is created in the OvO method is given by:

$$\text{Number of Models} = n \times \frac{n-1}{2} \tag{3.36}$$

where n is the number of classes in the original dataset.

On the other hand, in the (OvR) method every class in the original dataset is paired with all other classes in the original dataset. The number of binary models is created in the OvR method is equal to the number of classes in the original dataset. So, if the number of classes in the original dataset is 5, the number of created paired models is also 5.

In LR, there are several optimisers that can be used to find the coefficients that reduce the loss function (cost function), below is a list of the most common ones:

- **Newton Conjugate Gradient (Newton-CG)** [154] employees Conjugate Gradient (CG) method to solve the Newton equation. Newton method is a second-order iterative method used to minimise the cost function based on two derivatives; the first derivative which is the gradient descent and the second derivative which is Hessian. The Newton-CG approach uses two iterative layers: the CG method discovers the Newton step, and an outside technique adjusts the Newton step to ensure convergence. In contrast to the gradient descent, the CG methods take into consideration the history of the gradients and moves along the conjugate direction which leads to faster convergence.

- **Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS)** [155] is an optimisation method based on BFGS with limited memory utilisation. The BFGS is an iterative technique for solving a nonlinear optimisation problem that estimates the Newton method using an approximation of the Hessian function at every iteration, by accumulating all previous gradients. The L-BFGS stores a limited number of previous gradients based on predefined memory size, so it leaves a space for new gradients to be stores.

- **Stochastic Average Gradient (SAG)** [156] is a variance reduction method for SGD that incorporates the advantages of SGD and GD, it has the low iteration cost of SGD, but takes gradient step with respect to the approximation of the standard GD where it uses the gradient of each sample from the previous iteration.

- **Stochastic Average Gradient Augmented (SAGA)**[157] is an optimisation algorithm used to reduce the variance of unbiased estimation in SAG by using covariates (or "control variates"). Similar to SGA, SAGA utilises the GD as well as the SGD, but it overcomes the bias that presents in SAG by computing a number of iterations of SGD and the full GD alternatively.

In terms of regularisation, there are three common regularisations used with LR to prevent over-fitting, namely $L1$ (Lasso), $L2$ (Ridge), and Elastic Net [158]. $L1$ it works like a features selection technique by shrinking the less important features and removes some features, hence reduce the impact of unimportant features in prediction. However, $L1$ does not perform well with high-dimensional features. So, $L2$ is used because it shrinks unimportant features instead of removing them completely, but $L2$ is not good for feature reduction. However, $L1$, $L2$ are very useful when there is multicollinearity in the dataset.

$ElasticNet$ is used o overcome the limitations of $L1$, $L2$ by combining them together. where it can be used to remove or select features and at the same time to works with datasets with high-dimensional features.

3.1.6.5 K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a supervised non-parametric and lazy learning algorithm and is based on nearest training points in the feature space [159]. The main concept of the KNN algorithm is to find a predefined number of training instances closest in distance to the classified instance, and then to predict the class from these instances. To classify a new instance, KNN calculates the distance of neighbours, then it takes the K nearest neighbours as per the calculated distance, and counts the number of instances in each class among these K neighbours. Then it classifies the new instance to the class with the highest votes. The voting results are based either on the maximum number of neighbours (uniform), which means all neighbours have an equally weighted vote or based on the distance where every neighbour has a weighted vote based on its distance from the new instance to be classified. There are several methods to calculate the distance such as Euclidean distance, Manhattan distance, Minkowski distance, and Hamming distance, which are given as below:

- Euclidean distance:

$$Euclidean(a,b) = \sqrt{\sum_{i=1}^{K}(a_i - b_i)^2} \tag{3.37}$$

- Manhattan distance:

$$Manhattan(a,b) = \sum_{i=1}^{K} |(a_i - b_i)| \tag{3.38}$$

- Minkowski distance:

$$Minkowski(a,b) = \left(\sum_{i=1}^{K}(|a_i - b_i|)^p \right)^{\frac{1}{p}} \tag{3.39}$$

- Hamming distance:

$$Hamming_d = \sum_{i=1}^{K} 1_{(a_d \neq b_d)} \tag{3.40}$$

In the KNN algorithm, there are three main techniques to find nearest neighbours, Brute Force, K-D Tree and Ball Tree. Below is a brief description of these techniques:

- **Brute Force** computes the distances between all instances in the dataset. This technique is efficient and fast for small size datasets. However, it becomes infeasible for large datasets and the computation cost is very high.
- **K-D Tree** or k-dimension tree is used to address the limitation of Brute Force in terms of computation cost. K-D Tree is a binary tree-based method that stores each distance between one instance and the other in k-dimensional space (aggregate distance information for the instance) to avoid unnecessary measurements. For example, if instance x_1 is quite far from x_2, and instance x_2 is very close to instance x_3, then instances X_1 and x_3 are very distant. However, with high dimensions datasets, it becomes infeasible and inefficient.

- **Ball Tree** is used to address the limitation of KD Trees in higher dimensions datasets. It recursively partitions the dataset into hyperspheres (balls) with defined centroid and radius, and each ball contains a subset of instances that needed to be searched for the nearest neighbours. So, the number of candidate samples for a neighbour search is reduced and located within a hypersphere or a ball.

3.1.6.6 Artificial Neural Network Based on Multi-Layer Perceptron

Artificial Neural Network based on Multi-Layer Perceptron (ANN-MLP) is a feed-forward ANN that consists of multiple layers (input layer, one or more hidden layers, and output layer) [133]. The input layer or the visible layer receives the input from the dataset. Often the number of neurons in the input layer is equal to the number of features in the dataset. The output layer provides a judgement or prediction about the input, the number of neurons and the activation function in the output layer depends on the type of problem the neural network tries to solve. For example, A regression problem may have a single output neuron without an activation function. Binary classification problems may have a single output neuron with sigmoid activation function [160]. Multiclass classification problems may have multiple neurons in the output layer equal to the number of classes with softmax activation function [161]. The hidden layers serve as the computational engine between input and output layers. The ANN-MLP layers are fully connected, so each node in one layer is connected to every node in the following layer with different weights. ANN-MLP training implemented through supervised learning technique called backpropagation [162]. The backpropagation adjusts the connection weights to minimise the error between neural network predictions and the actual classes. In order to reduce the error, often called loss function or cost function (difference between the predicted output and the actual output), optimisers are used to modify neural network parameters such as weights and learning rate. There are various optimiser can be used in neural networks such as :

- **Gradient Descent(GD)** [163] is one of the most popular and basic optimisation algorithms which is an iterative optimisation algorithm to minimise the cost function by taking repeated steps (Learning Rate) in the opposite direction of the gradient or we can say in the opposite direction of the slope.

- **Stochastic Gradient Descent (SGD)** [163] is a variant of GD where is only one sample or a subset of samples is considered at each step instead of using the entire training dataset. Therefore, SGD reduces computation time.

- **SGD with Momentum** [164] is an accelerated version of SGD that takes into consideration the exponentially weighted average of the gradients, where the training process is faster by taking the gradient of the current step beside the gradient of the previous steps with exponential decay. This method calculates the gradient at the current location and then takes a big step in the direction of the updated accumulated gradient.

- **Nesterov Accelerated Gradient(NAG)** [163] is an enhanced version of SGD with momentum where it initially takes a big step towards the prior gradient and then calculates the gradient where it terminates and makes a correction to the gradient.

- **Adaptive Gradients (Adagrad)** [165] is an adaptive parametric optimiser which adjusts the learning rate to individual parameters (features), where it performs significantly higher updates for infrequent features and smaller updates for frequent features. Also, it eliminates the need to adjust learning rates manually and it is adjusted based on the accumulative sum of previous gradients. Adagrad works very well with sparse datasets where most of the values are zero.

- **Root Mean Squared Propagation (RMSprop)** [163] is an improvement to the Adagrad optimiser to reduce the aggressiveness decay of the learning rate by taking the exponential average of the previous gradients instead of the cumulative sum.

- **Adaptive Delta (Adadelta)** [166] is another improvement to the Adagrad optimiser to reduce the aggressiveness decay of the learning rate by reducing the accumulative sum of past gradients by taking the accumulative sum of fixed window size of previous gradients.

- **Adaptive Moment Estimation (Adam)** [167] is also an adaptive optimiser combines the properties of RMSprop and SGD momentum. In other words, Adam is a RMSprop with a momentum that uses previously squared gradients to calculate current gradients with a momentum to add fractions of previous gradients to the current one. In Adam, the gradients are updated inversely proportionally to the *l*2-*norm*, often called the Euclidean norm as it is calculated as the Euclidean distance.

- **Adaptive Moment Estimation Max(Adamax)** [167] is a variant of Adam that uses $l\text{-}infinity\text{-}norm$ where is only the largest previous gradients have effect on updated gradients.

- **Nesterov-accelerated Adaptive Moment Estimation (Nadam)** [168] is a combination of Adam and NAG. This method works like Adam but instead of using standard momentum, it uses Nesterov momentum.

The activation functions are used to activate or deactivate neurons, in other words the activation functions output determine whether the neuron's input is related or not to the process of prediction. The activation functions have a significant influence on the neural network capacity and performance, below are the most common activation functions:

- Sigmoid function [160] which is the function used in LR, see Section 3.1.6.4 for more details and it is given by:

$$Sigmoid(x) = \frac{1}{1+e^{-x}} \tag{3.41}$$

  where is $Sigmoid(x)$ is the activation function output, and x is the input.

- Hyperbolic Tangent (Tanh) is similar to sigmoid function in shape but the output range between $-1, 1$, and it can minimise the cost function faster, and it is given by:

$$Tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{3.42}$$

where is $Tanh(x)$ is the activation function output, and x is the input.

- Rectified Linear Unit (ReLU) [169] which is more efficient and faster than sigmoid and tanh function in training deep neural networks, and it is given by :

$$ReLU(x) = \begin{cases} x, & \text{if } x > 0. \\ 0, & \text{if } x \leqslant 0. \end{cases} \tag{3.43}$$

where is $ReLU(x)$ is the activation function output, and x is the input.

- Exponential Linear Units (ELU) [170] is similar to ReLU function which is used to accelerate the training of neural networks, also it can eliminate the vanishing gradient problem which can stop the training of neural network, and it is given by:

$$ELU(x) = \begin{cases} x, & \text{if } x > 0. \\ \alpha e^{x} - 1, & \text{if } x \leqslant 0. \end{cases} \tag{3.44}$$

where is $ELU(x)$ is the activation function output, α is the hyperparameter that controls the saturation point for negative inputs, and x is the input.

- Scaled Exponential Linear Unit (SELU) [171] is a variant of ELU function with self-normalisation that automatically normalise the output of hidden layers which is helpful to speed up the training process. The SELU is given by:

$$SELU(x) = \begin{cases} x, & \text{if } x > 0. \\ \lambda \alpha e^{x} - \alpha, & \text{if } x \leqslant 0. \end{cases} \tag{3.45}$$

where is $SELU(x)$ is the activation function output, α is the hyperparameter that controls the saturation point for negative inputs, λ is the hyperparameter that controls the saturation point for positive inputs, and x is the input.

- Softmax [161], also known as softargmax is similar to the sigmoid function, and is used to link the input (features) to the desired output (class) but for multiple classes. It calculates the relative probabilities for each class, and is given by:

$$P(c|p) = \underset{c\in C}{argmax}\left(\frac{e^{(V^{(L-1)}W^{L}+b^{L})}}{\sum_{k=1}^{N_c} e^{(V^{(L-1)}W_{k})}}\right) \tag{3.46}$$

where $P(c|p)$ is the output of softmax function, L is the last layer in the neural network, c is the desired class, C is the list of all classes, V^L is the features vector, N_c is the total number of classes, W is the weight of feature, and b is the bias.

3.1.7 Performance Metrics

The most frequently used metrics for evaluating the performance of classification algorithms are accuracy, precision, sensitivity (True Positive Rate), specificity (True Negative Rate) [89]. However, these metrics are subject to data distribution and insufficient to evaluate classifiers in imbalanced classification problems [88]. Sensitivity and precision do not take into consideration the true negatives. Hence, other metrics like F1-score [89] and geometric mean (Gmean) [91] are widely used to evaluate classifiers to balance between sensitivity and precision, as the ultimate goal of classifiers is to improve sensitivity without impacting precision [18]. Gmean and F1-score are excellent and accurate metrics because they are less influenced by the majority classes in the imbalanced data [172]. However, even Gmean and F1-score minimise the influence of imbalanced distribution, but they do not take into account the true negatives and classes contribution to overall performance [92]. Hence, in this thesis, advanced metrics are included in addition to these metrics, such as index of balanced accuracy (IBA) [92] and Area Under the Curve (AUC) [90]. Below is the list of metrics are used throughout this thesis:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.47}$$

$$Precision = \frac{TP}{TP + FP} \tag{3.48}$$

$$Sensitivity = TPR = \frac{TP}{TP + FN} \tag{3.49}$$

$$Specificity = TNR = \frac{TN}{TN + FP} \tag{3.50}$$

$$F1 = \frac{2 \cdot Precision \cdot Sensitivity}{Precision + Sensitivity} \tag{3.51}$$

$$Gmean = \sqrt{Sensitivity \cdot Specificity} \tag{3.52}$$

$$IBA_\alpha = (1 + \alpha \cdot (TPR - TNR)) \cdot GMean^2 \quad \text{,where} \quad 0 \leq \alpha \leq 1 \tag{3.53}$$

where TP, FP, TN, FN, TPR, TNR, and α, refer respectively to, true positive, false positive, true negative, false negative, true positive rate, true negative rate, and weighting factor.

The most appropriate metrics to evaluate imbalanced data are AUC and IBA. There are many advantages of using AUC to evaluate classifiers, particularly in medical fields [173]. AUC is independent of prevalence, and it can be used to compare comparison multiple classifiers and to compare classifier performance with different classes.

IBA is a performance metric that takes into consideration the contribution of each class to the overall performances so that high IBA is obtained when the accuracy of all classes are high and balanced. The IBA evaluates the relationship between TPR and TNR, which represents classes distribution. IBA can take any value between 0 and 1, and the best performance achieved when $TPR = TNR = 1$ with $\alpha = 1$, then $IBA = 1$.

3.1.8 Conclusions

This chapter presented common methods employed in this thesis to detect tremor severity, including signal preprocessing techniques such filtering which is used to eliminate non-tremor data or artefacts, segmentation of signal into chunks that represents the characteristics of tremor signal, features extraction in the time and the frequency domain that can be used to discriminate tremor severity, the resampling techniques that can be used to solve imbalanced dataset problem, the underlying working concept of various ML classifiers, as well as the performance metrics to evaluate the performance and robustness of these classification models. In addition, the dataset utilised in this thesis is also explained in this chapter. In the next chapters, these methods are explored along with other methods proposed in this research to overcome some of the limitations observed in the literature in terms of tremor severity estimation.

Chapter 4

Patients' and Healthcare Professionals' Perceptions on Wearable Devices and Assessment

4.1 Introduction

The focus of research is on developing an objective, cost-effective, reliable, responsive solution that quantifies tremors for people affected with Parkinson's disease that could be used by a patient, with or without the help of a caregiver. The research project was adopted through a sequential instrument design mixed-method approach [174]. Combining both qualitative methods and quantitative methods to develop a solution that is coherent in therapists' and patients' points of view to avoid the off-the-shelf solution [175]. The mixed-method allows collecting data that are more comprehensive by using both methods strengths to overcome individual weakness personal biases in every method to provide a broader perspective on the overall research problem. Also, assuring the validity and strength of research findings and conclusions through triangulation.

The implementation consists mainly of two phases that begin with a qualitative method to identify perspectives of healthcare professionals and patients of current assessment methods and to identify their preferences, needs and requirements of wearable devices. Then, the quantitative method used to quantify PD tremor kinematic data onto tremor severity ratings using ML algorithms, which will be discussed in Chapter 4.

Obtaining information from healthcare professionals and patients can play an important role in addressing the strengths and weaknesses of medical devices' design. Human Factors Engineering (HFE) is a discipline that focuses on human behaviour, capabilities, limitations, and characteristics in the design of interactive systems that involve humans to ensure effectiveness, safety, comfort and ease of use. [176]. Qualitative methods are commonly used to gather deep and rich information to assess medical design concepts and to identify users' preferences and requirements [177]. Therefore, this project employed qualitative methods to identify the perspectives of healthcare professionals and patients on current diagnosis and assessment methods. Furthermore, to identify their preferences, needs, and requirements of wearable devices to ultimately assess and monitor PD tremors. Also, to explore their expectations and outlooks on potential solutions.

4.2 Ethical Considerations

Ethics according to Word Medical Association "is the study of morality - careful and systematic reflection on and analysis of moral decisions and behaviour" [178], thus it is vital to address ethical considerations when conducting human research. The researcher should have an obligation to act ethically in order to protect the dignity, rights and well-being of research participants by avoiding causing any harm, physically, emotionally or psychologically and to protect participants' privacy, and confidentiality of personal information. Therefore, all research involving human beings should adhere to a set of ethical principles. The most commonly used and governing framework is principlism or ethical principle that developed by Beauchamp and Childress [179] for addressing ethical concerns in medical practice which is based on four pillars: autonomy, beneficence, non-maleficence, and justice. Below is a brief explanation of these principles.

- **Autonomy**: Participants should be treated as autonomous agents, and they have the right to make their own choices.

- **Beneficence or the doing of good**: The research is for participants' benefits.
- **Non-maleficence**: To abstain from doing harm and things that are against participants' interests. It reminds the researcher that he/she must consider the possible harm and avoid causing any harm, physically, emotionally or psychologically.
- **Justice**: Fairness and equality among participants, fair distribution of benefits and legal justice.

4.2.1 Ethical Approval

The work was conducted in accordance with the Declaration of Helsinki [180]. This work was approved by the Joint Inter-College Ethics Committee (JICEC) at Nottingham Trent University (reference JICEC1819-10). Prospective participants who expressed an interest in participating in this work were given information sheets. The Information Sheet detailed the research study and the methods for collecting and analysing relevant data, as well as the purposes for which these data would be used. The participants were informed that they had the right to withdraw from the study at any time, without giving a reason. The researcher provided the participants with written contact details if more information was needed. All participants provided written informed consent and were able and willing to participate.

4.2.2 Confidentiality and Data Protection

Legally, all rights and data usage and sharing and any legal concerns are protected by the EU Regulation (EU) 2018/1725. In order to achieve the anonymity, security and confidentiality of personal data collected for the project, especially any to be stored or processed off-site. The participants' identities have been coded; the consent form and personal data sheet have been maintained separately, so participants' personal information will not be disclosed. In addition, arrangements for retention, anonymisation and disposal of personal data at the end of the project.

4.3 Inclusion Criteria and Exclusion Criteria

The inclusion criteria for people aged between 18 and 85 years and who were clinically diagnosed with Parkinson's met the diagnostic guidelines of the United Kingdom Parkinson's Disease Society Brain Bank criteria [29], and happy to participate in the focus group discussion. Inclusion criteria for health care professionals required them to be aged 18 years and above, working closely with people affected with PD and aware of diagnosis and assessment procedures of PD, and they were happy to be interviewed. All participants were able to provide written informed consent. Participants were excluded if they were unable to provide informed consent and were unwell to participate, lack of capacity to consent.

4.4 Data Collection

In qualitative research, data can be collected using a range of methods, including observations, textual or visual analysis, individual interviews and focus group discussions [181]. However, interviews and focus groups discussions are the most widely used approaches, particularly in healthcare research [182]. There are three types of interviews: structured, semi-structured, and unstructured [177]. A structured interview is a series of predefined questions with little or no variation that are asked to every participant in exactly the same order. As a result, it is relatively quick and simple to conduct. This type of interview ensures a confident comparison among participants' responses and it is easy to test for reliability, but it lacks depth and detail and provides fewer opportunities for participants to engage with what is important to them [181].

An unstructured interview is a discovery interview that does not use any predefined questions and depends on social interaction between the researcher and the participant. It usually begins with an opening question and will progress based on the initial response. This sort of interview is flexible and useful for studies attempting to find patterns. However, it is time-consuming and difficult to administer and tends to divert from the topic. Moreover, it generates a significant volume of data making it difficult to classify and analyse [183].

The third type of interview is a semi-structured interview that comprises predefined structured questions (topic guide) with the flexibility to probe and ask questions following the participants' responses. The order and phrasing of the questions are modified by the interviewer to best fit the interview context. Also, it enables the interviewer to pursue new topics as needed and gives the participants the opportunity to freely share their opinions. In addition, it allows two-way communication and compares participants' responses and the reasons behind them. However, it is time-consuming difficult to analyse. Moreover, flexibility may compromise reliability [184].

Focus group discussion is a qualitative research method where a group of participants from similar backgrounds or experiences are interviewed together to discuss a specific topic of interest [185]. In focus group discussions participants are asked about their perspectives, beliefs, thoughts, or ideas. The discussions are predefined semi-structured interviews guided by a moderator and used to gain an in-depth understanding of the research subject. The role of the moderator is to ensure that the participants interact with each other [186]. A focus group can enable participants to disclose more freely. Also, it can generate a large amount of data on a topic in a short time. A focus group allows participants to agree or disagree with others so that it provides an insight into the participants' thinking and perspectives. Moreover, it increases validity due to the fact that some participants may feel more comfortable being with others. However, an inexperienced moderator may encounter several participants who are trying to influence the group and lead the discussion to an irrelevant subject. In addition, the collected data are more difficult to analyse [187].

In this thesis, a cross-validation or triangulation approach was adopted by first using exploratory semi-structured interviews, and later focus group discussions according to the procedure used by Lambert and Loiselle [188]. This approach was adopted to enrich the collected data and to increase the credibility and validity of the findings. The first step in this approach was to conduct four preliminary semi-structured individual interviews for healthcare professionals interviews topic guide, see Appendix A ; three with healthcare professionals (2 females, 1 male; age range: 52-61 years) were recruited from the Royal Derby Hospital, and one interview with 65 years female Parkinson's local supporter was recruited from Parkinson's UK, who closely works with PD patients and is understanding of their needs and requirements in different ways such as emotional support, informal discussions about patients' worries and experiences. Also, she arranges social events and invites healthcare professional speakers to provide information about PD. This voluntary work enriches her experience and knowledge about PD patients' requirements. These interviews were utilised to obtain further in-depth information on the current diagnosis and assessment processes. The results (themes) from the exploratory interviews were used to generate a discussion guide for the patient focus groups topic guide, see Appendix B. The interview with the local supporter was used to link healthcare professionals with patients' views and not to generate themes.

The second step was to conduct three focus group discussions involving 12 PD patients (5 females, 7 males; age range: 56-88 years) were recruited through Parkinson's UK, each lasting approximately 60 minutes. The discussions followed a semi-structured topic guide to allow a deeper insight into product design and development. The topic guide was designed to elicit general discussion with a more specific question as probes used if they were not raised or discussed by participants as per the recommendations of Braun and Clarke [124]. The demographic information of participants is summarised in Table 4.1.

Table 4.1: Participants demographic information.

Data source	Gender Male : Female	Number of participants	Age (year) Mean ± SD (range)	PD Duration (year) Mean ± SD (range)
Focus Group 1	2 : 1	3	73.83 ± 10.69 (56 - 88)	8.5 ± 7.29 (2 - 24)
Focus Group 2	3 : 3	6		
Focus Group 3	2 : 1	3		
Interviews	1 : 3	4	57.75 ± 6.29 (52-65)	NA

4.5 Data Analysis

Audio recordings of the healthcare professionals' interviews and patients' focus groups were transcribed verbatim by the lead researcher, which enabled initial familiarisation with the data. The data was then analysed using an inductive thematic approach following the six phases guideline as outlined by Braun and Clarke [124]. This involved familiarisation with the data by reading transcripts multiple times and annotating initial ideas and coding interesting elements in the data with a systematic approach by:

1. Organising data in meaningful forms.
2. Grouping codes based on potential themes.
3. Collecting data relevant to each theme.
4. Reviewing themes by refining themes or sub-themes by splitting, combining themes and find the relationship between themes.
5. Defining and writing up the themes.
6. Final analysis and connect analysis to the research question.

The results from the analysis of the interviews were utilised to design a focus group topic guide. Verbatim quotes are indicated using the following notation: **I** indicates Interview, **G** indicates focus group, **F** indicates female, **M** indicates Male, and the number indicates the focus group or interview number. For example, GF1 (Female from focus group number one).

4.6 Results and Discussions

This work investigated patients' and healthcare professionals' opinions toward the current diagnosis and assessment process together with their preferences toward wearable technology and their expectations and outlooks on potential solutions. Limitations of existing solutions and barriers to their use were explored alongside wearable technology design requirements and expected solution outcomes. Both groups of participants showed a unanimously positive response towards the use of wearable technology for remote continuous monitoring. The results from interviews identified the following relevant themes: (1) Current diagnosis and assessment are dubious art, (2) The role of aesthetics and design for acceptance and adoption (3) Patients and healthcare professionals want wearable technology that eases and refines treatments.

4.6.1 Current Diagnosis and Assessment Methods are Dubious Art

A common view amongst interviewees was that the current assessment is subjective, dependent on clinical expertise, and thus, inconsistent. Also, there is much scope for Type I (False Positive) and Type II (False Negative) errors in diagnosis.

"***The consultation thing can be a little bit subjective. We doubt each other's assessment.***" **(IM1)**

"***Sometimes we see patients who have been diagnosed with Parkinson's disease, but later it turns out to be not Parkinson's.***" **(IF3)**

"***So, it is about listening to their story, making your own assessment and discussing the options, which might help. It is very much an art.***" **(IF4)**

The interviews show the need for objective data for diagnosis and to distinguish between tremors and other abnormal movements such as ET. Probing questions about the reason behind this suggested that it is very difficult to differentiate PD tremor from ET tremor, especially in early disease stages, which is consistent with other research that has reported a high misdiagnosis rate of PD and ET to be approximately 25% of cases [189].

"**[New assessment method]** ***Will be able to tell the difference between those types of movements. So, you know, that would be, that will be helpful.***" **(IM1)**

The current assessment and monitoring processes are dependent on patients' memories and diaries which are not reliable, as previous studies suggested that 30-40% of PD patients will develop dementia [25], besides the fact that most patients are elderly and several lines of evidence suggest that memory decline occurs in individuals older than 60 years [190]. Also, many PD patients are unaware of their symptoms often cannot distinguish between PD signs and other abnormal activities [191].

"**[Patients]** ***Do not always remember precisely . . . they*** **[Neurologists]** ***might see them six months or even every 12 months . . . clinically we*** **[Neurologists]** ***asked them to video the movement that they are talking about.***" **(IM1)**

This work supports evidence from previous observations that current diagnosis and assessment scales are subjective, infrequent and depends on clinicians' skills and patients' recall [8, 45].

Sometimes consultants ask patients to video tremors they are experiencing at home for assessment, but elderly patients may often not be familiar with modern technology such as wearable and social computing as exemplified below, and also shown by [192].

"***They do not always know, and a lot of these patients are elderly, and they maybe get the instructions, or they will not know how to video.***" **(IM1)**

A highly surprising fact that emerged from the data was that the health care professionals reported scales or score systems were not commonly used for diagnosis and assessment, including the UPDRS. Probing questions about the reason behind this suggested that assessment scales are a time burden and need repetition. This result may be explained by the fact that UPDRS is primarily used for clinical trials and research.

" ***I do not give numbers*** **[scores for the symptoms]** ***they*** **[researchers]** ***might do in a research study.*** **" (IM1)**

"***We do not tend to use UPDRS routinely in the clinic.*** **"(IF4)**

The late diagnosis was a common concern amongst patients, due to late referral from general practitioners (GPs), as most GPs were not suspecting Parkinson's in the first visits. However, most of the patients were diagnosed correctly and quite quickly when they were examined by a neurologist. Another reported problem was that patients reported having symptoms years before they went to see a doctor or were referred to a specialist.

"***A pain in the shoulder, ... the doctor gave me a coat and an injection. That seemed to ease it, but then I needed them. I did not do anything about it for a moment. I just left it, and then I noticed that I was walking, and my right arm was not moving ...Then the doctor told me I had a clot on my brain.***"(GM1)

"***Dr*** [Name] ***diagnosed me and provided me a prescription for having some form of, what did he call it? spinal plates.*** " (GM3)

Concerns were expressed about infrequent assessment; the patients reported that assessments were typically carried out every six months. In addition, assessment depends on a patient's memory or diary and does not often involve examination or physical assessment. Instead, it mainly focuses on generic subjective patients reports (e.g., "how are you?" "have you got any issues?"). Consistently with healthcare professionals' reports, none of the participants mentioned the UPDRS motor examination. A common view amongst patient interviewees was that infrequent assessment affects their treatment by not taking the right medication and the right dose due to the long-time between appointments, and their symptoms might be controlled during the first few months post-assessment but not for the entire six months period. However, all patients reported that they can call nurses between appointments if they feel unwell or if they have any issues.

The patients, on the whole, demonstrated that they experience "peaks and troughs" from hour to hour and from day to day, and some of the symptoms appear at a specific time of the day or when they are doing specific tasks; these patterns are not picked up during clinic assessment.

> "***Because I have on and off*** **[Symptoms fluctuation]** ***I have started going on and off, so I am m sort of up and down, up and down.*** " **(GM1)**

4.6.2 The Role of Aesthetics and Design for Acceptance and Adoption

From healthcare professionals' point of view, wearable devices would be acceptable to most patients, particularly young patients. Device visibility may depend on the stage of the disease; healthcare professionals believed patients in more complex stages of PD would be more likely willing to wear visible devices than would newly diagnosed patients.

> "***So, it depends on what stage of Parkinson's you are talking about. So, if you are going to go to somebody newly diagnosed, they probably want something pretty discreet. But somebody who is in the more maintenance, more complex phase they may wear something that is a bit more visible.*** "**(IF3)**

Device design is one of the most important factors that determine whether patients are willing to wear the device. In line with previous studies [116–118, 120] healthcare professionals felt that the device must be comfortable, easy to use, non-invasive, and should easily be worn under clothes without catching/snagging. The device should also be water-resistant, washable, durable, and easy to fasten to minimise daily disruption.

One main objective of this work was to determine the most preferable and suitable part of the body to wear the device without affecting data quality. All healthcare professionals independently suggested the wrist would be most appropriate, with some focusing on reasons of patient comfort and others on detection of PD tremor characteristics. Given that the most typical tremor in PD is called a 'pill-rolling' rest tremor involving movement of the thumb and index finger, the wrist would be an appropriate location for the device as it could reliably pick up this type of tremor. These viewpoints match those reported in earlier studies [193, 194].

Healthcare professionals were familiar with what wearable devices are and their use and functionality, and they showed awareness of some commercial devices used for diagnosis and assessment, such as PKG. However, none had ever used these before because they are expensive, and the reports are difficult to interpret. All healthcare professionals reported that wearable devices that are available commercially have mostly been used for advanced treatments such as deep brain stimulation and advanced therapies.

" ***It is quite expensive. They do use it,*** **[Neurologist's name]** ***in*** **[City name]** ***he uses them a bit because he is doing, um, he has access to very expensive treatment and so he wants very objective data to give them the expensive treatment ... The software that they have developed to interpret the device finding is quite complicated*** " **(IM1)**

"***But I think the biggest problem with that is that you have to pay a certain amount of money for every report. Neurologists at*** **[Hospital name]** ***use it when they are thinking about advanced therapies*** " **(IF4)**

When asked about wearable technology, most patients have not heard about or used wearable technology. Following an explanation of the purposes of this technology as a part of this work, the majority of patients taking part in this work stated that they would be willing to use wearable technology and to be monitored

24/7, as long as the device is not invasive or on an undesired part of the body (e.g., neck or ankle). Patients stated the preferred part of the body was the wrist, as one would wear a watch; this was echoed by every participant in this work. This further supports healthcare professionals' points of view. Also, these results are in keeping with previous observational studies [193, 194].

Issues related to technology such as violation of privacy, difficulty in learning how to use technology, fear and discomfort of using technology and lack of human interaction were not particularly prominent in the discussion. Except for one patient who thought that the fear or dislike of modern technologies (Techno-phobia) could be a barrier for many elderly people, which echoes healthcare professionals' perspectives. What is interesting about this result is the conflicting perspectives between healthcare professionals and patients. In addition to the growth in the number and proportion of older people suggest that technology acceptance and adoption maybe is not related to techno-phobia, but due to different perspectives and lack of training on new technologies, in addition to technology designs that do not meet users' needs and requirements [195].

"***Well, I do not think it is an issue. I mean if it is if somebody is going to try it in the first place to see if it is going to work for everybody ... a lot of all the people do not trust modern technology***" (GM1)

Regarding wearable design aspects, there are a number of similarities between interviews and focus group discussions in terms of what patients would want, including for the device to be comfortable, non-invasive, waterproof, durable, small, and easy to fasten. Patients' discussions focused on the wearable hardware side more than did healthcare professionals. For example, patients discussed which materials are more comfortable, breathable, not sweaty/sticky, soft and spongy, including cloth over metal, leather, neoprene, reinforced material, stainless steel, rubber and silicon. Even though they held differing views about what material should be, the common ground was to use comfortable material. Moreover, all patients focused on the device wearing style, which is easy to fasten as the tremor affects their ability to fasten traditional clasps such as a buckle or flip styles. Velcro and elasticated straps were the most preferable styles to patients; for example, it was pointed out that Velcro straps could be tightened as needed. Some patients offered it would be better if they were involved in the development stages of the device. These findings are in line with previous studies [116–118, 120] in which wearable technology's acceptance was determined by appearance, comfortability, size, and ease of fitting.

"***What about these plastic, leathery straps because this all new stuff coming out. The only thing l would like is a nice easy clasp because that I have got very faster on it because this is my worst hand so if I have tried to do my watch, that is why I have gone facility like this* [He showed clasp of his watch] *that I can wear it easily, something sticks together***" **(GF3)**

"***I think it might be nice to have an insight into the development stages of whatever the devices that you are going to use so that we can have some input or whether you take it on board or not***" **(GM1)**

None of the patients had any concerns about device visibility. Perhaps somewhat remarkably, they wanted the device to identify them as people affected by PD, so it might indicate to the community they may need help. Indeed, it has been shown that psychological support would be helpful for people with PD since emotions affect the severity of the symptoms [196]. Also, it helps those affected avoid unwanted and uncomfortable situations if people know about their disorder, for example, the public might think they are drunk due to the nature of their symptoms. The feeling of embarrassment in public due to PD symptoms has also been described in other studies of PD [197, 198].

"***If it is designed to do with Parkinson's and in time people got to know if they saw that on you, you got Parkinson's, he might need help***" **(GF3)**

"***That happened to me when I was going through the park. People thought I am the person sitting on a bench though I was drunk they call back to me, told me to take more water with it, you know, just the usual gobby, which in fact that upset me. I never went out actually after that for a few days because it upset me that much***"**(GF2)**

4.6.3 Patients and Healthcare Professionals Want Wearable Technology That Eases and Refines Treatments

When healthcare professionals were asked about their expectations and outlooks from potential monitoring and assessment solutions, remarkably all interviewees shed new light on their expectations that the solution could lead to a better or new treatment, and as discussed in section 4.6.2 that objective data is needed for expensive treatment as reliable markers; therefore, the solution could help improve current treatments or lead to new ones. The expected solution could be used to evaluate treatments efficacy in terms of medications and rehabilitation.

"***If the nurse altered the medication, and you can detect if the medication reduced patient tremor a bit or knew it did not, or the tremor is a bit worse, and that is what you kind of go on, and that is enough, you know because then you can either try medication try relaxation techniques.***"**(IF3)**

There were some suggestions from healthcare professionals that the solution should be easier to use, provide "***very concise***" information, and be easy to interpret. As mentioned in section 4.6.2, one of the main reasons that interviewed healthcare professionals are not using devices available commercially is the complexity of interpreting their data and results.

"***It might be ideal to have something that measured tremor in some way or whether somebody was having an off and on, but it just needs to tell us that quickly and simply, without needing a degree in mastery the charts***" **(IF3)**

"***I think the information it would give would need to be very concise. We would not have time to be going through reams***" **(IF4)**

Concerns regarding lack of information on symptom fluctuations were widespread among healthcare professionals, supporting that symptoms fluctuate from time to time during the day and from day to day. Results show high interest from healthcare professionals in continuous monitoring and its importance in diagnosis and treatment decisions. This finding reflects evidence from previous studies that showed the benefits of continuous monitoring [199, 200].

"***Lots of patients come to us with issues at certain times of the day, whether it is tremor or slowness and stiffness or fatigue or, or you know, being sleepy or you know, all of which could be the symptom of Parkinson's***" **(IF4)**

"***A pattern to their off time and it might help the nurse how to adjust the medication ... If you had something [a solution] that was a bit more technical and a bit more useful in revealing symptoms.***" **(IF3)**

Consistent with the literature [12, 199], this research found that remote monitoring has a pivotal role in PD treatment as well as healthcare cost reduction, mainly for large geographical areas, and it could help patients who are not in the vicinity of a hospital to be diagnosed, particularly in cases of severe symptoms that make travel difficult. Also, it is easier and more efficient use of time for healthcare professionals to assess patients remotely rather than travelling to their homes.

"***Who is living way away, they might not be able to come to the clinic too often.*** **[Neurologist's name]** ***in*****[City name]** ***used it*** **[Wearable device name]*****because I think the geographical area that they cover is quite large***"**(IF3)**

Remote and continuous monitoring was a recurrent topic throughout the discussions; patients seemed to believe it would help clinicians identify their symptom fluctuations and patterns during the day, and if the medication does not help patients to manage their symptoms, continuous monitoring could enable clinicians to change medication in a timely way. Also, it was suggested to track symptom history over a period before a clinic visit, which would ensure that clinicians do not miss any information that might affect patients' treatment.

"***It would be better if it is 24/7 because then you go to get the full picture of the 24 hours, aren't you? . . . Because it is very tiring if you have got Parkinson's to travel a long way in a day***"**(GF1)**

"***I think if that goes through on that and is something the doctor picks up and appointment comes through to see quite quickly because something in your medication is not working as good as we think it could do.***"**(GF3)**

There were some suggestions to other functionalities to the device, such as fall detection and medication reminders, as this information could help healthcare professionals and patients alike, as forgetting medications and falls are common in PD patients/older age generally [27, 120].

"***Something that counted falls would be useful because patients do not remember how many falls they have ... Medication reminder would be useful and does not stop until they have taken it***"**(IF3)**

When asked patients about their expectations and outlooks from potentials monitoring and assessment solutions, a common view amongst patients was that early precise diagnosis and accurate frequent assessment could make their life easier and lead to better treatment, as mentioned earlier by **GF3** and below statement.

"***Something that could make our life easier***"(GM1)

"***Privacy sort of goes out the window***"(GM2) was the response if the patients have any concern regarding transferring the data over the internet or if someone sees their data or PD information, unlike those in earlier research [113, 118, 121]. The current patients seemed to be comfortable forgoing their privacy rights in terms of data access for the overall benefit of treatment. It was also found that patients do not have concerns about device visibility. On the contrary, patients made suggestions the device should be visible to identify them as people affected PD and that they may need help. This finding contrasts with previous studies which emphasize patients' preferences for the device to be discreet [113, 121].

"***I have Parkinson's. I am not bothered who knows I got Parkinson's***"(GM1)

"***Privacy, no, I think you lose all your privacy when you have got something like Parkinson's***"(GF2)

Supplementing the health care professionals' interviews results, it was suggested by patients to add a medication reminder option to the wearable devices and the applications as patients report facing issues with remembering to take the right medication on time. Additionally, patients also suggested a help call button would be useful on the device to summon an emergency contact when needed.

"***I take my medication at set times during the day. Well, this will be able to remind me which medication to take on time***"(GM2)

"I am wondering if you can put something on your watch and, say, I fall in the house or outside. If a press that button it goes straight through to my daughter's phone and I can speak to her" **(GF3)**

Previous research has established that user acceptance has a pivotal role in wearable technology adoption [116, 118, 119]; however, very few studies focused on wearable design methods, particularly the utilisation of UCD philosophy. The current work offers some key insights into user involvement in early design stages and to identify patients' and healthcare professionals' requirements and preferences, and the importance of patients and medical professional input was highlighted by participants in this work. Moreover, the prior disregard of such input (alongside the downsides of costliness and difficulty of interpretation of commercially available devices) may help explain the lack of commonplace adoption of wearable technologies. If users are involved in the design process, the device may better suit their needs and overcome any barriers to their use. Some options important to users may have previously been neglected by designers, such as medication reminders and fall detection, which may further interest patients and healthcare professionals to use such wearable devices.

In summary, these results show that current assessment and monitoring processes are subjective and depend on clinicians' skills and experiences. It was commented that "the consultation thing can be a little subjective". Also, they are not routinely applied in clinics because they are time-consuming and rely on patients recall, as one interviewee said: "We do not tend to use UPDRS routinely in the clinic". The participants' perceptions about using wearable devices to evaluate symptoms were supportive and suggested that this objective evaluation could make the current assessment easier and enhance current treatment. As one participant reported that "he wants very objective data to give them the expensive treatment". Another important finding was that no concern was raised about wearable devices visibility or private data conveyed through the internet. For example, one interviewee said "Privacy sort of goes out the window". Participants were interested in participating in device design, and they have proposed many design aspects and options that increase user acceptance and adoption. As one interviewee put it "I think it might be nice to have an insight into the development stages".

4.7 Limitations

This exploratory work has several limitations. First, The sample size may not be fully representative of the wider PD and healthcare provider population. Second, all participants were residing in the Nottingham area. Hence, perceptions may differ in other regions of the world which may limit the generalizability. However, qualitative research rarely seeks to generalize but to explore perceptions. Third, while interviewing participants, we have noticed different levels of knowledge and experience with technologies, so responses were likely based on previous experience with available technologies, such as wearable devices or smartwatches.

4.8 Conclusion

The results from the work undertaken in this chapter have found that current assessment and diagnosis methods are subjective and depend on healthcare professionals' skills, and this may lead to inconsistent assessment. Currently, there is no general agreement about a reliable, valid, sensitive and cost-effective solution to assess PD symptoms. Through this work, healthcare professionals' and patients' perspectives of wearable technology were positive and how it could be utilised to improve the current assessment process, thus increasing the chance of effective treatment. All interviewees shed new light on their expectations that the solution could lead to a better or new treatment, and the objective data is needed as reliable markers; therefore, the solution could help improve current treatments or lead to new ones. Moreover, the solution should be easier to use, provide very concise information, and be easy to interpret. Also, it should mimic the current scale. These results can help to design a solution with a high level of patients' and healthcare professionals' acceptance. In the chapters that follow, a machine learning approach integrated with signal processing are employed to develop an objective measurement solution easy to interpret and linked to MDS-UPDRS scale. The solution utilises commercially available devices that meet most of patients' and healthcare professionals' needs and requirements. Also, these devices can be modified to meet other requirements that have been discussed in this chapter.

Chapter 5

Enhanced Parkinson's Disease Tremor Severity Classification

5.1 Introduction

The Imbalanced data problem arises when the data classes are extremely skewed [134], where the number of samples in one or more classes (minority) is much lower than the number of samples in the other classes (majority). Therefore, the ML classifiers are much more inclined to the majority classes because of majority classes domination. However, in many applications detecting the minority classes is very important such as disease diagnosis, fault detection, spam detection, and sentiment analysis [18]. One of the most common solutions to this problem is the resampling dataset. There are three types of resampling techniques, over-sampling, under-sampling and hybrid resampling (combination of over- and under-sampling).

This chapter explores the use of various resampling techniques combined with signal processing and ML classifiers to enhance PD tremor severity estimation. Resampling techniques are integrated with well-known classifiers, such as ANN-MLP and RF. Advanced metrics are calculated to evaluate the proposed approaches such as AUC, Gmean and IBA.

The remainder of this chapter is structured as follows: Section 5.2 explains the proposed methodology, including dataset description, signal analysis, features extraction, applying different resampling techniques with classifiers, evaluation.

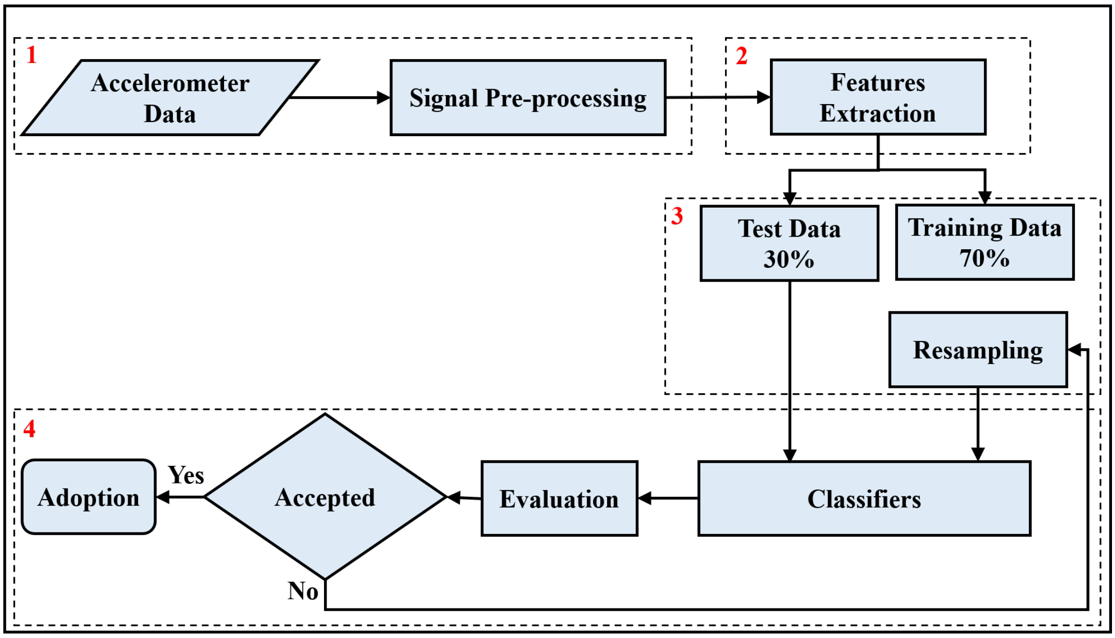


Figure 5.1: Proposed framework for tremor severity classification.

Followed by the results presented in Section 5.3. Section 5.4 concludes summarise the work have been undertaken in this chapter.

5.2 Materials and Methods

Figure 5.1 illustrates the proposed framework to classify imbalanced tremor severity dataset using resampling techniques. In the first step, the raw accelerometer signal is prepossessed to eliminate sensor orientation dependency, non-tremor data and artefacts. Set of tremor severity features extracted from the prepossessed signal in the second step. In the third step, data is split into training and test subsets and training data resampled to avoid classifiers bias. Finally, training and test data are passed into a classifier to estimate tremor severity and the results evaluated for adoption in the fourth step. Each step is described in detail in the subsequent sections.

5.2.1 Dataset

The work in this chapter makes use of the Dataset I (see Section 3.1.2), which is GENEActiv accelerometer tremor data that were collected on the first day. Table 5.1 shows classes (severities) distribution of (32414) instances (windows) segmented from collected data. It is clear how data distribution being skewed towards less severe tremor and this bias can cause significant changes in classification output, in this situation the classifier is more sensitive to identifying the majority classes but less sensitive to identifying the minority classes if they are eliminated.

5.2.2 Signal Processing

In order to extract meaningful features from accelerometer data, some preprocessing was performed to eliminate non-tremor data or artefacts. The vector magnitude of three orthogonal accelerations, namely A_X, A_Y, and A_Z has been calculated to avoid dependency on sensor orientation and to avoid processing signal in three dimensions. Also, since the work focus on the severity of any tremor type, and in order to remove low and high-frequency bands and retain tremors bands from the data as suggested by earlier work [26], a band-pass Butterworth filter with cut-off frequencies $3-6$ Hz for RT and $6-9$ Hz for PT and $9-12$ Hz for KT is applied.

The filtered signals were split into four seconds windows that can be labelled and used as inputs. Fixed-length sliding windows with 50% overlap was utilised, which has been shown in the literature to be effective in activity recognition [132].

Table 5.1: Imbalanced classes (severities) distribution.

Class (Tremor severity)	Instances (n=32414)
0	22584
1	6724
2	2195
3	874
4	37

5.2.3 Features Extraction

A wide range of commonly employed hand-crafted features are calculated to form the feature vector (see Section 3.1.4). The features were carefully selected to provide detailed and discriminatory information of signal characteristics and that are highly correlated with tremor severity, such as distribution, autocorrelation, central tendency, degree of dispersion, the shape of the data, stationarity, entropy measures and dissimilarity.

Tremor severity can be distinguished by amplitude, as the tremor amplitude showed a high correlation to the UPDRS score [201], as the amplitude increases when severity increases. Similarly, a previous study showed that tremor severity is highly correlated with frequency sub-bands [75], as every tremor severity or score appears within a specific frequency range, as shown in Table 5.2. Therefore, features such as mean, max, energy, number of peaks, number of values above and below mean are chosen besides median in case the values are not normally distributed. In addition, these features showed a high correlation with tremor severity classification in previous studies [79, 202]. In order to measure signal dispersion, the standard deviation is selected since it is found to be an effective measure to quantify tremor severity [86].

Skewness and kurtosis are chosen to measure data distribution. Kurtosis has been used in previous studies to detect tremors because tremor signals are spikier (high Kurtosis) than non-tremor signals [202]. Consequently, high severity tremor almost certainly has a high kurtosis value and vice versa. On the other hand, skewness measures the lack of symmetry, and it has been used to measure random movements to assess medication response, as while patients are on medication the tremor will decrease, thereby tremor signal skewness decrease

Table 5.2: Tremor severity vs frequency ranges.

Frequency range (Hz)	Tremor severity
0 - 0.50	0
0.5 - 0.9	1
0.9 - 1.8	2
1.8 - 3.4	3
> 3.4	4

[202]. Therefore, skewness is expected to decrease with less severe tremors and increase with high severe tremors. Spectral Centroid Amplitude (SCA), which is the weighted power distribution, and maximum weighted PSD are other features related to spectral energy distribution [203]. As every frequency sub-band represents a tremor severity [75], thus the maximum weighted power and the weighted power distribution can quantify tremor severity. The PD tremor is a rhythmical movement, hence sample entropy and autocorrelation have been chosen to measure regularity and complexity in time series data, as tremor's sample entropy and autocorrelation are significantly lower when compared to non-tremor movements which have been established by previous work [97, 204]. Other complexity measures that have been selected are the Complexity-Invariant Distance (CID) [205] and the Sum of Absolute Differences (SAD) [110]. SAD and CID measure time series complexity differently, as the more complex time series has more peaks and valleys, which increase the difference between two consecutive values in the window. Consequently, the tremor signal is more complex because tremor frequency and amplitude are higher than normal movements which increase the peaks and valleys in the signal. As a result, complexity is correlated with tremor severity.

Previous research has established that tremor intensity identifies tremor severity [26]. Therefore, tremor intensity at various frequencies can be quantified through PSD, and since tremor severity correlated with frequency ranges or bandwidth spread [75], thus three features are chosen; fundamental frequency, median frequency, and frequency dispersion. The fundamental frequency has the highest power among all frequencies in power the spectrum. The median frequency divides PSD into two parts equally. The frequency dispersion is the width of the frequency band which contains 68% of the PSD. In addition, guided by previous work, the difference between the fundamental frequency and the median frequency was extracted as an additional feature since tremor fundamental frequency could be different between patients [100].

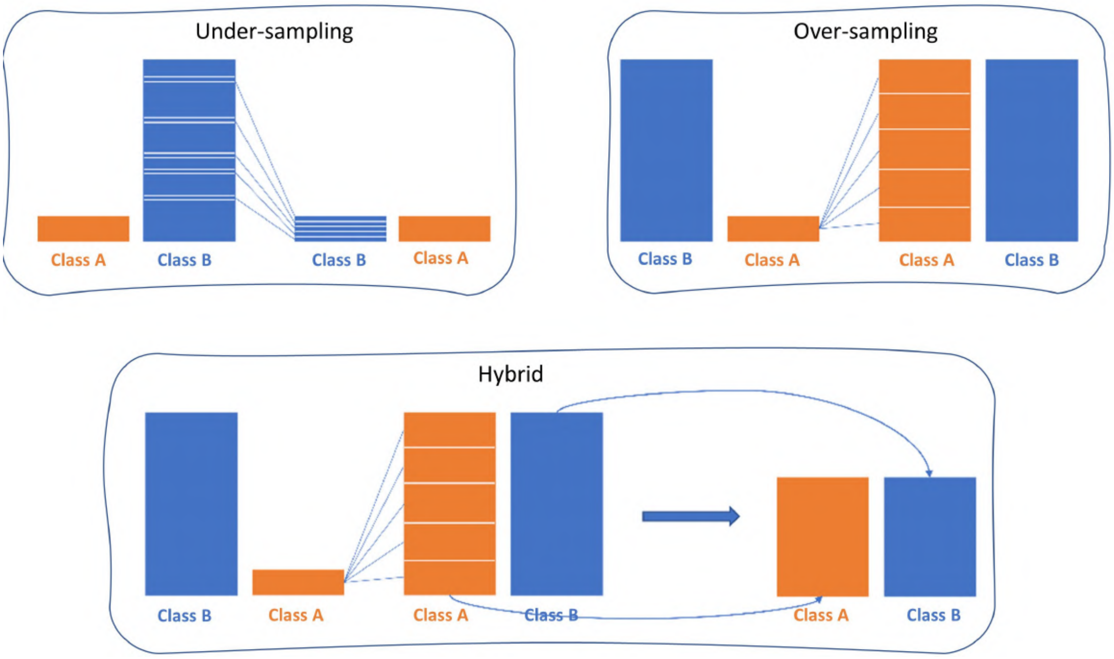


Figure 5.2: The working principle of under-sampling, over-sampling, and hybrid techniques.

5.2.4 Resampling Data

From the data in Table 5.1, it is apparent that the classes are unbalanced. Thus, different resampling techniques are described in Section 3.1.5 were utilised to eliminate the imbalanced class distribution effect. There are three approaches of resampling techniques as shown in Figure 5.2: Over-sampling techniques increase the number of instances in the minority classes, while under-sampling techniques remove samples from the majority classes. Hybrid techniques are over-sampling minority classes then under-sampling majority classes samples that overlap minority classes samples. Table 5.3 presents the resampling techniques used in this chapter.

Table 5.3: Resampling techniques used to enhance tremor severity estimation

Resampling Techniques		
Over-sampling	**Under-sampling**	**Hybrid**
SMOTE	TomekLinks	SMOTETomek
ADASYN	CNN	SMOTEENN
Borderline	AllKNN	
	IHT	
	NearMiss	

5.2.5 Classification and Evaluation

Two classifiers are considered for classification; ANN-MLP [133], and RF [148] (see Section 3.1.6). These classifiers were adopted based on previous studies that achieved high performance in the classification of different types of balanced and imbalanced datasets [206, 207].

In this work, the ANN-MLP was built using Keras [208] with TensorFlow [209] as the back-end. The neural network contains 102 nodes in the input layer (features vector shape), 200, 180, 180, 100 nodes in each of the four hidden layers respectively, and five nodes in the output layer correspond to the five classes (severities). Each hidden layer applied the ReLU [169] activation function since it is computationally efficient and tend to show better convergence performance than sigmoid function [160]. The output layer applied softmax activation function [161] to predict classes probabilities.

The RF classifier was built with 100 trees based on the suggestion of Oshiro et al. [210] that the number of trees should be between 64 and 128 trees. Gini impurity was selected as decision tree split criteria because it tends to split a node into one small pure node and one large impure node [211], and it can be computationally more efficient than entropy by avoiding log computation.

To evaluate the performance of the classification models, eight evaluation metrics are used, namely overall classification accuracy, precision, sensitivity, specificity, F1-score, Gmean, IBA and AUC. for more details see Section 3.1.7.

5.3 Results and Discussions

This section is presented in four parts. The first part will discuss the results of over-sampling techniques. The second part presents the results of under-sampling techniques. The third part presents the results of hybrid techniques. Finally, the best results obtained from each resampling technique are compared to determine the best resampling technique and the best classifier with the PD tremor dataset. In addition, the results without the resampling technique are presented as a baseline in order to evaluate resampling technique performance.

5.3.1 Over-sampling Results

Table 5.4 shows the performance of two classifiers RF and ANN-MLP, on the PD tremor severity dataset resampled using three over-sampling techniques, SMOTE, ADASYN and Borderline SMOTE. Overall, all the used over-sampling techniques improved classifiers performance significantly. Also, it can be observed that ANN-MLP classifier performed better than RF classifier with over-sampling, while the RF classifier achieved better results than ANN-MLP without over-sampling. The best results were achieved using ANN-MLP classifier combined with Borderline, with 95.04% overall accuracy, 96% Gmean, 93% IBA and 99% AUC. However, The AUC scores of both classifiers with all over-sampling techniques were 99%. Hence, it is important to evaluate classifiers with a different metric such as IBA, which shows slightly different performance between classifiers with different over-sampling techniques which support the discussion about imbalanced datasets performance metrics in Section 3.1.7.

The best performance is achieved using RF classifier with ADASYN technique and obtained 92.58% overall accuracy, 95% Gmean, 91% IBA and 99% AUC, while the worst performance of RF is using Borderline with 91.54% overall accuracy, 94% Gmean, 89% IBA and 99% AUC. On the other hand, ANN-MLP achieved the best performance using the Borderline technique and the worst with SMOTE. Both classifiers did not obtain the best performance with SMOTE, but it is still better than applying classifiers without over-sampling.

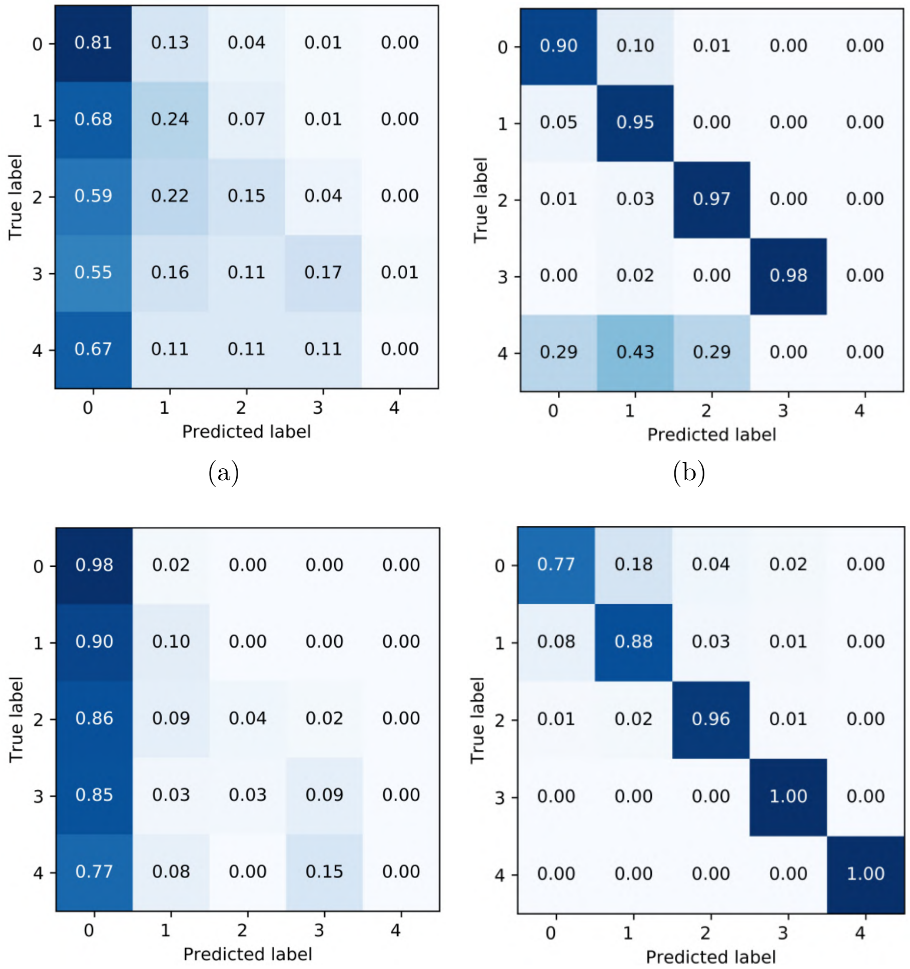
Interestingly, the worst performance among over-sampling techniques was obtained using the RF classifier with Borderline technique, while Borderline achieved the highest performance with ANN-MLP. This result shows that over-sampling techniques performance varies among classifiers, hence no assumption can be made about the best over-sampling technique, because every dataset, classifier and over-sampling technique has it is own characteristics, and different combinations could obtain different results.

Figure 5.3 shows the confusion matrices for ANN-MLP and RF classifiers

Table 5.4: Performance metrics with/without over-sampling for tremor severity classification with RF and ANN-MLP.

Classifier	Metric	Without Resampling	Over-Sampling Technique		
			SMOTE	ADASYN	Borderline
ANN-MLP	Accuracy	62.73%	92.28%	93.15%	95.04%
	Precision	60.00%	92.00%	93.00%	95.00%
	Recall	63.00%	92.00%	93.00%	95.00%
	Specificity	51.00%	97.00%	98.00%	98.00%
	F1-score	61.00%	92.00%	93.00%	95.00%
	G-Mean	50.00%	95.00%	95.00%	96.00%
	IBA	26.00%	89.00%	91.00%	93.00%
	AUC	87.00%	99.00%	99.00%	99.00%
RF	Accuracy	70.66%	92.23%	92.58%	91.54%
	Precision	66.00%	92.00%	93.00%	91.00%
	Recall	71.00%	92.00%	93.00%	92.00%
	Specificity	38.00%	98.00%	98.00%	97.00%
	F1-score	62.00%	92.00%	92.00%	92.00%
	G-Mean	32.00%	95.00%	95.00%	94.00%
	IBA	11.00%	90.00%	91.00%	89.00%
	AUC	92.00%	99.00%	99.00%	99.00%

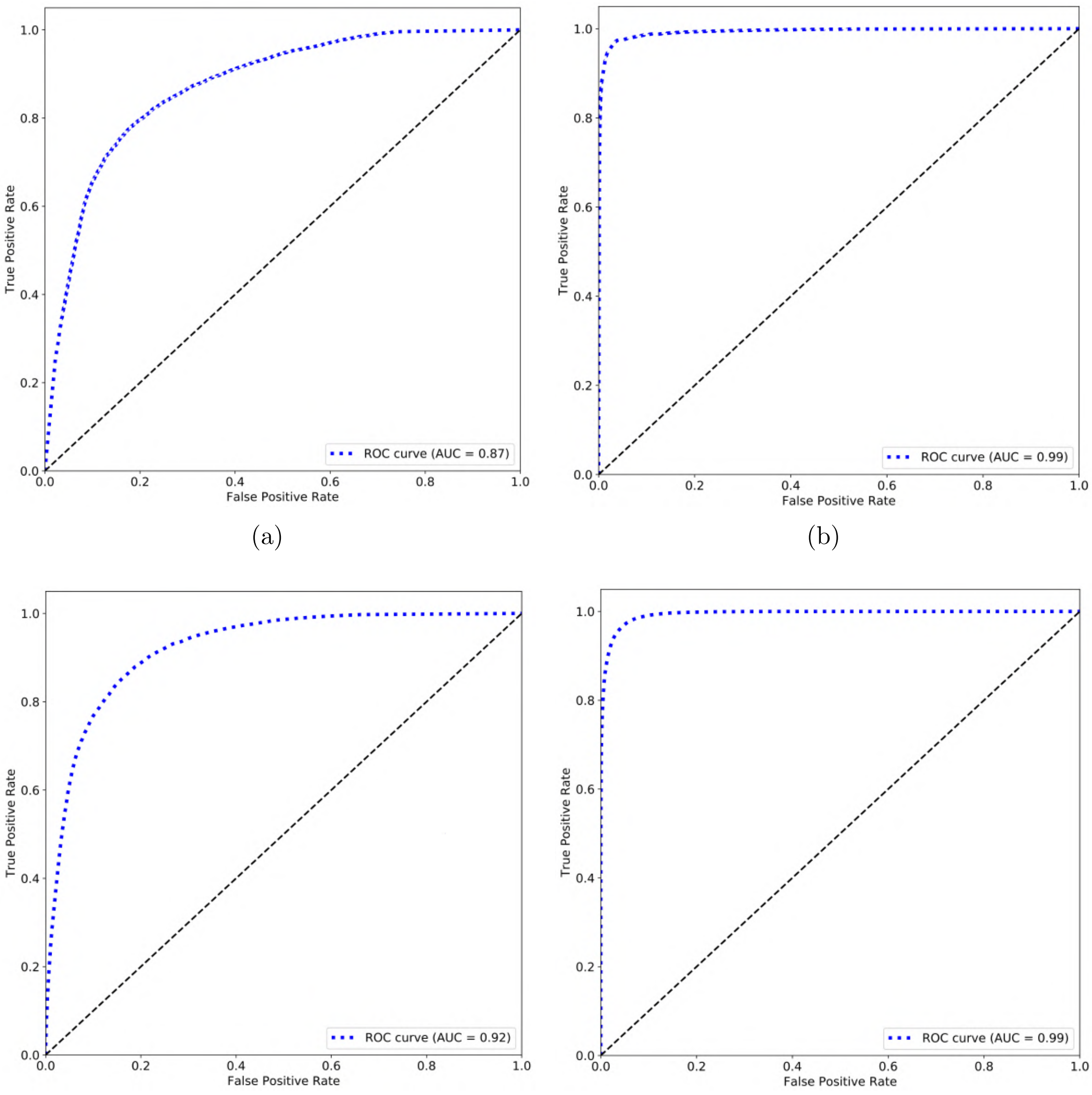
without resampling and with best oversampling techniques. It is clear from confusion matrices how oversampling techniques improved the prediction of all classes without any bias towards majority classes as shown in Figure 5.3b and Figure 5.3d. In contrast, both classifiers are biased to " class 0 " as shown in Figure 5.3a and Figure 5.3c. The associated ROC's for the same results are shown in Figure 5.4.



(a) (b)

(c) (d)

Figure 5.3: Normalised confusion matrices of ANN-MLP and RF without resampling and with best oversampling results; (a) ANN-MLP without resampling, (b) ANN-MLP with BorderlineSMOTE, (c) RF without resampling, (d) RF with ADASYN.



(a)

(b)

(c)

(d)

Figure 5.4: ROC of ANN-MLP and RF without resampling and with best oversampling results; (a) ANN-MLP without resampling, (b) ANN-MLP with BorderlineSMOTE, (c) RF without resampling, (d) RF with ADASYN.

5.3.2 Under-sampling Results

Table 5.5 shows the performance of two classifiers RF and ANN-MLP, on PD tremor severity dataset resampled using five under-sampling techniques, TomekLinks, CNN, AllKNN, IHT and NearMiss. It is clear that overall classifiers performance with under-sampling techniques is significantly worse than using over-sampling techniques. However, some under-sampling techniques improved classifiers performance. Both classifiers (ANN-MLP and RF) achieved the best performance with IHT under-sampling technique, but RF classifier achieved better results with 86.11% overall accuracy, 91.00% Gmean, 82.00% IBA, and 96.00% AUC. The worst performance of both classifiers is with CNN under-sampling technique and did not improve most metrics.

What is striking about the results in Table 5.5 is that most important metrics

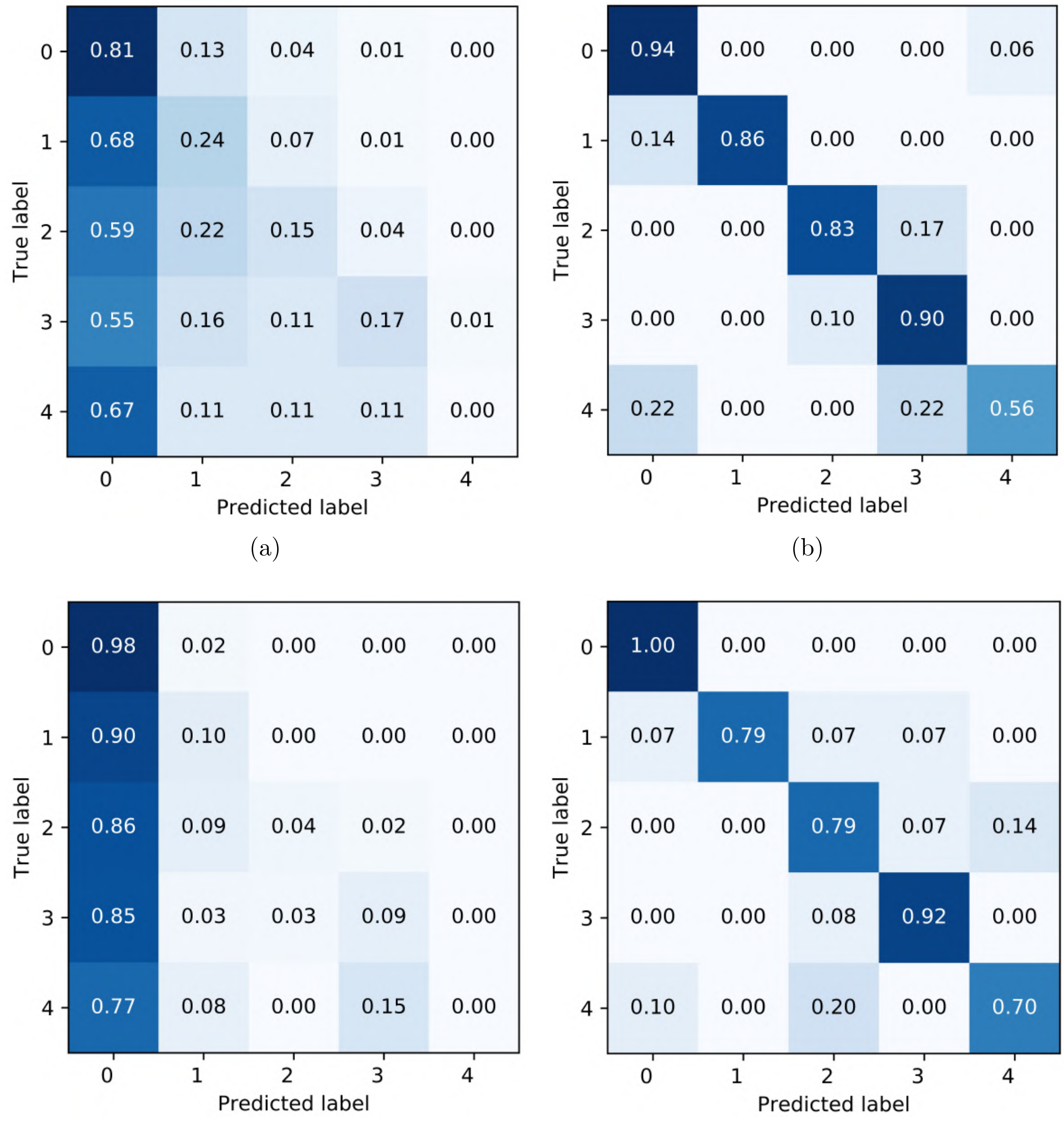
Table 5.5: Performance metrics with/without under-sampling for tremor severity classification with RF and ANN-MLP.

Classifier	Metric	Without Resampling	Under-sampling Technique				
			TomekLinks	CNN	AllKNN	IHT	NearMiss
ANN-MLP	Accuracy	62.73%	74.53%	38.71%	85.66%	70.00%	51.35%
	Precision	60.00%	70.00%	42.00%	83.00%	70.00%	51.00%
	Recall	63.00%	75.00%	39.00%	86.00%	71.00%	51.00%
	Specificity	51.00%	39.00%	79.00%	21.00%	90.00%	87.00%
	F1-score	61.00%	72.00%	39.00%	84.00%	70.00%	51.00%
	G-Mean	50.00%	46.00%	54.00%	34.00%	79.00%	66.00%
	IBA	26.00%	22.00%	29.00%	13.00%	63.00%	45.00%
	AUC	87.00%	92.00%	66.00%	96.00%	95.00%	76.00%
RF	Accuracy	70.66%	80.13%	39.57%	98.44%	86.11%	73.21%
	Precision	66.00%	72.00%	37.00%	87.00%	87.00%	77.00%
	Recall	71.00%	80.00%	40.00%	89.00%	86.00%	73.00%
	Specificity	38.00%	23.00%	78.00%	11.00%	96.00%	95.00%
	F1-score	62.00%	72.00%	37.00%	85.00%	86.00%	74.00%
	G-Mean	32.00%	17.00%	52.00%	9.00%	91.00%	83.00%
	IBA	11.00%	3.00%	28.00%	1.00%	82.00%	68.00%
	AUC	92.00%	76.00%	71.00%	77.00%	96.00%	93.00%

such as Gmean and IBA are very low and declined dramatically with some under-sampling techniques, despite that other metrics are improved. For example, when ALLKNN technique is applied with both classifiers the accuracy, precision and sensitivity improved significantly, while the IBA and Gmean declined. The IBA declined from 26% to 13% with ANN-MLP and from 11% to 1% with RF, Gmean declined from 50% to 34% with ANN-MLP and from 32% to 9% with RF. These results indicate that depending on standard metrics is not sufficient and appropriate for multi-class imbalanced dataset classification.

Similar to over-sampling techniques, some under-sampling techniques improved the performance of one classifier and deteriorated the other. For example, NearMiss improved RF classifier performance but deteriorate ANN-MLP performance, which supports the presented argument that resampling techniques do not perform similarly with different classifiers with different datasets.

Figure 5.5 shows the confusion matrices for ANN-MLP and RF classifiers without resampling and with best undersampling techniques. The associated ROC's for the same results are shown in Figure 5.6. It is clear from confusion matrices how undersampling techniques reduced the bias towards majority classes and enhanced minority classes prediction.



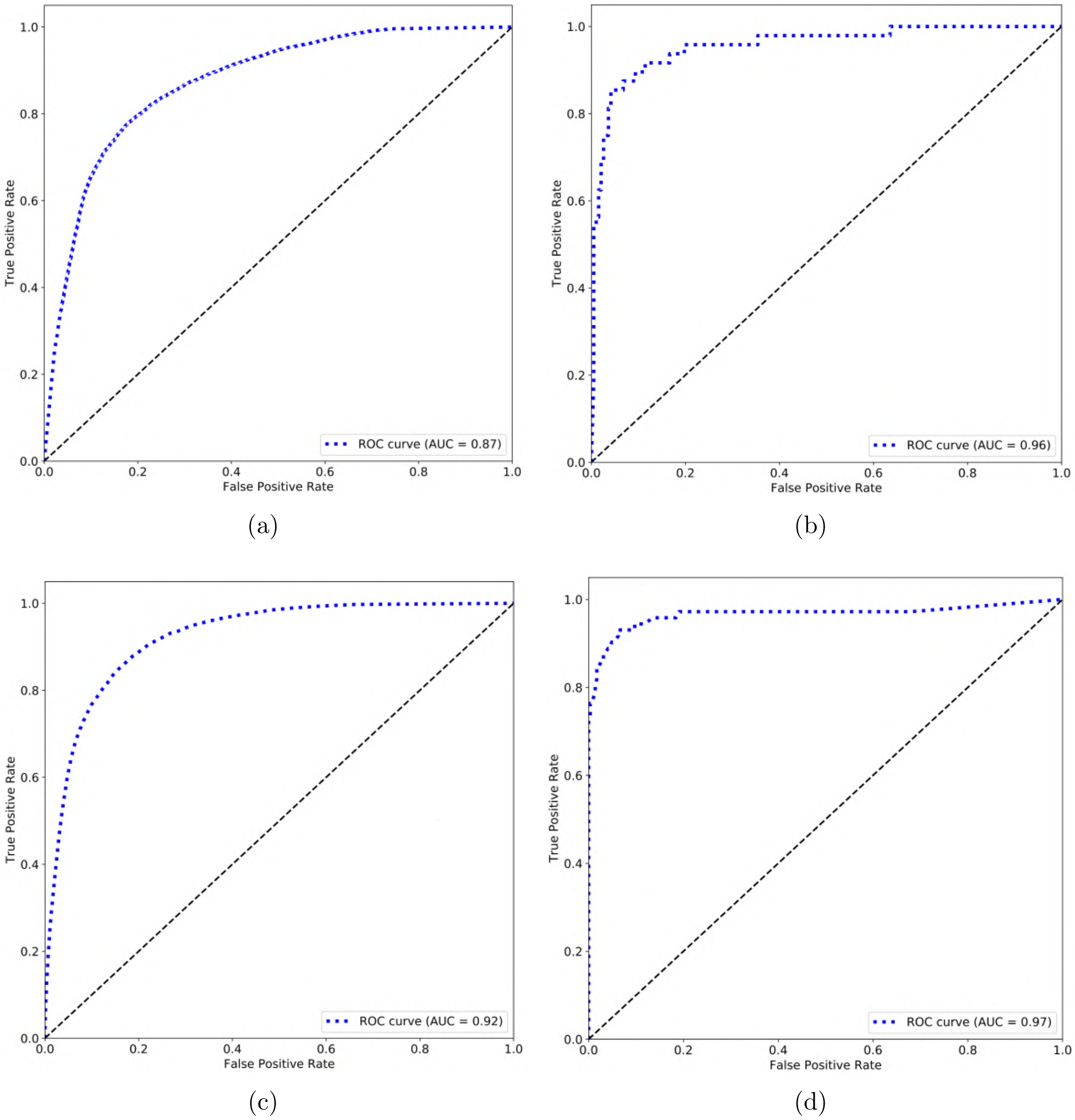
(a)

(b)

(c)

(d)

Figure 5.5: Normalised confusion matrices of ANN-MLP and RF without resampling and with best undersampling results; (a) ANN-MLP without resampling, (b) ANN-MLP with IHT, (c) RF without resampling, (d) RF with IHT.



(a) (b)

(c) (d)

Figure 5.6: ROC of ANN-MLP and RF without resampling and with best undersampling results; (a) ANN-MLP without resampling, (b) ANN-MLP with IHT, (c) RF without resampling, (d) RF with IHT.

5.3.3 Hybrid Results

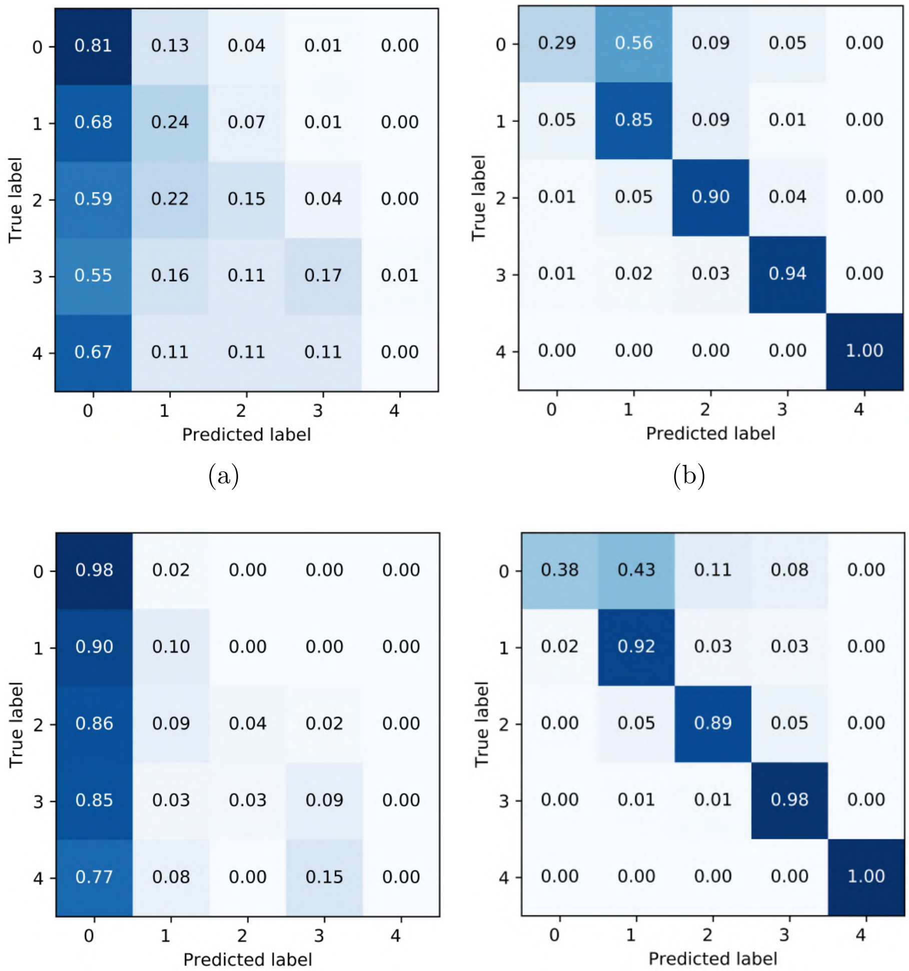
Table 5.6 shows the performance of two classifiers RF and ANN-MLP, on resampled PD tremor severity dataset using two hybrid techniques

SMOTETomek and SMOTEENN. In contrast to under-sampling techniques, both hybrid techniques improved classifiers performance significantly. But the SMOTEENN performance with both classifiers is better than SMOTETomek. However, SOMTEENN obtained the best results with RF classifier with 92.61% overall accuracy, 95% Gmean, 91% IBA and 99% AUC.

Table 5.6: Performance metrics with/without hybrid resampling for tremor severity classification with RF and ANN-MLP.

Classifier	Metric	Without Resampling	Hybrid Technique	
			SMOTETomek	SMOTEENN
ANN-MLP	Accuracy	62.73%	88.03%	90.24%
	Precision	60.00%	98.00%	90.00%
	Recall	63.00%	88.00%	90.00%
	Specificity	51.00%	96.00%	98.00%
	F1-score	61.00%	88.00%	90.00%
	G-Mean	50.00%	92.00%	94.00%
	IBA	26.00%	84.00%	88.00%
	AUC	87.00%	98.00%	98.00%
RF	Accuracy	70.66%	89.47%	92.61%
	Precision	66.00%	89.00%	93.00%
	Recall	71.00%	89.00%	93.00%
	Specificity	38.00%	97.00%	98.00%
	F1-score	62.00%	89.00%	92.00%
	G-Mean	32.00%	93.00%	95.00%
	IBA	11.00%	87.00%	91.00%
	AUC	92.00%	98.00%	99.00%

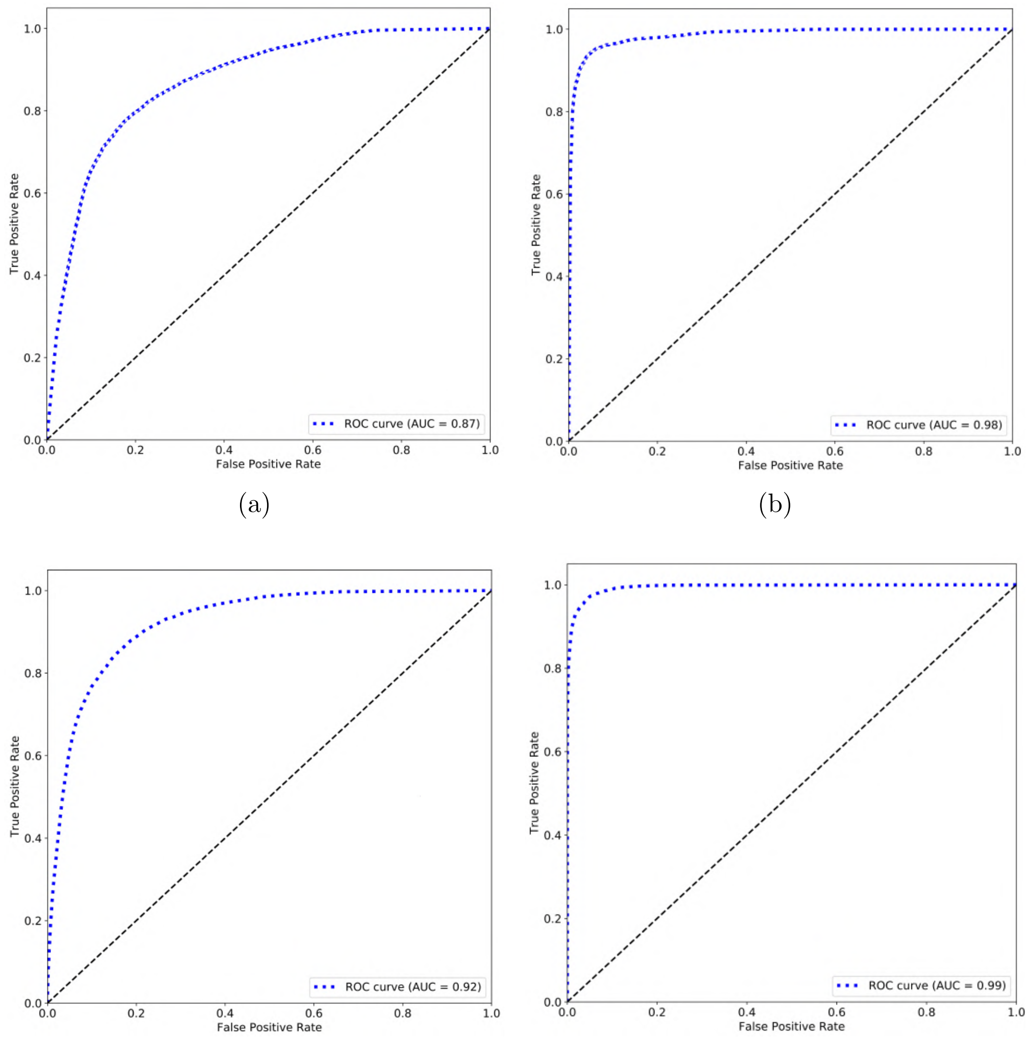
Figure 5.7 shows the confusion matrices for ANN-MLP and RF classifiers without resampling and with best hybrid resampling techniques. The associated ROC's for the same results are shown in Figure 5.8. Similar to the earlier discussed resampling techniques, hybrid resampling techniques reduce the bias or the domination of majority classes and improved the classification of minority classes.



(a) (b)

(c) (d)

Figure 5.7: Normalised confusion matrices of ANN-MLP and RF without resampling and with best hybrid resampling results; (a) ANN-MLP without resampling, (b) ANN-MLP with ENN, (c) RF without resampling, (d) RF with ENN.



(a) (b) (c) (d)

Figure 5.8: ROC of ANN-MLP and RF without resampling and with best hybrid sampling results; (a) ANN-MLP without resampling, (b) ANN-MLP with ENN, (c) RF without resampling, (d) RF withENN.

5.3.4 Performance Comparison

Table 5.7 shows the best results obtained from the two classifiers ANN-MLP and RF in combination with all resampling techniques. Among these results, the best performance was obtained from ANN-MLP classifier with Borderline and achieved 95.04% overall accuracy, 96% Gmean, 93% IBA and 99% AUC. While RF achieved

Table 5.7: Resampling techniques comparison for tremor severity classification with RF and ANN-MLP

	ANN			RF		
	Borderline	IHT	SMOTEENN	ADASYN	IHT	SMOTEENN
Accuracy	95.04%	70.00%	90.24%	92.58%	86.11%	92.61%
Precision	95.00%	70.00%	90.00%	93.00%	87.00%	93.00%
Recall	95.00%	71.00%	90.00%	93.00%	86.00%	93.00%
Specificity	98.00%	90.00%	98.00%	98.00%	96.00%	98.00%
F1-score	95.00%	70.00%	90.00%	92.00%	86.00%	92.00%
G-Mean	96.00%	79.00%	94.00%	95.00%	91.00%	95.00%
IBA	93.00%	63.00%	88.00%	91.00%	82.00%	91.00%
AUC	99.00%	95.00%	98.00%	99.00%	96.00%	99.00%

the best performance with ADASYN and SMOTEENN for all metrics, except the overall accuracy with very low difference (0.03%). However, both classifiers did not improve significantly with IHT in comparison with other resampling techniques, despite that RF performance was higher.

As mentioned in Section 3.1.7, the most important metrics are IBA and AUC, therefore the combinations of ANN-MLP with Borderline, RF with ADASYN and RF with SMOTEENN obtained the same results with 91% IBA and 99% AUC, and overall performance of these combinations achieved best results with a slight difference in some metrics. The worst improvement obtained among the best results is the combination of ANN-MLP with IHT then RF with IHT. So, the order of best combination from high to low is ANN-MLP with Borderline, RF with SMOTEENN, RF with ADASYN and finally ANN-MLP with SMOTEENN, as shown in Figure 5.9. It can thus be suggested that the best approaches to estimate tremor severity are over-sampling and hybrid approaches, while the worst is under-sampling approaches. This hypothesis is supported by the findings in Sections 5.3.1, 5.3.2 and 5.3.3.

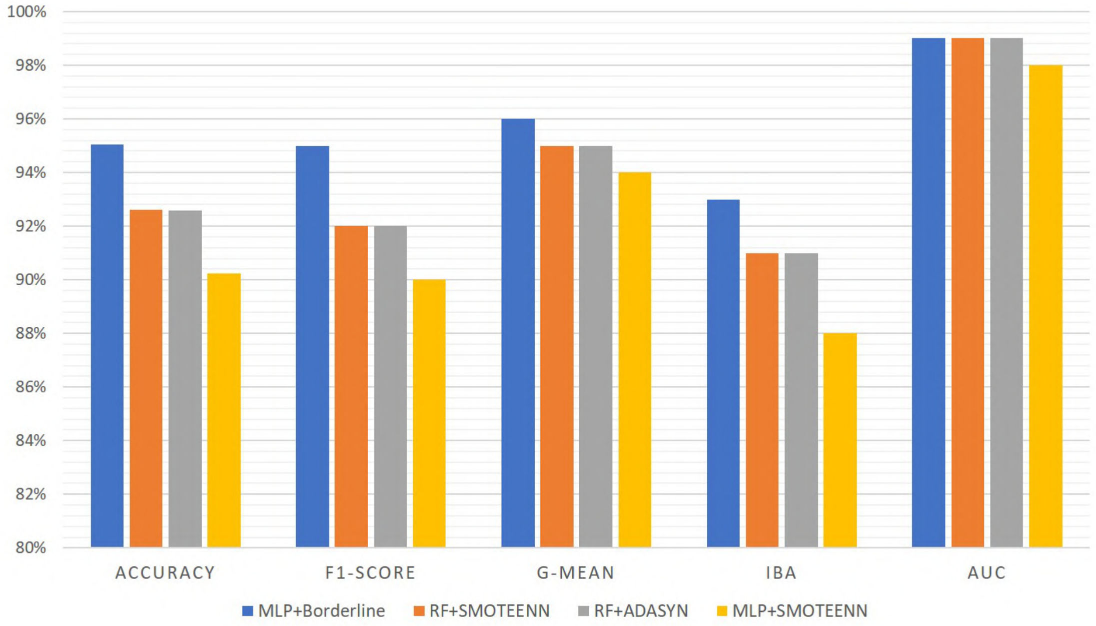


Figure 5.9: Best resampling techniques for tremor severity classification with RF and ANN-MLP.

5.4 Conclusion

In this chapter, a novel approach to enhance the tremor severity classification is proposed. The proposed approach is a combination of signal processing and resampling techniques; over-sampling, under-sampling and a hybrid combination. It can be concluded that that the proposed approach improves the classification process significantly. Classifiers with advanced metrics, such as AUC, Gmean and IBA that are not influenced by data distribution are evaluated. The results show that ANN-MLP with Borderline SMOTE is the best classification approach to identify tremor severity which has obtained 95.04% overall accuracy, 96% Gmean, 91% IBA and 99% AUC. also, the results show that over-sampling techniques performed better than under-sampling techniques and hybrid techniques. The results show that different resampling techniques achieved different results with different classifiers.

Chapter 6

Tasks Oriented Recommended System to Measure Tremor Severity

6.1 Introduction

To date, several objective methods have been proposed for measuring and quantifying PD tremors from data collected while patients performing scripted and unscripted tasks (see Chapter 2). However, up to now, the literature appears to focus on suggesting a tremor severity classification method without taking into consideration other aspects of tremor measurement such as data collection tasks and medication state. For example, in [110], authors reported tremor measurement of the left and the right hands. Even though motor examination of PD is a key aspect of tremor assessment [7], very few studies have explored the effect of the tasks used to collect tremor data on tremor classification and tremor severity detection. For example, in [112], the authors investigated two tasks (standing, sitting) effects on tremor measurement, and the results showed that the correlation with the clinical score is higher when patients were sitting. In addition, relatively little research has been carried out on medication state effects on tremor assessment. For example, In [87], the tremor severity were quantified under two conditions, while patients was on medication and off medication and showed that the correlation with the clinical score is higher when patients were on medication. This indicates a need to explore different aspects of tremor measurement that might improve the objective evaluation of PD tremors.

In order to propose a recommended system to measure tremor severity, it is essential to suggest and validate a method that includes a protocol of data collection including tasks where the tremor severity is highly distinguishable besides signal processing, features extraction, and classification algorithms. In addition, the importance of medication state is indisputable and should be explored as well. Given this, this chapter presents a novel comprehensive method to develop and validate a recommended system to measure and quantify PD tremor severity, including recommended tasks for data collection from different sensors, exploring various classifiers with exhaustive hyper-parameters tuning with and without resampling techniques. Moreover, it investigates medication state effects (ON and OFF) on tremor severity classification.

The remainder of this chapter is structured as follows: Section explains the proposed methodology, including dataset description, signal analysis, features extraction, resampling techniques, classifiers' hyper-parameter optimisation, performance metrics, recommended tasks framework, recommended classifiers and resampling techniques framework, evaluation and medication state effect. Followed by the results presented in Section 6.3. Section 6.4 concludes the work have been undertaken in this chapter.

6.2 Materials and Methods

To define a recommended system for PD tremor measurement, three main components should be identified, best task, best classifier, and best resampling technique. Figure 6.1 illustrates the proposed framework to find the recommended system(s) to detect tremor severity from four different sub-datasets.

Four sub-datasets were prepossessed independently in the first phase to eliminate reliance on sensor orientation and non-tremor data and artefacts. Various time and frequency domains features were extracted from the prepossessed data in the second phase. In the third phase, data was split into

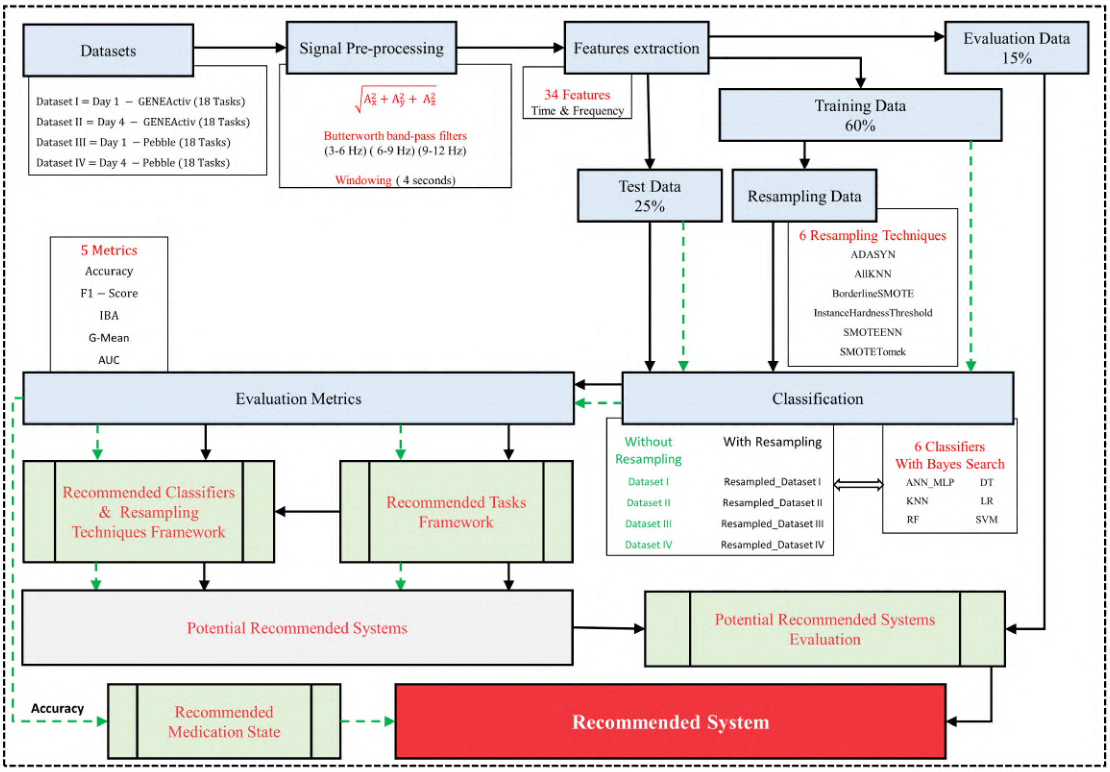


Figure 6.1: Proposed recommended system framework for tremor severity classification.

training, evaluation and test subsets. A copy of training data was resampled by six different resampling techniques independently, in the fourth phase. In the fifth phase, two copies of the training data (with resampling and without res-sampling), and the test data were applied to six different classifiers. The classification results were evaluated by five metrics in the sixth phase. In the seventh phase, the results of all five metrics are passed to recommended tasks framework, recommended classifier and resampling techniques framework. The recommended medication state is identified in the eighth phase utilising only the accuracy results of datasets without resampling. Each step is described in detail in the subsequent sections.

Table 6.1: Motor tasks for collected data

Tasks does involve direct wrist movement	**Code**	**Tasks does not involve direct wrist movement**	**Code**
Drawing and writing on a paper	drawg	Sitting	sittg
Take a glass of water and drink	drnkg	Standing	stndg
Folding towel	fldng	Stairs down	strsd
Finger to nose – left arm	ftnl	Stairs up	strsu
Finger to nose – right arm	ftnr	Sit to stand	ststd
Assembling nuts and bolts	ntblt	Walking while counting	wlkgc
Organizing sheets in a folder	orgpa	Walking through a narrow passage	wlkgp
Repeated arm movement – left arm	raml	Walking straight	wlkgs
Repeated arm movement – right arm	ramr		
Typing on a computer keyboard	typng		

6.2.1 Dataset

The work in this chapter makes use of four datasets, Dataset I, Dataset II, Dataset III and Dataset IV (see Section 3.1.2). The data were collected from 30 patients over four days using a Pebble Smartwatch [1] and GENEActiv accelerometer [2]. In this work, only labelled data that were collected on the first day and the fourth will be utilised. On the first day of data collection, patients came to the laboratory on their regular medication regimen (ON Medication) and performed set ADL tasks and tasks of motor examination of the MDS-UPDRS [7]. On the fourth day, the same procedures that were performed on the first day were performed once again, but the patients were OFF medication for twelve hours. For each task, on the first and the fourth days symptom severity scores (rated 0-4) were provided by a clinician.

The list of tasks performed can be categorised into two groups: Tasks involves direct wrist movement and tasks that does not involve direct wrist movement as shown in Table 6.1

Table 6.2 shows classes (severities) distribution of 103080 instances (windows) segmented from collected data on Day 1 and Day 4 with and without medication. It is clear how data distribution being skewed towards less severe tremors, and this bias can cause significant changes in classification output, in this situation the

[1] https://www.fitbit.com/pebble

[2] https://www.activinsights.com/products/geneactiv/

Table 6.2: Tremor severity distribution on Day 1 (On Medication) and Day 4 (Off Medication).

Tremor Severity	Day 1 (On Medication)		Day 4 (Off Medication)	
	GENEActiv	Pebble	GENEActiv	Pebble
0	18843	19389	16860	17215
1	5845	4491	6534	4421
2	2185	1357	2921	1112
3	845	117	676	103
4	43	11	53	59

classifier is more sensitive to identifying the majority classes but less sensitive to identifying the minority classes.

6.2.2 Signal Processing and Feature Extraction

As in the work presented in Chapter 5, the vector magnitude of mediolateral, vertical and anteroposterior accelerations is calculated and passed through three band-pass filters with cut-off frequencies $3-6$ Hz for RT and $6-9$ Hz for PT and $9-12$ Hz for KT. The filtered signals were segmented into 4 seconds windows using Fixed-length sliding technique with 50% overlap to isolate the tremor region, see Section 3.1.3 and Section 5.2.2. Various features in time and frequency domains were extracted from three frequency bands, $3-6$ Hz for RT, $6-9$ Hz for PT, and $9-12$ Hz for KT, to form a 102 features vector. Frequency domain features were extracted after transforming the signal to frequency domain using FFT, see Section 3.1.4 and Section 5.2.3 for more details.

6.2.3 Resampling Data

Based on the results of resampling techniques presented in Section 5.3, six resampling techniques are employed in the work in this chapter, which are AllKNN and IHT as under-sampling techniques; ADASYN and Borderline-SMOTE as over-sampling techniques; SMOTEENN and SMOTETomek as hybrid resampling techniques. These resampling techniques achieved the best performance. For more details about the working principles of these techniques see Section 3.1.5.

6.2.4 Classification and Hyper-parameter optimisation

Six different classifiers have been considered for classification; ANN-MLP [133], RF [148], SVM [150], DT [147], LR [152], and KNN [159], for more details see Section 3.1.6.

The six classifiers hyper-parameters have been optimised using the Bayesian optimisation algorithm [212]. The Bayesian optimisation algorithm utilises previous evaluations to predict the next set of hyper-parameters that are close to the optimum. Consequently, reducing the number of evaluations required to achieve the best score. In this work Bayes search method from Scikit-optimise [213] has been used with 32 iterations and cross-validation. Table 6.3 shows hyper-parameters search spaces that have explored in this chapter, see Section 3.1.6.

6.2.5 Performance Metrics

To evaluate the performance of the classification models, five evaluation metrics are used, namely overall classification accuracy, F1-score, Gmean, IBA and AUC. for more detail see Section 3.1.7. These metrics are employed in recommended tasks framework, recommended classifiers and resampling techniques framework, which will be discussed in more details in the following sections.

6.2.6 Recommended Tasks Framework

A key aspect of a recommended system is to identify the best tasks or activities performed by PD patients to detect tremor severity. Therefore, a recommended tasks framework is proposed, as shown in Algorithm 1. The algorithm basically utilises classification performance metrics of different classifiers with and without resampling of different tasks from different datasets to identify the best tasks.

Table 6.3: Classifiers' hyper-parameters search spaces.

Classifier	Hyperparameters Search Spaces
ANN-MLP	batch_size : [32, 64, 512] epochs : [200, 300] neurons : Integer (60, 100) optimizer : [SGD, RMSprop, Adam, Adadelta, Adagrad, Adamax, Nadam] activation : [relu, tanh, selu, elu, exponential]
KNN	n_neighbors : Integer (1, 20) weights : [distance, uniform] algorithm : [brute, ball_tree, kd_tree] metric : [minkowski, euclidean, manhattan] leaf_size : Integer (1, 20) p : Integer(1, 2)
RF	n_estimators : Integer(10, 250) max_features : Integer(1, 102) max_depth : Integer(5, 100) min_samples_split : Integer(2, 20) min_samples_leaf : Integer(1, 20) criterion : [gini, entropy]
DT	max_features : Integer(1, 102) max_depth : Integer(5, 100) min_samples_split : Integer(2, 20) min_samples_leaf : Integer(1, 20) criterion : [gini, entropy]
LR	penalty : [l2, none] C : [1e-2, 1e-1, 1e0, 1e1] solver : [newton-cg, lbfgs, sag, saga] max_iter : Integer(1, 1000)
SVM	C : [1, 2, 3, 4, 5, 6, 7, 8, 9, 10] gamma : [0.1, 0.01, 0.001] degree : (1, 5) kernel : [linear, poly, rbf, sigmoid]

Algorithm 1 Finding recommended task framework

```
$counter \leftarrow 0$
$metrics \leftarrow [Accuracy, AUC, G\text{-}mean, F1\text{-}score, IBA]$
$datasets \leftarrow [datasetI, datasetII, datasetIII, datasetIV, resampled\_datasetI,$
                   $resampled\_datasetII, resampled\_datasetIII, resampled\_datasetIV]$
$tasks \leftarrow [drawg, drnkg, fldng, ftnl, ftnr, ntblt, orgpa, raml, ramr, typng, sittg,$
             $stndg, strsd, strsu, ststd, wlkgc, wlkgp, wlkgs]$
$task\_above\_average\_counter \leftarrow [length(tasks)][length(tasks)]$
for $metric \in metrics$ do
    for $dataset \in datasets$ do
        $sum \leftarrow 0$
        $average \leftarrow 0$
        $dataset\_array \leftarrow [length(tasks)][length(tasks)]$
        for $task \in tasks$ do
            $max \leftarrow 0$
            for $metric\_value \in metric\_values$ do
                if $metric\_value > max$ then
                    $max = metric\_value$
                end if
            end for
            $sum = sum + max$
            add $(task, max)$ to $dataset\_array$
        end for
        $average = sum/length(tasks)$
        for $task, max \in dataset\_array$ do
            if $max > average$ then
                $task\_above\_average\_counter \leftarrow [task][counter + 1]$
            else
                $task\_above\_average\_counter \leftarrow [task][counter]$
            end if
        end for
    end for
end for
```

After classification, the performance metrics of all datasets were collected separately. After that, the following steps were performed for each collected metric result independently. The highest value of each metric of each task has been identified in two cases, the first case when the dataset was classified without resampling and the second case with resampling. Then an above-average rule has been applied for each dataset, where the values above average among all tasks has been selected. Following, the number of values above average counted for each task among all datasets.

In the final stage, the total number of all counters for all metrics for each task in all datasets is calculated and sorted in the descending order list. The list of tasks is grouped into three groups: recommended, neutral, and not recommended. Each group will contain six tasks from the datasets that have been performed during data collection.

6.2.7 Recommended Classifiers and Resampling Techniques Framework

After identifying the recommended tasks in the previous section, the results are used to identify the recommended classifier(s) and resampling technique(s). Figure 6.2 presents the proposed framework to identify which classifiers, hyper-parameters and resampling techniques achieved the highest accuracy for each task, and this will produce potential recommended systems, that will be evaluated in Section 6.2.8.

The first stage is to highlight the classifier(s) and hyper-parameters that achieved the highest accuracy with all resampling techniques, then selecting the most frequent classifier(s) that achieved the highest score. The second stage is to select the resampling technique(s) with the highest count with the selected classifier(s) in the first stage. If the count of the selected classifiers and the resampling techniques are more than one in the previous stage, the third stage was applied to filter the results based on the highest validation score then based on the lowest fit time. The potential recommended systems saved for evaluation, which will be explained in the following section.

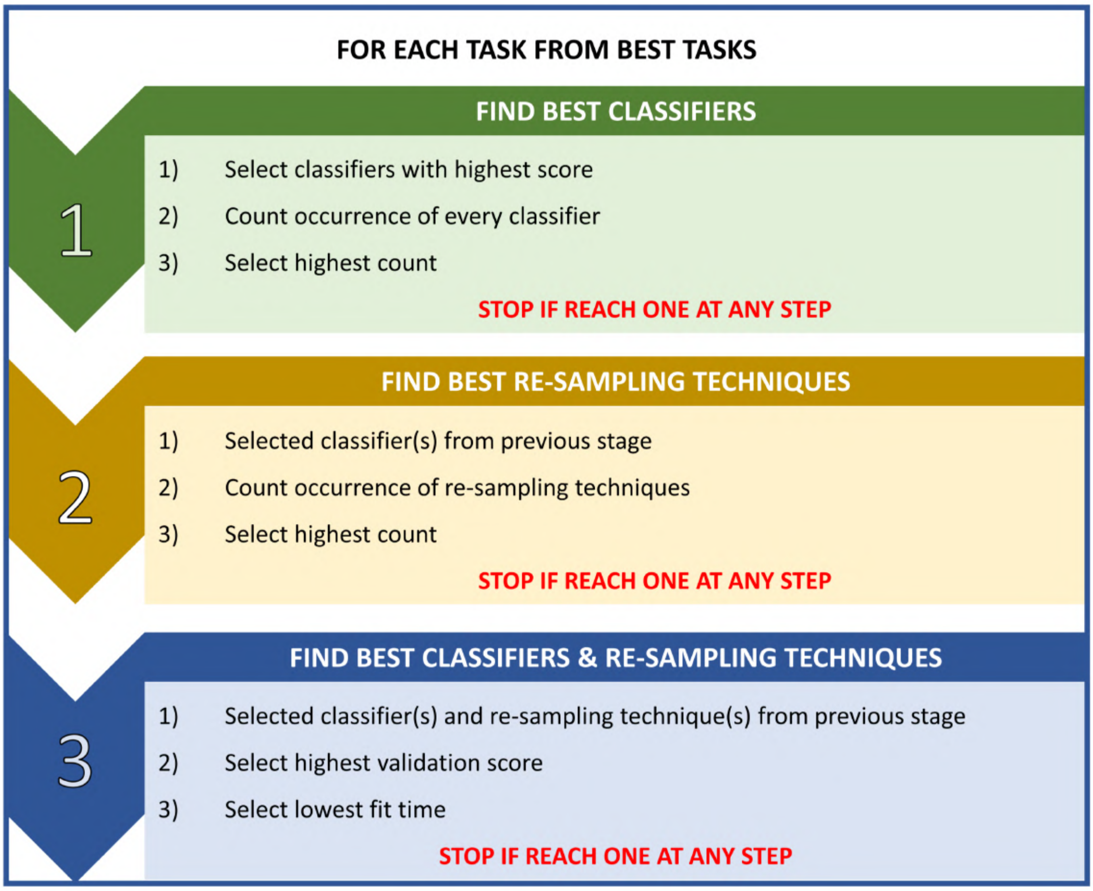

FOR EACH TASK FROM BEST TASKS
1
FIND BEST CLASSIFIERS
1) Select classifiers with highest score
2) Count occurrence of every classifier
3) Select highest count
STOP IF REACH ONE AT ANY STEP
2
FIND BEST RE-SAMPLING TECHNIQUES
1) Selected classifier(s) from previous stage
2) Count occurrence of re-sampling techniques
3) Select highest count
STOP IF REACH ONE AT ANY STEP
3
FIND BEST CLASSIFIERS & RE-SAMPLING TECHNIQUES
1) Selected classifier(s) and re-sampling technique(s) from previous stage
2) Select highest validation score
3) Select lowest fit time
STOP IF REACH ONE AT ANY STEP

Figure 6.2: Recommended classifiers and resampling techniques framework.

6.2.8 Potential Recommended Systems Evaluation

A number of saved potential recommended systems will be evaluated to determine the ideal system for deployment. The evaluation process utilised 15% of all datasets combined. As the recommended system should estimate tremor severity regardless of used data in this work and should work well if the data is collected using the same sensors while subjects are performing the recommended tasks found in this work. Evaluation data was split into two parts, 10% was evaluated through the metrics as described in Section 3.1.7 using the saved potential systems, and 5% was split into 20 samples used as external test data to be predicted as patient data.

The results of the first part of evaluation data, the 10%, was utilised to select top performance models (ideal models), and then the ideal models were tested and validated to predict the 5% external test data. The 5% test data was split into 20 separate samples to predict every sample overall tremor severity by calculating the value at which the probability mass function is the maximum.

6.2.9 Medication State Effect on Tremor Severity Estimation

In order to identify medication state effect and compare the sensors used to collect the data, the datasets are classified without resampling. Three classifiers are considered for classification; ANN-MLP, RF and Support Vector Machine SVM. In this section, an ANN-MLP with 102 nodes in the input layer corresponds to the number of extracted features, 180, 180, 100 nodes in each of the three hidden layers respectively based on prior explorative testing, and 5 nodes in the output layer match to the five tremor severities. A ReLU activation function is used in the hidden layers due to it is high convergence performance [160], and softmax activation function in the output layer to predict tremor severities probabilities.

Guided by previous work, the RF classifier was built with 100 trees [210], and Gini impurity as decision trees split criteria [211]. The SVM classifier is built with RBF kernel, and regularisation parameters $c = 10$ and $gamma = 0.1$.

6.3 Results and Discussions

The section is presented in three parts. The first part will discuss the recommended tasks. The recommended classifiers and resampling techniques are presented in the second part. The third part presents the potential recommended systems and the final recommended system.

6.3.1 Recommended Tasks

Table 6.4 shows the results of one metric (Accuracy) utilised to identify recommended tasks with resampling and without resampling, the highlighted values are above-average among each dataset, while the count above-average column shows values that above-average for datasets for each task. Full results of other metrics (AUC, F1-Score, G-Mean, and IBA) can be found in Appendix C (Tables C.1, C.2, C.3 and C.4) respectively. Closer inspection of all tables shows that resampling techniques improved all metrics significantly. However, classification metrics of all datasets follow the same trend when they resampled and when they did not resample.

Table 6.5 presents the results of count above-average of all metrics, and groups the 18 tasks performed during data collection into three groups; recommended, neutral, and not recommended. It can be observed that tasks that involve direct wrist movements have the lowest count (not recommended tasks), while tasks that do not involve direct wrist movements have the highest count (recommended tasks). The neutral tasks have a count less than the recommended task, but higher than the not recommended tasks. A likely explanation is that these tasks do not involve direct wrist movements similar to not recommended task. So, another possible area of future research would be to investigate these tasks in more detail with different patients.

Together these results provide important insights into tasks performed during data collection influence classification performance, therefore this work presents recommended tasks (stairs down, sitting, stairs up, walking straight, walking while counting and sit to stand) to be performed to measure tremor through wearable devices.

6.3.2 Recommended Classifiers and Resampling Techniques

After identifying the recommended task. The recommended classifier(s) and resampling technique(s) were identified following the framework which was described in Section 6.2.7. Figure 6.3 shows the results of first recommended task (strsd). In the first stage, two classifiers (ANN-MLP and SVM) have the highest count. In the second stage, three resampling techniques (ADASYN, BorderlineSMOT and SMOTETomek) have the highest count with both selected classifiers in the first stage. In the next stage, SVM achieved the highest

Table 6.4: Task highest accuracy of all classifiers and values above-average counts.

	Accuracy								
	Without Resampling				With Resampling				Count Above
	G-1	G-4	P-1	P-4	G-1	G-4	P-1	P-4	Average
drawg	66%	55%	88%	95%	93%	91%	95%	99%	**3**
drnkg	66%	58%	72%	79%	93%	93%	96%	97%	**0**
fldng	71%	63%	75%	80%	94%	91%	95%	96%	**0**
ftnl	77%	76%	65%	62%	97%	96%	95%	96%	**3**
ftnr	53%	68%	76%	86%	90%	98%	97%	99%	**3**
ntblt	71%	63%	71%	75%	95%	94%	95%	96%	**0**
orgpa	66%	75%	67%	77%	96%	98%	96%	97%	**2**
raml	77%	79%	68%	59%	96%	97%	98%	94%	**4**
ramr	68%	59%	82%	85%	96%	91%	98%	99%	**4**
typng	77%	71%	75%	67%	96%	93%	97%	96%	**1**
sittg	78%	75%	87%	93%	100%	98%	98%	99%	**8**
stndg	72%	65%	77%	76%	100%	98%	99%	97%	**3**
strsd	94%	81%	89%	90%	100%	100%	100%	100%	**8**
strsu	80%	86%	90%	100%	100%	100%	100%	100%	**8**
ststd	86%	79%	88%	81%	100%	99%	99%	100%	**7**
wlkgc	76%	74%	90%	83%	98%	96%	99%	98%	**7**
wlkgp	72%	73%	88%	84%	96%	97%	98%	98%	**6**
wlkgs	80%	79%	90%	88%	99%	98%	100%	99%	**8**
Average	**74%**	**71%**	**80%**	**81%**	**97%**	**96%**	**98%**	**98%**	

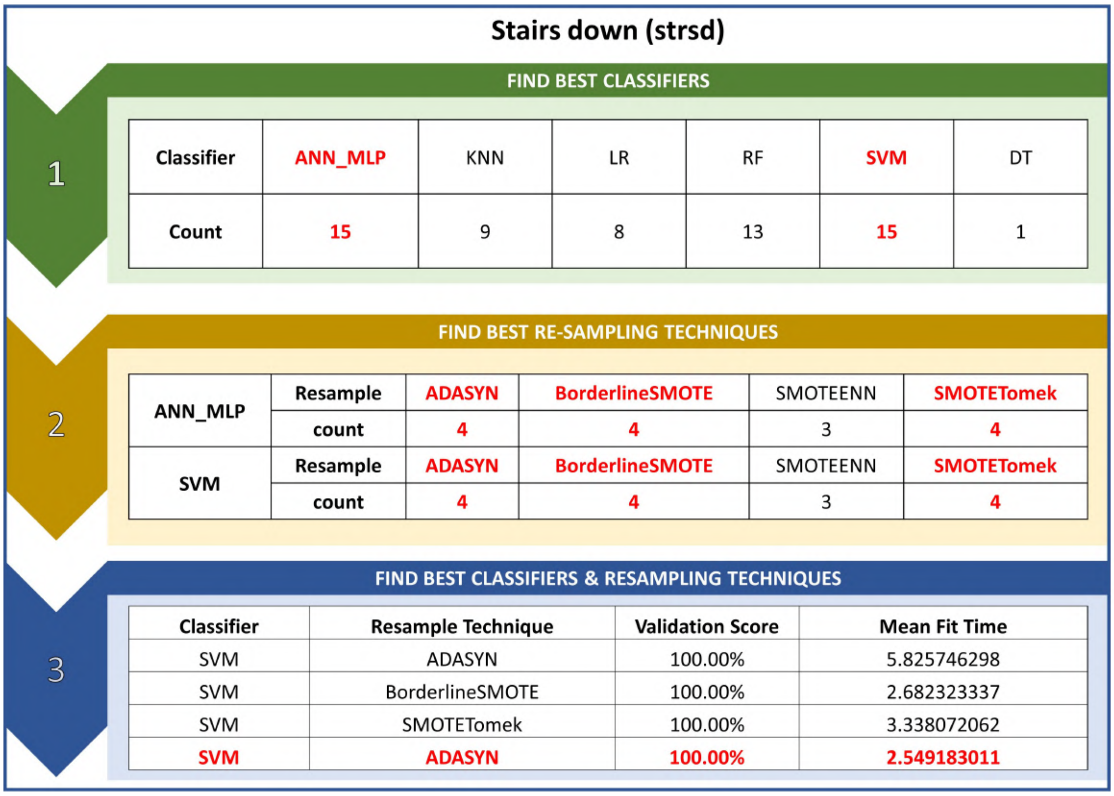
G-1 : GENEActiv - Day1, G-4 : GENEActiv - Day4, P-1 : Pebble - Day1, P-4 : Pebble - Day4

Table 6.5: Tasks above-average count for all metrics.

	Count Above Average						
	Task	**Accuracy**	**AUC**	**F1-score**	**G-Mean**	**IBA**	**Total**
Recommended Tasks	**strsd**	8	8	8	8	8	**40**
	sittg	8	7	8	8	8	**39**
	strsu	8	8	8	6	6	**36**
	wlkgs	8	8	8	6	6	**36**
	wlkgc	7	8	7	5	5	**3**
	ststd	7	7	7	5	4	**30**
Neutral Tasks	**ftnr**	3	6	4	6	5	**24**
	raml	4	6	3	6	5	**24**
	wlkgp	6	7	6	2	3	**24**
	ramr	4	5	4	5	5	**23**
	stndg	3	7	3	5	5	**23**
	ftnl	3	4	3	4	4	**18**
Not Recommended Taks	**orgpa**	2	6	2	2	2	**14**
	drawg	3	2	3	2	2	**12**
	typng	1	5	1	1	1	**9**
	fldng	0	4	0	2	2	**8**
	drnkg	0	3	0	1	1	**5**
	ntblt	0	1	0	0	0	**1**

validation score 100%. Finally, based on fit time, SVM combined with ADASYN was found to be the best model to classify tremor of strsd task, which is the first potential recommended system.The same procedure applied for all recommended tasks to produce six potential systems is presented in Table 6.6. What is interesting about the data in this table is that all potential recommended systems include SVM as a classifier. In addition, the most common kernel is ‘rbf', except system 4.

These findings suggest that SVM with over-sampling and hybrid resampling techniques (ADASYN, BorderlineSMOTE, SMOTETomek and SMOTEENN)



Stairs down (strsd)

1

FIND BEST CLASSIFIERS

Classifier	ANN_MLP	KNN	LR	RF	SVM	DT
Count	15	9	8	13	15	1

2

FIND BEST RE-SAMPLING TECHNIQUES

ANN_MLP	Resample	ADASYN	BorderlineSMOTE	SMOTEENN	SMOTETomek
	count	4	4	3	4
SVM	Resample	ADASYN	BorderlineSMOTE	SMOTEENN	SMOTETomek
	count	4	4	3	4

3

FIND BEST CLASSIFIERS & RESAMPLING TECHNIQUES

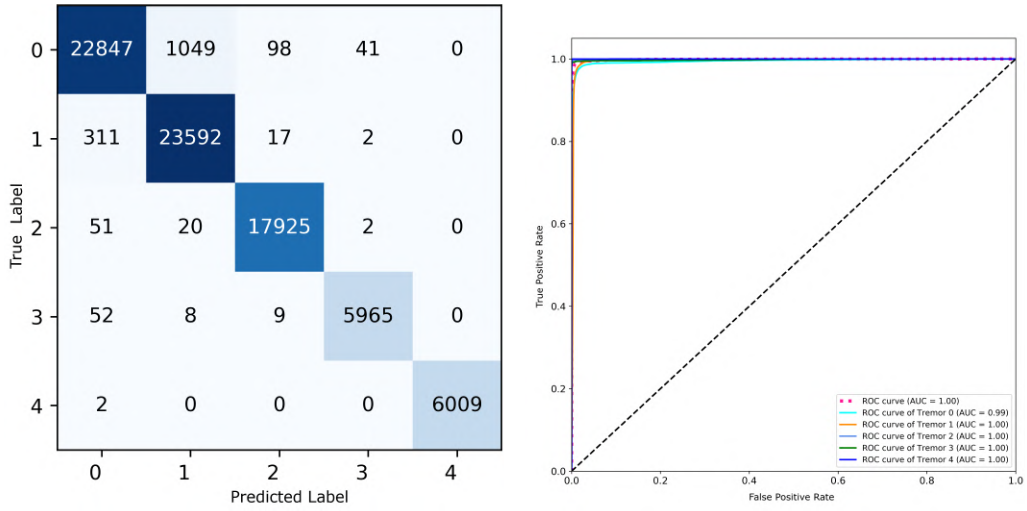
Classifier	Resample Technique	Validation Score	Mean Fit Time
SVM	ADASYN	100.00%	5.825746298
SVM	BorderlineSMOTE	100.00%	2.682323337
SVM	SMOTETomek	100.00%	3.338072062
SVM	ADASYN	100.00%	2.549183011

Figure 6.3: Recommended classifiers and resampling techniques (strsd).

performance is better than other classifiers and resampling techniques that have been investigated in this work. However, in order to identify a recommended system, the potential systems were evaluated as discussed in Section 6.2.8. The performance of potential systems on the evaluation data (10%) are presented in Table 6.7. It is apparent from this table, System 6 achieved the highest performance with 98% accuracy, 98% F1-Score, 98% G-mean, 97% IBA, and 99% AUC, While systems 4 and 5 achieved the worst performance. Systems 1, 2 and 3 performance is lower than System 6 but better than others. Therefore, top 4 systems were evaluated through the tremor severity prediction approach utilising the 5% (20 samples) external test data.

Table 6.6: Potential recommended systems.

System	Task	Classifier	Resample Technique	Validation Score	Hyper-Parameters	Mean Fit Time
System 1	strsd	SVM	ADASYN	100.00%	$C = 10$, $degree = 1$, $gamma = 0.1$, $kernel = rbf$	2.549183011
System 2	sittg	SVM	ADASYN	99.47%	$C = 6$, $degree = 5$, $gamma = 0.1$, $kernel = rbf$	5.469041586
System 3	wlkgs	SVM	ADASYN	98.34%	$C = 10$, $degree = 4$, $gamma = 0.1$, $kernel = rbf$	4.719249964
System 4	strsu	SVM	SMOTETomek	100.00%	$C = 1$, $degree = 5$, $gamma = 0.001$, $kernel = linear$	0.045000315
System 5	wlkgc	SVM	SMOTEENN	98.46%	$C = 10$, $degree = 1$, $gamma = 0.1$, $kernel = rbf$	1.642106652
System 6	ststd	SVM	BorderlineSMOTE	99.14%	$C = 3$, $degree = 5$, $gamma = 0.1$, $kernel = rbf$	6.840166569



(a) (b)

Figure 6.4: The confusion matrix and the ROC of the recommended system; (a) System 6 confusion matrix, (b) System 6 ROC

Table 6.7: Potential systems performance.

System	Classifier	Resample Technique	Accuracy	F1-Score	IBA	G-Mean	AUC
System 1	**SVM**	**ADASYN**	**97%**	**97%**	**96%**	**98%**	**99%**
System 2	**SVM**	**ADASYN**	**97%**	**97%**	**96%**	**98%**	**99%**
System 3	**SVM**	**ADASYN**	**97%**	**97%**	**96%**	**98%**	**100%**
System 4	SVM	SMOTETomek	96%	96%	94%	97%	99%
System 5	SVM	SMOTEENN	93%	93%	90%	95%	99%
System 6	**SVM**	**BorderlineSMOTE**	**98%**	**98%**	**97%**	**98%**	**99%**

The confusion matrix and the ROC curve of the recommended system (System 6) are presented in Figure 6.4.

Table 6.8 shows the predictions results of all 20 samples of the top 4 systems. Systems 2 and 6 predicted all samples correctly, while systems 1 and 3 misclassified sample 19. System one was not able to classify sample 19 exactly as it gives the same probability for severity 3 and 0, while the actual severity is 3. On the other hand, System 3 classified the same sample as 0. Hence, this work suggests System 6 as a recommended system, since it performed better on evaluation and test data and the second choice is System 2, then systems 1 and 3 respectively.

Table 6.8: Top four systems tremor severity predictions.

Sample Data	Actual Severity	Predicted Severity			
		System 1	System 2	System 3	**System 6**
Sample_01	0	0	0	0	0
Sample_02	1	1	1	1	1
Sample_03	2	2	2	2	2
Sample_04	3	3	3	3	3
Sample_05	4	4	4	4	4
Sample_06	0	0	0	0	0
Sample_07	1	1	1	1	1
Sample_08	2	2	2	2	2
Sample_09	3	3	3	3	3
Sample_10	4	4	4	4	4
Sample_11	0	0	0	0	0
Sample_12	1	1	1	1	1
Sample_13	2	2	2	2	2
Sample_14	3	3	3	3	3
Sample_15	4	4	4	4	4
Sample_16	0	0	0	0	0
Sample_17	1	1	1	1	1
Sample_18	2	2	2	2	2
Sample_19	3	**(3, 0)**	3	**(0)**	3
Sample_20	4	4	4	4	4

6.3.3 Medication State Effect on Tremor Severity Estimation

Fig. 6.5 shows the accuracy of three classifiers ANN-MLP, RF and SVM used to estimate tremor severity. The classifiers were evaluated with data have been collected using two sensors, GENEActive and Pebble smartwatch under two conditions: on medication and off medication.

It is clear that the accuracy is higher when patients were on medication with all classifiers and with both sensors. This finding was also reported by Zajki-Zechmeister et al. [87]

Overall, the performance of all classifiers with the Pebble smartwatch is higher than the GENEActive sensor. The best performance is achieved using the RF classifier when patients were on medication using the Pebble smartwatch and obtained 80% accuracy. On the other hand, the worst performance is achieved using the SVM classifiers when patients were off medication and obtained 63% accuracy.

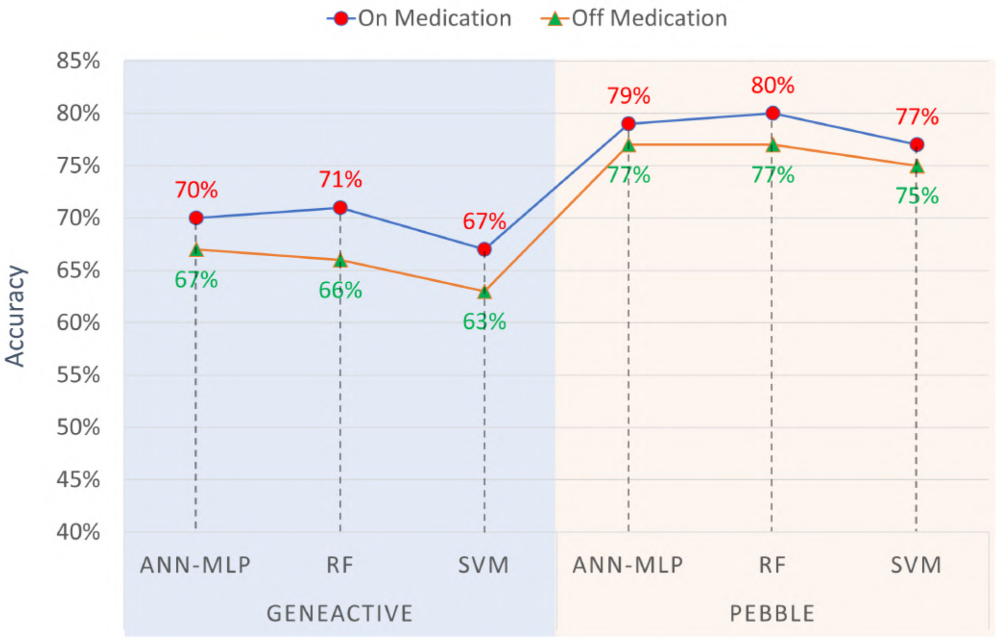


Figure 6.5: Classification accuracy of three classifiers (ANN-MLP, RF, SVM) with two sensors (GENEActive, Pebble smartwatch) under two conditions (ON and OFF medication).

Closer inspection of the figure shows that the RF classifier achieved higher accuracy than other classifiers with both sensors when patients were on medication and obtained 80% with the pebble smartwatch and 71% with the GENEActive sensor. On the other hand, the ANN-MLP achieved higher accuracy with both sensors when patients were off medication with 66% with the GENEActive sensor and 77% with the pebble smartwatch.

These findings suggest that medication state and type of sensor affect tremor severity objective measurement. In this work, tremor estimation is better when patients were on medication using the pebble smartwatch. However, with small sample size, caution must be applied, as the findings might not be applicable to all clinical settings and experiments setups. Moreover, the results show that different classifiers achieved different results with different sensors.

6.4 Conclusion

The main goal of the work in this chapter was to identify a recommended system that can be used to measure tremor severity using wearable devices combined with machine learning techniques. This work thoroughly examined the influence of tasks performed during data collection on classification performance. Furthermore, a comprehensive approach was used to identify the best classifiers, classifiers hyper-parameters, and resampling techniques in combination with signal processing and robust features extraction techniques. Different metrics, including accuracy, F1-score, G-mean, IBA and AUC, have been used to identify the recommended system using a novel algorithm to avoid bias. In general, ADL tasks that involve direct wrist movements are not suitable for tremor severity assessment such as drawing, writing, drinking, folding a towel, typing, organising sheets in a folder and assembling nuts and bolts. On the other hand, tasks that do not involve direct wrist movements achieved high performance of tremor severity classification. In addition, resampling techniques can improve classification performance. In this work, the recommended system has been suggested to evaluate tremor severity from data was collected using two types of wearable devices while patients are either on medication or off medication. The recommended system consists of three main components which are the classifier, the resampling technique, and the tasks to be performed during data collection. The findings of this work suggest that the best system is the SVM classifier combined with Borderline SMOTE over-sampling technique and the tasks are sitting, stairs up and down, walking straight, walking while counting, and standing. The suggested recommended system has been tested using evaluation data from two wearable devices and achieved 98% accuracy, 98% F1-score, 97% IBA, 98% G-mean and 99% AUC. In addition, it has been tested to predict tremor severity of test data from both wearable devices, and it was able to predict all samples correctly. Moreover, it has been shown that tremor quantification is better when patients are on medication. Also, it compared different classifiers with two different sensors and found that the Pebble smartwatch with the RF classifier is the best approach to evaluate tremor severity when patients are on medication and achieved 80% accuracy. On the other hand, the ANN-MLP and RF achieved the highest accuracy with the Pebble smartwatch when patients were off medication and obtained 77% accuracy.

Chapter 7

Conclusions and Future Work

7.1 Thesis Summary

The work undertaken in this thesis has presented a novel attempt to provide a comprehensive recommended system to enhance the tremor severity classification in Parkinson's Disease (PD) by developing an efficient, reliable, handy, and cross-platform sensor solution that objectively quantify tremor severity. Furthermore, to know and understand healthcare professionals' and patients' perspectives on current clinical assessment methods, as well as their preferences and requirements of wearable devices. The motivation behind this work is the current limitations in current clinical assessment methods, and there is no general agreement about a reliable and valid solution to measure PD tremor severity. In addition the limited adoption and implementation of wearable devices in PD assessment which has been discussed in Section 2.5.

To answer the research questions made in Section 1.3. Consequently, achieving project objectives and aim, a sequential instrument design mixed-method approach employed by combining both qualitative methods and quantitative methods to develop a solution that is coherent in clinicians and patients points of view to avoid the off-the-shelf solution. The work in this thesis is organised into two stages, starting from a qualitative approach, in which health professionals' and patients' perspectives and requirements are identified. Then a quantitative approach comprised signal processing and machine learning techniques to quantify PD tremor kinematic data onto a tremor severity rating scale. As a summary, Chapter 3 explored PD patients' and healthcare professionals' perspectives in terms of wearable devices design and implementation linked to current assessment methods. Chapter 5 enhanced PD tremor severity estimation of imbalanced data using resampling (data-level) approach. Chapter 6 explored tasks and medication state effects on tremor severity classification to develop a recommended system to quantify PD tremor.

7.1.1 Concluding Remarks

This thesis attempts to provide a recommended approaches to measure tremor severity in PD. The conclusions and major contributions for the various aspects of the project are presented below:

Patients' and Healthcare Professionals' Perceptions on Wearable Devices and Current Assessment

This thesis presents one of the first comprehensive qualitative studies to explore both patients' and healthcare professionals' perspectives. This is specifically linked to current diagnosis and assessment methods, wearable device design and materials, and the requirements and specifications of a combined PD monitoring solution. These findings will be of interest in the development of a monitoring solution that meets both clinicians' and patients' needs and requirements. A holistic approach was adopted by first using exploratory semi-structured interviews with healthcare professionals, followed by focus group discussions with patients affected with PD. The results reported in Chapter 4 identified the following relevant themes: (1) Current diagnosis and assessment are dubious art, (2) The role of aesthetics and design for acceptance and adoption (3) Patients and healthcare professionals want wearable technology that eases and refines treatments. Also, the solution should be easier to use, provide very concise information, and be easy to interpret. Also, it should mimic the current scale. The results have demonstrated that current assessment and diagnosis methods are subjective, inconsistent and depend on healthcare professionals' skills and their interpretations. The participants' perspectives were positive toward using wearable devices, particularly if they were involved in the early design stages. Patients emphasised that the devices should be comfortable, but they did not have any concerns regarding device visibility or data privacy transmitted over the internet when it comes to their health. In terms of wearing a monitor, the preferable part of the body for all participants was the wrist. Healthcare professionals stated a need for an economical solution that is easy to interpret. Some design aspects identified by patients included clasps, material choice, and form factor. In addition, the provision of additional features for the wearable device, like fall detection and medication alerts could be appealing to patients and has a pivotal role in terms of ultimate user acceptance.

We acknowledge that this work has several limitations. First, the sample size may not be fully representative of the wider PD and healthcare provider population. Second, all participants were residing in the Nottingham area. Hence, perceptions may differ in other regions of the world which may limit the generalisability. However, qualitative research rarely seeks to generalise but to explore perceptions. Third, while interviewing participants, it has been noticed different levels of knowledge and experience with technologies, so responses were likely based on previous experience with available technologies, such as wearable devices or smartwatches. Therefore, future research should attempt to include participants from different regions, with different experience and knowledge.

solving the imbalanced data problem by applying different resampling techniques with different classifiers. Also, it improved tremor severity detection significantly without neglecting minority classes. Moreover, it offers an important insights into advanced metrics and how standard metrics sometimes mislead classification results

Enhanced Parkinson's Disease Tremor Severity Classification

The work was undertaken in this chapter contributes to the existing knowledge on objective tremor severity quantification by solving the imbalanced data problem by applying different resampling techniques with different classifiers. Also, it has improved tremor severity detection significantly without neglecting minority classes. In addition, it offers important insights into advanced metrics and how standard metrics can mislead classification results. The proposed approach is a combination of signal processing and resampling techniques integrated with ANN-MLP and RF. Various resampling techniques were investigated; over-sampling (SMOTE, ADASYN, Borderline SMOTE), under-sampling (CNN, Tomek–links, AllKNN, IHT, NearMiss), and a hybrid combination (SMOTETomek, SMOTEENN). Advanced metrics that are not influenced by data distribution are used to evaluate the proposed approach such as AUC, Gmean and IBA, besides the common metrics such as accuracy, precision, sensitivity, specificity and F1-score. The results in Chapter 5 showed that the proposed approach improves the classification process significantly, and the ANN-MLP with Borderline SMOTE is the best classification approach to identify tremor severity which has obtained 95.04% overall accuracy, 96% Gmean, 93% IBA and 99% AUC. The combinations of ANN-MLP with Borderline, RF with ADASYN and RF with SMOTEENN obtained the same results with 91% IBA and 99% AUC, and overall performance of these combinations achieved the best results with a slight difference in some metrics. The worst improvement obtained among the best results is the combination of ANN-MLP with IHT then RF with IHT. So, the order of best combination from high to low is ANN-MLP with Borderline, RF with SMOTEENN, RF with ADASYN and finally ANN-MLP with SMOTEENN. It can thus be suggested that the best approaches to estimate tremor severity are over-sampling and hybrid approaches, while the worst is under-sampling approaches. Besides, the results showed that different resampling techniques achieved different results with different classifiers.

The generalisability of these results is subject to certain limitations. First, the sample size is small, and it's possible that it doesn't represent the entire PD population. Second, the data was gathered in a single environment. As a result, if the environment is changed, the outcomes may vary. Third, the proposed method should be tested on a variety of datasets

Tasks Oriented Recommended System to Measure Tremor Severity

This work provides one of the first attempts to discriminate tasks' effect on tremor severity detection by developing an efficient and unique metric rule-based algorithm to identify recommended and non recommended tasks to be performed for tremor data collection. The findings will be of interest for future research that investigates the measurement of tremor severity objectively by laying on a data collection protocol instead of the traditional trial and error approach. In addition, it explored medication state effect on tremor severity classification. The identified recommended system is based on the above-average rule of five advanced metrics (accuracy, F1-score, G-mean, IBA and AUC) results of four sub-datasets, six re-sampling techniques, six classifiers besides signal processing and robust features extraction techniques. The recommended system comprises recommended tasks, classifier, classifier hyper-parameters and resampling technique.

The results presented in Chapter 6 showed that SVM with over-sampling and hybrid re-sampling techniques (ADASYN, BorderlineSMOTE, SMOTETomek and SMOTEENN) performance is better than other classifiers and resampling techniques that have been examined in this chapter. These results support the findings in Chapter 5. Furthermore, it demonstrated that tasks that do not involve direct wrist movements are better than tasks that involves direct wrist movements for tremor severity measurements. In addition, resampling techniques improve classification performance significantly. The findings suggest a recommended system consists of Support Vector Machine (SVM) classifier combined with BorderlineSMOTE over-sampling technique and data collection while performing a set of recommended tasks which are sitting, stairs up and down, walking straight, walking while counting, and standing.

The identified recommended system has been evaluated using evaluation data from two wearable devices (Pebble smartwatch and GENEActiv) and obtained 98% accuracy, 98% F1-score, 97% IBA, 98% G-mean and 99% AUC. In addition, it has been tested to classify tremor severity of test data from both wearable devices, and it was able to predict all samples correctly. Moreover, it has been shown that tremor measurement is better when patients are on medication. Also, it compared different classifiers with the Pebble smartwatch and GENEActiv separately and found that the Pebble smartwatch with the RF classifier is the best approach to evaluate tremor severity when patients are on medication and achieved 80% accuracy. On the other hand, ANN-MLP and RF achieved the highest accuracy with the Pebble smartwatch when patients were off medication and obtained 77% accuracy.

This work was limited by the small sample size and the data collection in a single environment, besides the lack of inter and intra reliability of results validation from different researchers.

7.2 Answers to Research Questions

Based on the results achieved for the research objectives it is possible now to answer the research questions presented:

- **Research Question 1: What do PD patients and healthcare professionals want from wearable medical devices? How can they benefit from wearable technologies?** The results presented in Chapter 4 showed that wearable devices could be utilised to improve the current assessment process, and the objective data is needed as reliable markers; therefore, the solution could help improve current treatments or lead to new ones. Moreover, the solution should be easier to use, provide very concise information, and be easy to interpret. Also, it should mimic the current MDS-UPDRS scale.

- **Research Question 2: What is the role of wearable devices aesthetics and design for acceptance and adoption?** Again, the results presented in Chapter 4 found that healthcare professionals' and patients' perspectives of wearable technology were positive. Device design is one of the most important factors that determine whether patients are willing to wear the device. Healthcare professionals felt that the device must be comfortable, easy to use, non-invasive, and should easily be worn under clothes without catching/snagging. The device should also be water-resistant, washable, durable, and easy to fasten to minimise daily disruption.

- **Research Question 3: Can we perform automatic PD tremor severity classification from different wearable devices contains tri-axial accelerometer of multiple tasks using machine learning techniques?** The results presented in Chapter 5 and Chapter 6 showed that employing a wide range of handcrafted features that provide detailed and discriminatory information of signal characteristics and that are highly correlated with tremor severity was able to predict tremor severity with a very high level of accuracy.

- **Research Question 4: On which types of tasks must we focus while performing tremor severity assessment? In other words, Which tasks or types of tasks maximise inter-class separations?** The results in Chapter 6 showed that the significance of the tasks performed during data collection on classification performance. The results showed that ADL tasks that involve direct wrist movements are not suitable for tremor severity assessment such as drawing, writing, drinking, folding a towel, typing, Organising sheets in a folder and assembling nuts and bolts. On the other hand, tasks that do not involve direct wrist movements achieved high performance of tremor severity classification.

- **Research Question 5: How much improvement for imbalanced data classification can be achieved by employing resampling techniques?** The results in Chapter 5 showed that Resampling techniques can enhance classification performance significantly. Over-sampling techniques performed better than other resampling techniques, also hybrid techniques performed better than under-sampling techniques. Besides, it is found that different resampling techniques performed differently with different classifiers.

- **Research Question 6: Can we utilise advanced classification metrics to identify a cross-platform recommended system that comprises recommended tasks, the recommended classifier(s) and recommended resampling technique(s) to quantify tremor**

severity? The proposed approach in Chapter 6 was used to identify a recommended system that comprises the best tasks, the best classifiers, the best classifiers hyper-parameters, and the best resampling techniques based on advanced metrics instead of depending on one metrics, which make these systems more robust.

- **Research Question 7: What is the effect of medication state on tremor severity classification?** Medication state (ON or OFF) effect was investigated in Chapter 6 on objective measurement of tremor severity, and it has been shown that tremor quantification is better when patients are on medication. The findings will be of interest for future research that tries to measure tremor severity objectively.

7.3 Future Work

Following the work undertaken in this thesis, this section outlines the main directions for future work:

- The qualitative data was collected in Chapter 3 is relatively small. Therefore, an increasing number of participants to include participants from different regions, with different experiences and knowledge to obtain as many end-user views as possible would be very useful to design an acceptable prototype device. Moreover, to utilise other data collection methods such as surveys that allows massive information to be collected by a large number of individuals in a relatively short time. This would ensure that the views gathered from the PD patients and healthcare professionals are a true representation of the overall views.

- Design and develop a prototype wearable device based on the results are presented in Chapter 3 and based on UCD philosophy by involving PD patients and healthcare professionals in early design stages to ensure high acceptance and adoption. This could lead to improvements in the efficiency, comfort and aesthetics of the wearable device.

- The next stage of research should be testing the identified recommended system developed in Chapter 6 with different datasets that are collected from accelerometer sensors in a different environment. Also, to collect data based on the identified recommended tasks. These will increase the reliability and validity of the recommended system.

- In Chapters 5 and 6 various features have been extracted in both the time and frequency domains. However, classifier training time could be reduced and the results could be improved by analysing these features and selecting the most important features will be very useful. This includes adding or/and removing features.

- In this work, the data is processed offline in order to determine PD tremor severity. So, it would be interesting to make data collection and analysis in real-time. This can be achieved by utilising low cost and high resources cloud-based platforms which would improve the efficiency of the system.

- It would be interesting to build a user-friendly interface for patients and clinicians that eases the use of tremor severity estimation system. For example, to include instruction of data collection and tasks to be performed by patients with demonstration video and voice instruction. Also, to interpret and present the results of analysed data to clinicians visually and close to current scoring systems.

- In this thesis only one window size (4 seconds) with 50% overlap using a sliding window technique is investigated. Therefore, exploring different techniques with different windows sizes and different overlaps effects on tremor severity estimation can be studied as part of future work.

- The work reported in Chapter 6 investigated medication state effect on tremor severity measurement. Further research might explore medication response and efficacy including tremor progression and suspension.

- In this thesis, only accelerometer signals have been utilised. For future studies, gyroscope signals could be explored or the combination of both signals would be interesting to investigate tremor direction besides its amplitude.

References

[1] Joseph Jankovic, "Parkinson's disease: clinical features and diagnosis", *Journal of neurology, neurosurgery & psychiatry*, vol. 79, no. 4, pp. 368–376, 2008. 1, 2, 13, 14

[2] Daniel Weintraub, Cynthia L Comella, and Stacy Horn, "Parkinson's disease–part 1: Pathophysiology, symptoms, burden, diagnosis, and assessment", *Am J Manag Care*, vol. 14, no. 2 Suppl, pp. S40–S48, 2008. 1, 2, 13, 14, 16, 18, 21, 25, 30

[3] Ramón Cacabelos, "Parkinson's disease: from pathogenesis to pharmacogenomics", *International journal of molecular sciences*, vol. 18, no. 3, pp. 551, 2017. 1

[4] Parkinson's Foundation, "Statistics. http://parkinson.org/Understanding-Parkinsons/Causes-and-Statistics/Statistics", (accessed September 16, 2020). 1

[5] Parkinson's UK, "Facts and figures about parkinson's for journalists. https://www.parkinsons.org.uk/about-us/media-and-press-office", (accessed September 16, 2020). 1

[6] Kimberly D Seifert and Jonathan I Wiener, "The impact of datscan on the diagnosis and management of movement disorders: A retrospective study", *American journal of neurodegenerative disease*, vol. 2, no. 1, pp. 29, 2013. 1

[7] Christopher G Goetz, Barbara C Tilley, Stephanie R Shaftman, Glenn T Stebbins, Stanley Fahn, Pablo Martinez-Martin, Werner Poewe, Cristina Sampaio, Matthew B Stern, Richard Dodel, et al., "Movement disorder

society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): scale presentation and clinimetric testing results", *Movement disorders: official journal of the Movement Disorder Society*, vol. 23, no. 15, pp. 2129–2170, 2008. 2, 23, 24, 49, 118, 121

[8] Brian M Bot, Christine Suver, Elias Chaibub Neto, Michael Kellen, Arno Klein, Christopher Bare, Megan Doerr, Abhishek Pratap, John Wilbanks, E Ray Dorsey, et al., "The mpower study, parkinson disease mobile data collected using researchkit", *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016. 2, 19, 24, 85

[9] James M. Fisher, Nils Y. Hammerla, Thomas Ploetz, Peter Andras, Lynn Rochester, and Richard W. Walker, "Unsupervised home monitoring of Parkinson's disease motor symptoms using body-worn accelerometers", *Parkinsonism & Related Disorders*, vol. 33, 2016. 2, 24

[10] Minja Belić, Vladislava Bobić, Milica Badža, Nikola Šolaja, Milica Đurić-Jovičić, and Vladimir S Kostić, "Artificial intelligence for assisting diagnostics and assessment of parkinson's disease–a review", *Clinical neurology and neurosurgery*, p. 105442, 2019. 2

[11] Dongni Johansson, Kristina Malmgren, and Margit Alt Murphy, "Wearable sensors for clinical applications in epilepsy, parkinson's disease, and stroke: a mixed-methods systematic review", *Journal of neurology*, vol. 265, no. 8, pp. 1740–1752, 2018. 2, 26, 44

[12] Erika Rovini, Carlo Maremmani, and Filippo Cavallo, "How wearable sensors can support parkinson's disease diagnosis and treatment: a systematic review", *Frontiers in neuroscience*, vol. 11, pp. 555, 2017. 2, 92

[13] Hasan Hasan, Dilan S Athauda, Thomas Foltynie, and Alastair J Noyce, "Technologies assessing limb bradykinesia in parkinson's disease", *Journal of Parkinson's disease*, vol. 7, no. 1, pp. 65–77, 2017. 2, 43

[14] Alberto J Espay, Jeffrey M Hausdorff, Álvaro Sánchez-Ferro, Jochen Klucken, Aristide Merola, Paolo Bonato, Serene S Paul, Fay B Horak, Joaquin A Vizcarra, Tiago A Mestre, et al., "A roadmap for implementation

of patient-centered digital outcome measures in parkinson's disease obtained using mobile health technologies", *Movement Disorders*, vol. 34, no. 5, pp. 657–663, 2019. 2, 43

[15] Catarina Godinho, Josefa Domingos, Guilherme Cunha, Ana T. Santos, Ricardo M. Fernandes, Daisy Abreu, Nilza Gonçalves, Helen Matthews, Tom Isaacs, Joy Duffen, Ahmed Al-Jawad, Frank Larsen, Artur Serrano, Peter Weber, Andrea Thoms, Stefan Sollinger, Holm Graessner, Walter Maetzler, and Joaquim J. Ferreira, "A systematic review of the characteristics and validity of monitoring technologies to assess parkinson's disease", *Journal of NeuroEngineering and Rehabilitation*, vol. 13, no. 1, pp. 24, dec 2016. 2, 43

[16] Per Odin, K. Ray Chaudhuri, Jens Volkmann, Angelo Antonini, Alexander Storch, Espen Dietrichs, Zvezdan Pirtošek, Tove Henriksen, Malcolm Horne, David Devos, and Filip Bergquist, "Viewpoint and practical recommendations from a movement disorder specialist panel on objective measurement in the clinical management of parkinson's disease", *npj Parkinson's Disease*, vol. 4, no. 1, pp. 14, dec 2018. 2, 43

[17] Arthur G Money, Julie Barnett, Jasna Kuljis, Michael P Craven, Jennifer L Martin, and Terry Young, "The role of the user within the medical device design and development process: medical device manufacturers' perspectives", *BMC medical informatics and decision making*, vol. 11, no. 1, pp. 15, 2011. 2, 43, 44

[18] D Ramyachitra and P Manikandan, "Imbalanced dataset classification and solutions: a review", *International Journal of Computing and Business Research (IJCBR)*, vol. 5, no. 4, 2014. 3, 74, 98

[19] James Parkinson, "An essay on the shaking palsy", *The Journal of neuropsychiatry and clinical neurosciences*, vol. 14, no. 2, pp. 223–236, 2002. 12

[20] JA Obeso, Maria Stamelou, CG Goetz, Werner Poewe, AE Lang, D Weintraub, David Burn, Glenda Margaret Halliday, E Bezard, S Przedborski, et al., "Past, present, and future of parkinson's disease:

a special essay on the 200th anniversary of the shaking palsy", *Movement Disorders*, vol. 32, no. 9, pp. 1264–1310, 2017. 12

[21] Ole-Bjørn Tysnes and Anette Storstein, "Epidemiology of parkinson's disease", *Journal of Neural Transmission*, vol. 124, no. 8, pp. 901–905, 2017. 12, 16

[22] Stephen G. Reich and Joseph M. Savitt, "Parkinson's disease", *Medical Clinics of North America*, vol. 103, no. 2, pp. 337–350, 2019, Neurology for the Non-Neurologist. 12

[23] Hsiao-Chun Cheng, Christina M Ulane, and Robert E Burke, "Clinical progression in parkinson disease and the neurobiology of axons", *Annals of neurology*, vol. 67, no. 6, pp. 715–725, 2010. 13

[24] Joohi Shahed and Joseph Jankovic, "Motor symptoms in parkinson's disease", *Handbook of clinical neurology*, vol. 83, pp. 329–342, 2007. 13

[25] W Poewe, "Non-motor symptoms in parkinson's disease", *European journal of neurology*, vol. 15, pp. 14–20, 2008. 13, 18, 25, 85

[26] Paola Pierleoni, Lorenzo Palma, Alberto Belli, and Luca Pernini, "A real-time system to aid clinical classification and quantification of tremor in parkinson's disease", in *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2014, pp. 113–116. 13, 34, 39, 43, 51, 100, 102

[27] Natalie E Allen, Allison K Schwarzel, and Colleen G Canning, "Recurrent falls in parkinson's disease: a systematic review", *Parkinson's disease*, vol. 2013, 2013. 14, 93

[28] NICE, "NICE Guidance 2017", 2017. 16

[29] WR Gibb and AJ1033142 Lees, "The relevance of the lewy body to the pathogenesis of idiopathic parkinson's disease.", *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 51, no. 6, pp. 745–752, 1988. 16, 80

[30] Giovanni Rizzo, Massimiliano Copetti, Simona Arcuti, Davide Martino, Andrea Fontana, and Giancarlo Logroscino, “Accuracy of clinical diagnosis of parkinson disease: a systematic review and meta-analysis”, *Neurology*, vol. 86, no. 6, pp. 566–576, 2016. 16

[31] Thomas G Beach and Charles H Adler, “Importance of low diagnostic accuracy for early parkinson’s disease”, *Movement Disorders*, vol. 33, no. 10, pp. 1551–1554, 2018. 16

[32] Andrew Lees, “The bare essentials: Parkinson’s disease.”, *Practical neurology*, vol. 10, no. 4, pp. 240–246, 2010. 17

[33] NICE, “Parkinson’s disease — Management”, 2018. 17, 25

[34] CG Goetz, “Unified parkinson’s disease rating scale (updrs) and the movement-disorder society sponsored-unified parkinson’s disease rating scale (mds-updrs)”, *Encyclopedia of Movement Disorders*, 2010. 17

[35] Royal College of Physicians, “Consultant physicians working with patients, revised 5th edition (online update)”, 2013. 17

[36] Roongroj Bhidayasiri and Pablo Martinez-Martin, “Clinical assessments in parkinson’s disease: scales and monitoring”, *International review of neurobiology*, vol. 132, pp. 129–182, 2017. 17, 18, 21, 22, 24

[37] Robert A Hauser, Jeffrey Friedlander, Theresa A Zesiewicz, Charles H Adler, Lauren C Seeberger, Christopher F O’Brien, Eric S Molho, and Stewart A Factor, “A home diary to assess functional status in patients with parkinson’s disease with motor fluctuations and dyskinesia”, *Clinical neuropharmacology*, vol. 23, no. 2, pp. 75–81, 2000. 17

[38] Khalid Ali and Huw R Morris, “Parkinson’s disease: chameleons and mimics”, *Practical neurology*, vol. 15, no. 1, pp. 14–25, 2015. 17, 18

[39] C Vitale, MT Pellecchia, D Grossi, N Fragassi, T Cuomo, L Di Maio, and P Barone, “Unawareness of dyskinesias in parkinson’s and huntington’s diseases”, *Neurological Sciences*, vol. 22, no. 1, pp. 105–106, 2001. 18

[40] Arthur A Stone, Saul Shiffman, Joseph E Schwartz, Joan E Broderick, and Michael R Hufford, “Patient non-compliance with paper diaries”, *Bmj*, vol. 324, no. 7347, pp. 1193–1194, 2002. 18

[41] Lisa M Shulman, Ingrid Pretzer-Aboff, Karen E Anderson, Rashida Stevenson, Christopher G Vaughan, Ann L Gruber-Baldini, Stephen G Reich, and William J Weiner, “Subjective report versus objective measurement of activities of daily living in parkinson’s disease”, *Movement disorders*, vol. 21, no. 6, pp. 794–799, 2006. 18

[42] Douglas G Tincello, Kate S Williams, Miland Joshi, R Phillip Assassa, and Keith R Abrams, “Urinary diaries: a comparison of data collected for three days versus seven days”, *Obstetrics & Gynecology*, vol. 109, no. 2, pp. 277–280, 2007. 18

[43] Daniel Weintraub and David J Burn, “Parkinson’s disease: the quintessential neuropsychiatric disorder”, *Movement Disorders*, vol. 26, no. 6, pp. 1022–1031, 2011. 18

[44] Nima Toosizadeh, Jane Mohler, Hong Lei, Saman Parvaneh, Scott Sherman, and Bijan Najafi, “Motor performance assessment in parkinson’s disease: association between objective in-clinic, objective in-home, and subjective/semi-objective measures”, *PloS one*, vol. 10, no. 4, pp. e0124763, 2015. 18

[45] Ana Lígia Silva de Lima, Tim Hahn, Nienke M de Vries, Eli Cohen, Lauren Bataille, Max A Little, Heribert Baldus, Bastiaan R Bloem, and Marjan J Faber, “Large-scale wearable sensor deployment in parkinson’s patients: the parkinson@ home study protocol”, *JMIR research protocols*, vol. 5, no. 3, pp. e5990, 2016. 19, 24, 85

[46] Pablo Martinez-Martin, Carmen Rodríguez Blázquez, Maria João Forjaz, and Kallol Ray Chaudhuri, *Assesment Scales in Parkinson’s Disease*, Springer, 2014. 19, 23

[47] Claudia Ramaker, Johan Marinus, Anne Margarethe Stiggelbout, and Bob Johannes Van Hilten, “Systematic evaluation of rating scales for

impairment and disability in parkinson's disease", *Movement disorders: official journal of the Movement Disorder Society*, vol. 17, no. 5, pp. 867–876, 2002. 19, 20, 21, 22

[48] V Venkatesh and R Swarnalatha, "Rewiew and assessment of the rating scales of parkinson's disease", *IJAS [Internet]*, 2018. 19, 20

[49] The Cooperative Multicentric Group, P Martínez-Martin, A Gil-Nagel, L Morlán Gracia, J Balseiro Gómez, FJ Martínez-Sarriés, F Bermejo, T Del Ser Quijano, MC Macías, C Jiménez-Rojas, et al., "Intermediate scale for assessment of parkinson's disease. characteristics and structure", *Parkinsonism & related disorders*, vol. 1, no. 2, pp. 97–102, 1995. 21

[50] Margaret M Hoehn, Melvin D Yahr, et al., "Parkinsonism: onset, progression, and mortality", *Neurology*, vol. 50, no. 2, pp. 318–318, 1998. 21, 22

[51] Robert S Schwab, "Projection technique for evaluating surgery in parkinson's disease", in *Third symposium on Parkinson's disease.* E&S Livingstone, 1969, pp. 152–157. 21

[52] Andrea Ginanneschi, Francesco Degl'Innocenti, Maria T Maurello, Stefano Magnolfi, Paolo Marini, and Luigi Amaducci, "Evaluation of parkinson's disease: a new approach to disability", *Neuroepidemiology*, vol. 10, no. 5-6, pp. 282–287, 1991. 21

[53] Johan Marinus, Martine Visser, Anne M Stiggelbout, J Martin Rabey, Pablo Martínez-Martín, Ubaldo Bonuccelli, Peter H Kraus, and Jacobus J van Hilten, "A short scale for the assessment of motor impairments and disabilities in parkinson's disease: the spes/scopa", *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 75, no. 3, pp. 388–395, 2004. 22

[54] Christopher G Goetz, Werner Poewe, Olivier Rascol, Cristina Sampaio, Glenn T Stebbins, Carl Counsell, Nir Giladi, Robert G Holloway, Charity G Moore, Gregor K Wenning, et al., "Movement disorder society task force report on the hoehn and yahr staging scale: status and recommendations the movement disorder society task force on rating scales for parkinson's disease", *Movement disorders*, vol. 19, no. 9, pp. 1020–1028, 2004. 22

[55] Joel S Perlmutter, "Assessment of parkinson disease manifestations", *Current protocols in neuroscience*, vol. 49, no. 1, pp. 10–1, 2009. 22

[56] Janice L Palmer, Mary A Coats, Catherine M Roe, Shelly M Hanko, Chengjie Xiong, and John C Morris, "Unified parkinson's disease rating scale-motor exam: inter-rater reliability of advanced practice nurse and neurologist assessments", *Journal of advanced nursing*, vol. 66, no. 6, pp. 1382–1387, 2010. 24

[57] Ruirui Lu, Yan Xu, Xiaohui Li, Yongli Fan, Weiqi Zeng, Yang Tan, Kang Ren, Wenwu Chen, and Xuebing Cao, "Evaluation of wearable sensor devices in parkinson's disease: A review of current status and future prospects", *Parkinson's Disease*, vol. 2020, 2020. 25, 26

[58] K Harish, M Venkateswara Rao, Rupam Borgohain, A Sairam, and P Abhilash, "Tremor quantification and its measurements on parkinsonian patients", in *2009 International Conference on Biomedical and Pharmaceutical Engineering*. IEEE, 2009, pp. 1–3. 26

[59] Ramon G Garcia, Alejandro H Ballado, Arnold C Paglinawan, Charmaine C Paglinawan, Regina B Gavino, Bruce Aeron J Magcamit, Juan Carlo S Miranda, and Mario F Tiongson, "Hand tremor analyzer using accelerometry for preliminary diagnosis, classification and monitoring of selected movement disorders", in *2016 6th IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*. IEEE, 2016, pp. 392–396. 26

[60] Michelle Braybrook, Sam O'Connor, Philip Churchward, Thushara Perera, Parisa Farzanehfar, and Malcolm Horne, "An ambulatory tremor score for parkinson's disease", *Journal of Parkinson's disease*, vol. 6, no. 4, pp. 723–731, 2016. 26

[61] Francesco Bove, Giulia Di Lazzaro, Delia Mulas, Fabrizio Cocciolillo, Daniela Di Giuda, and Anna Rita Bentivoglio, "A role for accelerometry in the differential diagnosis of tremor syndromes", *Functional neurology*, vol. 33, no. 1, pp. 45, 2018. 26

[62] Bhuvan Molparia, Brian N Schrader, Eli Cohen, Jennifer L Wagner, Sandeep R Gupta, Sherrie Gould, Nelson Hwynn, Emily G Spencer, and Ali

Torkamani, “Combined accelerometer and genetic analysis to differentiate essential tremor from parkinson’s disease”, *PeerJ*, vol. 6, pp. e5308, 2018. 27

[63] Decho Surangsrirat, Chusak Thanawattano, Ronachai Pongthornseri, Songphon Dumnin, Chanawat Anan, and Roongroj Bhidayasiri, “Support vector machine classification of parkinson’s disease and essential tremor subjects based on temporal fluctuation”, in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2016, pp. 6389–6392. 27

[64] Alan Michael Woods, Mariusz Nowostawski, Elizabeth A Franz, and Martin Purvis, “Parkinson’s disease and essential tremor classification on mobile device”, *Pervasive and Mobile Computing*, vol. 13, pp. 1–12, 2014. 28

[65] Luay Fraiwan, Ruba Khnouf, and Abdel Razaq Mashagbeh, “Parkinson’s disease hand tremor detection system for mobile application”, *Journal of medical engineering & technology*, vol. 40, no. 3, pp. 127–134, 2016. 28

[66] R Arvind, B Karthik, N Sriraam, and J Kamala Kannan, “Automated detection of pd resting tremor using psd with recurrent neural network classifier”, in *2010 International Conference on Advances in Recent Technologies in Communication and Computing*. IEEE, 2010, pp. 414–417. 29

[67] Santosh Kumar Nanda, Wen-Yen Lin, Ming-Yih Lee, and Rou-Shayn Chen, “A quantitative classification of essential and parkinson’s tremor using wavelet transform and artificial neural network on semg and accelerometer signals”, in *2015 IEEE 12th International Conference on Networking, Sensing and Control*. IEEE, 2015, pp. 399–404. 29

[68] Nooshin Haji Ghassemi, Franz Marxreiter, Cristian F Pasluosta, Patrick Kugler, Johannes Schlachetzki, Axel Schramm, Bjoern M Eskofier, and Jochen Klucken, “Combined accelerometer and emg analysis to differentiate essential tremor from parkinson’s disease”, in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2016, pp. 672–675. 29

[69] Werner Poewe, “The natural history of parkinson’s disease”, *Journal of neurology*, vol. 253, no. 7, pp. vii2–vii6, 2006. 30

[70] Jose A Obeso, Maria C Rodriguez-Oroz, Christopher G Goetz, Concepcion Marin, Jeffrey H Kordower, Manuel Rodriguez, Etienne C Hirsch, Matthew Farrer, Anthony HV Schapira, and Glenda Halliday, “Missing pieces in the parkinson’s disease puzzle”, *Nature medicine*, vol. 16, no. 6, pp. 653–661, 2010. 30

[71] Charlotte A Haaxma, Bastiaan R Bloem, George F Borm, Wim JG Oyen, Klaus L Leenders, Silvia Eshuis, Jan Booij, Dean E Dluzen, and Martin WIM Horstink, “Gender differences in parkinson’s disease”, *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 78, no. 8, pp. 819–824, 2007. 30

[72] Daphne GM Zwartjes, Tjitske Heida, Jeroen PP Van Vugt, Jan AG Geelen, and Peter H Veltink, “Ambulatory monitoring of activities and motor symptoms in parkinson’s disease”, *IEEE transactions on biomedical engineering*, vol. 57, no. 11, pp. 2778–2786, 2010. 30

[73] O Martinez-Manzanera, Elizabeth Roosma, Martijn Beudel, RWK Borgemeester, Teus van Laar, and Natasha M Maurits, “A method for automatic and objective scoring of bradykinesia using orientation sensors and classification algorithms”, *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 5, pp. 1016–1024, 2015. 30

[74] Joseph P. Giuffrida, David E. Riley, Brian N. Maddux, Dustin A. Heldman, and Dustin A. Heldmann, “Clinically deployable kinesia™ technology for automated tremor assessment”, *Movement Disorders*, vol. 24, no. 5, pp. 723–730, apr 2009. 31, 36, 38

[75] Khalil Niazmand, Karin Tonn, Anastasios Kalaras, Stefan Kammermeier, Kai Boetzel, Jan-Hinnerk Mehrkens, and Tim C Lueth, “A measurement device for motion analysis of patients with parkinson’s disease using sensor based smart clothes”, in *2011 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops.* IEEE, 2011, pp. 9–16. 31, 32, 38, 101, 102

[76] Khalil Niazmand, Karin Tonn, Anastasios Kalaras, Urban M Fietzek, Jan-Hinnerk Mehrkens, and Tim C Lueth, "Quantitative evaluation of parkinson's disease using sensor based smart glove", in *2011 24th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2011, pp. 1–8. 32, 38

[77] George Rigas, Alexandros T Tzallas, Markos G Tsipouras, Panagiota Bougia, Evanthia E Tripoliti, Dina Baga, Dimitrios I Fotiadis, Sofia G Tsouli, and Spyridon Konitsiotis, "Assessment of tremor activity in the parkinson's disease using a set of wearable sensors", *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 3, pp. 478–487, 2012. 33, 38

[78] Avishai Wagner, Naama Fixler, and Yehezkel S Resheff, "A wavelet-based approach to monitoring parkinson's disease symptoms", in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5980–5984. 33, 39

[79] Hyoseon Jeon, Woongwoo Lee, Hyeyoung Park, Hong Ji Lee, Sang Kyong Kim, Han Byul Kim, Beomseok Jeon, and Kwang Suk Park, "Automatic classification of tremor severity in parkinson's disease using a wearable device", *Sensors*, vol. 17, no. 9, pp. 2067, 2017. 34, 36, 40, 101

[80] Houtao Deng, George Runger, and Eugene Tuv, "Bias of importance measures for multi-valued attributes and solutions", in *International conference on artificial neural networks*. Springer, 2011, pp. 293–300. 34

[81] Hyoseon Jeon, Woongwoo Lee, Hyeyoung Park, Hong Ji Lee, Sang Kyong Kim, Han Byul Kim, Beomseok Jeon, and Kwang Suk Park, "High-accuracy automatic classification of parkinsonian tremor severity using machine learning method", *Physiological measurement*, vol. 38, no. 11, pp. 1980, 2017. 34, 36, 40

[82] Jasmina Dj Novaković, Alempije Veljović, Siniša S Ilić, Željko Papić, and Tomović Milica, "Evaluation of classification models in machine learning", *Theory and Applications of Mathematics & Computer Science*, vol. 7, no. 1, pp. 39–46, 2017. 35

[83] Paolo Angeles, Yen Tai, Nicola Pavese, Samuel Wilson, and Ravi Vaidyanathan, "Automated assessment of symptom severity changes during deep brain stimulation (dbs) therapy for parkinson's disease", in *2017 International Conference on Rehabilitation Robotics (ICORR)*. IEEE, 2017, pp. 1512–1517. 35, 40

[84] Abdul Haleem Butt, Erika Rovini, Dario Esposito, Giuseppe Rossi, Carlo Maremmani, and Filippo Cavallo, "Biomechanical parameter assessment for classification of parkinson's disease on clinical scale", *International Journal of Distributed Sensor Networks*, vol. 13, no. 5, pp. 1550147717707417, 2017. 35, 36, 40

[85] Houde Dai, Guoen Cai, Zhirong Lin, Zengwei Wang, and Qinyong Ye, "Validation of inertial sensing-based wearable device for tremor and bradykinesia quantification", *IEEE Journal of Biomedical and Health Informatics*, 2020. 36, 37, 42

[86] Md Nafiul Alam, Benjamin Johnson, Jeffrey Gendreau, Kouhyar Tavakolian, Colin Combs, and Reza Fazel-Rezai, "Tremor quantification of parkinson's disease-a pilot study", in *2016 IEEE International Conference on Electro Information Technology (EIT)*. IEEE, 2016, pp. 0755–0759. 36, 40, 101

[87] Tibor Zajki-Zechmeister, Mariella Kögl, Kerstin Kalsberger, Sebastian Franthal, Nina Homayoon, Petra Katschnig-Winter, Karoline Wenzel, László Zajki-Zechmeister, and Petra Schwingenschuh, "Quantification of tremor severity with a mobile tremor pen", *Heliyon*, vol. 6, no. 8, pp. e04702, 2020. 36, 41, 43, 118, 135

[88] William Elazmeh, Nathalie Japkowicz, and Stan Matwin, "Evaluating misclassifications in imbalanced data", in *European Conference on Machine Learning*. Springer, 2006, pp. 126–137. 37, 74

[89] Haibo He and Edwardo A Garcia, "Learning from imbalanced data", *IEEE Transactions on knowledge and data engineering*, vol. 21, no. 9, pp. 1263–1284, 2009. 37, 74

[90] James A Hanley and Barbara J McNeil, “The meaning and use of the area under a receiver operating characteristic (roc) curve.”, *Radiology*, vol. 143, no. 1, pp. 29–36, 1982. 37, 74

[91] Jie Du, Chi-Man Vong, Chi-Man Pun, Pak-Kin Wong, and Weng-Fai Ip, “Post-boosting of classification boundary for imbalanced data using geometric mean”, *Neural Networks*, vol. 96, pp. 101–114, 2017. 37, 74

[92] Vicente García, Ramón Alberto Mollineda, and José Salvador Sánchez, “Index of balanced accuracy: A performance measure for skewed class distributions”, in *Iberian conference on pattern recognition and image analysis*. Springer, 2009, pp. 441–448. 37, 74

[93] Guillermina Vivar, Dora-Luz Almanza-Ojeda, Irene Cheng, Juan Carlos Gomez, Jose A Andrade-Lucio, and Mario-Alberto Ibarra-Manzano, “Contrast and homogeneity feature analysis for classifying tremor levels in parkinson’s disease patients”, *Sensors*, vol. 19, no. 9, pp. 2072, 2019. 37, 41

[94] Omid Bazgir, Seyed Amir Hassan Habibi, Lorenzo Palma, Paola Pierleoni, and Saba Nafees, “A classification system for assessment and home monitoring of tremor in patients with parkinson’s disease”, *Journal of medical signals and sensors*, vol. 8, no. 2, pp. 65, 2018. 37, 40

[95] George A Rigas, Alexandros T Tzallas, Dina A Baga, Themis P Exarchos, Christos D Katsis, Dimitra A Chaloglou, Spiros Th Konitsiotis, and Dimitrios I Fotiadis, “Perform: First steps in the assessment of patient motion status and support to treatment changes”, in *2009 9th International Conference on Information Technology and Applications in Biomedicine*. IEEE, 2009, pp. 1–4. 38

[96] Nathan D Darnall, Conrad K Donovan, Syeda Aktar, Han Yun Tseng, Paulo Barthelmess, Philip R Cohen, and David C Lin, “Application of machine learning and numerical analysis to classify tremor in patients affected with essential tremor or parkinson’s disease”, *Gerontechnology*, vol. 10, no. 4, pp. 208–219, 2012. 38

[97] Bryan T Cole, Serge H Roy, Carlo J De Luca, and S Hamid Nawab, “Dynamical learning and tracking of tremor and dyskinesia from wearable

sensors", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 5, pp. 982–991, 2014. 38, 102

[98] Alexandros T Tzallas, Markos G Tsipouras, Georgios Rigas, Dimitrios G Tsalikakis, Evaggelos C Karvounis, Maria Chondrogiorgi, Fotis Psomadellis, Jorge Cancela, Matteo Pastorino, María Teresa Arredondo Waldmeyer, et al., "Perform: a system for monitoring, assessment and management of patients with parkinson's disease", *Sensors*, vol. 14, no. 11, pp. 21329–21357, 2014. 38

[99] Di Pan, Rohit Dhall, Abraham Lieberman, and Diana B Petitti, "A mobile cloud-based parkinson's disease assessment system for home-based monitoring", *JMIR mHealth and uHealth*, vol. 3, no. 1, pp. e29, 2015. 39

[100] Omid Bazgir, Javad Frounchi, Seyed Amir Hassan Habibi, Lorenzo Palma, and Paola Pierleoni, "A neural network system for diagnosis and assessment of tremor in parkinson disease patients", in *2015 22nd Iranian Conference on Biomedical Engineering (ICBME)*. IEEE, 2015, pp. 1–5. 39, 102

[101] N Kostikis, Dimitris Hristu-Varsakelis, M Arnaoutoglou, and C Kotsavasiloglou, "A smartphone-based tool for assessing parkinsonian hand tremor", *IEEE journal of biomedical and health informatics*, vol. 19, no. 6, pp. 1835–1842, 2015. 39

[102] Houde Dai, Pengyue Zhang, and Tim C Lueth, "Quantitative assessment of parkinsonian tremor based on an inertial measurement unit", *Sensors*, vol. 15, no. 10, pp. 25055–25071, 2015. 39, 43

[103] Shivanthan AC Yohanandan, Mary Jones, Richard Peppard, Joy L Tan, Hugh J McDermott, and Thushara Perera, "Evaluating machine learning algorithms estimating tremor severity ratings on the bain–findley scale", *Measurement Science and Technology*, vol. 27, no. 12, pp. 125702, 2016. 39

[104] Gabriel Lugo, Mario Ibarra-Manzano, Fang Ba, and Irene Cheng, "Virtual reality and hand tracking system as a medical tool to evaluate patients with parkinson's", in *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*, 2017, pp. 405–408. 40

[105] Sara Soltaninejad, Andres Rosales-Castellanos, Fang Ba, Mario Alberto Ibarra-Manzano, and Irene Cheng, "Body movement monitoring for parkinson's disease patients using a smart sensor based non-invasive technique", in *2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom)*. IEEE, 2018, pp. 1–6. 40

[106] Guoen Cai, Zhirong Lin, Houde Dai, Xuke Xia, Yongsheng Xiong, Shi-Jinn Horng, and Tim C Lueth, "Quantitative assessment of parkinsonian tremor based on a linear acceleration extraction algorithm", *Biomedical Signal Processing and Control*, vol. 42, pp. 53–62, 2018. 41

[107] Han Byul Kim, Woong Woo Lee, Aryun Kim, Hong Ji Lee, Hye Young Park, Hyo Seon Jeon, Sang Kyong Kim, Beomseok Jeon, and Kwang S Park, "Wrist sensor-based tremor severity quantification in parkinson's disease using convolutional neural network", *Computers in biology and medicine*, vol. 95, pp. 140–146, 2018. 41

[108] Roberto López-Blanco, Miguel A Velasco, Antonio Méndez-Guerrero, Juan Pablo Romero, María Dolores Del Castillo, J Ignacio Serrano, Eduardo Rocon, and Julián Benito-León, "Smartwatch for the analysis of rest tremor in patients with parkinson's disease", *Journal of the neurological sciences*, vol. 401, pp. 37–42, 2019. 41

[109] Luis Sigcha, Ignacio Pavón, Nélson Costa, Susana Costa, Miguel Gago, Pedro Arezes, Juan Manuel López, and Guillermo De Arcas, "Automatic resting tremor assessment in parkinson's disease using smartwatches and multitask convolutional neural networks", *Sensors*, vol. 21, no. 1, pp. 291, 2021. 41

[110] N Kostikis, Dimitrios Hristu-Varsakelis, M Arnaoutoglou, and C Kotsavasiloglou, "Smartphone-based evaluation of parkinsonian hand tremor: Quantitative measurements vs clinical assessment scores", in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2014, pp. 906–909. 42, 43, 102, 118

[111] Adriano de Oliveira Andrade, Ana Paula Sousa Paixão, Ariana Moura Cabral, Amanda Gomes Rabelo, Luiza Maire David Luiz, Valdeci Carlos

Dionísio, Marcus Fraga Vieira, Janser Moura Pereira, Alice Rueda, Sridhar Krishnan, et al., "Task-specific tremor quantification in a clinical setting for parkinson's disease", *Journal of Medical and Biological Engineering*, vol. 40, no. 6, pp. 821–850, 2020. 42

[112] Jorrit I Hoff, Erik A Wagemans, and Bob J van Hilten, "Ambulatory objective assessment of tremor in parkinson's disease", *Clinical neuropharmacology*, vol. 24, no. 5, pp. 280–283, 2001. 42, 43, 118

[113] Enrica Papi, Athina Belsi, and Alison H McGregor, "A knee monitoring device and the preferences of patients living with osteoarthritis: a qualitative study", *BMJ open*, vol. 5, no. 9, 2015. 43, 45, 94

[114] Jeroen HM Bergmann, Vikesh Chandaria, and Alison McGregor, "Wearable and implantable sensors: the patient's perspective", *Sensors*, vol. 12, no. 12, pp. 16695–16709, 2012. 43, 45

[115] Marc Steen, Lottie Kuijt-Evers, and Jente Klok, "Early user involvement in research and design projects–a review of methods and practices", in *23rd EGOS Colloquium*, 2007, vol. 5, pp. 1–21. 44

[116] James M Fisher, Nils Y Hammerla, Lynn Rochester, Peter Andras, and Richard W Walker, "Body-worn sensors in parkinson's disease: Evaluating their acceptability to patients", *Telemedicine and e-Health*, vol. 22, no. 1, pp. 63–69, 2016. 44, 87, 90, 95

[117] Kathryn Mercer, Lora Giangregorio, Eric Schneider, Parmit Chilana, Melissa Li, and Kelly Grindrod, "Acceptance of commercially available wearable activity trackers among adults aged over 50 and with chronic illness: a mixed-methods evaluation", *JMIR mHealth and uHealth*, vol. 4, no. 1, pp. e7, 2016.

[118] Anneli Ozanne, D Johansson, Ulla Hällgren Graneheim, K Malmgren, F Bergquist, and M Alt Murphy, "Wearables in epilepsy and parkinson's disease—a focus group study", *Acta Neurologica Scandinavica*, vol. 137, no. 2, pp. 188–194, 2018. 45, 87, 90, 94, 95

[119] Anthony Santiago, James W Langston, Rita Gandhy, Rohit Dhall, Salima Brillman, Linda Rees, and Carrolee Barlow, "Qualitative evaluation of the

personal kinetigraph tm movement recording system in a parkinson's clinic", *Journal of Parkinson's disease*, vol. 9, no. 1, pp. 207–219, 2019. 44, 95

[120] Friederike JS Thilo, Sabine Hahn, Ruud JG Halfens, and Jos MGA Schols, "Usability of a wearable fall detection prototype from the perspective of older people–a real field testing approach", *Journal of clinical nursing*, vol. 28, no. 1-2, pp. 310–320, 2019. 44, 45, 87, 90, 93

[121] Elisa Bruno, Sara Simblett, Alexandra Lang, Andrea Biondi, Clarissa Odoi, Andreas Schulze-Bonhage, Til Wykes, Mark P Richardson, RADAR-CNS Consortium, et al., "Wearable technology in epilepsy: The views of patients, caregivers, and healthcare professionals", *Epilepsy & Behavior*, vol. 85, pp. 141–149, 2018. 45, 94

[122] Ben Beiske, *Research methods. Uses and limitations of questionnaires, interviews, and case studies*, GRIN Verlag, Munich, Germany, 2007. 45

[123] Mathers, N. and Fox, N. and Hunn,A., "Surveys and questionnaires. the nihr research design service for the east midlands/yorkshire & the humber", 2007. 45

[124] Virginia Braun and Victoria Clarke, *Successful qualitative research: A practical guide for beginners*, sage, 2013. 45, 82, 83

[125] Dax Steins, Helen Dawes, Patrick Esser, and Johnny Collett, "Wearable accelerometry-based technology capable of assessing functional activities in neurological populations in community settings: a systematic review", *Journal of neuroengineering and rehabilitation*, vol. 11, no. 1, pp. 1–13, 2014. 48

[126] Justin J Kavanagh and Hylton B Menz, "Accelerometry: a technique for quantifying movement patterns during walking", *Gait & posture*, vol. 28, no. 1, pp. 1–15, 2008. 48

[127] ACRMDOG Godfrey, Richard Conway, David Meagher, and Gearoid ÓLaighin, "Direct measurement of human movement by accelerometry", *Medical engineering & physics*, vol. 30, no. 10, pp. 1364–1386, 2008. 48, 49

[128] Matej Andrejašic, "Mems accelerometers", in *University of Ljubljana. Faculty for mathematics and physics, Department of physics, Seminar*, 2008. 49

[129] AK Bourke, JV O'brien, and GM Lyons, "Evaluation of a threshold-based tri-axial accelerometer fall detection algorithm", *Gait & posture*, vol. 26, no. 2, pp. 194–199, 2007. 49

[130] Michael J. Fox Foundation, "Data Sets: MJFF levodopa response study. https://www.michaeljfox.org/data-sets", 2019, (accessed September 16, 2020). 49

[131] Friedrich Foerster and Jochen Fahrenberg, "Motion pattern and posture: correctly assessed by calibrated accelerometers", *Behavior research methods, instruments, & computers*, vol. 32, no. 3, pp. 450–457, 2000. 51

[132] Stephen J Preece, John Yannis Goulermas, Laurence PJ Kenney, and David Howard, "A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data", *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 3, pp. 871–879, 2008. 51, 100

[133] Mehmed Kantardzic, *Data mining: concepts, models, methods, and algorithms*, John Wiley & Sons, 2011. 52, 70, 104, 123

[134] Gary M Weiss, "Mining with rarity: a unifying framework", *ACM Sigkdd Explorations Newsletter*, vol. 6, no. 1, pp. 7–19, 2004. 58, 98

[135] Bing Zhu, Bart Baesens, Aimée Backiel, and Seppe KLM Vanden Broucke, "Benchmarking sampling techniques for imbalance learning in churn prediction", *Journal of the Operational Research Society*, vol. 69, no. 1, pp. 49–65, 2018. 58, 59

[136] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer, "Smote: synthetic minority over-sampling technique", *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002. 59

[137] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning", in *2008 IEEE*

international joint conference on neural networks (IEEE world congress on computational intelligence). IEEE, 2008, pp. 1322–1328. 59

[138] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao, "Borderline-smote: a new over-sampling method in imbalanced data sets learning", in *International conference on intelligent computing*. Springer, 2005, pp. 878–887. 59

[139] Peter Hart, "The condensed nearest neighbor rule (corresp.)", *IEEE transactions on information theory*, vol. 14, no. 3, pp. 515–516, 1968. 61

[140] Ivan Tomek et al., "Two modifications of cnn", *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-6, no. 11, pp. 769–772, 1976. 61

[141] Ivan Tomek et al., "An experiment with the edited nearest-neighbor rule", *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-6, no. 6, pp. 448–452, 1976. 61

[142] Dennis L Wilson, "Asymptotic properties of nearest neighbor rules using edited data", *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-2, no. 3, pp. 408–421, 1972. 61

[143] Michael R Smith, Tony Martinez, and Christophe Giraud-Carrier, "An instance level analysis of data complexity", *Machine learning*, vol. 95, no. 2, pp. 225–256, 2014. 62

[144] Inderjeet Mani and I Zhang, "knn approach to unbalanced data distributions: a case study involving information extraction", in *Proceedings of workshop on learning from imbalanced datasets*, 2003, vol. 126. 62

[145] Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard, "A study of the behavior of several methods for balancing machine learning training data", *ACM SIGKDD explorations newsletter*, vol. 6, no. 1, pp. 20–29, 2004. 62

[146] Gustavo EAPA Batista, Ana LC Bazzan, Maria Carolina Monard, et al., "Balancing training data for automated annotation of keywords: a case study", in *WOB*, 2003, pp. 10–18. 62

[147] Wei-Yin Loh, “Classification and regression trees”, *Wiley interdisciplinary reviews: data mining and knowledge discovery*, vol. 1, no. 1, pp. 14–23, 2011. 63, 123

[148] Leo Breiman, “Random forests”, *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001. 64, 104, 123

[149] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, *An introduction to statistical learning*, vol. 112, Springer, 2013. 64

[150] Corinna Cortes and Vladimir Vapnik, “Support-vector networks”, *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995. 64, 123

[151] Arti Patle and Deepak Singh Chouhan, “Svm kernel functions for classification”, in *2013 International Conference on Advances in Technology and Engineering (ICATE)*. IEEE, 2013, pp. 1–9. 65

[152] Stephan Dreiseitl and Lucila Ohno-Machado, “Logistic regression and artificial neural network classification models: a methodology review”, *Journal of biomedical informatics*, vol. 35, no. 5-6, pp. 352–359, 2002. 66, 123

[153] Vladislav V Levshinskii, “Multiclass classification in the problem of differential diagnosis of venous diseases based on microwave radiometry data”, *Program Systems: Theory and Applications*, 2021. 66

[154] Jianke Yang, “Newton-conjugate-gradient methods for solitary wave computations”, *Journal of Computational Physics*, vol. 228, no. 18, pp. 7007–7024, 2009. 67

[155] Dong C Liu and Jorge Nocedal, “On the limited memory bfgs method for large scale optimization”, *Mathematical programming*, vol. 45, no. 1, pp. 503–528, 1989. 67

[156] Mark Schmidt, Nicolas Le Roux, and Francis Bach, “Minimizing finite sums with the stochastic average gradient”, *Mathematical Programming*, vol. 162, no. 1-2, pp. 83–112, 2017. 67

[157] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien, "Saga: A fast incremental gradient method with support for non-strongly convex composite objectives", *arXiv preprint arXiv:1407.0202*, 2014. 67

[158] Joseph O Ogutu, Torben Schulz-Streeck, and Hans-Peter Piepho, "Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions", in *BMC proceedings*. Springer, 2012, vol. 6, pp. 1–6. 68

[159] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas, "Supervised machine learning: A review of classification techniques", *Emerging artificial intelligence applications in computer engineering*, vol. 160, no. 1, pp. 3–24, 2007. 68, 123

[160] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks", in *Advances in neural information processing systems*, 2012, pp. 1097–1105. 70, 72, 104, 128

[161] John S Bridle, "Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition", in *Neurocomputing*, pp. 227–236. Springer, 1990. 70, 73, 104

[162] Henry Leung and Simon Haykin, "The complex backpropagation algorithm", *IEEE Transactions on signal processing*, vol. 39, no. 9, pp. 2101–2104, 1991. 70

[163] Sebastian Ruder, "An overview of gradient descent optimization algorithms", *arXiv preprint arXiv:1609.04747*, 2016. 70, 71

[164] Ning Qian, "On the momentum term in gradient descent learning algorithms", *Neural networks*, vol. 12, no. 1, pp. 145–151, 1999. 71

[165] Jeffrey Dean, Greg S Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Quoc V Le, Mark Z Mao, Marc'Aurelio Ranzato, Andrew Senior, Paul Tucker, et al., "Large scale distributed deep networks", in *Neural Information Processing Systems*, 2012. 71

[166] Matthew D Zeiler, "Adadelta: an adaptive learning rate method", *arXiv preprint arXiv:1212.5701*, 2012. 71

[167] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization", *arXiv preprint arXiv:1412.6980*, 2014. 72

[168] Timothy Dozat, "Incorporating nesterov momentum into adam", in *International Conference on Learning Representations*, 2016. 72

[169] Xavier Glorot, Antoine Bordes, and Yoshua Bengio, "Deep sparse rectifier neural networks", in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 2011, pp. 315–323. 73, 104

[170] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)", *arXiv preprint arXiv:1511.07289*, 2015. 73

[171] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter, "Self-normalizing neural networks", *arXiv preprint arXiv:1706.02515*, 2017. 73

[172] Ricardo Barandela, José Salvador Sánchez, V Garca, and Edgar Rangel, "Strategies for learning in class imbalance problems", *Pattern Recognition*, vol. 36, no. 3, pp. 849–851, 2003. 74

[173] Karimollah Hajian-Tilaki, "Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation", *Caspian journal of internal medicine*, vol. 4, no. 2, pp. 627, 2013. 75

[174] John W Creswell, Michael D Fetters, and Nataliya V Ivankova, "Designing a mixed methods study in primary care", *The Annals of Family Medicine*, vol. 2, no. 1, pp. 7–12, 2004. 77

[175] Paul A Nutting, Kathryn Rost, Miriam Dickinson, James J Werner, Perry Dickinson, Jeffrey L Smith, and Beth Gallovic, "Barriers to initiating depression treatment in primary care practice", *Journal of General Internal Medicine*, vol. 17, no. 2, pp. 103–111, 2002. 77

[176] Vaishali Hegde, “Role of human factors/usability engineering in medical device design”, in *2013 Proceedings Annual Reliability and Maintainability Symposium (RAMS)*. IEEE, 2013, pp. 1–5. 78

[177] Dilanthi Amaratunga, David Baldry, Marjan Sarshar, and Rita Newton, “Quantitative and qualitative research in the built environment: application of “mixed” research approach”, *Work study*, 2002. 78, 80

[178] World Medical Association, “Medical ethics manual. https://www.wma.net/wp-content/uploads/2016/11/Ethics_manual_3rd_Nov2015_en.pdf”, 2015, (accessed May 3, 2021). 78

[179] Tom L Beauchamp, James F Childress, et al., *Principles of biomedical ethics*, Oxford University Press, USA, 2001. 78

[180] John R Williams, “The declaration of helsinki and public health”, *Bulletin of the World Health Organization*, vol. 86, pp. 650–652, 2008. 79

[181] Paul Gill, Kate Stewart, Elizabeth Treasure, and Barbara Chadwick, “Methods of data collection in qualitative research: interviews and focus groups”, *British dental journal*, vol. 204, no. 6, pp. 291–295, 2008. 80

[182] R Legard, J Keegan, and K Ward, “In-depth interviews in: Qualitative research practice: A guide for social science students and researchers”, *Eds: Jane Ritchie and Jane Lewis*, 2003. 80

[183] Ann Bowling, *Research methods in health: investigating health and health services*, McGraw-hill education (UK), 2014. 80

[184] Alan Bryman, *Social research methods*, Oxford university press, 2016. 81

[185] Monique M Hennink, *Focus group discussions*, Oxford University Press, 2013. 81

[186] Tobias O. Nyumba, Kerrie Wilson, Christina J Derrick, and Nibedita Mukherjee, “The use of focus group discussion methodology: Insights from two decades of application in conservation”, *Methods in Ecology and evolution*, vol. 9, no. 1, pp. 20–32, 2018. 81

[187] Karen L Then, James A Rankin, and Elena Ali, "Focus group research: what is it and how can it be used?", *Canadian journal of cardiovascular nursing*, vol. 24, no. 1, 2014. 81

[188] Sylvie D Lambert and Carmen G Loiselle, "Combining individual interviews and focus groups to enhance data richness", *Journal of advanced nursing*, vol. 62, no. 2, pp. 228–237, 2008. 81, 175, 180

[189] Kurt A Jellinger, Giancarlo Logroscino, Giovanni Rizzo, Massimiliano Copetti, Simona Arcuti, Davide Martino, and Andrea Fontana, "Accuracy of clinical diagnosis of parkinson disease: A systematic review and meta-analysisauthor response", *Neurology*, vol. 87, no. 2, pp. 237–238, 2016. 84

[190] Scott A Small, "Age-related memory decline: current concepts and future directions", *Archives of neurology*, vol. 58, no. 3, pp. 360–364, 2001. 85

[191] Martina Amanzio, Silvia Monteverdi, Alessandra Giordano, Paola Soliveri, Paola Filippi, and Giuliano Geminiani, "Impaired awareness of movement disorders in parkinson's disease", *Brain and Cognition*, vol. 72, no. 3, pp. 337–346, 2010. 85

[192] Lars Tore Vassli and Babak A Farshchian, "Acceptance of health-related ict among elderly people living in the community: A systematic review of qualitative evidence", *International Journal of Human–Computer Interaction*, vol. 34, no. 2, pp. 99–116, 2018. 85

[193] Mehmet Gövercin, Y Költzsch, M Meis, S Wegel, M Gietzelt, J Spehr, S Winkelbach, M Marschollek, and E Steinhagen-Thiessen, "Defining the user requirements for wearable and optical fall prediction and fall detection devices for home use", *Informatics for health and social care*, vol. 35, no. 3-4, pp. 177–187, 2010. 88, 89

[194] Halley P Profita, James Clawson, Scott Gilliland, Clint Zeagler, Thad Starner, Jim Budd, and Ellen Yi-Luen Do, "Don't mind me touching my wrist: a case study of interacting with on-body technology in public", in *Proceedings of the 2013 International Symposium on Wearable Computers*, 2013, pp. 89–96. 88, 89

[195] United Nations, ". 89

[196] Gina S Charlton and Corinne J Barrow, "Coping and self-help group membership in parkinson's disease: an exploratory qualitative study", *Health & social care in the community*, vol. 10, no. 6, pp. 472–478, 2002. 90

[197] Marianne Caap-Ahlgren, Lena Lannerheim PhD, and MD Ove Dehlin, "Older swedish women's experiences of living with symptoms related to parkinson's disease", *Journal of advanced nursing*, vol. 39, no. 1, pp. 87–95, 2002. 90

[198] Björg Thordardottir, Maria H Nilsson, Susanne Iwarsson, and Maria Haak, ""you plan, but you never know"–participation among people with different levels of severity of parkinson's disease", *Disability and rehabilitation*, vol. 36, no. 26, pp. 2216–2224, 2014. 90

[199] Majd Alwan, Devon Wiley, and Jeremy Nobel, "State of technology in aging services", *Center for Aging Services Technology (CAST)*, 2007. 92

[200] Martina Mancini, Mahmoud El-Gohary, Sean Pearson, James McNames, Heather Schlueter, John G Nutt, Laurie A King, and Fay B Horak, "Continuous monitoring of turning in parkinson's disease: rehabilitation potential", *NeuroRehabilitation*, vol. 37, no. 1, pp. 3–10, 2015. 92

[201] Jenna E Thorp, Peter Gabriel Adamczyk, Heidi-Lynn Ploeg, and Kristen A Pickett, "Monitoring motor symptoms during activities of daily living in individuals with parkinson's disease", *Frontiers in neurology*, vol. 9, pp. 1036, 2018. 101

[202] Murtadha D Hssayeni, Michelle A Burack, Joohi Jimenez-Shahed, and Behnaz Ghoraani, "Assessment of response to medication in individuals with parkinson's disease", *Medical engineering & physics*, vol. 67, pp. 33–43, 2019. 101, 102

[203] MSA Megat Ali, MN Taib, N Md Tahir, and AH Jahidin, "Eeg spectral centroid amplitude and band power features: A correlation analysis", in *2014 IEEE 5th Control and System Graduate Research Colloquium*. IEEE, 2014, pp. 223–226. 102

[204] A Yu Meigal, SM Rissanen, MP Tarvainen, SD Georgiadis, PA Karjalainen, O Airaksinen, and M Kankaanpää, "Linear and nonlinear tremor acceleration characteristics in patients with parkinson's disease", *Physiological measurement*, vol. 33, no. 3, pp. 395, 2012. 102

[205] Gustavo EAPA Batista, Eamonn J Keogh, Oben Moses Tataw, and Vinicius MA De Souza, "Cid: an efficient complexity-invariant distance for time series", *Data Mining and Knowledge Discovery*, vol. 28, no. 3, pp. 634–669, 2014. 102

[206] Farrokh Manzouri, Simon Heller, Matthias Dümpelmann, Peter Woias, and Andreas Schulze-Bonhage, "A comparison of machine learning classifiers for energy-efficient implementation of seizure detection", *Frontiers in systems neuroscience*, vol. 12, pp. 43, 2018. 104

[207] Ramin Ghorbani and Rouzbeh Ghousi, "Comparing different resampling methods in predicting students' performance using machine learning techniques", *IEEE Access*, vol. 8, pp. 67899–67911, 2020. 104

[208] François Chollet et al., "Keras: The python deep learning library", *Astrophysics Source Code Library*, 2018. 104

[209] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems", *arXiv preprint arXiv:1603.04467*, 2016. 104

[210] Thais Mayumi Oshiro, Pedro Santoro Perez, and José Augusto Baranauskas, "How many trees in a random forest?", in *International workshop on machine learning and data mining in pattern recognition*. Springer, 2012, pp. 154–168. 104, 128

[211] Richard A Berk, "Classification and regression trees (cart)", in *Statistical learning from a regression perspective*, pp. 1–65. Springer, 2008. 104, 128

[212] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas, "Taking the human out of the loop: A review of bayesian

optimization", *Proceedings of the IEEE*, vol. 104, no. 1, pp. 148–175, 2015. 123

[213] Tim Head, Gilles Louppe MechCoder, Iaroslav Shcherbatyi, et al., "scikit-optimize. https://scikit-optimize.github.io/stable/", 2018, (accessed September 16, 2020). 123

Appendix A

This appendix presents the healthcare professionals' interview topic guide that is used to conduct the preliminary semi-structured individual interviews with healthcare professionals according to the procedure used by Lambert and Loiselle [188]. This topic guide discuss three main ideas: PD diagnosis, current PD monitoring and assessment methods and wearable technology.

Healthcare Professional Interview Topic Guide

Aims:

1. To identify perspectives of healthcare professionals about current methods of diagnosing and monitoring of Parkinson's disease.

2. To identify if healthcare professionals prefer monitored to be done at the clinic or home environment.

3. To identify if healthcare professionals prefer wearable or non-wearable devices.

Guidance notes:

The participant will be reminded they can terminate their participation at any time, without giving a reason.

Introduction:

1. **The interviewer introduce himself.**

   Thank you for giving your time to this research study. My name is

Ghayth AlMahadin and I am a PhD student in the department of Science and Technology at Nottingham Trent University, researching ways to best measure Parkinson’s symptoms.

2. **Explain the purpose of study.**

   In this interview, I will be asking you about your views on current ways to assess and measure tremor in people with Parkinson’s disease, and to explore new ways of assessment.

3. **Remind the interviewee of the participant information sheet he/she has received and ensure that they have read, understood.**

   - I expect the interview to last about 30-45 minutes, is that ok?
   - The interview will be recorded and then I will transcribe it to use as data for my research.
   - If there are any questions you do not want to answer then that is ok.
   - You can stop the interview anytime if you wish.
   - The questions will be flexible and open-ended to allow you the chance to raise the issues or bring up ideas that you feel are important.
   - Your responses will be anonymised in any findings we publish.
   - There are no ‘right’ or ‘wrong’ answers. We are interested in knowing your opinion.
   - Is there anything you would like to ask me before we begin? **(The interviewer will answer all questions)**.

4. **Ensure that he/she signed the interview consent form.**

5. **I am now going to start the recording.**

First, I will ask you about the diagnosis and monitoring processes. I would then like to get your opinion about using technology to assist with diagnosis and monitoring of Parkinson’s tremor.

Parkinson’s disease diagnosis

1. Are you trained for PD diagnostic? When did you complete your training?
2. How many years of experience do you have in diagnosis of PD?
3. How often are you involved in providing the first diagnosis of PD?
4. Can you explain the current methods or process you are using for PD diagnosis? How long does the assessment take?
5. How many healthcare professionals are typically involved in the diagnosis process at your practice?
6. What do you think about these methods? Advantages and disadvantages **(probe a few times for disadvantages: "Any others? ")**? What would make it better?
7. How accurate are these methods, in your opinion? Why?
8. Do you think these methods are subjective or objective? Why?
9. In your opinion, what alternatives can be offered for diagnosis? A few years in the future, how would you like to see PD being diagnosed and monitored?
10. What is the most noticeable symptom of PD in your opinion?
11. What is the most common initial symptom in your opinion?

Current monitoring approach

1. Can you describe how PD patients are currently monitored? How often? Where? How long does each monitoring session take?
2. How useful do you find current methods? Why?
3. In your opinion, do you think these sessions are enough? Should there be more or fewer?
4. What is your opinion on the importance of monitoring PD symptoms? Is it related to patients' treatment, and if so, how?

5. What do you think about the current monitoring process? How could it be made it better?

6. Can you think of any alternatives that could be developed in the future to assess and/or monitor PD?

7. If there was technology that could easily be used for assisting monitoring and/or diagnosis, what would be your opinion on health care professionals' adoption of this?

Wearable technology

The interviewer will explains wearable technology and answer any questions.

1. How would you feel about monitoring patients' conditions at home using a wearable device? Would you be interested in this kind of technology? Would you use it? Why or why not?

2. If you could monitor patients' tremors using a wearable device, how do you think the device should look and feel? why?

3. Are there any alternative options to these wearable devices in your opinion? What are they?

4. What part of body would be best to wear the device for tremor measurement, and why?

5. Do you have any concerns about such device? Explain please

6. Do you think it would better to use the technology for monitoring or diagnosis? Why?

7. What is your opinion on the device collecting data all day and night 24/7? If not why? If yes, what do you think about sending data to the clinic over internet? **If they have any concerns (e.g. Security of data), probe for that.**

8. What type of data or information could help you to assess the tremors?

9. Would this kind of technology interest you? Would you use it? Why?
10. What improvements could be made to the device that could make it more useful?
11. Do you think the patients will engage with this technology and wear the device? Why or why not? (you may need to explain the level of engagement that would be needed by the patient)

Closing

Is there anything else you would like to say about what we have discussed? Do you have any questions?

Thank you for your time and useful participation

Appendix B

This appendix presents patients' interview topic guide that is used to collect data from PD patients according to the procedure used by Lambert and Loiselle [188]. This topic guide discuss three main ideas: PD diagnosis, current PD monitoring and assessment methods and wearable technology.

Patients Focus Group Topic Guide

Aims:

1. To identify perspectives of patients with current methods to diagnosis and monitor of Parkinson's disease.
2. To identify if patients prefer to be monitored within clinic or at home.
3. To identify if patients prefer wearable or non-wearable devices.

Guidance notes:

The participant will be reminded they can terminate their participation at any time, without giving a reason.

Introduction:

1. **The interviewer introduce himself.**
   Thank you for giving your time to this research study. My name is **Ghayth AlMahadin** and I am a PhD student in the department of Science and Technology at Nottingham Trent University, researching ways to best measure Parkinson's symptoms.

2. **Explain the purpose of study.**
   In this group we will discuss your views on current ways to assess and measure tremor in people with Parkinson's disease, and to explore new ways of assessment.

3. **Ensure that all participants have received, read and understood the participant information sheet.**

   - I expect the discussion to last about 100 minutes, and **you can leave at any time with no reason given**. is that ok?
   - The discussions will be recorded and then I will transcribe it to use as data for my research.
   - You can withdraw from the discussion anytime if you wish.
   - The questions will be flexible and open-ended to allow you the chance to raise the issues or bring up ideas that you feel are important.
   - Your responses will be anonymised in any findings we publish.
   - There are no 'right' or 'wrong' answers. We are interested in knowing your opinion.
   - Is there anything you would like to ask me before we begin? (**The interviewer will answer all questions**).

4. **Ensure that all participants signed the consent form.**

5. **I am now going to start the recording.**

First, I will ask you about the diagnosis and monitoring processes. I would then like to get your opinion about using technology to assist with diagnosis and monitoring of Parkinson's tremor.

Parkinson's disease diagnosis

1. What was the first symptom of PD you noticed?

2. Can you describe the diagnosis process and your experience of it?

3. What do you think about diagnosis process? What would make it better?
4. In your opinion, what alternatives could or should be offered for diagnosis?

Current monitoring approach

1. Can you describe how is your condition is currently monitored? How often? Where?
2. Would you like to attend more or less often than you do now? Why? Are there any issue you face when you attend these sessions?
3. What do you think about the current monitoring process? How could it be made better?
4. What alternatives would you like to see in the future? why might these be useful?
5. If there was technology that could easily be used for assisting monitoring and/or diagnosis, what would be your opinion on health care professionals' adoption of this?

Wearable technology

The researcher will explain wearable technology and answer any questions.

1. How would you feel about monitoring your conditions at home using a wearable device? would you be interested in this kind of technology? Would you use it? Why/why not?
2. If the doctors can monitor your condition using a wearable device, how do you think the device should look and feel? why?
3. What part of body will you prefer to wear the device (wrist, hand, arm) and why?
4. Are there any alternative options to wearable devices in your opinion? What are they?

5. Do you have any concerns about device visibility?

6. How long do you think you would be willing to wear such a device, and why? What if the device looks and feels like a watch?

7. What is your opinion on the device collecting data all day and night 24/7? If not, why? If yes, what do you think about sending data to the clinic over internet? **If they have any concerns (e.g. Security of data).**

Closing

Is there anything else you would like to say about what we have discussed? Do you have any questions?

Thank you for your time and useful participation

Appendix C

This appendix presents the best results of metrics (AUC, F1-Score, G-Mean, and IBA) that has been used to identify the best task to measure tremor severity. The highlighted values are above-average among each dataset, while the count above-average column shows values that above-average for datasets for each task. The total count of above-average of all metrics is utilised to identify the best task, i.e. all metrics contributes to identify the best tasks that could be used to collect tremor data.

Table C.1: Tasks highest AUC of all classifiers and values above average counts.

	AUC								
	Without Re-sampling				**With -Re-sampling**				**Count Above Average**
	G-1	**G-4**	**P-1**	**P-4**	**G-1**	**G-4**	**P-1**	**P-4**	
drawg	84%	76%	96%	71%	99%	98%	99%	100%	**2**
drnkg	84%	85%	89%	92%	99%	99%	100%	100%	**3**
fldng	92%	89%	91%	93%	100%	99%	99%	100%	**4**
ftnl	93%	93%	83%	79%	100%	100%	99%	99%	**4**
ftnr	74%	91%	92%	96%	97%	100%	100%	100%	**6**
ntblt	88%	83%	87%	60%	99%	99%	100%	99%	**1**
orgpa	90%	92%	90%	90%	100%	100%	100%	100%	**6**
raml	95%	93%	90%	89%	100%	100%	100%	100%	**6**
ramr	89%	88%	96%	95%	100%	99%	100%	100%	**5**
typng	93%	91%	91%	86%	100%	99%	100%	100%	**5**
sittg	94%	93%	96%	81%	100%	100%	100%	100%	**7**
stndg	93%	88%	92%	92%	100%	100%	100%	100%	**7**
strsd	99%	94%	98%	97%	100%	100%	100%	100%	**8**
strsu	91%	97%	97%	100%	100%	100%	100%	100%	**8**
ststd	96%	90%	66%	94%	100%	100%	100%	100%	**7**
wlkgc	95%	92%	98%	96%	100%	100%	100%	100%	**8**
wlkgp	89%	90%	96%	95%	100%	100%	100%	100%	**7**
wlkgs	94%	93%	97%	98%	100%	100%	100%	100%	**8**
Average	**91%**	**90%**	**91%**	**89%**	**100%**	**100%**	**100%**	**100%**	

G-1 : GENEActiv - Day1, G-4 : GENEActiv - Day4, P-1 : Pebble - Day1, P-4 : Pebble - Day4

Table C.2: Task highest F1-score of all classifiers and values above average counts.

	F1-Score								
	Without Re-sampling				With Re-sampling				Count Above Average
	G-1	G-4	P-1	P-4	G-1	G-4	P-1	P-4	
drawg	65%	53%	83%	92%	93%	91%	95%	99%	**3**
drnkg	61%	56%	69%	74%	93%	93%	96%	97%	**0**
fldng	65%	55%	71%	73%	94%	91%	95%	96%	**0**
ftnl	71%	68%	64%	61%	97%	96%	95%	96%	**3**
ftnr	52%	68%	72%	83%	90%	98%	97%	99%	**4**
ntblt	62%	57%	63%	68%	95%	94%	95%	96%	**0**
orgpa	60%	69%	64%	68%	96%	98%	96%	97%	**2**
raml	69%	75%	64%	55%	96%	97%	98%	94%	**3**
ramr	64%	58%	78%	80%	96%	91%	98%	99%	**4**
typng	71%	64%	72%	65%	96%	93%	97%	96%	**1**
sittg	76%	71%	84%	92%	100%	98%	98%	99%	**8**
stndg	69%	58%	72%	76%	100%	98%	99%	97%	**3**
strsd	92%	78%	86%	88%	100%	100%	100%	100%	**8**
strsu	74%	86%	88%	100%	100%	100%	100%	100%	**8**
ststd	81%	72%	83%	76%	100%	99%	99%	100%	**7**
wlkgc	71%	74%	85%	79%	98%	96%	99%	98%	**7**
wlkgp	66%	67%	84%	80%	96%	97%	98%	98%	**6**
wlkgs	76%	76%	86%	85%	99%	98%	100%	99%	**8**
Average	**69%**	**67%**	**76%**	**78%**	**97%**	**96%**	**98%**	**98%**	

G-1 : GENEActiv - Day1, G-4 : GENEActiv - Day4, P-1 : Pebble - Day1, P-4 : Pebble - Day4

Table C.3: Task highest G-mean of all classifiers and values above average counts.

	G-Mean								
	Without Re-sampling				**With -Resampling**				**Count Above**
	G-1	**G-4**	**P-1**	**P-4**	**G-1**	**G-4**	**P-1**	**P-4**	**Average**
drawg	64%	54%	24%	45%	95%	92%	95%	99%	**2**
drnkg	49%	57%	58%	48%	95%	95%	97%	98%	**1**
fldng	56%	36%	56%	49%	95%	93%	96%	97%	**2**
ftnl	48%	40%	68%	68%	98%	98%	96%	97%	**4**
ftnr	62%	74%	54%	58%	92%	99%	98%	100%	**6**
ntblt	40%	47%	43%	44%	96%	95%	96%	96%	**0**
orgpa	46%	53%	62%	43%	97%	99%	97%	98%	**2**
raml	38%	58%	62%	61%	98%	98%	99%	95%	**6**
ramr	61%	66%	52%	62%	97%	94%	99%	99%	**5**
typng	37%	42%	57%	53%	97%	94%	98%	97%	**1**
sittg	71%	58%	56%	61%	100%	99%	99%	99%	**8**
stndg	68%	53%	48%	72%	100%	98%	99%	98%	**5**
strsd	64%	66%	62%	68%	100%	100%	100%	100%	**8**
strsu	46%	78%	46%	100%	100%	100%	100%	100%	**6**
ststd	35%	53%	53%	45%	100%	99%	99%	100%	**5**
wlkgc	53%	76%	52%	53%	99%	97%	100%	99%	**5**
wlkgp	50%	54%	50%	47%	97%	98%	98%	99%	**2**
wlkgs	55%	64%	47%	53%	99%	99%	100%	99%	**6**
Average	**52%**	**57%**	**53%**	**57%**	**98%**	**97%**	**98%**	**98%**	

G-1 : GENEActiv - Day1, G-4 : GENEActiv - Day4, P-1 : Pebble - Day1, P-4 : Pebble - Day4

Table C.4: Task highest IBA of all classifiers and values above average counts.

	IBA								
	Without Re-sampling				**With Re-sampling**				**Count Above Average**
	G-1	**G-4**	**P-1**	**P-4**	**G-1**	**G-4**	**P-1**	**P-4**	
drawg	43%	30%	6%	19%	89%	85%	91%	98%	**2**
drnkg	25%	33%	34%	24%	90%	91%	94%	95%	**1**
fldng	36%	14%	34%	23%	91%	86%	93%	94%	**2**
ftnl	25%	17%	46%	46%	96%	95%	92%	95%	**4**
ftnr	38%	54%	30%	34%	85%	97%	96%	99%	**5**
ntblt	15%	23%	21%	18%	92%	91%	93%	93%	**0**
orgpa	22%	30%	38%	17%	95%	97%	94%	95%	**2**
raml	15%	36%	39%	38%	95%	96%	97%	91%	**5**
ramr	37%	43%	28%	37%	94%	88%	97%	98%	**5**
typng	15%	18%	32%	29%	94%	89%	96%	94%	**1**
sittg	53%	36%	33%	39%	100%	97%	97%	99%	**8**
stndg	47%	29%	25%	53%	99%	97%	99%	96%	**5**
strsd	43%	46%	37%	48%	100%	100%	100%	100%	**8**
strsu	23%	62%	23%	100%	100%	100%	100%	100%	**6**
ststd	15%	27%	27%	22%	100%	98%	98%	100%	**4**
wlkgc	29%	57%	29%	29%	98%	95%	99%	98%	**5**
wlkgp	26%	30%	25%	24%	94%	96%	97%	97%	**3**
wlkgs	32%	42%	21%	30%	98%	98%	99%	99%	**6**
Average	**30%**	**35%**	**29%**	**35%**	**95%**	**94%**	**96%**	**97%**	

G-1 : GENEActiv - Day1, G-4 : GENEActiv - Day4, P-1 : Pebble - Day1, P-4 : Pebble - Day4