The Development of Digital Biomarkers for Multiple Sclerosis from Remote Smartphone- and Smartwatch-Based Assessments



Andrew P. Creagh

St. Cross College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2020

The Development of Digital Biomarkers for Multiple Sclerosis from Remote Smartphone- and Smartwatch-Based Assessments

Andrew P. Creagh
St. Cross College
University of Oxford
Trinity 2020

Abstract

Neurodegenerative diseases, such as multiple sclerosis (MS), follow subtle and unpredictable disease courses, patient-to-patient, and over time, ensuring that it is notoriously difficult to quantify effective therapeutic interventions and disease management techniques. The gold-standard methods to measure the impact of these diseases on daily-life rely on a combination of subjective clinician-rated scores and patient-reported outcomes (PROs), that are rife with inherent limitations and may only be assessed just once or twice a year. As a result, there is an urgent need to revolutionise clinical examination strategies to integrate methods that accurately and remotely monitor disease-related changes and deterioration, that certainly occur at-home and between clinical visits.

In recent years, the emergence of digital health technologies have opened the possibility of developing rich, continuous, and objective measures that can be administered remotely and out-of-clinic. The ubiquity of smartphone and smartwatch consumer devices, and the wealth of digital sensors they contain, offer the perfect tool to develop digital and objective disease biomarkers: *digital biomarkers*.

This thesis aimed to establish how digital patterns of MS disease could be captured remotely, and longitudinally, through smartphone- and smartwatch-based sensor data collected from the Draw-a-Shape (DaS) test, an assessment of upper-extremity (UE) function, and the Two-Minute Walk Test (2MWT), which assessed ambulation in in people with multiple sclerosis (PwMS), both as part of the FLOODLIGHT PoC study. This work demonstrates how hand-crafted features from touch-screen and inertial sensors characterised UE and ambulatory function in PwMS, to remotely classify and predict disease severity using feature-based machine learning (ML) techniques. Deep networks offer more data-driven approaches to feature learning. However, remote digital health data is often sparse and heterogeneous. To overcome these difficulties, a transfer learning (TL) framework enabled a deep convolutional neural network (DCNN) to capture a better representation of gait function in PwMS, especially for those subjects with subtle and milder symptoms. Despite these improvements, the lack of transparency of "black-box" deep networks remains a large stumbling block to the wider acceptance of deep learning for clinical applications. Ensuing work therefore aimed to visualise DCNN decisions using relevance heatmaps to explain the gait patterns that were reflective of those who were healthy versus PwMS. From these robust and interpretable models, it was demonstrated how surrogate continuous model severity outcomes could be developed to longitudinally monitor PwMS. Smartphone-based sensor outcomes successfully stratified healthy and PwMS groups, tracked the disease course of PwMS, and even demonstrated a sensitivity to capture MS relapse events.

This work, with on-going future studies, demonstrates the clinical utility of objective and transparent digital sensor-based outcomes to monitor PwMS remotely and out-of-clinic. The development of *digital biomarkers* through sensor-based outcomes may enable clinicians to understand and learn new facets of habitual MS disease, to identify disease progression earlier, and to help develop more rapid and appropriate treatment decisions for PwMS.

This thesis is dedicated to Peadar Creagh.

For if it wasn't for you,
I would never have made it this far.

I d'óige oscail do mheabhair is bailigh an fhoghlaim leat.

Acknowledgements

First, and foremost, I would like to thank my supervisor Professor Maarten De Vos. Maarten, thank you for having faith in my abilities, for your guidance and support throughout my D.Phil. I will always be appreciative and grateful towards this opportunity you gave to me.

Secondly, I would like to express my gratitude to Professor Lionel Tarassenko for his supervision, support, and shepherding my thesis through to submission.

Without Dr. Michael Lindemann, Dr. Florian Lipsmeier and Dr. Christian Gossens, this thesis would not have been possible. Michael and Christian, thank you for your support and vision for this project and collaboration. Digital biomarkers have become much bigger since we started all of this! To Florian, I want to thank you for your guidance throughout my time in Oxford and in Basel. It has been thoroughly enjoyable working and collaborating with you; you challenged me to learn, and I have learned a lot. I look forward to more beers by the Rhine sometime! A special thanks to Alf Scotland, and all those in F. Hoffmann-La Roche who I've had the pleasure of collaborating with; thanks for all the advice and expertise that you provided throughout.

It is also important to acknowledge all of the individuals who contributed data towards my thesis, and to all the people in the world affected by multiple sclerosis.

Thank you to the IBME, and all the those who have helped support me during my D.Phil. To John Prince for reading this thesis from various different parks across San Francisco. You were there from the start at Tyndale Road and you were there at the end. To Ollie, Kirubin, Navin, Stefan, Maddy and Eoin. You were not only colleagues, but all friends for life, and I will always appreciate your support.

Thank you to all my friends in Oxford, and those friends who supported me from home, whether it be Batterstown, Dunboyne, Drumcondra or, Trinity.

To my family, Peadar, Pauline and Dearbhle (not forgetting Charlotte, Hester and Felix *et al.*). Most of all, your love and undying support throughout my life is the reason I am where I am right now.

And finally, thank you from the bottom of my heart to Abbie for all your love. We've had such fun times together and we've kept each other going through the hardest of times. *Is é Binn Ghulbain an chéad sliabh eile a dhreapfaimid le chéile!*

The year 2020 did not make it easy to complete my D.Phil, but I am proud of myself for reaching the finish line.

Contents

List of Figures **xi**

List of Tables **xiv**

Recurring Abbreviations **xv**

1 Introduction **1**
1.1 Motivation 1
1.2 Research Aims 4
1.3 Thesis structure & key contributions 5

2 A Background to Multiple Sclerosis: Diagnosis and Monitoring **8**
2.1 Multiple Sclerosis 9
2.1.1 Pathology and Epidemiology 9
2.1.2 Pathophysiology 11
2.1.3 Gait and Ambulatory Function 12
2.1.4 Upper-Extremity and Hand-Motor Function 13
2.2 Assessing Multiple Sclerosis Disability 13
2.2.1 Limitations Assessing Multiple Sclerosis Disability 17
2.3 Digital Biomakers 18
2.3.1 Developing Digital Biomarkers from Digital Sensor Data . . 20
2.3.2 Remote Digital Biomarker Assessments 20
2.3.3 Digital Sensor Data to Measure Upper Extremity Function . 21
2.3.4 Digital Sensor Data to Measure Ambulatory Function 22
2.3.5 Smartphones and Smartwatches to Characterise Ambulatory Function 24
2.3.6 Using Machine Learning to Construct Digital Biomarkers . . 25
2.4 Background Summary 26

3 Essential Methodology & Technical Background **28**
3.1 Introduction 28
3.2 Building a modular, scalable and parallelizable feature extraction pipeline 29

3.3 Exploratory Data Analysis 30
3.3.1 Data Visualisation 30
3.3.2 Correlation Analysis 31
3.3.3 Test-Retest Reliability 33
3.4 Confirmatory Data Analysis 34
3.4.1 Hypothesis Testing 34
3.4.2 Parametric versus Non-parametric hypothesis testing 34
3.4.3 Correcting for multiple hypothesis testing 40
3.5 Linear Methods for Regression and Classification 41
3.5.1 Ordinary Least Squares 41
3.5.2 Iteratively Re-Weighted Least Squares 42
3.5.3 Logistic Regression and Classification 43
3.6 Regularisation & Feature Selection 45
3.6.1 ℓ_1 & ℓ_2 Regularisation 45
3.6.2 Ridge Regression 45
3.6.3 Least Absolute Shrinkage and Selection Operator 46
3.7 Non-Linear Methods for Regression and Classification 47
3.7.1 Support Vector Machines 47
3.7.2 Application of Kernels to Support Vector Machines 50
3.7.3 Support Vector Regression 51
3.7.4 Classification and Regression Trees 52
3.8 Gradient-Based Optimisation 54
3.9 Deep Learning and Neural Networks 56
3.9.1 Convolutional Neural Networks 58
3.9.2 Training a Deep Network 59
3.9.3 Improving Network Generalisability and Robustness 60
3.10 Visualising Deep Networks 61
3.10.1 Layer Wise Relevance Propagation 62
3.11 Evaluation Approaches and Metrics 64
3.11.1 Cross Validation 64
3.11.2 Class Imbalance Problems 65
3.11.3 Classification Evaluation 65
3.11.4 Regression Evaluation 67

4 Definition of Datasets **70**
4.1 Introduction 70
4.2 The FLOODLIGHT Proof-of-Concept Study 71
4.2.1 Study Material and Equipment 72
4.2.2 Study Participants 73

4.2.3 Clinical Assessment . . . 73
4.3 Study Activities . . . 76
4.3.1 The Draw-a-Shape Test . . . 76
4.3.2 The Two-Minute Walk Test . . . 77
4.4 Human Activity Recognition Datasets . . . 78
4.4.1 UCI HAR . . . 78
4.4.2 WISDM . . . 79

5 Smartphone-based Remote Characterisation of Upper Extremity Function using the Draw a Shape Test 80
5.1 Introduction . . . 80
5.2 Background . . . 82
5.2.1 The Nine-Hole Peg Test . . . 82
5.2.2 Digitally Augmenting Manual Dexterity Assessments . . . 83
5.3 Methods . . . 83
5.3.1 Dataset . . . 83
5.3.2 Feature Extraction . . . 84
5.3.3 Regression Model to 9HPT . . . 88
5.4 Results . . . 91
5.4.1 Feature Demonstration . . . 92
5.4.2 Feature Evaluation . . . 93
5.4.3 Model Evaluation . . . 95
5.5 Discussion . . . 100
5.5.1 Feature Discussion . . . 101
5.5.2 Model Discussion . . . 102
5.5.3 Limitations and Considerations . . . 104
5.6 Conclusion . . . 106

6 Smartphone- and Smartwatch-Based Remote Characterisation of Ambulation during the Two-Minute Walk Test 108
6.1 Introduction . . . 108
6.2 Methods . . . 110
6.2.1 Dataset . . . 110
6.2.2 Feature Extraction . . . 111
6.2.3 Exploratory Analysis . . . 119
6.2.4 Model Construction . . . 119
6.3 Results . . . 121
6.3.1 Feature Analysis . . . 121
6.3.2 Classification Analysis . . . 128
6.4 Discussion . . . 131

6.4.1 Feature Evaluation . . . 131
6.4.2 Classification Evaluation . . . 133
6.4.3 Limitations and Considerations . . . 135
6.5 Conclusion . . . 137

7 Deep Learning for the Remote Characterisation of Ambulation in Multiple Sclerosis using Smartphones 139
7.1 Introduction . . . 139
7.2 Background . . . 141
7.2.1 Deep Networks Applied to Inertial Sensor Data . . . 141
7.2.2 Deep Transfer Learning . . . 143
7.2.3 Leveraging Human Activity Recognition for Transfer Learning 145
7.3 Methods . . . 145
7.3.1 Pre-Processing . . . 146
7.3.2 Model Construction . . . 147
7.3.3 Deep Transfer Learning . . . 147
7.3.4 Model Parametrisation . . . 149
7.3.5 Dataset Construction and Model Evaluation . . . 149
7.4 Results . . . 151
7.4.1 HAR Evaluation . . . 151
7.4.2 HAR Evaluation on FLOODLIGHT . . . 151
7.4.3 FLOODLIGHT Evaluation . . . 153
7.5 Discussion . . . 157
7.5.1 Limitations and Considerations . . . 159
7.6 Conclusion . . . 160

8 Developing Interpretable Smartphone-based Sensor Models for Remote Subject Monitoring 162
8.1 Introduction . . . 163
8.2 Background . . . 165
8.2.1 Explaining Neural Network Decisions . . . 165
8.2.2 Layer-wise Relevance Propagation . . . 166
8.2.3 Towards Longitudinal Remote Subject Monitoring . . . 168
8.3 Methods . . . 169
8.3.1 Developing Interpretable Smartphone Sensor Models . . . 169
8.3.2 Estimating Disease Severity . . . 170
8.3.3 Longitudinal Trend Monitoring of Remote Smartphone-Based Outcomes . . . 171
8.4 Results . . . 173
8.4.1 Visualising Remote Sensor Data . . . 173

8.4.2 Interpreting HAR decisions . . . 175
8.4.3 Interpreting MS Remote Sensor Data . . . 177
8.4.4 Digitally Estimated Severity Outcome . . . 182
8.4.5 Longitudinal Characterisation of Digitally Estimated Severity Outcomes . . . 182
8.5 Discussion . . . 186
8.5.1 Interpretation of Activity-based Characteristics . . . 186
8.5.2 Interpretation of MS-influenced Characteristics . . . 188
8.5.3 Longitudinal Trend Monitoring of Remote Smartphone-Based Outcomes . . . 190
8.5.4 Limitations and Considerations . . . 193
8.6 Conclusion . . . 195

9 Conclusions and Future Work 197
9.1 Summary of Contributions . . . 198
9.2 Limitations Leading to Future Work . . . 203
9.2.1 Study Design and Considerations . . . 203
9.2.2 The Migration to Unobtrusive and Free-Living Ambulatory Monitoring of PwMS . . . 205
9.2.3 Improving Severity Outcome Estimates . . . 206
9.2.4 Longitudinal Remote Monitoring of PwMS . . . 208

Appendices

A Methods: Additional Distributions, Algorithms and Features 211
A.1 Additional Distributions . . . 211
A.1.1 The z-distribution . . . 211
A.1.2 The Student's t-distribution . . . 212
A.1.3 The Chi-Square distribution χ^2 . . . 213
A.1.4 The F-distribution . . . 213
A.2 Algorithms . . . 214
A.2.1 Principal Component Analysis (PCA) . . . 214
A.2.2 t-distributed Stochastic Neighbour Embedding . . . 216
A.2.3 The SVM Dual Representation . . . 218
A.2.4 Piecewise Linear Interpolation . . . 219
A.2.5 Shape-Preserving Piecewise Cubic Hermite Interpolation . . . 220
A.2.6 The Box-Cox Transformation . . . 220
A.2.7 Orientation Independent Transformation . . . 221
A.3 Signal Decompositions . . . 222
A.3.1 The Discrete Wavelet Transform (DWT) . . . 222

A.3.2 Empirical Mode Decomposition (EMD) 224
A.4 Feature Extraction 227
A.4.1 Chapter 6: Gait Feature Space 227
A.4.2 Chapter 5: The Draw a Shape Feature Space 235

B Supplementary Results **239**
B.1 Chapter 5: The Distribution of All Clinical 9HPT Predictions . . . 239
B.2 Chapter 7: DCNN Model Parameterisation 240
B.3 Chapter 7: DCNN Transfer Learning 241
B.4 Chapter 7: Activity Predictions in FL 242
B.5 Chapter 7 & 8: Disease Representation in FL 243

References

List of Figures

2.1 The demyelination of a nerve affected by multiple sclerosis. 10
2.2 Typical disease course for various Multiple Sclerosis phenotypes. 11
2.3 Demonstration of the clinically assessed Nine-Hole Peg Test (9HPT). 16
2.4 Digital outcomes for remote MS-patient monitoring. 19

3.1 Overview of exploratory data analysis methodologies. 32
3.2 Diagram comparing ℓ_1 (left) and ℓ_2 (right) regularisation approaches. 46
3.3 Diagram illustrating the SVM classifier. 48
3.4 Simple diagram describing the construction of a Random Forest (RF) model. . 53
3.5 Illustration of a commonly used architecture in deep-learning. 57
3.6 Schematic describing the computation of Layer-wise Relevance Propagation (LRP) 63
3.7 A binary confusion matrix 66

4.1 The FLOODLIGHT PoC test suite 71
4.2 The FLOODLIGHT PoC (NCT02952911) study design 72
4.3 Smartphone and smartwatch device inertial sensor coordinates. 73
4.4 Distribution of Nine-Hole Peg Test (9HPT) times. 75
4.5 Distribution of Expanded Disability Status Scale (EDSS) scores. 75
4.6 Demonstration of the Draw-a-Shape test. 76
4.7 Example of Two Minute Walk Test (2MWT) smartphone-based inertial sensor data. 77
4.8 The smartphone-based remote Two Minute Walk test (2MWT). 77
4.9 Survival curves depicting participant adherence rates 78
4.10 Example of smartphone-based inertial sensor data in UCI HAR. 79

5.1 Example illustrations of *circle* shape drawn by HC, nPwMS, and aPwMS . . . 85
5.2 Example illustrations of *figure-8-shapes* shape drawn by HC, nPwMS, and aPwMS. 86
5.3 Pixel density heat map of *spiral* shapes drawn by by HC, nPwMS, and aPwMS. 88
5.4 Example of *square* shapes drawn by HC, nPwMS and aPwMS subjects. 89
5.5 Pairwise linear Pearson's (r) correlation matrix. 93
5.6 Out-of-sample test MAE [s] as features are cumulatively added to an IRLS model. 96
5.7 Comparison of the out-of-sample test MAE [s] across shapes. 97
5.8 Scatter plot of average, Y, versus predicted, $\hat{Y}$, Nine-Hole Peg Test (9HPT) times. 98
5.9 Scatter plot of mean 9HPT [s] per subject against the standard deviation (SD). 99

6.1 Schematic of the 2MWT gait processing pipeline . . . 112
6.2 Inertial sensor data pre-processing steps. . . . 113
6.3 Continuous Wavelet Transform (CWT) time-frequency representation. . . . 116
6.4 The scale-dependent CWT energy density E_s distribution. . . . 117
6.5 SVM optimisation strategy. . . . 121
6.6 Pairwise gait-feature Spearman's ρ correlation matrix. . . . 122
6.7 Example of PCA performed on top 25 smartphone and smartwatch gait features. 126
6.8 Comparison of step count-based biomechanical features. . . . 127
6.9 The agreement between two step counting algorithms. . . . 128
6.10 Comparison of out-of-sample subject classification performance. . . . 130

7.1 Schematic of proposed transfer learning (TL) approach. . . . 146
7.2 Schematic of deep transfer learning approach. . . . 148
7.3 Confusion matrices for $\mathcal{T}_S$ HAR classification results . . . 151
7.4 Visualisation of UCI HAR trained DCNN features using t-SNE. . . . 152
7.5 Visualisation of UCI HAR trained DCNN features evaluated in FL using t-SNE. 152
7.6 Visualisation of FL trained DCNN features using t-SNE. . . . 153
7.7 Confusion matrix for subject classification outcomes. . . . 156

8.1 Schematic illustrating the clinical utility of interpretable models. . . . 164
8.2 Panel plot illustrating the DCNN MS gait-signal explanations using LRP. . . . 174
8.3 Panel plot illustrating example performance of UCI HAR DCNN model. . . . 175
8.4 Illustrative examples of static versus movement predicted UCI HAR DCNN epochs. 176
8.5 Panel plot illustrating movement activity predicted from a UCI HAR DCNN. . 177
8.6 Comparison of Layer-Wise Relevance Propagation visualisation for HC, PwMSmild and PwMSmod examples. . . . 178
8.7 Visualisation of the average gait signal through the LRP-CWT analysis framework. 181
8.8 Continuous disease severity outcome estimate, EDSS and subject group. . . . 182
8.9 Panel plot illustrating the longitudinal severity estimate outcomes. . . . 183
8.10 The average longitudinal trajectory for HC, PwMSmild and PwMSmod. . . . 185

A.1 The standard z-distribution. . . . 212
A.2 The t-distribution. . . . 212
A.3 The χ^2-distribution. . . . 213
A.4 The $F-$distribution. . . . 214
A.5 The Discrete Wavelet Transform (DWT) filterbank . . . 223
A.6 The multi-resolution decomposition of a gait-signal . . . 224

B.1 Distribution of actual Y and predicted $\hat{Y}$ NHPT times [s] per subject. . . . 239
B.2 Classification performance over varying number of daily tests sampled per subject. 240
B.3 Comparison of DCNN transfer learning approaches. . . . 241

B.4 Visualisation of UCI HAR trained CNN features extracted on FL using *t*-SNE . 242
B.5 Visualisation of the CNN features trained on FLOODLIGHT (FL) using *t*-SNE. 243
B.6 Confusion matrix for subject classification outcomes for unified multi-class models. 243
B.7 (d) Comparison of incorrectly classified HC participant using LRP. 245

List of Tables

4.1 Demographics for healthy controls (HC) and people with multiple sclerosis (PwMS). 74

5.1 Demographics and characteristics of the Draw-a-Shape population subset. . . . 84
5.2 Comparison of top DaS features between handed models. 94
5.3 Mean absolute error (MAE) in IRLS predictions $\hat{Y}$ of average Y 9HPT times [s] 97
5.4 Performance metrics of average Y versus predicted $\hat{Y}$ 9HPT times. 100

6.1 Population Demographics for Two-Minute Walk Test subset 111
6.2 Comparison of step count and top features between HC and PwMSmod. . . . 124
6.2 Comparison of step count and top features between HC and PwMSmod. . . . 125
6.3 Smartphone-based step count classification outcomes by EDSS grouping 128
6.4 Smartphone-based subject classification outcomes for various classifiers. 130

7.1 Overview of source $\mathcal{D}_S$ and target domain $\mathcal{D}_T$ datasets. 150
7.2 HAR classification $\mathcal{T}_S$ performance for each $\mathcal{D}_S$ 151
7.3 Classification results between various models for each task subset, $\mathcal{T}$. 155

A.1 Description of the features extracted from the Two-Minute Walk Test 227
A.2 Description of features extracted from the Draw-a-Shape Test 236

B.1 Parameters that maximised each target classification task, $\mathcal{T}_T$ 240
B.2 Breakdown in the distribution of predicted HAR activities in the 2MWT . . . 242

Recurring Abbreviations

2MWT, 6MWT . . Two- or Six-Minute Walk Test
9HPT . . . . . Nine-Hole Peg Test
BF . . . . . Brown-Forsythe Test
CNN . . . . . Convolutional Neural Network
CNS . . . . . Central Nervous System
CSF . . . . . Cerebrospinal Fluid
CV . . . . . Cross-Validation
CWT . . . . . Continuous Wavelet Transform
DaS . . . . . Draw-a-Shape Test
DCNN . . . . . Fully-Connected Deep Convolutional Neural Network
DL . . . . . Deep-Learning
DNN . . . . . Deep Neural Networks
DWT . . . . . Discrete Wavelet Transform
EDSS . . . . . Expanded Disability Status Scale
EMD . . . . . Empirical Mode Decomposition
FL . . . . . FLOODLIGHT PoC study
FN . . . . . False Negative
FP . . . . . False Positive
GLM . . . . . Generalised Linear Model
HAR . . . . . Human Activity Recognition
HC . . . . . Healthy Control
HMF . . . . . Hand-Motor-Function
ICC . . . . . Intraclass Correlation Coefficient
IMF . . . . . Intrinsic Mode Functions
IMU . . . . . Inertial Measurement Unit
IRLS . . . . . Iteratively Re-Weighted Least Squares
KWt . . . . . Kruskal-Wallis test
LASSO . . . . . Least Absolute Shrinkage and Selection Operator
lassoglm . . . . . LASSO for GLMs
LR . . . . . Logistic Regression
LRP . . . . . Layer-Wise Relevance Propagation
MAE . . . . . Mean-Absolute Error

MF1 . . . . . Macro F1-score
ML . . . . . Machine Learning
MRI . . . . . Magnetic Resonance Imaging
MS . . . . . Multiple Sclerosis
MSE . . . . . Mean Squared Error
MsEn . . . . . Multiscale Entropy
MSFC . . . . . Multiple Sclerosis Functional Composite
OLS . . . . . Ordinary-Least Squares
PCA . . . . . Principal Componet Analysis
PoC . . . . . Proof-of-Concept
PRO . . . . . Patient-reported Outcome
PwMS . . . . . People with Multiple Sclerosis
aPwMS . . . . . PwMS with presumably abnormal HMF, 9HPT >22.7 [s]
nPwMS . . . . . PwMS with presumably normal HMF, 9HPT ≤22.7 [s]
PwMSmild . . . . . Mildly disabled PwMS, EDSS [0–3.5]
PwMSmod . . . . . Moderately disabled PwMS, EDSS [3.5–6]
PwMSsevere . . . . Severely disabled PwMS, EDSS > 6
RBF . . . . . Gaussian Radial Basis Function
RFR . . . . . Random Forest Regression
RMSE . . . . . Root-Mean Squared Error
RRMS . . . . . Relapsing–Remitting Multiple Sclerosis
SI . . . . . Subject Identification
SNR . . . . . Signal-to-Noise Ratio
SPMS . . . . . Secondary-Progressive Multiple Sclerosis
SVM . . . . . Support Vector Machines for Classification
SVR . . . . . Support Vector Machines for Regression
T25FW . . . . . Timed-25 ft. Walk Test
TKEO . . . . . Teager-Kaiser Energy Operator
TL . . . . . Transfer Learning
TN . . . . . True Negative
TP . . . . . True Positive
***t*-SNE** . . . . . t-Distributed Stochastic Neighbour Embedding
UE . . . . . Upper-Extremity

Common Notation

a, A	Scalar value, a or A (variable, parameter or constant)
$\mathbf{x}$	Column or row vector, $\mathbf{x}$ or $\boldsymbol{x}$; see[1]
$\mathbf{X}$	Matrix, $\mathbf{X}$
x_i	The i^{th} element of $\mathbf{x}$ or i^{th} instance of x
x_{ij}	The $(i \times j)^{th}$ element of $\mathbf{X}$ or $(i, j)^{th}$ instance of x
$x^{(k)}$	The k^{th} example, input, or layer of $\mathbf{x}$ or x
$\bar{\mathbf{x}}$	The mean value of elements in $\mathbf{x}$ or across instances of x
$\mathbf{x}^\top$, $\mathbf{X}^\top$	The transpose of vector $\mathbf{x}$, or matrix $\mathbf{X}$
$\mathbf{X}^{-1}$	Matrix inverse of $\mathbf{X}$
$\mathbf{X}^+$	Moore-Penrose pseudo-inverse of $\mathbf{X}$
$\mathbb{R}^N$	Column vector of real number, or length N
$\mathbb{R}^{1\times P}$	Row vector of real numbers, of length P
$\mathbb{R}^{N\times P}$	Matrix of real numbers, of dimensions $(N \times P)$
$\frac{dx}{dt}$	The first derivative of $\mathbf{x}$ with respect to t
$\frac{d^2x}{dt}$	The second derivative of $\mathbf{x}$ with respect to t
$\frac{\partial f}{\partial \mathbf{w}}$	The partial derivative of f with respect to $\mathbf{w}$
$\nabla_{\mathbf{w}} f$	The gradient of f with respect to $\mathbf{w}$: $\nabla_{\mathbf{w}} f = \left(\frac{\partial f}{\partial \mathbf{w}_1}, \frac{\partial f}{\partial \mathbf{w}_2}, \cdots, \frac{\partial f}{\partial \mathbf{w}_P}\right)$
$\nabla_{\mathbf{w}} f$	Jacobian Matrix $\mathbf{J} \in \mathbb{R}^{m\times n}$ of $f : \mathbb{R}^n \to \mathbb{R}^m$
$\nabla^2_{\mathbf{w}} f(\mathbf{w})$	Hessian, $\mathbf{H}$ of f at $\mathbf{w}$: $\nabla^2_{\mathbf{w}} f = \left(\frac{\partial^2 f}{\partial \mathbf{w}_1^2}, \frac{\partial^2 f}{\partial \mathbf{w}_2^2}, \cdots, \frac{\partial^2 f}{\partial \mathbf{w}_P^2}\right)$
$E(\cdot)$	An objective function
$\mathcal{L}(\cdot)$	A loss function

Common Notation

$\in$, $\notin$	Belongs to a set, does not belong to a set
$\forall$	for all
:	where
$\sim$	is distributed by
$f : \mathbb{A} \to \mathbb{B}$	The function f with domain $\mathbb{A}$ and range $\mathbb{B}$
$f(\mathbf{x}; \mathbf{w})$	A function of $\mathbf{x}$, paramaterised by $\mathbf{w}$. Most often written as $f(\mathbf{x})$ and $\mathbf{w}$ is omitted for brevity
$\ln x$	Natural logarithm of x
$\lvert\mathbf{x}\rvert$	Absolute value of $\mathbf{x}$
$\lvert\lvert\mathbf{x}\rvert\rvert_p$	ℓ_p norm of $\mathbf{x}$
$\lvert\lvert\mathbf{x}\rvert\rvert$	ℓ_2 norm of $\mathbf{x}$
$\mathcal{D}$	A domain or dataset
$\mathcal{T}$	A specific task, such a classification or regression
$P(\cdot)$	A probability distribution over a discrete variable
$p(\cdot)$	A probability distribution over a continuous variable or a variable whose type has not been specified
$p(\cdot\lvert\cdot)$	Conditional density distribution
$\Phi(\cdot)$, $\phi(\cdot)$	A Kernel or non-linear function
$\mathcal{N}(\mu, \Sigma)$	A Normal Gaussian distribution, with mean μ and covariance Σ
$\Gamma(\cdot)$	Gamma function
$\arg\min_{\mathbf{w}} f(\mathbf{w})$	The value of $\mathbf{w}$ for which $f(\mathbf{w})$ attains it's minimum
$\arg\max_{\mathbf{w}} f(\mathbf{w})$	The value of $\mathbf{w}$ for which $f(\mathbf{w})$ attains it's maximum
$H(\boldsymbol{x})$	Entropy of random variable $\boldsymbol{x}$

[1] I generally reserve $\mathbf{x}$ for feature values, and $\boldsymbol{x}$ as the raw data, or undefined random variables.

Associated Publications

Creagh, A. P., Simillion, C., Scotland, A., Lipsmeier, F., Bernasconi, C., Belachew, S., van Beek, J., Baker, M., Gossens, C., Lindemann, M. and de Vos, M. (2020). "Smartphone-based remote assessment of upper extremity function for multiple sclerosis using the Draw a Shape Test". *Physiological Measurement* 41(5): 054002. `https://doi.org/10.1088/1361-6579/ab8771`. [1].

Creagh, A. P., Simillion, C., Bourke, A., Scotland, A., Lipsmeier, F., Bernasconi, C., van Beek, J., Baker, M., Gossens, C., Lindemann, M. and de Vos, M. (2020). "Smartphone- and Smartwatch-Based Remote Characterisation of Ambulation in Multiple Sclerosis during the Two-Minute Walk Test". *IEEE Journal of Biomedical and Health Informatics*, 25(3), pp.838-849. `https://doi.org/10.1109/JBHI.2020.2998187` [2].

Creagh A.P., Lipsmeier F., Lindemann M. and de Vos, M. (2021). "Interpretable Deep Learning for the Remote Characterisation of Ambulation in Multiple Sclerosis using Smartphones". *Nature Scientific Reports*, 11(1), pp.1-14. `https://doi.org/10.1038/s41598-021-92776-x` [3].

Creagh A.P., Lipsmeier F., Lindemann M. and de Vos, M. (2022). "Longitudinal Remote Characterisation of Ambulation in Multiple Sclerosis using Smartphones". *IEEE open journal of engineering in medicine and biology*, 3, pp. 202-210. `https://doi.org/10.1109/OJEMB.2022.3221306` [4]

Creagh, A.P. (2020). "AI and the Digital Healthcare Revolution Uniting Disciplines". *Phenotype*. Oxford, UK: pp. 14-16. [5].

Cheng, W.-Y., Lipsmeier F., **Creagh, A.P**, Scotland A., Kilchenmann T., Jin L., Schjodt-Eriksen J., Wolf D., Zhang-Schaerer Y.-P., Garcia I. F., Siebourg-Polster J., Soto J., Verselis L., Facklam M. M., Boess F., Koller M., Grundman M., Monsch A. U., Postuma R., Ghosh A., Kremer T., Taylor K. I.,Czech C., Gossens C. and Lindemann M. (2018). "Large-Scale Continuous Mobility Monitoring of Parkinson's Disease Patients Using Smartphones". In *International conference on Wireless Mobile Communication and Healthcare* (pp. 12-19). `https://doi.org/10.1007/978-3-319-98551-0_2`. [6].

Cheng, W.-Y., Lipsmeier F., Scotland A., **Creagh, A.P.**, Kilchenmann T., Jin L., Schjodt-Eriksen J., Wolf D., Zhang-Schaerer Y.-P., Garcia I. F. Siebourg-Polster J., Soto J., Verselis L., Facklam M. M., Boess F., Koller M., Grundman M., Monsch A. U., Postuma R., Ghosh A., Kremer T., Taylor K. I.,Czech C., Gossens C. and Lindemann M, (2017). "Smartphone-based continuous mobility monitoring of Parkinsons disease patients reveals impacts of ambulatory bout length on gait features". *IEEE Life Sciences Conference (LSC)*. `https://doi.org/10.1109/LSC.2017.8268169`. [7].

1 Introduction

1.1 Motivation

Multiple sclerosis (MS) is a chronic and progressive neurodegenerative disease that affects around 1 in 1000 people and more than 2 million worldwide [8, 9]. MS is considered an autoimmune disorder which causes irreversible demyelination of the central nervous system (CNS) [10]. This neurodegeneration is typically diagnosed in young adults and continues to cause progressive, varied and unpredictable physical and mental disability over time [9]. The causes of MS are not well understood; there are currently no curative treatments and most therapeutic interventions are aimed at managing the symptoms.

People with MS (PwMS) typically exhibit impairments to physical function, most commonly during ambulation [11, 12], but can also demonstrate upper-extremity (UE) disability [13], visual impairments [14], sensory disturbances [15], and cognitive dysfunction [16]. MS follows an unpredictable and heterogeneous disease course patient-to-patient, and over time, which limits the effectiveness of disease management and monitoring [8]. Patients with early forms of MS initially experience sudden acute symptoms developing (a "relapse") over days before generally plateauing over weeks or months ("remission") [8, 17]. Other

forms of MS or later stages are characterised by insidious worsening of symptoms until severe disability or death [17, 18].

The primary and gold-standard means of assessing PwMS are a combination of clinician-administered rating scales and Patient-Reported Outcomes (PROs) that remain unchanged for years. These assessments include the Expanded Disability Status Scale (EDSS) [19], and other supplementary scores: namely the Multiple Sclerosis Functional Composite (MSFC) [20], which encompass tests aimed at specific domains of MS dysfunction, such as ambulation and UE abilities. Assessments include the gait-based Timed-25 ft. Walk Test (T25FW) [21] and Two-Minute Walk Test (2MWT) [22], as well as the UE-based Nine-Hole Peg Test (9HPT) [23]. Patient-reported outcomes, for instance, the Multiple Sclerosis Impact Scale-29 (MSIS-29) and Multiple Sclerosis Walking Scale-12 (MSWS-12) also play a significant role in determining MS disease activity [24],

Standard MS clinical assessments and self-reported scales have several well documented limitations [25, 26]. These problems include a low intra- and inter-rater reliability, rater or self-rater subjectivity, as well as an inherent lack of granularity needed to measure MS symptom severity precisely and consistently.

Magnetic Resonance Imaging (MRI) is a developing field alternatively offering objective imaging measures as surrogates for MS disease activity [27, 28], yet have limitations also. For example, the pathology of MS is still poorly understood; MRI outcomes are currently only validated to clinically determine the presence of MS and are not yet accepted as surrogates for primary clinical outcome measures in clinical trials (CT) [29, 30]. To boot, conducting frequent routine MRI assessments are impractical and heavily resource intensive.

Any in-clinic tool, PRO or imaging outcome otherwise all innately suffer from the same limitations due to infrequent assessment. PwMS are generally only ever assessed every 3 to 6 or even every 12 months in trials or in routine care practice respectively [25, 31]. Consequently, the current tools at our disposal are coarse-gained, and may certainly miss episodic manifestations of the disease due to relapses in PwMS. As such, it has been alternatively proposed to develop

out-of-clinic assessments aimed at monitoring the signs and symptoms of PwMS' habitual state [32]. Developing continuous and objective measures that can be administered remotely at high-frequency could greatly augment current in-clinic assessment pictures, informing better disease management techniques, and enabling the development of better therapeutic interventions. Importantly, earlier identification of changes in PwMS impairment has been critical to deliver better therapeutic strategies [33]. Furthermore, out-of-clinic outcomes may even help uncover previous unseen characteristics of the MS disease.

Digital health technologies, such as apps, smartphones and wearables that can collect a variety of biological, physical, or behavioural measurements from the human body [34]. These tools could be used to develop *digital biomarkers*: objective, rich, and sensitive digital health-related outcome measurements that are ecologically valid and can be collected at high frequency and remotely [35–37]. Nowadays, consumer wearable sensors, such as smartphone and smartwatches specifically, are ubiquitous in everyday life. These devices therefore present the perfect tool to develop digital biomarker outcomes to assess PwMS.

The use of digital technologies to develop digital biomarkers in MS is a nascent field, where the current assessments and metrics evaluated are only in the early proof-of-concept (PoC) stage of development. This thesis is therefore motivated by an attempt to explore, build, and establish how objective digital sensor data from smartphones and smartwatches can measure and monitor MS disease activity. The subsequent work presented throughout this thesis will aim to demonstrate how MS physical dysfunction can be characterised using remotely collected smartphone and smartwatch sensor data obtained from the FLOODLIGHT PoC (FL) study. To achieve this, data collected from the Draw-a-Shape (DaS) Test, an assessment of upper-extremity (UE) function based on tracing various pre-defined shapes on a smartphone touch-screen will first be investigated. Next, the characterisation of ambulatory and gait-related function will be explored in data collected from smartphone and smartwatch inertial sensors during a remotely conducted Two-Minute Walk Test (2MWT). It is hoped that the work presented in this thesis

will help catalyse methodological approaches to develop —and the capability of— remote digital outcomes to monitor PwMS.

1.2 Research Aims

This thesis aims to evaluate the utility of digital technologies, specifically consumer-grade smartphones and smartwatches with embedded touch-screen and inertial sensors to objectively evaluate people with multiple sclerosis (PwMS) to create *digital biomarkers*. Through recording sensor data obtained from remotely administered tests of physical function in PwMS, which can be modelled using machine learning (ML) techniques, it is hoped to capture the underlying patterns of this disease outside of routine and infrequent clinical assessments. This thesis focuses on answering the following broad research questions:

1. Can objective digital biomarkers of MS physical function be captured from remotely acquired smartphone and smartwatch sensor data?
    - Can remotely collected digital health data augment current gold-standard clinical assessments?
    - How do objective smartphone and smartwatch measures capture aspects of PwMS physical function?
    - Can these measures distinguish disease status versus healthy controls (HC)?
    - Can these measures stratify across PwMS levels of disability?
2. Can these objective markers be interpreted with regard to a clinical context and do they agree with the current clinical understandings of MS? How do these measures further our understanding to the characterisation of MS from remotely obtained sensor data, especially for sub-clinical manifestations in PwMS?
3. Can objective digital technology outcome measures also be used to perform remote longitudinal disease monitoring?

- If so, how do these longitudinal patterns characterise varying PwMS sub-groups versus healthy controls?
- Can remote outcome measures detect high-frequency symptom variation or the presence of disease related events (i.e. relapses)?

1.3 Thesis structure & key contributions

- **Chapter 1** introduces the overall scope, motivations, and key goals of the research presented in this thesis.
- **Chapter 2** expands deeper into the clinical background on multiple sclerosis: the pathology and epidemiology of MS, as well as the pathophysiology characterising the symptoms of disease, with specific focus on the physical impairments in upper-extremity function and ambulation. This chapter explores the limitations of the current in-clinic methods to measure disease progression, introducing instead how smartphones and smartwatches (or wearables) can be used to objectively measure the physical and cognitive symptoms of MS disability, even outside of clinic. Within this, the concept of digital biomarkers will be introduced and we discover how previous literature has catalysed this shift towards utilising wearables to objectively evaluate PwMS, thus framing the motivations behind this thesis.
- **Chapter 3** provides a technical background of the general methodology implemented in this thesis, including the conceptual exploratory and confirmatory statistical analysis approaches undertaken, as well as the general machine learning methodology, frameworks, and evaluation techniques which are crucial to understanding this work. Subsequent, and more chapter specific methodological approaches are later introduced in each respective chapter.
- **Chapter 4** presents the FLOODLIGHT PoC (FL) dataset which this thesis' work is built upon. In this, the smartphone- and smartwatch-based data

collection applications are detailed, including an outline of the remote assessments encompassing the FL testing battery and further detail on the routine in-clinic assessments for PwMS. The clinical attributes of the participants in FL are summarised, including the population demographics and subject data contributions over the 24-week study.

- **Chapter 5** establishes the foundation for how digital sensor-based assessments may enable an out-of-clinic objective augmentation of traditional rater-administered assessments, specifically focusing on upper-extremity function as measured by the Draw-a-Shape test (DaS). This chapter introduces how DaS features that are motivated by disease pathology and UE function can be extracted from various shapes (*circle*, *figure-8-shape*, *spiral*, *square*) traced on smartphone touchscreen. The ability of these measures to characterise HC and PwMS UE function will be demonstrated through the prediction of the clinical gold-standard 9HPT assessment using multivariate machine learning modelling.

- **Chapter 6** shifts focus to examine how ambulatory function in PwMS versus HC can be remotely distinguished using smartphone- and smartwatch-based inertial sensor data. This chapter again introduces how (gait) characteristics of PwMS can be represented by features extracted from the raw sensor data during a remote 2MWT, and expands the framework introduced in chapter 5 to now directly differentiate subgroups of PwMS and HC as dichotomous classification tasks. This chapter also aims to investigate the important considerations and analysis techniques for evaluating remotely collected smartphone and smartwatch sensor data in PwMS.

- **Chapter 7** dives further into the characterisation of ambulatory function in PwMS during the 2MWT in FL. Building on the limitations of feature-based methodologies, this chapter will aim to explore how deep networks offer data-driven approaches to feature learning, and will investigate how these features may capture a better representation of gait function in PwMS collected from

inertial sensor data. Within this chapter, a deep transfer learning framework will be proposed that may prove beneficial to learn useful characteristics of PwMS disease in the presence of sparse and heterogeneous data.

- **Chapter 8** aims to overcome some of the inherent transparency issues surrounding complex deep networks by using state-of-the-art visualisation techniques. The work presented here aims to decode a network's decision (from chapter 7) using relevance heatmaps, thus highlighting important segments of an input which influenced a model's decision. This interpretation technique will then be applied towards healthy and MS ambulatory examples, aiming to understand the characteristics that may be reflective of disease. Finally, this chapter will demonstrate how surrogate disease outcomes can longitudinally monitor the signs and symptoms of MS disease, which can be cross-sectionally interpreted using the visualisation techniques described above.

- **Chapter 9** finishes this thesis with a final summary of the research, contributions, and limitations introduced throughout this work. Lastly, these concluding remarks are used to motivate a discussion towards potential avenues of future work needed to advance the field of *digital biomarkers*.

2

A Background to Multiple Sclerosis: Diagnosis and Monitoring

Contents

2.1 Multiple Sclerosis 9
2.2 Assessing Multiple Sclerosis Disability 13
2.3 Digital Biomakers 18
2.4 Background Summary 26

This background chapter aims to provide the reader with a brief overview of multiple sclerosis (MS) and an essential introduction to the diagnosis and monitoring of people with MS (PwMS). First, this chapter will outline the important characteristics of MS, including the pathology and epidemiology of this disease, along with the pathophysiology and symptoms typically observed in PwMS. The current gold-standard assessment techniques to diagnose and monitor PwMS will be introduced, along with the inherent limitations of those in-clinic assessments. Currently, the onset and the severity of MS is manually assessed in-clinic by a physician or via imaging modalities such as the magnetic resonance imaging (MRI). Such practices yield infrequent measurements, require in-clinic visits, and are potentially costly. Thanks to the advancement in sensor technology, digital devices are promising tools to assess PwMS. This overview is intended to frame

how alternative methods, such as sensor-based outcomes in PwMS can help characterise and understand this disease. The current sensor-based literature in MS is then reviewed in order to motivate how digital tools, such as smartphones and smartwatches, may present a unique opportunity to characterise and monitor PwMS, particularly outside of the clinic.

2.1 Multiple Sclerosis

2.1.1 Pathology and Epidemiology

Multiple sclerosis (MS) is a chronic neurodegenerative disease that typically affects young adults causing irreversible physical and mental disability and a progressive neurological deterioration over time [8].

MS is the leading cause of non-traumatic disability among young and middle-aged people in many developed countries [9] and affects more than 2 million people worldwide [8] with an incidence of around 1 in 1000 [9]. MS is an autoimmune disorder which causes a chronic and progressive demyelination of the central nervous system (CNS). It is thought that that unknown agent(s) trigger T cell–mediated inflammatory attacks in the CNS which causes irreparable damage to the myelin and the axons [10]. While the causes of MS are still unknown, there appears to be a degree of genetic susceptibility [38–40]. There however also appears to be non-genetic stimulation's to develop MS, such as a latent virus [41], a person's metabolism [42], or links to other environmental factors [43]. This generic susceptibility of MS has manifested in striking epidemiological consequences. For one, MS is particularly prevalent in Caucasians of Northern Europe and in temperate (middle latitude) zones [45]. Some have also linked this to solar or ultraviolet exposure differences among and within populations, as well as linking vitamin D levels (which are typically produced in the skin in response to sunlight) as factors modulating against the risk of MS [46, 47]. However, many of these associations are still poorly understood and constitute quantifying environmental factors that are difficult to measure [9, 43, 45]. Importantly, MS seems to relate to gender and age factors,

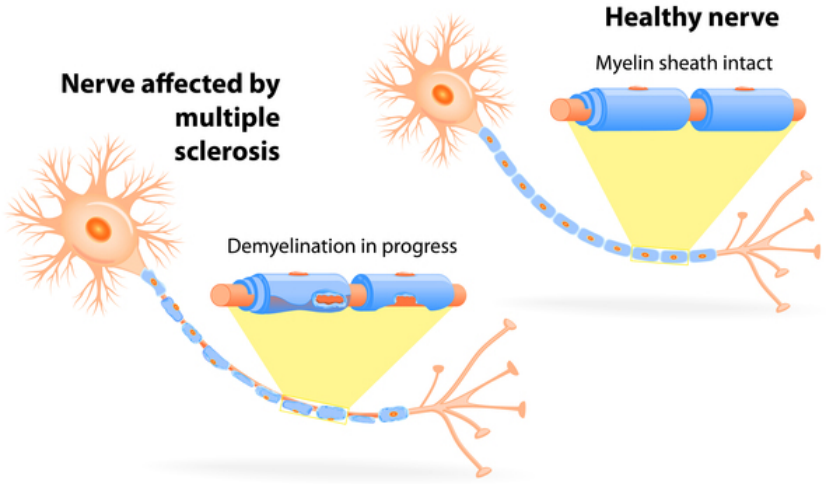


Figure 2.1: The demyelination of a nerve affected by multiple sclerosis versus that of a healthy axon. The myelin sheath (a mixture of lipids and proteins) encases the axonal region of nerve cells in the part of the brain called white matter. MS is an autoimmune disorder, where a person's own immune cells, known as "T-cells" attack the myelin in the central nervous system (CNS), causing destructive and irreparable damage to the nerves and their ability to conduct electrical signals. Image Credit: NIH, National Library of Medicine (US), Genetics Home Reference [44].

where the disease affects twice as many women than as men [9, 45] and typically manifests during the ages of 20-45 in both sexes [8, 9].

Although MS follows a highly heterogeneous and subject-specific disease course, the disease profiles can be grouped into four clinical phenotypes which are based on disease progression [48]: the majority of PwMS will initially experience Relapsing–remitting MS (RRMS), a state dominated by sudden acute symptoms developing (a "relapse") over days before generally plateauing over weeks or months [8], termed "remission". RRMS generally affects 85% of PwMS. RRMS disease activity typically occurs acutely at a sub-clinical level. Secondary-progressive MS (SPMS) can occur in some RRMS patients, where the disease course continues to worsen with or without periods of remission. After 10 years, around half RRMS patients will go onto develop SPMS [17, 49, 50]. Those experiencing consistent but worsening symptoms can be thought of as having Primary-progressive MS (PPMS) [8, 17, 49], affecting roughly 10% of PwMS [50]. Progressive-relapsing MS is more rare (affecting fewer than 5% of PwMS); it occurs from diagnoses as a progressive

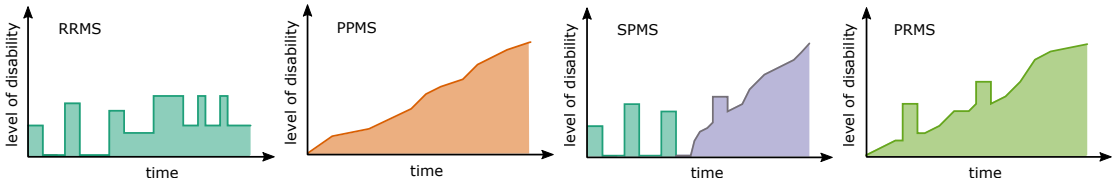


Figure 2.2: Typical disease course for various Multiple Sclerosis phenotypes: Relapsing-remitting MS (RRMS); Primary-progressive MS (PPMS); Secondary-progressive MS (SPMS) and Progressive-relapsing MS (PRMS) [48].

disease course, with periods of relapse, but without any remission periods. Figure 2.2 illustrates the four categories of disease course of MS.

At present, there are a few ways to diagnose MS; the clinical diagnosis of MS is based on presence of one or more evidential presentations such as numerous episodes of the disease (at least two), chronic inflammation of the CNS based on examination of the cerebrospinal fluid (CSF) [8]. Magnetic resonance imaging (MRI) has recently become the gold-standard to diagnose MS to determine the presence of lesions (at least two) in the white matter of the CNS [8, 27, 28].

2.1.2 Pathophysiology

MS is a highly mutable and heterogeneous disease, which affects multiple physical, cognitive, and behavioural aspects of human function[1]. Crucially, although the disease may manifest in multiple functional domains in a population, individual disease profiles are idiosyncratic, presenting symptom in a subset or all of the functional domains, and may fluctuate over time.

Impairments to physical functional are the most common manifestations of MS, such as upper- and lower-limb disability [11] as well as sensory disturbances [15, 51], and cognitive dysfunction [16]. PwMS often have moderate-to-severe spasticity, mostly in the legs [52]. Common sensory disturbances include paresthesias (limb numbness and tingling), dysesthesias (painful burning, prickling, or aching), and bladder (urinary sphincter) disturbances [8]. Some PwMS can have optic neuritis [14], causing double vision or even partial or complete loss of vision. Thanks to the unique sensitivity of the wearable devices in capturing the physical signals

[1]Pathophysiology refers to the functional changes that accompany a disease.

related to hand and gait functions, in this thesis I will focus on two areas of physical functional impairment: (1) gait or ambulatory Function; and (2) upper-extremity (UE) function or hand-motor function (HMF).

2.1.3 Gait and Ambulatory Function

Ambulatory function have been perceived as the most prominent physical impairments in PwMS [12] with approximately 85% MS patients reporting mobility impairments [53]. Many studies suggest that alterations during ambulation (gait) is predictive of MS [54–56]. PwMS can often have postural instability and difficulties with balance [57, 58]. Simple gait metrics such as reduced gait velocity (slow gait) can identify and isolate PwMS from healthy controls (HC), [59, 60] . More complex measures such as stride-to-stride fluctuations and ambulatory bout variability have expanded this classification [54, 59, 60]. Gait variability, characterised by the variance in biomechanical parameters such as hip, knee and ankle joints rotations derived from laboratory-based motion capture systems distinguished PwMS from healthy control (HC) participants [60]. Importantly, work by Crenshaw *et al.* (2006) link fatigue in PwMS with slow gait [60]. Other studies have associated fatigue in PwMS to affect cadence (steps/minute) and daily step count, which were significantly different between MS subgroups determined by the severity of their disability [61, 62]. Recent work by Shema-Shiratzky and Hausdorff *et al.* (2019) also link cadence, and other parameters of gait, such as stride time variability, stride regularity, step regularity and gait complexity, to gait fatigability in PwMS [63]. Comber *et al.* (2016) offer a substantial review of gait deficits in people with MS [11]. This meta-analysis found a large effect of MS disease on biomechanical gait parameters, such as stride length, velocity, double support duration step length and swing phase duration. Other biomechanical gait parameters, such as step width, stride time and cadence were altered in PwMS, although these were non-significant. Faster walking conditions amplified the indication of disability and all effects reported were apparent for those with low MS severity scores (see section 2.2) – indicating that gait was impaired even in those with lower disability. In fact,

studies by Sosnoff *et al.* demonstrate that even those with minimal impairment due to MS showed greater variability in short bouts of gait compared to age-matched healthy controls [56, 64]. Together, these findings and similar ones of those by Kaipust *et al.* (2012) [65] suggest that increases in gait variability are seen during early stages of disease and progress as the disability worsens, hinting that even subtle alterations of gait may be predictive of MS.

2.1.4 Upper-Extremity and Hand-Motor Function

The impairment of upper-extremity (UE) function and manual dexterity resulting from sensory and motor deficits is widely reported across all subtypes of MS, although progressive MS is associated with higher prevalence of UE dysfunction and greater impairment of manual dexterity [13, 66]. Cerebellar Upper Limb Ataxia (CULA), various tremor types, spasticity and generally reduced hand-motor function (HMF) are impairments in PwMS [67–69]. Fatigue is one of the most common symptoms in PwMS [70, 71], and can be perceived by clinicians to manifest as UE function impairment [72, 73]. Furthermore, impaired vision and ocular symptoms may affect HMF and coordination [14, 74]. UE dysfunction impacts PwMS ability to perform activities of daily-living, affecting their independence, work retention and quality of life [73, 75]. The latter study by Yozbatiran *et al.* (2006) found cognitive function was associated with upper-extremity function, and impairment thereof. It should also be considered that impaired cognition, cognitive overload, stress and depression are all symptomatic of Multiple Sclerosis [16] (see Rao *et al.* (1991), Bobholz *et al.* (2003) and Chiaravalloti *et al.* (2008) for a comprehensive reading on the topic [16, 76, 77]). Furthermore, symptoms of fatigue and cognitive function have been explicitly linked in PwMS [78].

2.2 Assessing Multiple Sclerosis Disability

A number of in-clinic assessments can be employed to evaluate PwMS. MRI is nearly always used to clinically determine the presence of MS [27, 28, 79] (see section 2.1.1 above), is useful for monitoring disease progression [80], and the effects of

treatment in clinical trials [8, 81–84]. Various MRI outcome measures have been proposed as surrogates of disease activity [29, 82–84]. Measures such as the total volume of brain lesions on T2-weighted MRI [84], specific lesions on T1-weighted images are typical MS trial outcomes (surrogate endpoints) [85]. While imaging is playing a significant role in the diagnosis and prognosis of PwMS, MRI metrics have not been accepted as surrogates for primary clinical outcome measures in clinical trials, particularly in Phase 3 trials [29]. Imaging surrogates have been disputed as *"not adequate to explain the complex pathophysiology resulting in disability, disease progression, and clinical relapses in MS"* [29, 30]. Other studies have proposed optical coherence tomography (OCT) (an imaging technique using near-infrared light) to measure structures within the eye [86–88], particularly the nonmyelinated axons, and macular volume of the retinal nerve fiber layer (RNFL) [29]. For instance, RNFL thinning has been proposed as a surrogate measure for optic nerve and brain axonal loss [89] and longitudinal brain atrophy in PwMS [87], where OCT measures have also been shown to correspond well to MRI measures [90]. However, much like MRI surrogates, OCT measures have not been validated as a surrogate for disability or disease progression in MS trials [29]. OCT may fail to characterise demyelination or capture other relevant pathways that represent the manifestation of the disease needed for a clinical trial outcome, such as quantifying disability, relapses or an ability to capture conversions of people to clinically-definite MS (CDMS). It must also be considered that conducting frequent routine MRI or OCT assessments are impractical and heavily resource intensive.

Instead, primary clinical trial outcomes rely on a combination of clinician rated scales, such as the Expanded Disability Status Scale (EDSS) [19] and the Multiple Sclerosis Functional Composite (MSFC) [20], as well as event reporting, such as the number of relapses or even CDMS [29]. For CDMS, if the patient has been recruited into a trial after one clinically isolated syndrome (CIS) (such a demyelinating event), the first relapse in the trial will then correspond to the second clinical attack and confirmation of CDMS.

At present, clinician-administered rating scales are the primary and gold-standard means of assessing PwMS disease activity and disability. The most widely used clinical-administered rating scale is the EDSS. It assigns PwMS on a scale from 0 (no disability) to 10 (death due to MS), in increments of 0.5, based on the rated degree of perceived disability and handicap by a neurologist rating across various functional systems of the CNS. In essence, lower scale values represent impairments based on neurological examination and higher scale values ($>$ 6 EDSS) measure severe disability of PwMS (PwMSsevere). Importantly for this thesis, the determination of EDSS 3.5–6 is heavily reliant on aspects of walking ability. By definition, gait disorders begin to become prominent in subjects with EDSS$\geq$3.5, and those PwMS with EDSS 3.5–6 are often denoted as moderately disabled (PwMSmod). Subjects with EDSS$<$3.5 are then subsequently considered mildly impaired with respect to gait function (PwMSmild).

Administering the EDSS will nearly always be supplemented by the Multiple Sclerosis Functional Composite (MSFC) [20, 91]. The MSFC is an additional composite of assessments aimed to evaluate important domains of overall PwMS impairment [20]. The MSFC has predictive validity to the EDSS, to quantitative imaging measurements such as brain atrophy, distinguish PwMS disease categories (e.g. SPMS) as well as correlating to PwMS self-reported measures [20, 91]. The MSFC includes the assessment of maximum walking speed using the Timed 25-Foot Walk (T25FW), recording the time taken (in seconds [s]) for a participant to (safely) walk a pre-defined distance (i.e. 25 feet) as fast as possible [20, 21]. The Nine-Hole Peg Test (9HPT) is typically used as a standardized upper-extremity outcome measure in the MSFC [20]. The 9HPT requires participants to repeatedly place and then remove nine pegs into nine holes on a board, one at a time, as quickly as possible[2]. A demonstration of the 9HPT is given in figure 2.3. Generally, the 9HPT is the most frequently used measure of manual dexterity in MS research, clinical trials and clinical practice, both independently and in conjunction with other rater-assessments [23, 91–94]. Together with the EDSS, the T25FW and the

[2]The 9HPT forms a fundamental aspect of chapter 5, and is discussed there in more detail

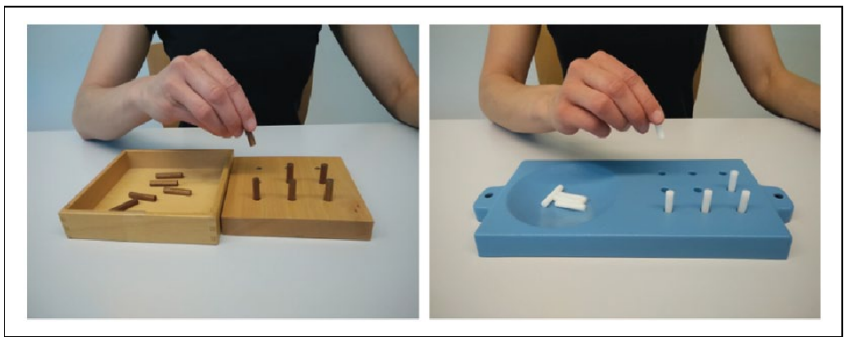


Figure 2.3: Demonstration of the clinically assessed Nine-Hole Peg Test (9HPT). The 9HPT requires participants to repeatedly place and then remove nine pegs into nine holes, one at a time, as quickly as possible. The original wooden square version (left) and the most commonly used commercially available plastic version (right) of the 9HPT are depicted. Image credit: [23].

9HPT also form part of the composite Confirmed Disability Progression (CDP) which has been proposed as a specific endpoint in MS drug trials [95]. The 9HPT and T25FW are sensitive to detect MS disability, Kragt *et al.* (2006) for example demonstrated that worsening of PwMS' T25FW or 9HPT (i.e. longer completion times [s]) by even 20% represented a clinically relevant increase in PwMS's self-reported perceived disability [94].

Cognitive function is assessed in the MSFC by means of the Paced Auditory Serial Addition (PASAT) test, a concentration-based assessment [96].

Other independent, specific functional assessments have been developed for PwMS [97]. For example the 6-Minute Walk Test (6MWT) has been proposed as a measure to capture gait function and fatigability in PwMS [22, 63]. The 6MWT consists of prolonged walking (i.e. for 6 minutes) over a fixed course (or a treadmill) and is typically recorded as walking distance (in metres [m]) [98]. Recently, many MS clinical assessments have preferred to implement the shorter Two-Minute Walk Test (2MWT) instead [22] promoted by Gijbels *et al.* (2011), who found that PwMS' performance is sufficiently captured with the 2MWT, at much lower patient burden [99]. Similar physical assessments include the Timed Up and Go test (TUG) measuring combined mobility, balance, and gait ability as PwMS stand-up from a chair and walk a short course under as fast, but safe, as they can [22]. The

Berg Balance Scale (BBS) and five-times sit-to-stand test (FTSST) commonly supplement the clinical scores introduced above [100] [22, 101].

Many Patient-reported outcomes (PROs) are also used in clinical practice; for instance, the Multiple Sclerosis Impact Scale-29 (MSIS-29) and Multiple Sclerosis Walking Scale-12 (MSWS-12) [24].

2.2.1 Limitations Assessing Multiple Sclerosis Disability

Standard MS clinical assessments and self-reported scales, however, have a number of well documented limitations [25, 26]. For instance, the intra- and inter-rater reliability of the EDSS [102] has been criticised [103], especially for milder EDSS scores (1.0—3.5) that demonstrate greater variability [104]. Clinicians often disagree on the appropriate interval scales (i.e. each step up the scale should be represented by 0.5 or 1) [105]. The sensitivity of the EDSS has been reported to be inhomogeneous across scales and over time [25, 105], for instance changes have been observed to occur much more frequently for milder EDSS scores compared to a severe EDSS score of 6 [106]. While the MSFC has been generally deemed to be reliable, contentions exist regarding the interpretability of the composite score [105]. As mentioned previously, some studies have determined that an approximate 20% worsening in the T25FW and 9HPT portions of the MSFC indicate a clinically meaningful change in MS disability [94, 107, 108], while others studies have specified that this change needs to be sustained over a 3 month period [109], or determined that this threshold is not sufficient to detect disease progression [110].

Specifically within the MSFC, the inter-rater and test-retest reliability of the 9HPT is generally high across a range of disability levels; however most studies focus on more disabled PwMS populations [23]. One large-scale study with a healthy population, for example, reports high inter-rater reliability but only moderate test-retest reliability [111]. Regardless, the 9HPT has satisfactory discriminative and ecological validity in PwMS [94]. Similarly, while the T25FW is a simple assessment procedure with a high-face validity to capture gait-related function [21], it is of limited value. Besides differences in clinical instructions influencing the

T25FW execution [112], a ceiling effect exists for those with normal walking speeds [112–115]; therefore the T25FW is less sensitive for detecting differences among PwMS with a mild disability [114, 116]. Furthermore, like the 9HPT, the T25FW is reported only in seconds and is in nature an inherently blunt assessment tool.

Besides the limitations of the specific MS clinical assessments, the infrequent in-clinic administration of these assessments, typically every 3 to 6, or 12 months, in MS clinical trials or in routine care practice respectively, may miss episodic manifestations of the disease due to relapses [25, 31]. Additionally, these assessments may not be suitable for early and sensitive detection of insidious and slowly evolving change of MS disability as seen in progressive MS. In a recent recommendations for the diagnosis, management and treatment in multiple sclerosis (MS), Giovannoni *et al.* (2016) highlight the importance of regularly monitoring MS disease activity to better inform disease management technique and develop better therapeutic interventions [32]. The authors further emphasise the importance of early interventional treatments for PwMS. For instance, earlier identification of changes in PwMS impairment are important to identify and provide better therapeutic strategies [33].

2.3 Digital Biomakers

The expression *"biomarker"*, a portmanteau of "biological marker", refers to a broad subcategory of medical signs —-that is, objective indications of medical state observed from outside the patient— which can be measured accurately and reproducibly [117]. Biomarkers are often diagnostic tools and surrogate endpoints in clinical research [118–120]. Traditionally, biomarkers are generally paraclinical and directly measure biological properties indicative or predictive of health or disease status (for example, blood tests). Specifically in PwMS, some examples of potential biomarkers include the measurement of neurofilament levels in cerebrospinal fluid (CSF), which could indicate axonal degeneration [121] [25]. As mentioned previously in section 2.2, MRI T1 and T2 white matter lesions, or brain volume loss (atrophy) could be considered biomarkers of disease progression in MS, as well as thinning of the RNFL viewed using OCT.

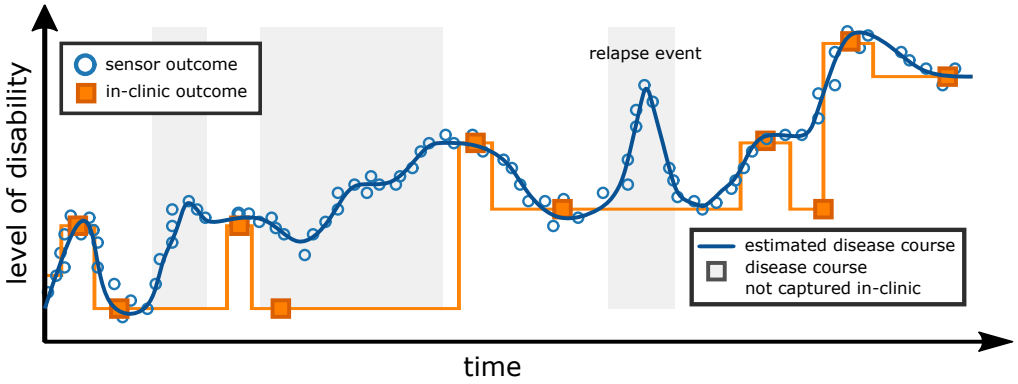


Figure 2.4: Digital outcomes for remote MS-patient monitoring. Digital sensor-based outcomes created from digital health technologies, such as smartphone-based sensor outcomes, enable clinicians to continuously monitor the disease course of people with Multiple Sclerosis (PwMS) outside of the clinic. These objective sensor-based measures can be administered remotely and at high-frequency to PwMS. This granularity of assessment may help capture subtle manifestations of habitual MS, as well as the disease course and relapse events that are currently missed by infrequent and less sensitive in-clinic measures. Diagram adapted from: [128].

Digital health technology tools are sensor-based technologies, such as apps, smartphones and wearables that can collect a variety of biological (e.g. blood glucose), physical (e.g. heart rate), pseudo-physical (gait, dexterity) or behavioural (e.g. GPS[3]) measurements from the human body [34]. *Digital biomarkers* can therefore be defined using digital technologies as: *"objective, quantifiable, physiological, and behavioural measures that are collected by means of digital devices that are portable or wearable"* [35–37].

The potential assessment of neurodegenerative diseases, such as MS [122], Parkinson's disease (PD) [123], Huntington's disease (HD) [124] and Alzheimer's disease (AD) [125] could be advanced by the emerging field of digital biomarkers [37]. Importantly, digital biomarkers offer the unique opportunity to overcome some of the limitations of current clinical outcome measures for neurodegenerative diseases. Those limitations with respect to MS assessments are introduced in section 2.2.1, such as intra- and inter-rater variability, infrequent administration, and are coarse grained. The significant advantages of digital biomarker outcomes are thus the ability to collect objective, rich, and sensitive health-related measurements that can be collected at high frequency and remotely [34, 126, 127].

[3]Global Positioning System.

2.3.1 Developing Digital Biomarkers from Digital Sensor Data

To develop digital biomarkers from digital technologies, objective and ecologically valid assessments, and subsequent measurements calculated from these, must be first developed. "Active tests" represent a standardised task which can be executed remotely and recorded by digital sensor data [31, 34]. These measurements can be used to derive digital outcomes which are predictive of a disease status or progression. Passive data could also be recorded, which does not require the individual's explicit active participation in a task [34].

Consumer wearable sensors (such as smartphone and smartwatches) could be used to administer active tests, record passive data, and create digital outcome measures [34]. These digital health technologies could be used to monitor disease activity ubiquitously, more subtly, and remotely in PwMS [31, 122, 129–133].

2.3.2 Remote Digital Biomarker Assessments

Remote digital biomarker outcomes using smartphone and smartwatches can window a patient's habitual disease state, giving clinicians the ability to both actively monitor a patient's disease symptoms in an objective way during a routine of active tests, and to passively monitor a patient over sustained periods unobtrusively [34, 128]. These new technologies offer new data-driven metrics for clinical decision-making during in-clinic visits, and facilitate out-of-sample assessment and prediction in out-of-clinic settings [122]. Figure 2.4 illustrates the clinical utility of digital outcomes for remote MS-patient monitoring. A practical implementation of smartphone-based monitoring was exemplified in a recent investigation by Bove *et al.* (2015) [130]. In this study, PwMS (n=38) were asked to complete daily tasks on a smartphone over a year period. The results from this study suggested that sensor data captured useful representations of PwMS disability and demonstrated the feasibility of administering out-of-clinic assessments using digital devices.

2.3.3 Digital Sensor Data to Measure Upper Extremity Function

Manual dexterity assessments are easy to implement using smartphone touch screen sensors. Many standard in-clinic tests can be augmented to a digital format, easily allowing for a remote assessment of dexterity, and can potentially extract a richer characterisation of hand-motor function performance than using manual tests alone. For instance the drawing of shapes, and specifically the drawing of spirals (spirography), is of particular interest since it represents a direct quantitative test transformation of a commonly employed qualitative clinical method to analyse patient's handwriting on paper [69]. Digital spirals, typically drawn using a stylus, are a common method used to analyse impairment in subjects with neurodegenerative diseases [19, 131, 134–139]. By digitising these shapes, not only can the visual structure of the shape drawn be analysed (such as how accurate the shape drawn was), but because a smartphone touch screen samples finger movement continuously, a higher resolution picture of *how* the shape was drawn, both spatially and temporally can also be recorded.

Prior efforts to develop digital drawing platforms for clinical use have been mainly focused on assessing patients with Parkinson's disease through spirography [134, 137, 139–145]. However, Longstaff *et al.* (2006) assessed spiral drawing performance in PwMS [138] and Feys *et al.* (2007) used digital circle, square and spiral drawings to quantify MS intention tremor [136]. Vianello *et al.* (2017) have developed a smartphone-based application in healthy elderly subjects using a variety of shapes that subjects had to draw [146]. Recently, digital spirals have been described as part of a validation study for the self-assessment of PwMS versus the MSFC, which composed of a larger smartphone-based phone application suite also testing gait, cognition and vision [131]. Other studies in PwMS have implemented similar digital manual dexterity tasks using smart tablet-based devices [147].

A major limitation of these studies however is that drawing tasks are often performed in-clinic under clinical supervision. As such there is a need to examine the capability of digital sensor data to measure upper extremity function in PwMS

outside of the clinic; the reliability of remotely administered UE assessments, the ability to capture MS UE disease symptoms, as well as the relationship to MS UE in-clinic assessments.

2.3.4 Digital Sensor Data to Measure Ambulatory Function

Digital sensor data can be applied to measure ambulatory gait function in a range of neurological disorders and specifically in PwMS [122]. Besides the observer- and stopwatch-based T25FW, 6MWT and 2MWT clinical assessments, a richer characterisation of ambulation in PwMS is typically conducted using specialised in-clinic gait technologies. These range from observer-based video-assessments [58] to high-resolution camera based motion sensors [59, 148, 149] or footfall sensor-based walkways (such as Gaitrite) [56, 64] and can capture a high-fidelity understanding of gait function in PwMS. Many of the MS gait-related clinical characteristics introduced in section 2.1.3 are investigated using these systems. In recent years there has been a shift towards the adoption of body worn sensors instead to objectively evaluate gait performance, circumventing the need for resource-intensive and expensive gait laboratory equipment [55, 122, 150–155]. Often termed inertial measurement units (IMU), or inertial sensors, these wearable devices generally contain an accelerometer, gyroscope, and magnetometer.

Recent studies by Spain *et al.* (2012) have demonstrated how wearable IMUs affixed to the wrists, shank and the trunk captured upper and lower body characteristics related to dynamic balance during ambulation [55]. These IMU-based measures were found to significantly differentiate PwMS and HCs ($p<0.05$) compared to standard stop-watch timed tests such as the T25FW and Timed-Up-and-Go (TUG) test. Further studies by Spain *et al.* (2014) followed PwMS over an 18-month period, applying discrete IMU-based gait assessments every 6 months. Results from this study again demonstrated that IMU gait features can stratify across HC, PwMSmild and PwMSmod groups, and indicated that PwMS have a higher inertial sensor gait feature variability than HC. This is akin to results presented by Huisinga et al (2013) who observed accelerometery gait features to be more variable in PwMS than

in HC [54]. Comparable work by Greene *et al.* (2015) have likewise demonstrated that PwMS can be distinguished from HC by modelling gait features from shank mounted inertial sensors using a cross-sectional analysis of the TUG test [153].

Many studies employing wearable inertial sensors in PwMS have attempted to characterise biomechanical properties of gait, for example: stride, swing and support times, or range of motion angles [55, 153–155]. Some studies have proposed extracting specific parameters of gait from inertial sensors to examine in PwMS [156]. These features are related to now well-defined global domains of gait function, such as rhythm, variability, and balance and coordination [157–160].

In parallel, other studies have demonstrated how signal-based features (i.e. not directly attempting to calculate a biomechanical property, such as step length) can instead be used as proxies to measure gait function from inertial sensors in PwMS [63, 153, 161, 162]. Features can range from the simple moments of the sensor signal (e.g. the mean or variance) [153] to more complex signal abstractions such as the kernel density estimations [161, 163]. Related follow-up work has additionally validated how fatigue-related gait features can be extracted from inertial sensors affixed to PwMS during the 6MWT [164]. Other proxies to capture gait-related parameters in PwMS IMU data have proposed entropy-based signal complexity metrics [63, 165].

Despite this, the majority of any wearable-based studies of MS gait in the literature have explored ambulatory ability using multiple inertial sensors during fixed lengths of controlled walking, in-clinic. These studies have focused on collecting sensor data from PwMS in highly controlled environments, under clinical supervision. As introduced in section 2.3.2, assessments and tasks that can be administered remotely may potentially capture unseen characteristics of PwMS disease. One of the key advantages of wearable IMU devices are that they offer the opportunity to measure ambulatory function remotely and out-of-clinic. As such, the work presented in this thesis will aim to evaluate the capability of remote out-of-clinic assessments using consumer wearable devices, such as a single smartphone or smartwatch to monitor PwMS ambulation. Recent work by Shah *et al.* (2020) has also suggested that out-of-clinic (free-living) gait characteristics may be especially pertinent of

PwMS pathology [166]. In another study, some of those authors have demonstrated that PwMS walked at a slower pace in daily living than in the laboratory, and suggested that gait parameters estimated from short walking bouts during free-living may be the most informative methodology to quantify level of PwMS disability [167]. Separately, other authors have demonstrated the accuracy of remotely captured IMU gait features to calculate at-home walking speed, which they determined was reflective of in-clinic walking speed, validated using T25FW times [168].

2.3.5 Smartphones and Smartwatches to Characterise Ambulatory Function

Smartphone and smartwatches are uniquely equipped with IMUs, thereby enabling these technologies to capture gait, activity, and physical function related inertial sensor-data [169], in- and out-of-clinic. Although advances have been made, there is a relative research vacuum in terms of utilising smartphone- and smartwatch-based IMU data to capture gait function characteristics in PwMS. Much of the existing literature has explored the use of these devices to quantify gait function in Parkinson's disease patients. [6, 170–172]. Specifically, in PwMS, Balto *et al.* (2015) validate that smartphone and smartwatches can be used to record steps, and that some step counts correlate to disability scores, albeit weakly [129]. Related studies indicate, that instead, a daily remote step count monitoring using smartwatches and wearables can distinguish levels of PwMS disability [173, 174]. These studies however have focused on holistic activity levels, which can be considered as *macro-gait*, rather than quantifying individual characteristics of PwMS during ambulation (*micro-gait*). As introduced in sections 2.1.3 and 2.3.4, it is *micro-gait* ambulatory patterns (for example, cadence or step length) instead that are examined during in-clinic MS-gait assessments. As such, this work will aim to examine the capability of remotely executed ambulatory assessments to capture PwMS *micro-gait*.

Smartphones and smartwatches have already been applied as digital tools for other useful ambulatory-related tasks. Human activity recognition (HAR) using smartphones (i.e. discerning between ambulatory activities such as walking, jogging,

climbing stairs and from other static tasks such as sitting down) is a rapidly growing field [175–181]. Another exploding research area is associated with subject identification and gait pattern recognition tasks using smartphones [182–184]. Both HAR and gait pattern recognition problems are related fields of interest to this thesis to help develop digital biomarkers for PwMS, and are expanded further in chapter 7.

2.3.6 Using Machine Learning to Construct Digital Biomarkers

Machine learning (ML) and statistical modelling have remarkably advanced the development of digital biomarkers [5]. Models can be built on digitally generated data to infer and predict present and future properties of disease such as the likelihood and severity, as well as the longitudinal trajectory of the disease course. The application of ML to develop digital biomarkers from smartphone and wearable-generated data in Parkinson's disease (PD) is beginning to flourish [6, 170, 171, 185–191]. Studies have already successfully used ML to distinguish PD from HC subjects [186, 189, 190], assess PD symptom severity [170, 188] and the effect of therapeutic treatments [171, 191], as well as longitudinally characterising PD performance [187].

The use of ML to develop digital biomarkers in PwMS is relatively a nascent space, however. The majority of studies integrating ML, and specifically deep learning (DL)[4], have been applied within the clinical imaging domain. Many studies have successfully automated MS lesion segmentation in MRI images, for instance using supervised [192] and unsupervised [193] convolutional neural network (CNN) architectures. Other studies have demonstrated the ability of CNN networks to identify the presence of MS from MRI images alone [194]. Similar preceding work by Zhang *et al.* (2016) established MRI imaging features to classify PwMS from HC [195].

The capabilities to monitor PwMS have been explored by Zhao *et al.* (2017), who report that ML-based models can (modestly) predict a change (i.e. a clinical

[4]Considered a sub-field of ML, deep learning will be introduced later as part of chapter 7.

deterioration or absence thereof) in EDSS at a 5-year follow-up from baseline using a combination of demographic, clinical and MRI outcome data [196].

Focusing on time-series based measurements of physical function in PwMS, one recent study has indicated that healthy participants could be distinguished from milder and moderately disabled PwMS [197]. This study, by Gong *et al.*, utilised CNN-based deep networks to learn a representation of gait function from accelerometry data captured using multiple sensors during a prescribed in-clinic walking assessment. Related studies, by Sun *et al.* (2019), demonstrated how PwMS could also be distinguished from HCs using ML-modelling of sway measurements extracted from an instrumented force plate [165]. Other recent work has established how PwMS at risk of falls can be identified from wearable accelerometer data through modelling performed using DL [198].

2.4 Background Summary

The following summarises the background information necessary for the work presented in this thesis:

- Multiple sclerosis is a heterogeneous, progressive and mutable disease which can affect multiple domains of physical and cognitive function.
- The current methods to measure disease progression can be infrequently assessed, coarse-grained, or resource intensive;
- As digital devices become more ubiquitous worldwide in daily life, it is promising to utilise these technologies to digitally augment specific analogue in-clinic assessments, out-of-clinic, as well as passively monitoring activity level characteristics;
- Smartphones and smartwatches (or wearables) can be used to objectively measure the physical and cognitive symptoms of PwMS disability. This thesis will focus on the former aspect, specifically in the domains of manual dexterity and ambulatory function;

- Key advantages of remote digital assessments are that they can be (1) administered at high-frequency and (2) measured longitudinally in out-of-clinic environments. This may extract a richer representation of disease-based features in PwMS and capture sub-clinical characteristics of MS which may occur outside of the clinic;
- Developing objective digital outcome measures, or *digital biomarkers*, to characterise, predict and monitor disease progression in PwMS is a nascent field. This thesis aims to help build the foundations to develop suitable digital biomarkers to advance the diagnosis, estimation and longitudinal prediction of MS in a digital, remote, and objective fashion.

3

Essential Methodology & Technical Background

Contents

3.1 Introduction 28
3.2 Building a modular, scalable and parallelizable feature extraction pipeline 29
3.3 Exploratory Data Analysis 30
3.4 Confirmatory Data Analysis 34
3.5 Linear Methods for Regression and Classification . . . 41
3.6 Regularisation & Feature Selection 45
3.7 Non-Linear Methods for Regression and Classification 47
3.8 Gradient-Based Optimisation 54
3.9 Deep Learning and Neural Networks 56
3.10 Visualising Deep Networks 61
3.11 Evaluation Approaches and Metrics 64

3.1 Introduction

The key concept underpinning machine learning (ML) lies in the estimation of some variable or parameter, $\hat{\mathbf{y}} \in \mathcal{Y}$, from some measured data, $\mathbf{x} \in \mathcal{X}$, using a function (the model), such that $\hat{\mathbf{y}} = f(\mathbf{x})$, where $\mathcal{Y}$ and $\mathcal{X}$ denote the properties of the estimated variable and measured data (termed feature space), respectively. Throughout this thesis, typically y_i will refer to the i^{th} labelled observation in

$\mathbf{y} \in \mathbb{R}^{N\times 1}$ whereas $\hat{y}_i$ will refer to the i^{th} prediction in $\hat{\mathbf{y}} \in \mathbb{R}^{N\times 1}$; $\mathbf{x}_i$ are the i^{th} observations in $\mathbf{X} \in \mathbb{R}^{N\times P}$; where N are the number of observations and P are the number of variables (features). The parameters of the model $f(\mathbf{x})$ are "learned" from the data by minimising or maximising an objective function $E(\mathbf{w})$[1] over some training data, in the form:

$$E(\mathbf{w}) = \alpha \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(y_i, f(\mathbf{x}_i, \mathbf{w})) \tag{3.1}$$

where $\mathbf{w} \in \mathbb{R}^{1\times P}$ are the weights or parameters of a model. It is typical to minimise equation (3.1) over N observations with respect to some loss function $\mathcal{L}(y_i, \hat{y}_i)$, which aims to minimise the error between the true label and estimate, depending on $\mathcal{Y}$; α represents a constant which is often added for mathematical convenience when optimising over $E(\mathbf{w})$[2]; multiplying $1/N$ makes the loss invariant to N[3]. The example above describes *supervised*— ML where the outcomes of the observations $\mathbf{y}$ are known, whereas *unsupervised*— ML, refers to when $\mathbf{y}$ are unknown and $f(\mathbf{x})$ is optimised based on some statistic of $\hat{\mathbf{y}}$ instead, for example the variance.

3.2 Building a modular, scalable and parallelizable feature extraction pipeline

Smartphone-based raw physiological and behavioural signals contain a wealth of information which can be utilised to monitor disease severity and progression. The raw data by itself however is often highly dimensional and not directly useful, where it must be first transformed, summarised and described using *features* or *digital biomarkers*. Typically, the feature space $\mathcal{X}$ is collected from multiple observations and assembled into a design matrix, commonly denoted as $\mathbf{X} \in \mathbb{R}^{N\times P}$, with a corresponding response vector $\mathbf{y} \in \mathbb{R}^{N\times 1}$, where P are the number of features and N are the number of observations (see above 3.1).

[1] The nomenclature "objective function", "error function", "cost function" and "loss function" are often used interchangeably in the literature. In this thesis I refer to a loss function as measuring a specific penalty on a single prediction versus the ground truth, whereas an objective function is the function to be optimised during training, with an error / cost function being an objective function that is specifically minimised.

[2] For example adding $1/2$ to the least squares estimate of $E(\mathbf{w})$ (see section 3.5.1)

[3] (however is often omitted from notation for brevity).

This thesis outlines the extraction of multiple physiological and behavioural features relating to physical function using the data collected from the Draw-a-Shape (DaS) test and the Two-Minute Walk Test (2MWT) as part of the FLOODLIGHT PoC (FL) study assessments (see later in chapter 4 where this study, the assessments and datasets are introduced). Feature extraction is performed over a *'big'* quantity of data from multiple participants contributing repeated measurements over 24 weeks from multiple sources (i.e. various assessments, captured with smartphone and smartwatch devices, from multiple sensors). As such, this thesis lays the foundations to perform large, scalable, modular and parallelizable feature extraction.

The feature space, $\mathcal{X}$, for Draw-a-Shape (chapter 5) and the Gait 2MWT (chapter 6) are introduced in each respective chapter, where $\mathcal{X}$ for each modality are described in full in appendix A.4. The two main pillars of statistical data analyses are: (1) exploratory and (2) confirmatory approaches.

3.3 Exploratory Data Analysis

First, exploratory data analysis is performed to characterise the data, generally *qualitatively*, to motivate further analysis and techniques thereof.

3.3.1 Data Visualisation

Exploratory data analysis is especially pertinent when developing a useful feature space and should involve visualisation of the distributions within and the relationships between different components, features and measurements of the dataset. Besides standardised techniques to investigate the distributions of the data (histograms, boxplots etc), methods such as Principal Component Analysis (PCA) are especially useful to understand the underlying feature structure [199, 200]. Briefly, PCA applies an orthogonal linear transformation to a multivariate feature space, $\mathbf{X} \in \mathbb{R}^{N \times P}$, representing $\mathbf{Z}$ as orthogonal latent features that are ordered according to the proportion of the variance in $\mathbf{X}$ that they represent. Visualising these first few (principal) components gives an insight into the structure and variance of the feature space, as demonstrated in example figure 3.1c, figure

3.1d and figure 3.1e. PCA is often used as a dimensionality reduction transformation and is detailed more in (appendix A.2.1). Another useful approach to understand the structural relationship of the feature space is t-distributed Stochastic Neighbour Embedding (t-SNE) [201]. The concept is to embed high-dimensional points in low dimensions (typically 2-D) that represents and preserves the similarities between points. Explicitly, t-SNE allows a visualisation of highly dimensional data as a representation of similarities between data points as joint probabilities. This technique attempts to minimise a distance metric (for instance, the Kullback-Leibler (KL) divergence [202]) between these joint probabilities of the low-dimensional embedding and the high-dimensional feature space. t-SNE is described in more detail in (appendix A.2.2) and an example of which can be found later in figure 7.4.

3.3.2 Correlation Analysis

Correlation analysis is a useful method to determine the association or dependence between sets of random variables, such as the dependence between features, or for assessing a features' clinical utility by measuring the association to an established clinical metric.

3.3.2.1 Pearson's Correlation Coefficient

The Pearson's correlation coefficient [203, 204], r, (an eponym from mathematician, *K. Pearson*) is used to quantify the linear strength of correlation between two sets of random variables $\mathbf{x} \in \mathbb{R}^{N\times 1}$, and $\mathbf{y} \in \mathbb{R}^{N\times 1}$ is defined as:

$$r(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{N}(x_i-\bar{\mathbf{x}})(y_i-\bar{\mathbf{y}})}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{\mathbf{x}})^2}\sqrt{\sum_{i=1}^{N}(y_i-\bar{\mathbf{y}})^2}} \tag{3.2}$$

where $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are the sample means of $\mathbf{x}$ and $\mathbf{y}$ respectively. The values of r are bounded between $[-1, 1]$ signifying perfect negative (-1) or positive $(+1)$ correlation, where values close to zero signify weak or no correlation between $\mathbf{x}$ and $\mathbf{y}$.

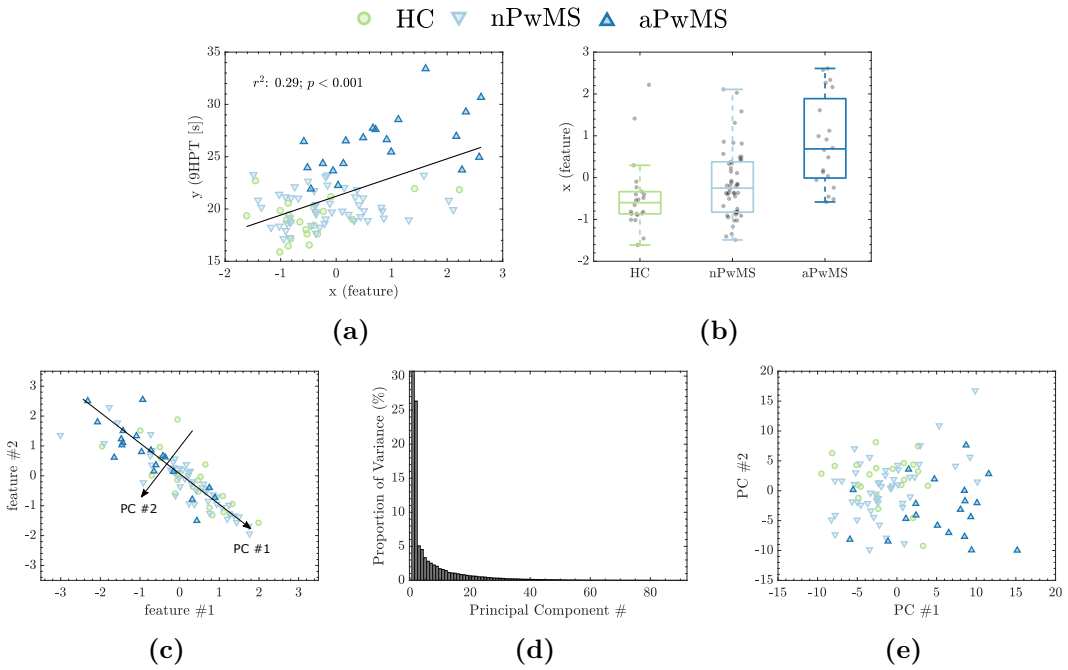


Figure 3.1: Overview of exploratory data analysis methodologies. Example data are taken from the Draw a Shape Test features as part of chapter 5, for $n = 93$ subjects, representing known subject groups: HC, nPwMS and aPwMS. Panel **(a)** illustrates the univariate correlation of feature $\mathbf{x}$ to response $\mathbf{y}$, the clinical Nine-Hole Peg Test (9HPT). A line of best fit is depicted in black, as well as the coefficient of determination, r^2 and corresponding $p-$value. Panel **(b)** describes the boxplot distribution of $\mathbf{x}$ individually for each group (KWt, $p < 0.001$). Boxplots summarise the distribution in $\mathbf{x}$, depicting the median, the 25^{th} to 75^{th} percentiles (termed the interquartile range). The whiskers extend the data points that are less than 1.5 times the interquartile range away. Data points beyond this are considered outliers. Panel **(c)** depicts the relationship between two correlated features ($r = -0.87$, $\rho = -0.90$, p< 0.001) with individual data-points coloured by known subject groups. Principal Component Analysis (PCA) identifies the two directions (PC #1 and PC #2) along which this 2-dimensional example have the largest variance. Panel **(d)** describes the variance of the principal components when PCA is performed on a large set of correlated features $\mathbf{X} \in \mathbb{R}^{93\times144}$. PCA aims to maximise the variance explained by the data and subsequently orders the components by the proportion of variance explained, thus the first 2 components explain $> 50\%$ of the variance in $\mathbf{X}$. Panel **(e)** represents the PCA with samples plotted in two dimensions using their projections onto the first two principal components, coloured by subject group. Panel **(a)** represents the 1^{st} feature and panel **(c)** represents the 2^{nd} and 4^{th} features, receptively, as described in the non-dominant feature ranking table 5.2, introduced later. The large set of (n=144) correlated features are also described later in figure 5.5. Note: feature values have been normalised to have unit variance using their respective mean μ and standard deviation σ.

3.3.2.2 Spearman's Rank Correlation Coefficient

An extension of Pearson's correlation r, is the non-linear Spearman's rank correlation coefficient, ρ, (an eponym from statistician, *C. Spearman*) which determines the monotonic dependence between two sets of two sets of random variables $\mathbf{x} \in \mathbb{R}^{N\times1}$,

and $\mathbf{y} \in \mathbb{R}^{N\times 1}$ [204, 205], and is defined by:

$$\rho(\mathbf{x}, \mathbf{y}) = 1 - \frac{6 \cdot \sum_{i=1}^{N} d_i^2}{N \cdot (N^2 - 1)} \tag{3.3}$$

where d_i is the difference between the two ranks of each x_i and y_i observation; $\rho \in [-1, 1]$ and can be interpreted in the same way as r.

3.3.3 Test-Retest Reliability

The intra-class correlation coefficient (ICC) is a widely used metric to quantify the test-retest reliability of repeated observations in the biomedical field [206–209]. The reliability of the feature values can be inferred if we consider each feature value over repeated test observations per subject. Using the ANOVA framework (to be introduced in section 3.4.2.7) specifically for random effects models, let $i = 1, ..., N$ subjects have $j = 1, ..., k$ repeated measures of $\mathbf{x}$ (i.e. a repeated feature value). In this work the $ICC(3, k)$ is implemented (Strout and Fleiss convention [206]) also denoted the $ICC(C, k)$ (Mc Graw and Wong convention [209]) which considers the two-way random average measures with k repeated measurements. Let:

$$x_{ij} = \bar{x} + r_j + c_j + \epsilon_{ij} \tag{3.4}$$

where x_{ij} is the i^{th} subject observation of the j^{th} repeated measurement; $\bar{\mathbf{x}}$ is the overall mean of $\mathbf{x}$; r_j and c_j are the unobserved random row-wise and column-wise effects for the j^{th} repeated measurement where $r_j \sim \mathcal{N}(0, \sigma_r^2)$ and $c_j \sim \mathcal{N}(0, \sigma_c^2)$; c_j is fixed however such that $\sum_j c_j = 0$ and can be ignored. The model errors are represented by ϵ_{ij}, where $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$. Therefore let the ICC be:

$$ICC = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_\epsilon^2/k} \tag{3.5}$$

For more information and expansion on how to compute the ICC we refer the readers to [206, 209]. Feature reliability can then be categorised as either poor (ICC<0.5), moderate (ICC=0.5-0.75), good (ICC=0.75-0.9) or excellent (ICC>0.9).

3.4 Confirmatory Data Analysis

Confirmatory data analysis is a process where explicit hypotheses are tested using statistical tools to *quantitatively* measure properties of the data such as inference, significance, and confidence.

3.4.1 Hypothesis Testing

A hypothesis defines an understanding about a population that is tested, for instance that two samples lie on the same distribution [204, 210, 211]. The *null hypothesis* (H_0) in essence represents an absence of a characteristic or the lack of an effect which forms the *alternative hypothesis* (H_1). The statistic compared varies with the type of test, although the distribution of the test statistic under H_0 should be be known (or assumed). The p-value represents the supremum probability of a *test statistic* being greater than a critical value under the null hypothesis than a value computed randomly from the population. If the p-value of a test is less than a pre-defined threshold, the *significance level* (typically $\alpha = 0.05$), then the H_0 is rejected in favour of H_1. The significance level α is often set through the number of permitted type I (rejections of true null hypotheses) or type II (failures to reject false null hypotheses).

3.4.2 Parametric versus Non-parametric hypothesis testing

Hypothesis tests can be divided into parametric versus non-parametric. Parametric tests make assumptions about the data, such as the assumption of Gaussianity, independence within the data, known variances or that the data is continuous [204, 211]. The most commonly known test statistic is the "t-statistic" as part of the *Student's* t-test, which was developed by William Sealy Gosset, a statistician working at the Guinness Brewery at St. James' Gate in Dublin, Ireland [212]. The t-distribution was developed to test if various batches of stout were different, however to circumvent Guinness's ban on publishing, Gosset reported his work under the pseudonym *"Student"* [213], hence the name: *"Student's* t-test".

3.4.2.1 *One-sample* and *Paired-sample* t-test:

The *one-sample* t-test investigates whether the sample(s), $\boldsymbol{x}$, comes from a normal distribution with mean equal to zero and unknown variance [213]. The test statistic is:

$$t = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \tag{3.6}$$

where $\bar{x}$ is the sample mean, μ is the hypothesised population mean, σ is the sample standard deviation, and n is the sample size. Under the null hypothesis, a p-value is calculated from the test statistic that has *"Student's* t-distribution" with $n-1$ degrees of freedom (using look-up tables). For more information on the *Student's* t-distribution, see appendix A.1.2. A *paired-sample* t-test follows (eq. 3.6) and tests whether two dependent (*"paired"*) samples $\boldsymbol{x}$ and $\boldsymbol{y}$ come from a normal distribution with mean equal to zero and unknown variance.

3.4.2.2 *Two-sample t-test*:

Tests H_0 whether two independent random samples, $\boldsymbol{x}$ and $\boldsymbol{y}$, come from normal distributions with equal means and equal (but unknown) variances. The test statistic is:

$$t = \frac{\bar{x} - \bar{y}}{\sigma\left(\frac{1}{n} + \frac{1}{m}\right)^{1/2}} \tag{3.7}$$

where $\bar{x}$ and $\bar{y}$ are the sample means, for n and m respective sample sizes, and the sample standard deviations are represented by σ.

Equal Variances (*Student's* t-distribution assumption): Under the assumption that the two data samples $\boldsymbol{x}$ and $\boldsymbol{y}$ are from populations with equal variance, σ is represented by the the pooled standard deviations, and under the assumption H_0 has a *Student's* t-distribution with degrees of freedom ($d = n + m - 2$):

$$\sigma = \left(\frac{(n-1)\sigma_x^2 + (m-1)\sigma_y^2}{n+m-2}\right)^{1/2} \tag{3.8}$$

where σ_x and σ_y are the specific sample standard deviations.

Unequal Variances (*Welch's* correction): Often termed *Welch's* t-test [214], in cases where it is not assumed that the two data samples are from populations with equal variance the test statistic is defined using equation 3.7:

$$t = \frac{\bar{x} - \bar{y}}{\left(\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right)^{1/2}} \tag{3.9}$$

The distribution of the test statistic under H_0 is approximated as a *Student's t*-distribution with a number of degrees of freedom given by Satterthwaite's approximation:

$$d \approx \frac{\left(\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right)^2}{\frac{\sigma_x^4}{n^2(n-1)} + \frac{\sigma_y^4}{m^2(m-1)}} \tag{3.10}$$

where d are the approximated number of degrees of freedom.

3.4.2.3 Chi-Square Goodness-of-Fit Test

The Chi-Square Test can be used to determine if the frequency between groups or categories in a population are different. By grouping the data into K distinct categories (or bins):

$$\chi^2 = \sum_{k=1}^{K} \frac{O_k - E_i}{E_k} \tag{3.11}$$

Here, O_k is the observed frequency for category k and E_k is the expected frequency for the k^{th} category. To calculate the expected frequencies for category k, $E_k = N \times p_k$, where p_k is the hypothesised proportion of observations in bin k. The test statistic follows an approximate chi-square (χ^2) distribution when the counts N are sufficiently large, for $n \to \infty$. For more details on the χ^2-distribution, see appendix A.1.3.

3.4.2.4 Kolmogorov-Smirnov test

The Kolmogorov-Smirnov (KS) test is a non-parametric test of the null hypothesis that the population cumulative distribution function (cdf) of the data $\boldsymbol{x} \in \mathbb{R}^{N\times 1}$ is equal to a hypothesised cdf [215]. The test statistic is the maximum absolute difference between the empirical cdf calculated from $\boldsymbol{x}$ and the hypothesised cdf:

$$D^* = \max_{\boldsymbol{x}} \left(|\hat{F}(\boldsymbol{x}) - G(\boldsymbol{x})| \right) \tag{3.12}$$

where $\hat{F}(\boldsymbol{x})$ is the empirical cdf and $G(x)$ is the cdf of the hypothesised distribution, for instance a Gaussian, $\boldsymbol{x} \sim \mathcal{N}(\mu, \sigma^2)$. The test statistic, D^* of the null hypothesis is then evaluated against critical values obtained from a table at a specified α significance value. The two-sample KS test can be evaluated through substituting $G(x)$ for $\hat{F}_2(\boldsymbol{y})$, the empirical cdf of a second sample data $\boldsymbol{y} \in \mathbb{R}^{N\times 1}$.

3.4.2.5 *One-sample* and *Paired-sample* Wilcoxon signed rank test:

The Wilcoxon signed rank test (WSRT) is a non-parametric equivalent to the t-test, which utilises the median of $\boldsymbol{x}$ based on ranks [210, 211]. The sum of the ranks (in ascending order) of the positive differences between each observation and a hypothesised median value $\mu_{1/2}$ (typically 0) is defined by W. For paired samples W can be represented as:

$$W = \sum_{i=1}^{N_r} sign(d_i) \cdot R_i \tag{3.13}$$

where d_i are the difference between each pair of samples, the rank is denoted as R_i and have been sorted in ascending order according to $|d|$. N_r signifies the number of paired samples when pairs resulting in $|d| = 0$ have been removed.
The p-value can then approximated from the $z-$statistic:

$$z = \frac{W - n(n+1)/4}{\left(\frac{n(n+1)(2n+1)}{24}\right)^{1/2}} \tag{3.14}$$

where n is the sample size. See algorithm implementation for further reading, including the methodology for tie adjustments [210]. For more information on the z-distribution see appendix A.1.1.

3.4.2.6 Wilcoxon rank sum test or Mann-Whitney U test:

The Wilcoxon rank sum test, which is equivalent to the Mann-Whitney U test [216], is a non-parametric test for the equality in population medians of two independent samples $\boldsymbol{x}$ and $\boldsymbol{y}$ with sample sizes n and m [210, 211]. For cases when $n < m$, the test statistic is performed on the first sample and vice-versa. The two independent samples $\boldsymbol{x}$ and $\boldsymbol{y}$ are first ordered, where the Mann-Whitney U-statistic, U, is defined

as the number of times a y_i precedes an x_i through the ordering. Equivalence of U to the Wilcoxon rank sum statistic for one-sample is defined by:

$$U = W - \frac{n(n+1)}{2} \tag{3.15}$$

where n defines the sample size and W is defined by equation 3.13. The $p-$value then is approximated using the z-statistic (when $n < m$):

$$z = \frac{W - \left(\frac{n \cdot m + n(n+1)}{2}\right)}{\left(\frac{n \cdot m(n+m+1)}{12}\right)^{1/2}} \tag{3.16}$$

See algorithm implementation for further reading, including the methodology for tie adjustments [216]. For more information on the z-distribution see appendix A.1.1.

3.4.2.7 One-way analysis of variance

The one-way analysis of variance (ANOVA) tests for differences between the group means of multiple groups [217]. For $i = 1, 2, ...N$ total observations comprised from $j = 1, 2, ...k$ groups, y_{ij} is an independent observation for the j^{th} group; with a group mean of $\bar{y}_j$ and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ are the random errors, ANOVA can be represented as a generalised linear model (see section 3.5):

$$y_{ij} = \bar{y}_j + \epsilon_{ij} \tag{3.17}$$

ANOVA tests for differences between-group means by partitioning the total sum of squares ($\mathrm{SS_{tot}}$) into sum of squares due to between-groups effect ($\mathrm{SS_R}$) and sum of squared errors ($\mathrm{SS_{err}}$), $\mathrm{SS_{tot}} = \mathrm{SS_R} + \mathrm{SS_{err}}$, such that:

$$\mathrm{SS_{tot}} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y})^2 = \sum_{j=1}^{k} n_j(\bar{y}_{.j} - \bar{y}_{..})^2 + \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_{.j})^2 \tag{3.18}$$

where $\bar{y}_{.j} - \bar{y}$ is the variation between groups; $\bar{y}_{..}$ is the overall sample mean; $y_{ij} - \bar{y}_{.j}$ estimates the within group variation; and n_j is the sample size for the j^{th} group; The test statistic for ANOVA then follows an F-distribution with $(k-1, N-k)$ degrees of freedom:

$$F = \frac{(N-k)\sum_{j=1}^{k} n_j(\bar{y}_{.j} - \bar{y})^2}{(k-1)\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_{.j})^2} \tag{3.19}$$

For more information on the F-distribution see appendix A.1.4.

3.4.2.8 Kruskal-Wallis Test

The Kruskal-Wallis (KWt) test by ranks is a non-parametric version of classical one-way ANOVA (3.18), and an extension of the Wilcoxon rank sum test (3.15) to more than two groups whereby it compares the medians of the groups of data [218]:

$$H = (N-1)\frac{\sum_{j=1}^{k} n_j(\bar{r}_j - \bar{r})^2}{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(r_{ij} - \bar{r})} \tag{3.20}$$

where n_j is the sample size for the j^{th} group; r_{ij} is the rank of the i^{th} observation of the j^{th} group over all observations; $\bar{r}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} r_{ij}$ is the average rank of all observations in group j; $\bar{r} = \frac{N+1}{2}$ is the average of all r_{ij}; and N is the overall sample size. If the data contains no ties (3.20) reduces to

$$H = \frac{12}{N+1}\sum_{j=1}^{k}\frac{\bar{r}_j}{n_j} - 3(N+1) \tag{3.21}$$

where k are the number of groups. In the case of ties, a correction can be applied by dividing H by:

$$1 - \frac{\sum(t^3 - t)}{N^3 - N} \tag{3.22}$$

where t designates the number of tied observations in a group of tied value. The F-statistic used in classical one-way ANOVA is replaced by a chi-square (χ^2) statistic for the KWt.

3.4.2.9 Brown-Forsythe Test

The Brown-Forsythe (BF) test by medians is used to investigate if various groups of data have been drawn with equal variances [219]. The (non-parametric) BF test statistic is the F-statistic derived from a one-way ANOVA (3.18) on the absolute deviation of the groups from their respective medians:

$$z_{ij} = |y_{ij} - \tilde{y}_j| \tag{3.23}$$

where $\tilde{y}_j$ defines the median of the j^{th} group. The F-statistic can therefore be defined under:

$$F = \frac{(N-k)\sum_{j=1}^{k} n_j(\bar{z}_{j.} - \bar{z})^2}{(k-1)\sum_{j=1}^{k}\sum_{i=1}^{n_j}(z_{ij} - \bar{z}_{j.})^2} \tag{3.24}$$

where N are the total number of observations; k are the number of groups; n_i are the number of observations in the j^{th} group; $\bar{z}_{j\cdot}$ are the group means of z_{ij}, and $\bar{z}$ is the overall mean of z_{ij}.

3.4.3 Correcting for multiple hypothesis testing

When performing multiple hypothesis testing (i.e. making multiple comparisons using the same test) the likelihood of type I errors increases, which can be characterised in H_0 by the false discovery rate (FDR) [220, 221]. The FDR is arguably a more flexible control of type I errors contrasted to family-wise error rate (FWER) controlling (such as the harsh Bonferroni correction [222]) [220, 221, 223]. This thesis therefore uses methods to control the FDR at level α (see 3.4.1) using the linear step-up procedure introduced by Benjamini and Hochberg (BH) [223]. Briefly, for a N number of multiple hypothesis tests performed, $H_0^1, H_0^2, ..., H_0^N$, and corresponding p-values, p_1, p_2, ...p_N, which are ranked in ascending order, the BH procedure controls the FDR at α. For a given α, it does so by finding the largest k such that $p_{(k)} \leq \alpha \cdot (\frac{k}{N})$, where all $H_0^{(i)}$ for $i = 1, ..., k$ are now rejected to report the *a priori* probabilities after controlling for the FDR.

3.5 Linear Methods for Regression and Classification

A regression model explicitly describes a relationship between predictor(s) $\mathbf{X} \in \mathbb{R}^{N \times P}$ and continuous response variables $\mathbf{y} \in \mathbb{R}^{N}$, the most basic of which is *linear regression*[224]. For an i^{th} observation row of $\mathbf{X}$, $\mathbf{x} \equiv \mathbf{x_i} \in \mathbb{R}^{1 \times P}$:

$$\hat{y} = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_P x_P + \epsilon \tag{3.25}$$

$$= w_0 + \sum_{j=1}^{P} w_j x_j + \epsilon \tag{3.26}$$

$$= \mathbf{w}^\top \mathbf{x} + b \tag{3.27}$$

$$= \mathbf{w}^\top \mathbf{x} \tag{3.28}$$

where w_j values denote the slope (weights, or regression coefficients) of the x_j features; w_0 is the intercept term; and ϵ denote the residual (model) errors term, which are assumed to be normally distributed with constant variance, $\epsilon \sim \mathcal{N}(0, \sigma^2)$ [200, 224]. Often a linear model is described in vector notation (3.27), $\mathbf{w} = [w_1, w_2, ..., w_P]$, where w_0 is denoted as the "bias", b term, and the ϵ-term is often omitted. More succinctly, a linear model can be also represented by the equation (3.28) where $\mathbf{w} = [w_0, w_1, w_2, ..., w_{P+1}]$ is a vector of regression coefficients, including w_0 (or b) as the first value and $\mathbf{x} \in \mathbb{R}^{1 \times (P+1)}$ which is first concatenated with a vector column of ones to account for w_0 (or b) in $\mathbf{w}$ at the first index.

3.5.1 Ordinary Least Squares

Ordinary Least Squares (OLS) is the most common method of estimating the parameters $\mathbf{w}$ of (3.26) to find $\hat{\boldsymbol{w}}$ [200, 224, 225]. OLS minimises the error $E(\mathbf{w})$ (3.1), with a loss $\mathcal{L}(y_i, \hat{y}_i)$ represented by the residual sum of squares (SS_{res}), or the squared "residual errors" (e):

$$\mathcal{L}(y_i, \hat{y}_i) = \sum_{i=1}^{N} |y_i - \hat{y}_i|^2 = \sum_{i=1}^{N} e_i^2 \tag{3.29}$$

$$\hat{\boldsymbol{w}} = \arg\min_{\mathbf{w}, w_0} \{ \frac{1}{2} \sum_{i=1}^{N} |y_i - w_0 \sum_{j=1}^{P} w_j x_{ij}|^2 \} \tag{3.30}$$

where $\hat{\boldsymbol{w}} \in \mathbb{R}^P$ are estimated regression coefficients (containing both w_0 and w) for N samples and P number of features which can be solved using linear algebra[4][200]. This is akin to maximising the log-likelihood which we will discuss in section 3.5.3. In the literature, models can also frequently be represented in linear algebra (matrix) notation: $\mathbf{y} = \mathbf{W}\mathbf{X} + \boldsymbol{\epsilon}$, where $\mathbf{y} \in \mathbb{R}^{N\times 1}$, $\mathbf{X} \in \mathbb{R}^{N\times(P+1)}$ and $\mathbf{W}, \boldsymbol{\epsilon} \in \mathbb{R}^{1\times(P+1)}$. Least squares models can then be solved computationally, typically using the *Moore–Penrose* inverse[5] $\mathbf{X}^+ = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\top$, where: $\hat{\boldsymbol{W}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$.

3.5.2 Iteratively Re-Weighted Least Squares

In OLS it is typically assumed that errors follow a normal distribution that are free from outliers, $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Considering that the squared residuals are penalised (3.29), OLS is therefore sensitive to the presences of outliers [200]. A robust least-squares regression using iteratively re-weighted least-squares (IRLS) can minimise biases caused by outliers in the data [226, 227]. Estimation with IRLS is also more robust than OLS when there is a strong suspicion of heteroscedasticity within the data. First, the IRLS algorithm at step $t = 0$ involves solving the weighted linear least squares:

$$\hat{\boldsymbol{w}}^{(t+1)} = \arg\min_{\mathbf{w}}(\sum_{i=1}^{N} \alpha_i^{(t)}|y_i - w_0\sum_{j=1}^{P} w_j x_{ij}|^2) \tag{3.31}$$

where α is the diagonal matrix of weights[6], usually with all elements set initially to: $\alpha_i^{(0)} = 1$. Next, the initial least-squares residuals e_i are determined to compute the updated weights $\alpha_i^{(t)}$; in this case a "bisquare weighting" method is used, given by:

$$\alpha_i^{(t)} = \begin{cases} (1 - (e_i')^2)^2 & |e_i'| < 1 \\ 0 & |e_i'| \geq 1 \end{cases} \tag{3.32}$$

where e_i' are given by the Studentised adjusted residuals:

$$e_i' = \frac{e_i}{K\sigma\sqrt{1 - h_{ii}}} \tag{3.33}$$

[4]The 1/2 serves for mathematical convenience to solve the equation using the derivative of $E(\mathbf{w})$, however will often be omitted from notation for brevity.

[5]Coincidentally, Roger Penrose jointly won the 2020 Nobel Prize in Physics during the writing of this chapter, for his work demonstrating that black hole formations are a robust prediction of the general theory of relativity!

[6]To avoid confusion I term w_i as the regression coefficients and α_i as the wights for IRLS.

where e_i are the least-squares residuals (3.30); K is a tuning constant equal to 4.685; σ is an estimate of standard deviation of the error term given by $\sigma = MAD/0.6745$, where MAD is the median absolute deviation of the residuals where the constant 0.6745 makes the estimate unbiased for a normal distribution [227]. Leverages, h_{ii}, are a measure of the influence of a given observation x_i on $\hat{y}_i$. The leverages are computed from Gram–Schmidt QR decomposition[7] using the projection matrix H (called the hat matrix, see [200]); in (3.33) h_{ii} adjust the residuals by reducing the weight of large h_{ii} for x_i. The iterations are performed and weights updated at $t = [1, 2, ...\tau]$ for τ iterations or if the fit converges before then. The IRLS models can be fit following the same formation as in section 3.5.1.

3.5.3 Logistic Regression and Classification

Generalised linear models (GLMs) are extensions of linear regression models that can have non-linear outputs [228]. GLMs utilise canonical link functions, ϕ, to transform the outputs of a linear regression: $\varphi = \mathbf{w}^\top \mathbf{x}$ (3.28) to another distribution, such as with a logistic $\phi = \sigma(\varphi)$ link function (or inversely the logit, representing the log-odds) which will be used to form Logistic Regression for binary classification tasks [200, 229]:

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

in this case ϕ is sigmoidal and is bounded between $[0, 1]$, therefore the output of σ can be interpreted as the probability of $y = 1$:

$$p(\mathbf{x}; \mathbf{w}) = \frac{1}{1 + e^{-(\mathbf{w}^\top \mathbf{x})}} \tag{3.35}$$

A threshold can be applied to the probabilistic output p to determine a classification prediction $\hat{y}$ for a Logistic Regression model; threshold values are typically chosen as 0.5, but this can be altered based on the use case.

Logistic regression models are usually trained by Maximum Likelihood Estimation (MLE) by maximising a likelihood function [200]. First for N observations,

[7] For more information we refer the reader to pg. 55 of Hastie *et al.* [200].

where y_i is the i^{th} response, with a conditional probability of $y_i = 1$ given by $p(\mathbf{x_i}; \mathbf{w})$, the likelihood is defined by:

$$p(\mathbf{y}|\mathbf{w}) = \prod_{i=1}^{N} p(\mathbf{x_i}; \mathbf{w})^{y_i} \{1 - p(\mathbf{x_i}; \mathbf{w})\}^{(1-y_i)} \tag{3.36}$$

The error function $E(\mathbf{w})$ can be defined by the negative logarithm of the likelihood (log-likelihood or often referred to as the cross-entropy) for actual $\mathbf{y}$ and predicted classes $\hat{\mathbf{y}}$ yielding[8]:

$$E(\mathbf{w}) = -\ln(p(\mathbf{y}|\mathbf{w})) = -\sum_{i=1}^{N} \{y_i \ln(\hat{y_i}) + (1 - y_i) \ln(1 - \hat{y_i})\} \tag{3.37}$$

$$= -\sum_{i=1}^{N} \{y_i \ln(p(\mathbf{x_i}; \mathbf{w}) + (1 - y_i) \ln(1 - p(\mathbf{x_i}; \mathbf{w}))\} \tag{3.38}$$

$$= -\sum_{i=1}^{N} \{y_i (\mathbf{w}^\top \mathbf{x}_i) - \ln(1 + e^{\mathbf{w}^\top \mathbf{x}_i})\} \tag{3.39}$$

To maximise the log-likelihood, we take the gradient of the error function with respect to $\mathbf{w}$ and set its derivatives to zero:

$$\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} = \nabla E(\mathbf{w}) = \sum_{i=1}^{N} \mathbf{x}_i (p(\mathbf{x}_i; \mathbf{w}) - y_i) \tag{3.40}$$

With linear regression discussed in section 3.5.1, maximising the likelihood, (or conversely minimising the SS_{res}), will lead to a closed-form solution, whereas for logistic regression this is no longer possible due to the non-linearity of the logistic sigmoid function σ. We therefore use the Newton-Raphson algorithm to solve and update $\nabla E(\mathbf{w})$, also known as the Jacobian $\mathbf{J}$ [230], (which is still convex). This also requires the second-derivative of $E(\mathbf{w})$ (often termed the Hessian matrix, $\mathbf{H}$):

$$\mathbf{H} = \frac{\partial^2 E(\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}^\top} = \nabla^2 E(\mathbf{w}) = \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^\top p(\mathbf{x}_i; \mathbf{w})(1 - p(\mathbf{x}_i; \mathbf{w})) \tag{3.41}$$

This iterative approach updates the weighting matrix $\mathbf{w}$, which $\mathbf{H}$ depends on and can be thought of as following the same weighted least squares problem proposed in section 3.5.2.

[8] Following the same nomenclature as section 3.5.1.

3.6 Regularisation & Feature Selection

3.6.1 ℓ_1 & ℓ_2 Regularisation

Many statistical and machine learning models can easily overfit to the training data, resulting in poorer estimations, models that are not generalisable or too complex. Regularisation is often introduced to mitigate against this. For example large coefficient values in a regression can be penalised by adding a regularisation term to a loss function, or through reducing the number of parameters or features used in a model. The most common regularisers use the ℓ_p-norm defined by [200, 231]:

$$||\mathbf{x}||_p = \left(\sum_{i=1}^{N} |x_i|^p\right)^{1/p} \tag{3.42}$$

for any $\mathbf{x} \in \mathbb{R}^{N\times 1}$, where the real number $p \geq 1$ defines the ℓ_p space. Combining equations (3.1) with (3.42) gives a generalised Lagrangian form[9] of an objective function $E(\mathbf{w})$ with ℓ_p-norm regularisation:

$$\arg\min_{\mathbf{w}} / \max \{E(\mathbf{w}) + \lambda||\mathbf{w}||_p\} \tag{3.43}$$

$$\arg\min_{\mathbf{w}} \left\{\alpha \sum_{i=1}^{N} \mathcal{L}(y_i, f(\mathbf{x_i}, \mathbf{w})) + \lambda||\mathbf{w}||_p\right\} \tag{3.44}$$

where $\lambda \geq 0$ is a tuning parameter defining how much the model is regularised; the larger the value of λ, the greater the amount of shrinkage. Figure 3.2 compares the regularisation of $\mathbf{w}$ using the $||\mathbf{w}||_1$ (ℓ_1) and the $||\mathbf{w}||_2$ (ℓ_2).

3.6.2 Ridge Regression

Ridge Regression (Tikhonov Regularisaion) [200, 231] is a technique which utilises the ℓ_2-norm to impose a penalty on the size of the coefficients in a linear regression (3.30) such that:

$$\hat{\boldsymbol{w}} = \arg\min_{\mathbf{w}} \left\{\sum_{i=1}^{N} |y_i - w_0 \sum_{j=1}^{P} w_j x_{ij}|^2 + \lambda \sum_{j=1}^{P} w_j^2\right\} \tag{3.45}$$

This is equivalent to minimising the sum of squares with a constraint of the form: $||w||_2 = \sum_j^N w_j^2 \leq t$ (where t controls the amounts of shrinkage; there is an exact

[9] (to be either minimised or maximised)

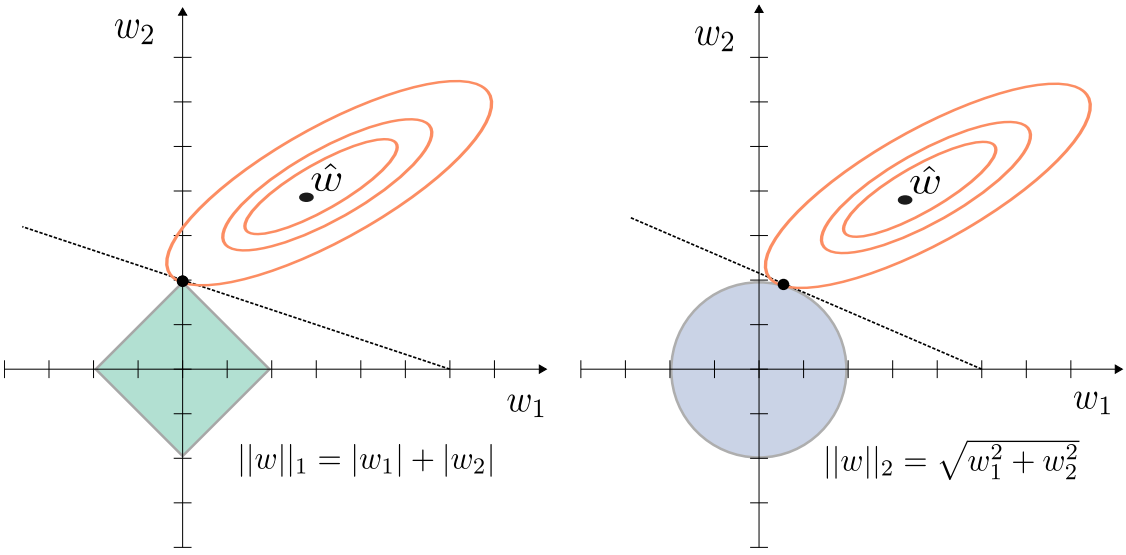


Figure 3.2: Diagram comparing ℓ_1 (left) and ℓ_2 (right) regularisation approaches for a two dimensional constraint. The solid areas represent the constraint regions for ℓ_1 (LASSO) and ℓ_2 (Ridge regression) respectively, whereas $\hat{w}$ and surrounding contours depict the usual (unconstrained) least-squares error function. The intersection of the error function and the optimal parameterisation of the model are highlighted with black dots, where it is evident that ℓ_1 regularisation yields a sparse solution where $w_2 = 0$. Diagram adapted from [231].

relationship between t and corresponding λ (3.44) which is data dependent). Figure 3.2 demonstrates the shrinkage induced by the LASSO penalty.

3.6.3 Least Absolute Shrinkage and Selection Operator

The Least Absolute Shrinkage and Selection Operator (LASSO) [231, 232] is a technique that conversely solves the ℓ_1-penalised sum of squares (3.30) in a linear regression such that:

$$\hat{\boldsymbol{w}} = \arg\min_{\mathbf{w}} \left\{ \sum_{i=1}^{N} |y_i - w_0 \sum_{j=1}^{P} w_j x_{ij}|^2 + \lambda \sum_{j=1}^{P} |w_j| \right\} \tag{3.46}$$

This is equivalent to minimising the sum of squares with a constraint of the form: $||w||_1 = \sum_j^N |w_j| \leq t$. Because of the form of the ℓ_1-penalty, LASSO both shrinks coefficients but also encourages sparsity in a model's parameters and thus inherently forms feature selection, shrinking non-important features to zero. Figure 3.2 demonstrates the sparsity induced by the LASSO penalty.

The LASSO can also be extended to perform feature selection for classification by substituting a canonical link function (such as the logistic $\phi = \sigma(\varphi)$ 3.34)

and following the same procedure outlined in section 3.5.3, essentially performing regularised-logistic regression (denoted *lassoglm* in this work) [231].

3.7 Non-Linear Methods for Regression and Classification

Methods for regression and classification introduced in this thesis thus far have considered the linear combination of features for regression and classification. The following section outlines non-linear classification and regression methodologies.

3.7.1 Support Vector Machines

Support Vector Machines (SVM) build upon the foundations introduced in section 3.5 by leveraging the simple linear relationship from equation (3.27) for binary classification tasks. Known as the Perceptron Classifier, when $y \in \{-1, +1\}$, if the classes are separable, a linear classification boundary (hyperplane) can be given by [200, 230, 233, 234]:

$$f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b \tag{3.47}$$

for an i^{th} observation row, $\mathbf{x}_i \in \mathbb{R}^{1\times P}$, $\mathbf{w} \in \mathbb{R}^{1\times P}$; the response variable is denoted as y_i, where $\mathbf{y} \in \mathbb{R}^{N\times 1}$, and where P are the number of features and N are the number of observations. The predicted class is denoted by $sign(f(\mathbf{x}_i))$; therefore $y_i f(\mathbf{x}_i) > 0, \forall i$ defines the function (3.47) which we can train to maximise the margin, M, between the two classes. Importantly, we can choose the normalisation of $\mathbf{w}$ such that the lines $\mathbf{w}^\top \mathbf{x}_+ + b = +1$ and $\mathbf{w}^\top \mathbf{x}_- + b = -1$ draw the boundaries or *support vectors* defining the positive and negative classes, where the total margin between the two classes is therefore, $2M = \frac{2}{||\mathbf{w}||}$ (demonstrated in figure 3.3). Learning a binary SVM can be formulated as an optimisation to maximise the margin such that:

$$\begin{aligned} &\arg\max_{\mathbf{w},b} \left\{ \frac{2}{||\mathbf{w}||} \right\} \\ &\text{subject to} : \begin{cases} \mathbf{w}^\top \mathbf{x}_i + b \geq 1 & \text{if } y_i = +1,\ \forall i \\ \mathbf{w}^\top \mathbf{x}_i + b \leq 1 & \text{if } y_i = -1,\ \forall i \end{cases} \end{aligned} \tag{3.48}$$

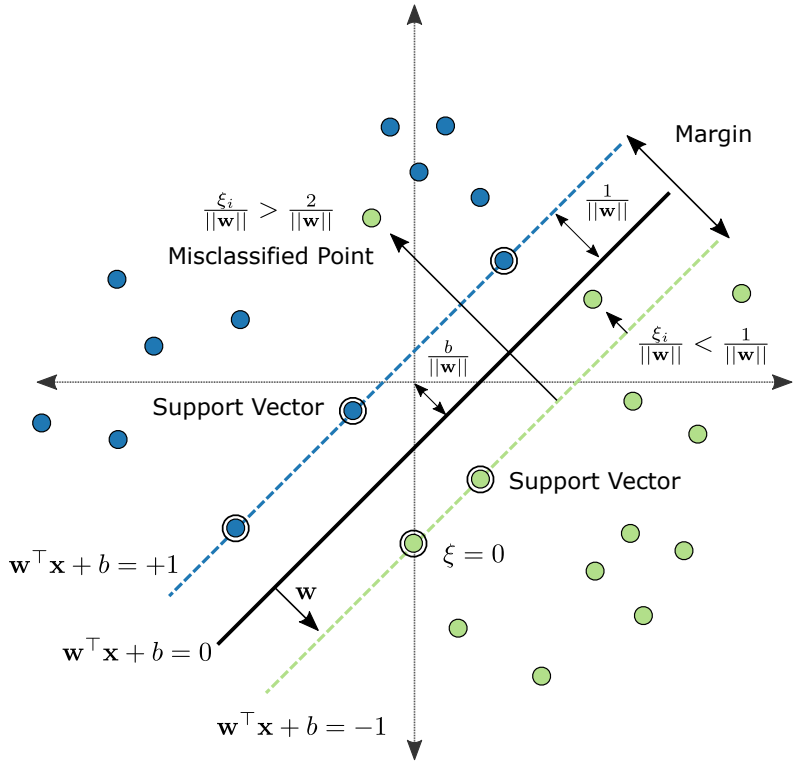


Figure 3.3: Diagram illustrating the SVM classifier for a two class problem visualised in two-dimensional feature space. The decision boundary is represented by a solid line, while the dashed lines represent the maximal margin width $2M = 2/||\mathbf{w}||$ where the support vectors along these margins are highlighted. Margin violations (i.e. on the wrong side of the respective class margins but in the correct side of the hyper-plane) are denoted by a point that is $0 < \xi_i \leq 1$. Misclassified points are represented by $\xi_i > 1$.

However, cases of completely separable linear classes in real-world data is very idealised. To account for this we introduce *slack variables*, ξ, which allow for some classes to overlap in the feature space; the idea is that the largest margin maximises the space between the classes (but allows some violations, hence "soft-margin") and will generalise better to real-world and test data. To find the hyperplane that creates the biggest soft-margin M:

$$\begin{aligned} &\arg\max_{\mathbf{w},b,\xi_\mathbf{i}} \left\{ \frac{2}{||\mathbf{w}||} + C \sum_{i=1}^{N} \xi_i \right\} \\ &\text{subject to :} \quad \xi_i \geq 0,\ y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq M - \xi_i,\ \forall i \end{aligned} \tag{3.49}$$

and maximising $||\mathbf{w}||^{-1}$ is equivalent to minimising $||\mathbf{w}||^2$ yielding[10]:

$$\begin{aligned} &\arg\min_{\mathbf{w},b,\boldsymbol{\xi}_\mathbf{i}} \left\{ \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N}\xi_i \right\} \\ &\text{subject to}: \quad \xi_i \geq 0,\ y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i,\ \forall i \end{aligned} \tag{3.50}$$

which is now a convex optimisation problem, where C is the regularisation cost parameter defining how "soft" the margin is yet allowing the maximal margin (i.e. how many violations (ξ_i) are permitted, where $\sum_i \xi_i \leq C$; points violate their margin by $M\xi_i$, whereas points on the correct side of the margin represent $\xi_i = 0$). This is a convex optimisation problem over $\mathbf{w}$, b, and $\boldsymbol{\xi}_\mathbf{i}$ which can be written in the (***primal***) Lagrangian form as:

$$\arg\min_{\mathbf{w},b} \left\{ C\frac{1}{N}\sum_{i=1}^{N}\max\left(0, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)\right) + \frac{1}{2}||\mathbf{w}||^2 \right\} \tag{3.51}$$

where the for each sample, satisfying the two constraints above allow ξ_i to be represented as the ***hinge loss*** in the form:

$$\mathcal{L}(y_i, \mathbf{x}_i; \mathbf{w}) = \max\left(0, 1 - y_i(f(\mathbf{x}_i)\right) = \begin{cases} \xi_i \geq 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b), & \forall i \\ \xi_i \geq 0, & \forall i \end{cases} \tag{3.52}$$

The SVM can also be expressed as a sum over it's *support vectors*, $\boldsymbol{\alpha}$, which is known as the dual form:

$$\textbf{Primal:} \qquad f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b \tag{3.53}$$

$$\textbf{Dual:} \qquad f(\mathbf{x}) = \sum_{i=1}^{N}\alpha_i y_i(\mathbf{x}_i^\top \mathbf{x}) + b \tag{3.54}$$

The SVM can therefore be formulated to learn a linear classifier by solving an optimisation problem over α_i for $\boldsymbol{\alpha} \in \mathbb{R}^N$, which provides us with a solution for the (***dual***) Lagrangian form:

$$\begin{aligned} &\arg\max_{\boldsymbol{\alpha}>\mathbf{0}} \left\{ \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{j,k=1}^{N}\alpha_j\alpha_k y_j y_k \mathbf{x}_j^\top \mathbf{x}_k \right\} \\ &\text{subject to}: \quad \sum_{i=1}^{N}\alpha_i y_i = 0,\ 0 \leq \alpha_i \leq C,\ \forall i \end{aligned} \tag{3.55}$$

[10]Note that $M = 1/||\mathbf{w}||$; and that $1/2$ is added for mathematical convenience used for optimisation as outlined in 3.8.

Learning a linear SVM *dual* classifier (3.55) is much more efficient than the *primal* (3.51), as it only requires learning N parameters to solve over $\boldsymbol{\alpha}$ and $(\mathbf{x}_j^\top \mathbf{x}_k)$, where $\boldsymbol{\alpha}$ is also sparse as only the support vectors, defined by $\alpha_i \neq 0$, determine the class margins needed for computation. The mathematical formulation of the *dual* representation of the SVM is described in appendix section A.2.3.

3.7.2 Application of Kernels to Support Vector Machines

Often data will not be linearly separable in a lower dimensional feature space $\mathbf{x} \in \mathbb{R}^d$, but may be separable in an alternative[11] or a higher dimension $\mathbf{x} \in \mathbb{R}^D$ [233, 234]. A kernel function $k(\cdot)$ can be used to create a non-linear feature space mapping $\Phi : \mathbf{x} \rightarrow \Phi(\mathbf{x}), \quad \mathbb{R}^d \rightarrow \mathbb{R}^D$, for instance, transforming $\mathbf{x}$ to a higher dimension. This is known as the "kernel trick", defined by [230, 234]:

$$k(\mathbf{x}, \mathbf{x}') = \Phi(\mathbf{x})^\top \Phi(\mathbf{x}') \tag{3.56}$$

where a kernel $k(\cdot)$ is a symmetric function of its arguments such that: $k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}', \mathbf{x})$. Therefore substituting (3.56) into the SVM primal (3.53) and SVM dual (3.54) yields:

$$\textbf{Primal:} \qquad f(\mathbf{x}) = \mathbf{w}^\top \Phi(\mathbf{x}) + b \tag{3.57}$$

$$\textbf{Dual:} \qquad f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i y_i \Phi(\mathbf{x_i})^\top \Phi(\mathbf{x}) + b \tag{3.58}$$

$$= \sum_{i=1}^{N} \alpha_i y_i k(\mathbf{x_i}, \mathbf{x}) + b \tag{3.59}$$

Again, the dual classifier can be easily learned in transformed feature space (by substituting the kernel trick 3.56 into 3.55) where only the N dimensional α is needed to be solved. Therefore using the kernel trick, an SVM classifier can be learnt for high dimensional features spaces, without actually having to map the points into the high dimensional space. There are many different kernel functions which can be utilised, however this thesis only investigated the two most common kernels typically implemented: a linear kernel:

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{x}' \tag{3.60}$$

[11] For instance where data is separable in polar coordinates rather than in Cartesian coordinates.

and a Gaussian radial basis function (RBF) kernel:

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(\frac{-||\mathbf{x} - \mathbf{x}'||^2}{2\sigma^2}\right) \quad \text{for } \sigma > 0, \tag{3.61}$$

where we denote $\gamma = 1/2\sigma^2$.

3.7.3 Support Vector Regression

Support Vector Machines can also be utilised to learn a regression model (denoted SVR) in linear or non-linear feature space (see the "kernel trick" discussed in 3.7.2) [230]. This can be achieved by introducing a ϵ-insensitive error measure (V_ϵ) to the objective function (3.1) (which can also be regularised, using equation 3.44):

$$E(\mathbf{w}) = \sum_{i=1}^{N} V_\epsilon(y_i - f(x_i)) + \frac{1}{2}||\mathbf{w}||^2 \tag{3.62}$$

V_ϵ gives zero error if the absolute difference between the prediction and response value, $|y_i - f(x_i)|$, (i.e. the ℓ_1 loss (3.42) is less than some ϵ, where $\epsilon > 0$, such that:

$$V_\epsilon(y_i - f(x_i)) = \begin{cases} 0 & \text{if } |y_i - f(x_i)| < \epsilon, \\ |y_i - f(x_i)| - \epsilon, & \text{otherwise} \end{cases} \tag{3.63}$$

This can also be written as $V_\epsilon(y_i - f(x_i)) = \max(0, |y_i - f(x_i)| - \epsilon)$

As before in section 3.7.1, slack variables can be introduced for points that violate the ϵ-insensitive error measure and learning is performed by optimising over (3.49):

$$\begin{aligned} &\min_{\mathbf{w}, \xi_i, \hat{\xi}_i} \left\{ C \sum_{i=1}^{N} (\xi_i + \hat{\xi}_i) + \frac{1}{2}||\mathbf{w}||^2 \right\} \quad \text{subject to :} \\ &y_i \leq f(x_i, \mathbf{w}) + \epsilon + \xi_i, \quad y_i \geq f(x_i, \mathbf{w}) - \epsilon - \hat{\xi}_i, \\ &\xi_i \geq 0,\ \hat{\xi}_i \geq 0,\ \forall i = \{1, ..., N\} \end{aligned} \tag{3.64}$$

where ξ_i & $\hat{\xi}_i$ represent slack variables defining whether $y_i > f(x_i, \mathbf{w}) + \epsilon$ or $y_i < f(x_i, \mathbf{w}) - \epsilon$, for $\xi_i, \hat{\xi}_i > 0$ respectively. Again, this *primal* SVR can be formulated and solved as a *dual* problem similar to (3.55) using the methodology demonstrated in section 3.7.1, as well as kernel transformed using the *kernel trick* (3.56).

3.7.3.1 SVM Parameter Tuning

In this thesis, SVM parameter tuning was performed via grid-search over internal cross validation (see section 3.11.1) to determine the optimal values of kernel parameter γ (3.61), the penalty parameter C, and the ℓ_1 soft-margin regularisation parameter ϵ for SVR models. These values were all investigated over the log-scaled range $[10^{-3}, ..., 10^{3}]$, selecting the pairs that gave the lowest CV misclassification error (see section 6.2.4) [234, 235].

The generalisation performance of SVM decisions can be greatly affected by high dimensional feature spaces $\mathcal{X}$ (i.e. a large number of P features) however, especially if there exists a high redundancy in $\mathcal{X}$ (let alone being computationally inordinate to run a model with a large dimensional $\mathcal{X}$) [236]. It is therefore practically important to perform feature selection prior to SVM classification or regression, especially if $P > N$ [235]. As you will see in chapters 5 and 6 we utilise LASSO to determine parsimonious sets of features prior to building SVM models.

3.7.4 Classification and Regression Trees

Linear and non-linear regression and classification techniques introduced in this thesis thus far have relied on linear methods to select suitable features in order to make a decision or prediction. Classification and Regression Trees (CART) specifically Random Forests (RF) are a multi-functional, non-linear method capable of performing regression[12], classification and feature selection [237]. Unlike the linear filter-based methods of feature selection introduced thus far, for example, LASSO (3.46) (which are also often used to (linearly) pre-select features prior to application in (non-linear) methods, such as SVMs), RFs incorporate non-linear feature selection as part of the model methodology. Random Forests consist of a large ensemble of decision trees arranged in a hierarchical structure, as depicted in figure 3.4. To build an individual tree, we recursively descend through the hierarchy, performing binary splits (decisions) at each level in the structure (a node, j) using a single feature $x_j \in \mathcal{X}^p$ based on a threshold value (splitting criterion) s_j,

[12]Random Forests specifically for regression are denoted as RFR throughout this thesis.

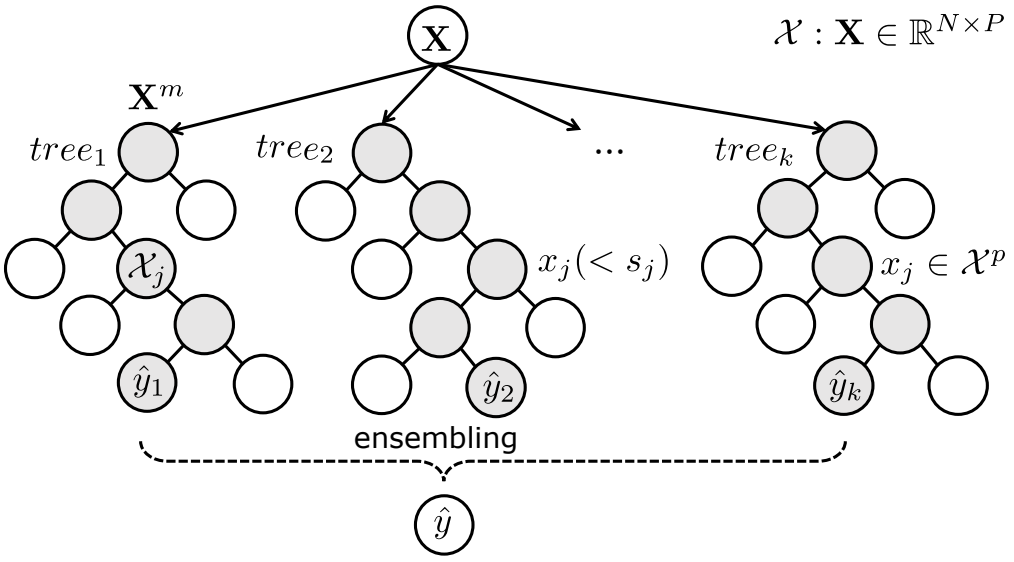


Figure 3.4: Simple diagram describing the construction of Classification and Regression Trees (CART), specifically a Random Forest (RF) model. Starting from the root node, k number of decision *trees* are grown and evaluated using $m < N$ random observations ($\mathbf{X}^m$) for each *tree*. At each j^{th} node, p random subsets of the features ($\mathcal{X}^p$) are considered to evaluate a split. A final prediction $\hat{y}$ is ensembled (for instance by majority voting) across the $\hat{y}_1, \hat{y}_2, ..., \hat{y}_k$ weaker decision *tree* predictions. In this example, the feature space is recursively partitioned into $\mathcal{X}_j$ sets by determining the optimal split criterion s_j from a feature $x_j \in \mathcal{X}^p$ at each j^{th} node.

sub-partitioning the feature space $\mathcal{X}_j$ at each node. A tree is typically expanded until all leaves are pure (i.e each partition $\mathcal{X}_j$ represents only one class) or until all leaves contain less than the minimum number of samples in a partition $\mathcal{X}_j$ required to split a node. While decision trees are susceptible to over-fitting, the advantage of a RF is that multiple trees can be learned, introducing a variability between the trees. An individual tree selects $m < N$ random subset of observations (with replacement), and each node considers a random subset of p features for each split. Importantly, the same feature can be selected for multiple nodes in the tree, and can have different associated s_j values at each node. Once all trees have been grown, each of the 'weaker' decisions are aggregated (or *ensembled*) creating a robust final prediction. Classification predictions are deduced as the majority class label of the observations present in each final partition $\mathcal{X}_j$, whereas for continuous prediction (i.e. regression) the mean of the (continuous) responses would be calculated instead. To determine the optimal split criterion s_j for each x_j to create each $\mathcal{X}_j$ we evaluate the *Gini* importance, which quantifies the average gain of purity (i.e. the presence

of one class) caused by splits of a given variable. For regression the mean decrease in MSE is assessed instead (equation 3.29). RFs have relatively little hyperparamater tuning, which typically only include the number of trees to build ($k = 1500$ in this thesis) and the number of input variables chosen at each node (p). Values of p as suggested in [238, 239] are: $p \in \{\sqrt{P}, 2\sqrt{P}, \sqrt{P}/2\}$.

3.8 Gradient-Based Optimisation

Gradient descent (GD) (or steepest descent) is a method used to find the mimima of an objective function $E(\mathbf{w})$ by calculating the derivative of $E(\mathbf{w})$ with respect to the weights, the gradient $\nabla_{\mathbf{w}} E(\mathbf{w})$, such that the optimisation repeatedly takes a step t in the direction of steepest decrease of $E(\mathbf{w})$ [202, 230]:

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta \nabla_{\mathbf{w}_t} E(\mathbf{w}_t) \tag{3.65}$$

where η is the learning rate, $\eta > 0$, determining the size of the step. This is similar to the concept of optimising (3.40 and 3.41) using the Newton-Raphson methodology (introduced section 3.5.3), however Newton-Raphson is computationally expensive as it requires solving the second derivative Hessian $\mathbf{H}$, which can be intractable and attracts saddle points in higher dimensions [240, 241].

Specifically, to minimise the objective function $E(\mathbf{w})$ for an SVM (3.51 or 3.55), the hinge loss (3.52) is not differentiable. Instead, the gradient descend must be modified to compute the sub-gradient of $E(\mathbf{w})$ at each t, where the iterative update is as follows:

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \nabla_{\mathbf{w}_t} E(\mathbf{w}_t) \tag{3.66}$$

$$\leftarrow \mathbf{w}_t - \eta_t \frac{1}{N} \sum_{i=1}^{N} (\lambda \mathbf{w}_t + \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{x}_i, y_i; \mathbf{w}_t)) \tag{3.67}$$

and $\lambda = 2/(NC)$ in equation (3.51). This is known as the Primal Estimated Sub-Gradient Solver (*Pegasos*) algorithm [242], where at each iteration, t, cycles through the data with the following updates:

$$\begin{aligned} \mathbf{w}_{t+1} &\leftarrow \mathbf{w}_t - \eta_t(\lambda \mathbf{w}_t - y_i \mathbf{x}_i) \quad \text{if } y_i(f(x_i)) < 1 \\ &\leftarrow \mathbf{w}_t - \eta_t(\lambda \mathbf{w}_t) \qquad\qquad \text{otherwise} \end{aligned} \tag{3.68}$$

where in *Pegasos* the learning rate is set at $\eta_t = \frac{1}{\lambda t}$ [242].

Stochastic gradient descent (SGD) is a modification to gradient descent (3.65) which considerably speeds up a convergence [202, 243]. SGD replaces the gradient with an unbiased estimate of the gradient, computing the average gradient over a subset of sample(s), a *mini-batch*, of m examples drawn i.i.d from the original dataset. Therefore an estimate of the gradient from equation (3.65) is determined as such:

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta \frac{1}{m} \nabla_{\mathbf{w}_t} \sum_{i=1}^{m} E(\mathbf{w}_t) \tag{3.69}$$

where m is the number of observations in the subset. It is also common to decrease the learning rate linearly following: $\eta_t = (1-\alpha)\eta_0 + \alpha\eta_\tau$ with $\alpha = \frac{t}{\tau}$ until iteration $t = \tau$ where it remains constant. Despite the improvements over GD, learning with SGD can sometimes be slow typically caused by poor conditioning of the Hessian ($\mathbf{H}$) matrix and variance in the stochastic gradient [202]. The concept of *momentum* is introduced to overcome this, whereby an exponentially decaying average of the negative gradient, the velocity $\boldsymbol{v}$, accumulates the gradient elements $\frac{1}{m}\nabla_{\mathbf{w}_t} \sum_{i=1}^{m} E(\mathbf{w}_t)$, driving the direction of convergence [202, 244, 245]. Finally, *Adam* is a robust and computationally efficient variant of SGD which incorporates momentum, as well as producing an adaptive estimation of the first-order and second-order moments, which are controlled by decay rates β_1 and β_2 respectively [246]. Importantly, Adam also includes bias corrections to these moment estimates yielding state-of-the-art results.

3.9 Deep Learning and Neural Networks

Deep learning (DL) is a broad overarching term given to *representation learning* (i.e. to allow an algorithm to learn features directly) "by introducing representations that are expressed in terms of other, simpler representations" [202], which can learn different levels of abstraction through the combination of multiple stacked (hence *deep*) non-linear model layers [247]. Each layer "transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level" [247]. Deep learning typically describes Artificial Neural Networks (ANN) that are stacked in cascading architectural arrangements, i.e. Deep Neural Networks (DNN). Deep learning can also refer to convolution neural networks (CNN), or combined CNN and fully-connected DNN architectures (DCNN) [202, 247]. Other common network architectures include recurrent neural networks (RNNs), particularly long short-term memory networks (LSTMs), which can learn sequential and related relationships, however are not explored in this thesis. The feed-forward[13] ANN—in its simplest case, often called the *multi-layer perceptron* (MLP)—can be thought of as the building blocks for deep networks, acting as a a function consisting of a number of (often non-linear) functions successively layered together, which map an input $\mathbf{x}$ to some output y [248]. Considering first a single perceptron introduced in (3.27): $\hat{y} = \mathbf{w}^\top \mathbf{x} + b$, which as we know (3.26) can be written in the form: $\hat{y} = \sum_{i=1}^{P} w_i x_i + b$, the output of the j^{th} hidden unit (i.e. the units not treated as outputs) can be obtained by first forming a linear combination of the P input values (features in this case) and is given as:

$$a_j = \sum_{i=1}^{P} w_{ji}^{(1)} x_i + b_j^{(1)} \tag{3.70}$$

$$z_j = \phi(a_j) \tag{3.71}$$

$$a_k = \sum_{j=1}^{M} w_{kj}^{(2)} z_j + b_k^{(2)} \tag{3.72}$$

where $j = 1, ..., M$ denote the number of connections in a layer, the superscript (1) indicates that the corresponding parameters are in the "first" layer of the

[13] i.e. there are no feedback loops, allowing the network outputs to be calculated as explicit functions of the inputs, weights, biases.

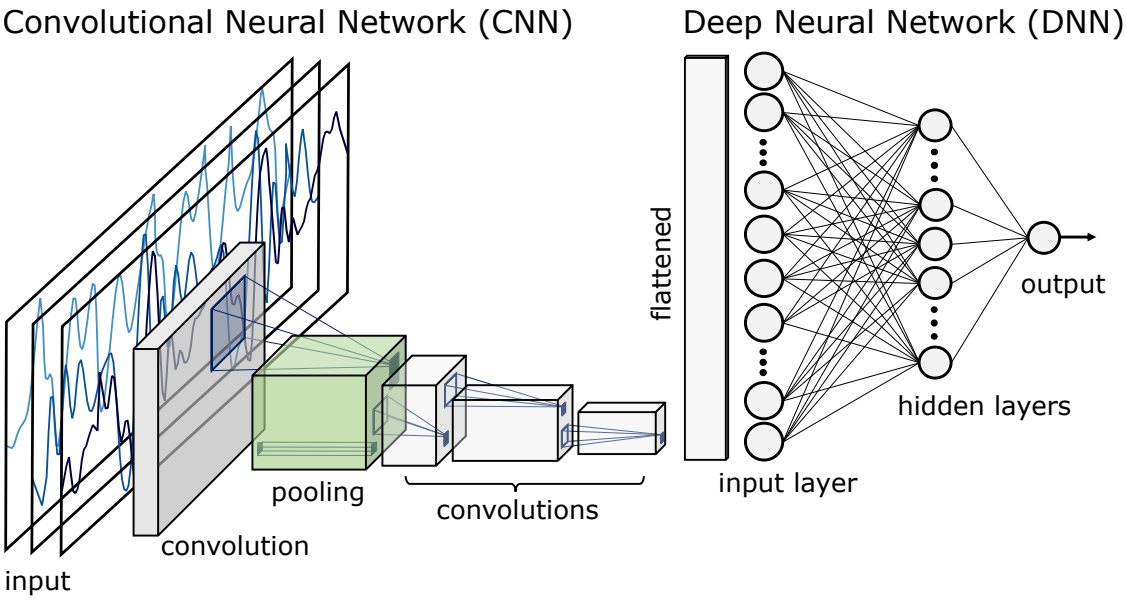


Figure 3.5: Illustration of a commonly used architecture in deep-learning, that are inspired by the connectivity patterns in the animal visual cortex. Convolution Neural Networks (CNN) are a class of algorithm which can learn patterns (termed "features") within the data. In this example, an 3-axis accelerometer signal can be sequentially transformed (left-to-right) using a series of convolution and pooling operations into flattened vectors. A simple Deep Neural Network (DNN) can be stacked together consisting of hidden layers (hence a "deep" artificial neural network, denoted by the circles), at the DNN input layer. As data flows through each node of the network, it will again be non-linearly transformed. The elements of the final output layer (in this case a classification "sigmoid" layer) represent the probabilities of the presence of disease versus healthy for example. During the training process, the internal parameters of the network layers are adjusted to improve the model accuracy. Different output layers can perform separate tasks such as classification or regression. CNN and DNN models are typically combined, as fully-connected Deep Convolutional Neural Networks (DCNNs) in deep-learning approaches. Figure adapted from [5].

network. Here, $w_{ji}^{(1)}$ denote the weights going from i to hidden unit j, where b_j are the corresponding biases. The activation of the hidden unit j, z_j, is obtained by transforming the linear sum in (3.70) using a positive and monotonically increasing activation function ϕ. These outputs are then linearly combined to give output unit activation in (3.72), where $k = 1, ..., K$ total number of outputs. The choice of activation function is determined by the task [248], for example a final output y_k can be determined for a regression task as the identity $y_k(\mathbf{x}, \mathbf{w}) = a_k(\mathbf{x}, \mathbf{w})$, or for a binary classification task the logistic function (3.34) $\phi = \sigma : y_k(\mathbf{x}, \mathbf{w}) = \sigma(a_k(\mathbf{x}, \mathbf{w}))$ could be used. For multi-class targets, the softmax function can be used, where binary target variables $y_k \in 0, 1$ denote a 1-of-K coding scheme indicating the class k (where in this case the number of outputs from (3.72) should equal the number

of classes in a task). The network outputs are thus interpreted as $\hat{y}_k = p(y_k = 1|\mathbf{x})$, where the $\sum_k \hat{y}_k = 1$; the softmax is defined as:

$$\hat{y}_k(\mathbf{x}, \mathbf{w}) = \frac{\exp(a_k(\mathbf{x}, \mathbf{w}))}{\sum_j \exp(a_j(\mathbf{x}, \mathbf{w}))} \tag{3.73}$$

Rectified Linear Unit (ReLU) activation functions ($z_k = \max(0, a_j)$) are nearly always used in hidden layers, and help overcome vanishing gradient problems commonly experienced with other hidden layer activation functions [202, 249]. Considering a more generalised notation for later in this thesis, a DCNN model, mapping only *feed-forward* architecture for some input i, each unit comprises a weighted sum of its input which can be transformed through a non-linear activation $\phi(\cdot)$, can therefore be described by:

$$a_j = \phi\left(\sum_i w_{ji} z_i + b_j\right) \tag{3.74}$$

3.9.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a class of ANN algorithm that follow a grid like topology and are especially robust displaying *translation invariance*[14] [202, 250]. CNNs make use of convolutional operators, described by:

$$s(t) = (\boldsymbol{x} * w)(t) = \sum_{a=-L}^{L} \boldsymbol{x}(a) w(t - a) \tag{3.75}$$

where $\boldsymbol{x}$ is the raw input data, such as a time-series, of length $2L$ and $w(\cdot)$ is the weighting function or *kernel*. The output $s(t)$ at time index t is often considered as a representation of the features learned (a *feature map*). The number of convolutional kernels applied to an input, as well as the width and height of $w(\cdot)$ are pre-defined.

CNNs have been shown to act as unsupervised feature extractors and can be applied directly onto the raw signals to learn temporal and spatio-temporal information [202, 250, 251]. The features extracted by convolutional layers can then be arranged to create a final output (classification or regression) through feed forward fully connected dense (DNN) layers.

[14] *Translational invariance* refers to a representation that is approximately invariant to small translations of the input. For instance, when presented with inputs of elements that are in a different order or have been translated, a model output will not vary [202].

3.9.2 Training a Deep Network

Following the same training scheme introduced in section 3.5.3, for a classification task, the cross-entropy error function (3.37) can be minimised using the stochastic gradient-based optimisation approaches outlined in section 3.8. The cross entropy error function (3.37) can be expanded for multiple classes (K) by simply using the softmax function (3.73), such that:

$$E(\mathbf{w}) = -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ki} \ln(\hat{y}_{ki}(\mathbf{x}_i, \mathbf{w})) \tag{3.76}$$

3.9.2.1 Error Backpropagation

As we learned in 3.8, the derivatives of an error function with respect to the weights $E(\mathbf{w})$ must be calculated to optimise any model. *Error Backpropagation* is the essential method to compute the partial derivatives $\nabla E(\mathbf{w})$, by propagating backwards through the network, whereby those derivatives are then used to update the network and adjust the weights [248]. Firstly, considering a simple feed-forward network (as in section 3.9) with inputs i, hidden layers j and output layer k, during *forward propagation* the network activations, weights and biases are determined, following the generalised equation (3.74). Backpropagation calculates the partial derivatives of the error with respect to each weight w in the network, which can be represented as a summation over all n connection pathways in a network:

$$\frac{\partial E}{\partial w_{ji}} = \sum_{n} \frac{\partial E^n}{\partial w_{ji}} \tag{3.77}$$

Considering each connection pathway, the error depends on w_{ji} via the summed input a_j to unit j. Therefore we can apply the chain rule for partial derivatives to give:

$$\frac{\partial E^n}{\partial w_{ji}} = \frac{\partial E^n}{\partial a_j}\frac{\partial a_j}{\partial w_{ji}} \tag{3.78}$$

Considering that each unit comprises of a weighted sum of its input in the form given by (3.74) before activation ϕ (the bias term can also be omitted, see [248]). We can therefore solve:

$$\frac{\partial a_j}{\partial w_{ji}} = z_i \tag{3.79}$$

The errors with respect to each neuron, $\frac{\partial E^n}{\partial a_j}$ can be represented using the chain rule again as:

$$\frac{\partial E^n}{\partial a_j} = \sum_k \frac{\partial E^n}{\partial a_k} \frac{\partial a_k}{\partial a_j} \tag{3.80}$$

summing over all $j \to k$ unit connections. It can be seen that only variations in a_j represent variations in the error function only through variations in a_k. Therefore using the nomenclature again from (3.74), we can represent the back-propagation errors as:

$$\frac{\partial E^n}{\partial a_j} = \phi'(a_j) \sum_k w_{kj} \frac{\partial E^n}{\partial a_k} \tag{3.81}$$

which we obtain for each unit in the network, substituting into (3.81) and (3.79) into (3.78), where $\phi'(\cdot)$ is the derivative of the activation function. The weights are then updated in an iterative process based on these derivatives using the gradient-based optimisation methodology for $\nabla E(\mathbf{w})$, such as SGD (3.69) as outlined in section 3.8.

3.9.2.2 Practical Implementation of Deep Networks

The entire pipleline to train a deep network involves performing multiple forwards and backwards propagations through the network during the training phase to update the weights of the model. For each update, a *batch* set of N samples are processed independently, and in parallel, to compute the weight updates. A sufficient *batch* sample is needed to adequately approximate the distribution of the data, yet needs to be small enough for computational efficiency. An *epoch* generally signifies one training iteration over the entire dataset, where after each *epoch* the network is evaluated, usually against the validation set error. This allows the best model to be logged, or for the tuning of hyperparameters such as scheduling learning rate changes [202].

3.9.3 Improving Network Generalisability and Robustness

3.9.3.1 Pooling

Pooling functions replace the z_j of a network at certain locations with a summary statistic (e.g. mean, max, sum) based on the outputs in a specific region (e.g. the

maximum in a square region, denoted *max-pooling*, see figure 3.5). Pooling helps network robustness by reducing the number of parameter overheads as well as making the representations learned approximately *translationally invariant*[15] [202].

3.9.3.2 Batch Normalisation

Batch Normalisation (BN) is an adaptive re-parametrisation method which injects additive and multiplicative noise in the hidden units by standardising the activations of a previous layer at each batch with unit variance (see A.1.1) [202, 252]. This has both a regularising effect and drastically improves optimisation, as during the learning phase (i.e during back-propagation) the gradient never converges towards any operation that will increase the variability or mean of any z_j (3.71).

3.9.3.3 Network Regularisation

Extending the regularisation approaches introduced in (3.6.1), penalties such as ℓ_1 and ℓ_2 can also be added to the same objective function $E(\mathbf{w})$ in neural networks (as in equation 3.44) [202, 248]. Regularisation penalties can also be applied on a per-layer basis, such as applying weight decay to upstream of the convolutional layer's kernel $w(\cdot)$, or to the layer's output z_j (3.71).

3.9.3.4 Dropout

Dropout regularisation consists of randomly omitting (dropping) a percentage of units z_j in a neural network to decrease the network's sensitivity to noise and has proven effective to prevent over-fitting [202, 253].

3.10 Visualising Deep Networks

The field of explaining the internal decisions, or rationale behind those decisions, in deep networks is rapidly developing. Attribution is a method whereby relevance or contribution of the original input sample towards a network's decision are calculated [254]. This contribution is then typically projected back onto the input

[15] See footnote 14.

sample as heatmaps, in order to visually interpret the regions, and therefore the characteristics, that influenced a deep network's decision.

3.10.1 Layer Wise Relevance Propagation

Layer Wise Relevance Propagation (LRP) is an attribution method that has been proposed to interpret deep learning models by back-propagating through a network to decompose the final output decision, $f(\boldsymbol{x})$ [255–258]. Firstly, a trained model's activations, weights and biases are obtained in a forward pass through the network. Secondly, during a backwards pass through the model, LRP attributes relevance to the individual input nodes, layer by layer. For example R_k denotes the relevance for neuron k in layer $^{(l+1)}$, and $R_{j\leftarrow k}$ defines the share of R_k that is redistributed to neuron j in layer $^{(l)}$. The fundamental concept underpinning LRP is that the conservation of relevance per layer such that: $\sum_j R_{j\leftarrow k} = R_k$ and $R_j = \sum_k R_{j\leftarrow k}$. These two equations, when combined, also ensure a relevance conservation property between layers $\sum_j R_j = \sum_j \sum_k R_{j\leftarrow k} = \sum_k \sum_j R_{j\leftarrow k} = \sum_k R_k$. Conservation of relevance in the redistribution process also holds globally, so that we can write the chain of qualities where we combine to:

$$\sum_{i=1}^{N} R_i = \cdots = \sum_j R_j = \sum_k R_k = \cdots = f(x) \tag{3.82}$$

Together, both above relevance decomposition and pooling steps ensure a local relevance conservation property, where i is an input neuron in layer $l - n$. The conservation of total relevance per layer can can also be denoted as:

$$\sum_j R_{j\leftarrow k}^{(l,l+1)} = R_k^{(l+1)} \tag{3.83}$$

Propagation rules are implemented to withhold this conservation property. Considering a DNN model, which can be described by previously introduced the neuron activation (3.74): $a_k = \phi\left(\sum_j a_j w_{jk} + b_k\right)$, which consists of a_j, the activations from the previous layer, and w_{jk}, b_k, the weight and bias parameters of the neuron. The function ϕ is a positive and monotonically increasing activation function. In case of a component-wise operating non-linear activation, e.g. a

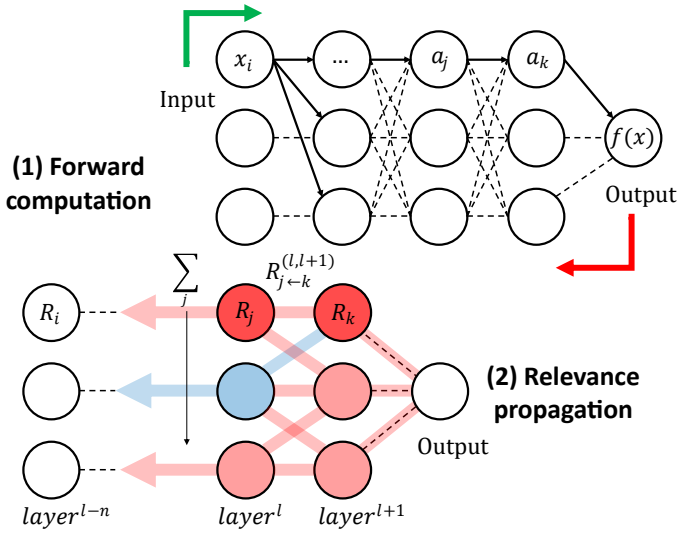


Figure 3.6: Schematic describing the computation of Layer-wise Relevance Propagation (LRP). LRP decomposes the network's output score into the individual contributions of the input neurons, while conserving the total amount of relevance across layers. a are the neuron activations in a given DNN layer. R represent the corresponding relevance propagation values at each node, within each layer l, during back-propagation. Positive relevance is shown in red whereas negative relevance is shown in blue. Relevance values $R_i > 0$ signify components x_i which represent the presence of the predicted class, while conversely. $R_i < 0$ contradict the prediction of that class. $R_i \approx 0$ indicate inputs x_i which have little or no influence to the model's decision. Adapted from [257].

ReLU ($\forall j = k : x_k = \max(0, x_j)$) then $\forall j = k : R_j = R_k$, since the top layer relevance values R_k only need to be attributed towards one single respective input j for each output neuron k. Figure 3.6 schematically describes the computation of LRP from a forward pass through the network to the redistribution of relevance (R_i) to the inputs x_i. The $\alpha\beta$-rule for LRP has been shown to work well at decomposing a model's decisions:

$$R_j = \sum_k \left(\alpha \frac{a_j w_{jk}^+}{\sum_j a_j w_{jk}^+} - \beta \frac{a_j w_{jk}^-}{\sum_j a_j w_{jk}^-} \right) R_k, \tag{3.84}$$

where each term of the sum corresponds to a relevance propagation $R_{j \leftarrow k}$, where $()^+$ and $()^-$ denote the positive and negative parts respectively, and where the parameters α and β are chosen subject to the constraints $\alpha - \beta = 1$ and $\beta \geq 0$.

The LRP algorithm initiates at the model output and iterates over all layers in the model in a backwards pass until the relevance scores R_i for all inputs of x_i are computed. Relevance values $R_i > 0$ signify components x_i which represent the presence of the predicted class, while conversely $R_i < 0$ contradict the prediction of

that class. $R_i \approx 0$ indicate inputs x_i which have little or no influence to the model's decision. Typical $\alpha\beta$ values used in previous studies are: $\alpha = 1$ & $\beta = 0$ ($\alpha_1\beta_0$-rule), which allows the filtering of only *positive* relevance values. Other parametrisations used in this work are $\alpha = 2$ & $\beta = 1$ ($\alpha_2\beta_1$-rule), which allows for the attribution of *negative relevance* contributions of x_i [257].

A signed small stabilising term can be added to the denominator (termed ϵ-rule):

$$R_j = \sum_k \frac{a_j w_{jk}}{\epsilon + \sum_j a_j w_{jk}} R_k, \tag{3.85}$$

The ϵ-rule has been shown to filter noisy heatmaps by absorbing some relevance when the contributions to the activation of neuron k are weak or contradictory [255, 259]. For larger values of ϵ only the most prominent explanation factors are retained, yielding a more sparse and less noisy explanation. In accordance with [255, 260], $\alpha\beta$-rules were applied to convolutional layers and the ϵ-rule ($\epsilon = 0.01$) to dense layers.

3.11 Evaluation Approaches and Metrics

3.11.1 Cross Validation

In order to determine the generalisability of machine learning models, most evaluation methods follow a structured approach known as cross validation for splitting the data in order to prevent over-fitting and mitigate against biases [235, 261, 262]. K-fold cross validation (CV) is the most common methodology, whereby the data is split into K disjoint subsamples (or folds), chosen randomly but with roughly equal size. Training is performed on the $K - 1$ fold (in-sample) and testing on the K^{th} left out fold (out-of-sample). The training set is often further split into smaller sets for parameter selection (validation) using an internal K-fold CV approach. In this thesis we employ a 5-fold, subject-wise, cross-validation (CV) approach. Stratifying refers to splitting each fold with equal proportions of classes where possible. For datasets which contain repeated measurements (such as multiple tests) from the same subject, it is important to partition the data by *subject*. Subject-wise CV ensures that there is no "leakage" of subject information between training and testing partitions where models can learn individual subjects' characteristics

rather than the task characteristics (such as disease recognition) [263–265]. CV is often repeated N times with new random partitions in order to reduce bias in re-sampling and dataset splitting.

3.11.2 Class Imbalance Problems

Many real-world datasets, especially healthcare related, are susceptible to the class-imbalance problem, where one group is over represented; or extensions of that, where subjects contribute varying numbers of recordings each. To overcome any biases, re-sampling schemes by bootstrapping can be employed such as: randomly up-sampling the minority class, randomly down-sampling the majority class, penalising a loss function to account for any imbalance [266, 267] or even creating synthetic samples like Synthetic Minority Over-Sampling Technique (SMOTE) [268]. In this thesis, a conservative class balancing approach was adopted by:

- randomly sampling m tests per subject, with replacement;
- balancing the classes in the training and validation sets by randomly up-sampling the minority class, with replacement;
- reporting out-of-sample test set performance metrics which account for class imbalances (see section 3.11.3).

While it is acknowledged that biases may still occur from duplication of subjects' tests caused by re-sampling [266, 269], this simple approach heuristically minimises the discarding of data, while retaining as many unique participants as possible.

3.11.3 Classification Evaluation

A confusion matrix is depicted in figure 3.7 in its simplest form as a binary classification problem. By summing the total number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), a number of

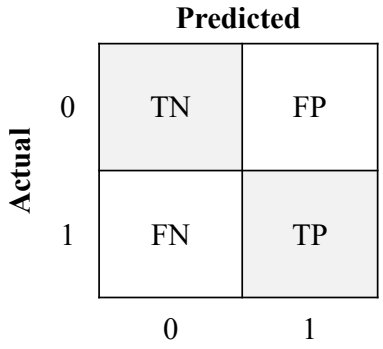


Figure 3.7: Schematic of a binary classification confusion matrix (CM). The CM summarises the agreement between predictions $\hat{\mathbf{y}}$ and the ground truth $\mathbf{y}$.

summary measures can be defined [270]:

$$Acc = 100 \cdot \frac{\mathrm{TN} + \mathrm{TP}}{\mathrm{TN} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}} \tag{3.86}$$

$$precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{3.87}$$

$$recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3.88}$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{3.89}$$

Accuracy (Acc) is the simplest metric and corresponds to the percentage of correct predictions. However, in cases of imbalanced data (see section 3.11.2), it is more representative to report metrics which are more robust to imbalances, such as precision (the proportion of true positives from those positive predictions), recall (the proportion of actual positives that are correctly classified) and the F_1 score, which accounts for class imbalances by unifying both precision and recall using the harmonic mean [266, 270]. For multi-class problems, the confusion matrix can simply be expanded and the *macro-average* of each metric (i.e. for each binary classification task) can be reported instead (e.g the macro F1-score (MF1)). Cohen's kappa (κ) is also a robust classification measure which accounts for imbalanced data, as well as chance agreement, and can be expanded to multiple classes (see [271] for more details).

3.11.4 Regression Evaluation

Regression methods, such as introduced in section 3.5, model continuous estimates and can be evaluated by the coefficient of determination, r^2, which measures the proportionate amount of variation in the response variable $\mathbf{y}$ explained by the models predictions $\hat{\mathbf{y}}$ [224]:

$$r^2 = 1 - \frac{\mathrm{SS}_{\mathrm{res}}}{\mathrm{SS}_{\mathrm{total}}}, \quad \text{where } \mathrm{SS}_{\mathrm{total}} = \sum_{i=1}^{N} |y_i - \bar{y}|^2 \tag{3.90}$$

and $\bar{y}$ is the average $\mathbf{y}$; $\mathrm{SS}_{\mathrm{res}}$ are the least squares estimates (3.30). Furthermore regression estimates can also be evaluated based on mean absolute error (MAE) and root-mean squared error (RMSE):

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{3.91}$$

$$\mathrm{RMSE} = \frac{1}{N} \sqrt{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \tag{3.92}$$

where $\hat{y}$ are the model predictions and N are the number of observations. MAE penalises the absolute deviation of predicted and estimate y values (i.e. the ℓ_1-penalty, see 3.42), whereas RMSE penalises the squared deviations (i.e. the ℓ_2-penalty, see 3.42) [272]. Importantly, both metrics are reported in the original units of $y \in \mathcal{Y}$, greatly aiding interpretability of model performance.

Table 3.1: Summary of Methodology

Methodology	Task	Comment(s) / Assumption(s)	Application Example(s)	Chapter(s)
Exploratory Data Analysis				
PCA	Dimensionality reduction		Expressing the structure of a large set of features in terms of a smaller number of latent principal components	5, 6, 7
t-SNE	Embedding high-dimensional data for visualisation in a low-dimensional space		Visualisation of DCNN features clustered by outcome group	7
r	Linear correlation between continuous variables	Continuous linear relationship	Association between features and clinical measures	5, 6, 8
ρ	Non-linear correlation between continuous variables	Continuous non-linear relationship (non-parametric)		5, 6, 8
ICC	Test Re-Test Reliability	The validly of the ANOVA framework assumptions	Feature reliability and reproducibility	5, 6
Confirmatory Data Analysis				
t-test: (one-sample)	Assess whether the sample(s): $\boldsymbol{x} \sim \mathcal{N}(0, \sigma^2)$	Normality: $\bar{x} \sim \mathcal{N}(\mu, \sigma^2)$, where σ^2/n follows a normal distribution, where $\boldsymbol{x}$ are i.i.d.		
(paired-sample)	Tests whether two *dependent* samples $\boldsymbol{x}$ & $\boldsymbol{y}$ come from the same distribution	Normality; equal variances between the two samples, $\sigma_x^2/n \approx \sigma_y^2/n$ (homoscedasticity)	Assessing handed differences between DaS feature values	5, 6
(two-sample)	Tests whether two *independent* random samples, $\boldsymbol{x}$ & $\boldsymbol{y}$, come from the same distribution	Normality, $\sigma_x^2/n \approx \sigma_y^2/n$ (homoscedasticity);	Assessing demographic differences between independent HC and MS groups e.g. age	5, 6
(Welch's)		Normality; $\sigma_x^2/n_1 \neq \sigma_y^2/n_2$ and are corrected against	Assessing differences in estimated severity outcomes between clinical visits	8
χ^2 test	Assess whether two categorical variables are related or not (independence)	Test statistics follow a χ^2 distribution when the observations are independent and normally distributed	To determine differences in the proportion of male/female ratios between HC and MS groups	4, 5, 6
KS test	Non-parametric (one-sample) test of the null hypothesis that the population cumulative distribution function (cdf) of $\boldsymbol{x}$ is equal to the hypothesised distribution	Test decision for the null hypothesis that the data in vector $\boldsymbol{x} \sim \mathcal{N}(\mu, \sigma^2)$	To assess the normality of features	4, 5, 6, 7, 8
WSRT	Non-parametric version of the one-sample and paired-sample t-test	Non-parametric, without any necessity for normality, or known and equal variances	Assessing dominant & non-dominant handed differences between paired 9HPTs or feature values	5, 6

Continued on Next Page...

Methodology	Task	Comment(s) / Assumption(s)	Application Example(s)	Chapter(s)
U test	Non-parametric version of the two-sample t-test	Non-parametric, without any necessity for normality, or known and equal variances, or equal group sizes	Assessing feature differences between independent HC and MS groups	4, 5, 6, 8
ANOVA	Tests the null hypothesis that the data in multiple categorical groups comes from the same distribution.	Normality; independence of observations; homoscedasticity within groups	Assessing feature differences between multiple groups simultaneously	5, 6, 8
KWt	Tests the null hypothesis that the data in multiple categorical groups comes from the same distribution.	The KWt is a non-parametric version of classical one-way ANOVA	Assessing feature differences between multiple groups simultaneously	5, 6, 8
BF test	To test if multiple data samples have equal variances	Non-parametric BF assumes non-normality, utilising the median	To assess the if the variance in clinical Nine Hole Peg Tests (9HPT) are equal across HC and MS groups.	5, 8
Methods for Regression and Classification				
OLS	Regression	Linear model	Predicting the Nine Hole Peg Test (9HPT)	5
IRLS		Linear model robust to outliers and data heteroscedasticity		5
LR	Classification	Linear model through a non-linear link function for classification	Classification of HC and PwMS subgroups	6
LASSO	Regression / Classification	Linear model incorporating feature selection through ℓ_2 regularisation	Feature selection	5, 6
SVM	Classification	Linear model incorporating a non-linear feature transformation	Classification of HC and PwMS subgroups	6, 8
SVR	Regression	Non-linear model incorporating a non-linear feature transformation	Predicting the Nine Hole Peg Test (9HPT)	5
RF / RFR	Classification / Regression	Non-linear model incorporating non-linear feature selection	Classification of HC and PwMS subgroups / Predicting the Nine Hole Peg Test (9HPT)	5, 6
DCNN	Classification	Non-linear model incorporating representation (feature) learning from the raw data	Classification of HC and PwMS subgroups	7
LRP	Attribution	Attribution through heatmap visualisation of deep networks' decisions	Healthy versus MS gait pattern explanation	8

PCA: Principal Component Analysis (A.2.1); t-SNE: t-distributed Stochastic Neighbour Embedding (A.2.2); r: Pearson's correlation coefficient (3.2), ρ: Spearman's rank correlation coefficient (3.3); ICC: Intra-class correlation coefficient (3.3.3); t-test: Student's t-test (3.4.2.1); χ^2: Chi-Square Goodness-of-Fit test (3.4.2.3); KS test: Kolmogorov-Smirnov test (3.4.2.4); WRST: Wilcoxon signed rank test (3.4.2.5); U test: Mann-Whitney U test (3.4.2.6); ANOVA: One-way analysis of variance (3.4.2.7); KWt: Kruskal-Wallis test (3.4.2.8); BF test: Brown-Forsythe test (3.4.2.9); OLS: Ordinary Least Squares (3.5.1); IRLS: Iteratively Re-weighted Least-squares (3.5.2); LR: Logistic Regression (3.5.3); LASSO: Least Absolute Shrinkage and Selection Operator (3.6.3); SVM: Support Vector Machines for Classification (3.7.1); SVR: Support Vector Machines for Regression (3.7.3); RF: Random Forest for Classification (3.7.4); RFR: Random Forest for Regression; DCNN: Deep Convolutional Neural Network (3.9.1); LRP: Layer-wise Relevance Propagation (3.10.1).

4
Definition of Datasets

Contents

4.1 Introduction 70
4.2 The FLOODLIGHT Proof-of-Concept Study 71
4.3 Study Activities 76
4.4 Human Activity Recognition Datasets 78

4.1 Introduction

This chapter briefly introduces the study underpinning the work explored in this thesis. The "Monitoring of Multiple Sclerosis (MS) Participants With the Use of Digital Technology (Smartphones and Smartwatches) - A Feasibility Study" (NCT02952911), was a proof-of-concept (PoC) study to assess the feasibility of remote patient monitoring, using smartphones and smartwatch devices, applying a range of testing modalities in people with Multiple Sclerosis (PwMS) and healthy controls (HC) over 24 weeks [31, 273]. This was the first deployment of the smartphone application, known as "FLOODLIGHT", developed by F-Hoffmann La Roche, Basel, to assess PwMS. The subsequent feasibility study of FLOODLIGHT, NCT02952911, is referred to in this thesis as the FLOODLIGHT PoC (FL).

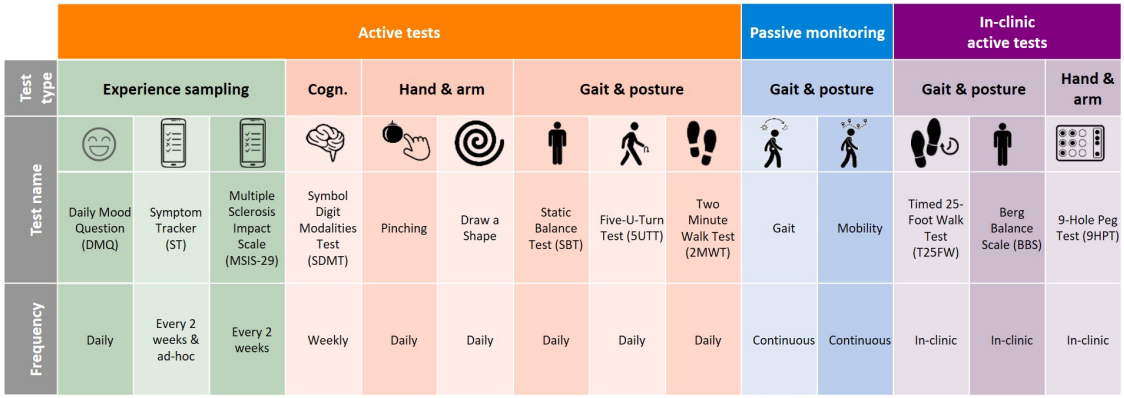


Figure 4.1: The FLOODLIGHT PoC test suite. Diagram indicating the active tests, patient-reported outcomes (PRO), clinical assessments and passive monitoring conducted as part of the FLOODLIGHT PoC testing suite. Cogn.: Cognition-based assessments. Diagram adapted from: [31].

The analysis presented in this thesis focuses on the physical functional assessments as part of the FLOODLIGHT PoC: the Draw-a-Shape Test (DaS) and the remote Two-Minute Walk Test (2MWT).

4.2 The FLOODLIGHT Proof-of-Concept Study

The FLOODLIGHT PoC Study aimed to evaluate participant adherence to smartphone- and smartwatch-based assessments, as well as exploring the ability of sensor-based outcomes to capture aspects of the symptoms of MS remotely. As introduced in chapter 2, the use of digital technologies to assess PwMS has gathered significant momentum in recent years [130, 173, 274, 275]. High-frequency monitoring assessments may be more accurate than conventional outcomes recorded at periodic visits in detecting subtle progressive sub-clinical changes, that may predict disease activity or long-term disability in PwMS [130]. Furthermore, it has been shown that an earlier identification of changes in PwMS impairment are important to identify and provide better therapeutic strategies [33]. The NCT02952911 study combined continuous sensor data captured with smartphones and smartwatches and standard clinical outcome measures [31]. Participants were asked to perform a variety of digital assessments (ranging from daily to bi-weekly) over a 24 week duration of the study. These assessments aimed to evaluate various domains of physical and

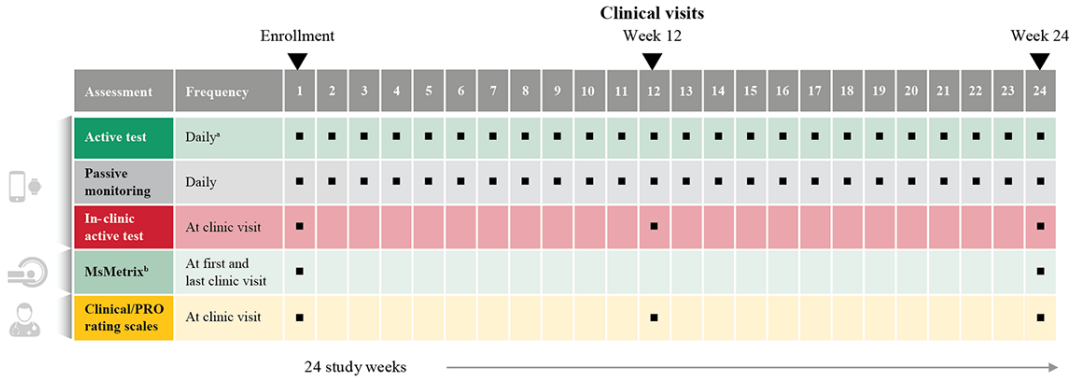


Figure 4.2: The FLOODLIGHT PoC (NCT02952911) study design. PRO: patient-reported outcome. [a] indicates that active tests were administered weekly or every two weeks. [b] indicates that brain magnetic resonance imaging was performed in people with multiple sclerosis. Taken from: [31].

cognitive function, including upper-extremity (UE) function or hand-motor function (HMF), ambulatory or gait function, as well as self-reported outcomes based on a subject's perception of their daily symptoms (or lack thereof) . Each smartwatch and smartphone contained pre-installed apps that prompted the user to perform various assessments and protocols, known as "active tests". Active tests included the Draw-a-Shape Test (DaS), a remote Two-Minute Walk Test (2MWT), a static balance test (SBT), a 5-U-Turn Test (5UTT), as well as an electronic version of the Symbol Digit Modalities Test (SDMT). The app also passively recorded sensor data during daily-life, referred to as "passive monitoring" [31]. Figure 4.1 outlines the FLOODLIGHT testing suite, comprised of active tests, passive tests, as well as those assessments performed in-clinic, and their corresponding administering frequency. This frequency of assessment is detailed further in the study design depicted in figure 4.2 . For more information on the NCT02952911 study we refer the reader to [31].

4.2.1 Study Material and Equipment

Subjects were provided with a Samsung Galaxy S7 smartphone and Motorola 360 Sport smartwatch. Both smartphone and smartwatch devices contain 3-axis accelerometer ($\boldsymbol{a}_x, \boldsymbol{a}_y, \boldsymbol{a}_z$) and gyroscope ($\boldsymbol{g}_x, \boldsymbol{g}_y, \boldsymbol{g}_z$) sensors which were sampled at 50 Hz. Each subject was also provided with a waist-worn running belt and instructed to attach the smartphone to the anterior of their waist. The smartwatch could be

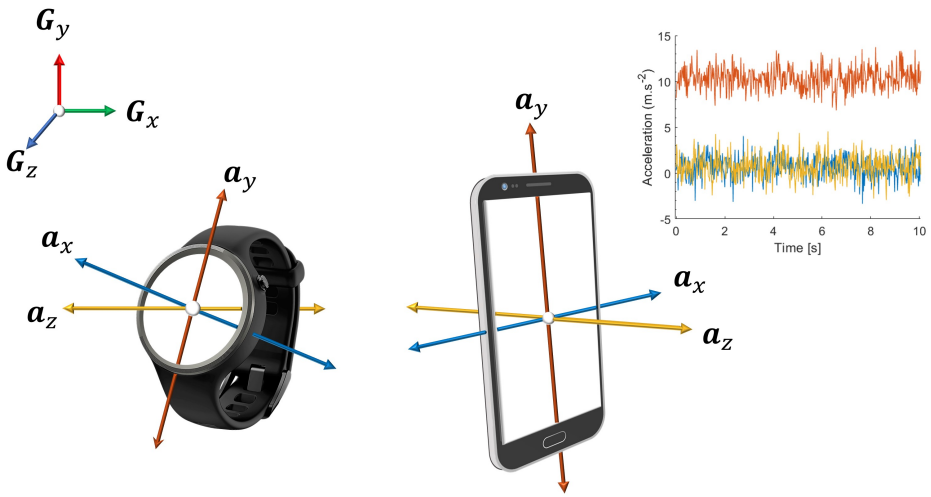


Figure 4.3: Smartphone and smartwatch device inertial sensor coordinates. Mock up depicts the Samsung Galaxy S7 smartphone and Motorola 360 Sport smartwatch provided to participants in FLOODLIGHT. Both devices contain a 3-axis accelerometer sensor, $(\boldsymbol{a}_x, \boldsymbol{a}_y, \boldsymbol{a}_z)$, which at any instantaneous time point t is incident to the direction of gravity (vertical), defined globally as $(\boldsymbol{G}_x = 0, \boldsymbol{G}_y = -9.81\ m.s^{-2}, \boldsymbol{G}_z = 0)$.

worn on either wrist. Figure 4.3 depicts the sensor coordinates for smartphone and smartwatch devices, as well as the coordinate reference frame with respect to gravity.

4.2.2 Study Participants

People with multiple sclerosis and healthy controls, preferentially partners or cohabitants, were enrolled in the NCT02952911 study. Among eligibility criteria included a baseline EDSS score [0–5.5] and age 18–55 [31]. A total of n=25 HC and n=76 PwMS were recruited, where the overall demographic and clinical information is detailed in table 4.1. Note: Subsets of the NCT02952911 study data were used in the subsequent results chapters 5 and 6 based on subjects that contributed sufficient data for each respective analysis and are detailed more in the respective chapters: Draw-a-Shape (chapter 5) and gait 2MWT (chapter 6).

4.2.3 Clinical Assessment

At each scheduled in-clinic visit (baseline[1], week 12, and week 24), a number of clinical assessments were performed. This thesis will focus on specific clinical

[1]Often denoted as week 0.

Table 4.1: Demographics for healthy controls (HC) and people with multiple sclerosis (PwMS). Results reported as the mean ± standard deviation across the population; [s] indicates measurement in seconds; PPMS, Primary Progressive Multiple Sclerosis; SPMS, Secondary-Progressive Multiple Sclerosis; RRMS, Relapse Remitting Multiple Sclerosis; EDSS, Expanded Disability Status Scale; T25FW, the Timed 25-Foot Walk; 9HPT, 9-Hole Peg Test; NA, not applicable.

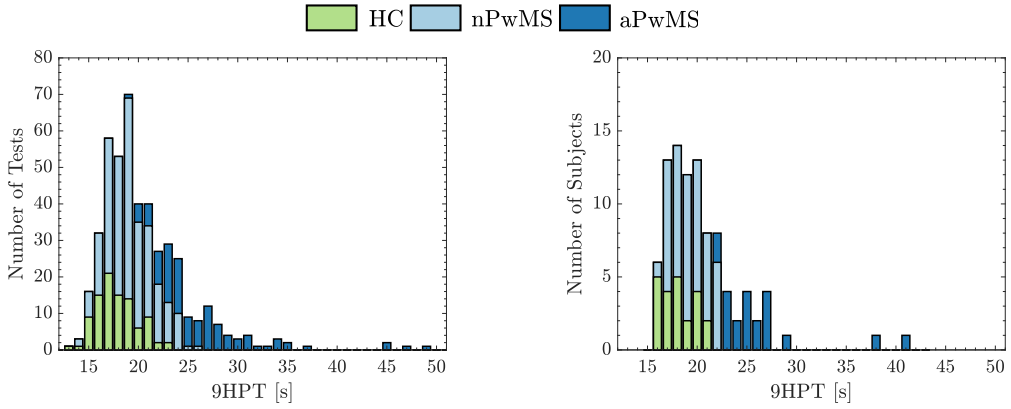
	HC (n=25)	PwMS (n=76)
Age (years)	34.9 ± 7.9	39.5 ± 7.9
Female, n (%)	7 (28%)	53 (70%)
MS diagnosis, n (%)		
PPMS	NA	3 (4%)
SPMS	NA	4 (5%)
RRMS	NA	69 (91%)
Time since MS symptom onset (years)	NA	11.3 ± 7.0
Relapse in the past year[1], n (%)	NA	18 (24%)
EDSS	NA	2.4 ± 1.4
T25FW [s]	5.0 ± 1.0	6.0 ± 2.1
9HPT [s]		
dominant hand	18.9 ± 2.1	22.1 ± 4.6
non-dominant hand	19.5 ± 2.0	22.8 ± 4.9

[1] Proportion of PwMS with ≥ 1 relapse in the past year.

assessments, such as the Expanded Disability Status Scale (EDSS) [19], the 9-Hole Peg Test (9HPT)[2] [23] and the Timed 25-Foot Walk (T25FW) test [21]. These three assessments form an integral part of the Multiple Sclerosis Functional Composite (MSFC) (see background section 2.2). Other reference clinical tests were also performed as well as a number of patient reported outcomes (PRO) [24], such as the Multiple Sclerosis Impact Scale–29 (MSIS-29) [276, 277]. For more information on the various means of assessing MS disability, as well as an appraisal thereof, we refer the reader back to sections 2.2 and 2.2.1 respectively. The distribution of 9HPT times and EDSS scores in FLOODLIGHT PoC are depicted in figure 4.4 and figure 4.5, respectively. Both figures represent the distribution of all available clinical scores (baseline, week 12, and week 24), as well as the average score per subject in the study. Histograms are coloured by the respective PwMS groups which form part of the analysis in chapter 5 and chapters 6–8, where various PwMS groups are introduced in more detail.

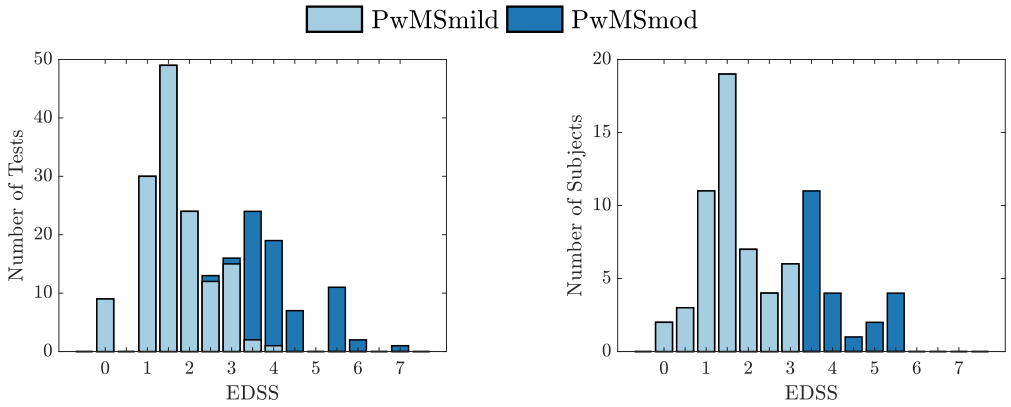
Participants were also asked to report relapses on the smartphone app, termed the Multiple Sclerosis Symptom Tracker (MSST) [31]. In this, *"patients were asked*

[2] The 9HPT forms an integral part of chapter 5, and will be discussed in more detail.



(a) All 9HPTs **(b)** Average 9HPT

Figure 4.4: Distribution of Nine-Hole Peg Test (9HPT) times. Figure **(a)** depicts the distribution of 9HPT times [s] from all available clinical visits, whereas **(b)** depicts the average 9HPT per subject. 9HPTs are pooled for dominant and non-dominant handed tests. PwMS were grouped into presumably normal (nPwMS) (9HPT $\leq$ 22.7 [s]) versus abnormal (aPwMS) hand/arm function based on a upper limit of normal range (9HPT >22.7 [s]), defined as the average 9HPT times for HCs plus two standard deviations from normative data. See chapter 5 for more details.



(a) All EDSS scores **(b)** Average EDSS score

Figure 4.5: Distribution of Expanded Disability Status Scale (EDSS) scores. Figure **(a)** depicts EDSS scores from all available clinical visits, whereas **(b)** depicts the average EDSS per subject. PwMS subjects were grouped into those with presumably mild and moderate symptoms as defined by the EDSS. By definition, gait disorders begin to become prominent in subjects with EDSS$\geq$3.5 and subjects with EDSS<3.5 are considered mildly impaired. See chapter 6 for more details.

if they experienced new or significantly worsening symptoms during the last 2 weeks. If yes, onset of the symptoms and the patients' perception to whether they think they experienced a relapse (yes, no, or unsure) are recorded". For more information on the study assessments in the FLOODLIGHT PoC we refer the reader to [31].

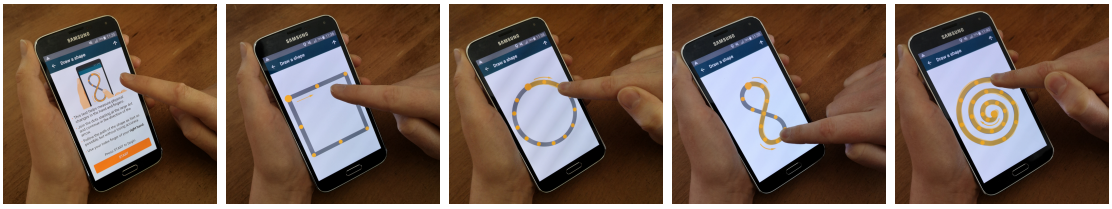


Figure 4.6: Demonstration of the Draw-a-Shape test performed by a healthy participant using a smartphone. Each participant was instructed to draw six different shapes presented on the smartphone screen as fast and as accurate as possible, within a maximum time of 30 seconds per each attempted shape: a diagonal *line* bottom left to top right and a diagonal *line* top right to bottom left (not depicted), a *square*, a *circle*, a *figure-8-shape*, and a *spiral*.

4.3 Study Activities

The FLOODLIGHT application aimed to asses a range of domains which could be affected in PwMS through the various tests collected at specific and continuous intervals. This thesis aims to focus on the assessments of physical function, namely upper-extremity function and lower-extremity function (or ambulation). The DaS test and the 2MWT are prime candidate assessments respectively to expand and understand from the FLOODLIGHT PoC study. The DaS test aimed to capture upper-extremity function measures as participants traced various pre-defined shapes on a smartphone touch-screen. Conversely, the remotely performed 2MWT aimed to capture lower-limb, or ambulatory and gait-related function recorded from inertial sensors within smartphone and smartwatch devices.

4.3.1 The Draw-a-Shape Test

The Draw-a-Shape test (DaS) instructed all study participants daily to draw six different shapes presented on the smartphone screen as fast and as accurate as possible, within a maximum time of 30 seconds per each attempted shape. The six shapes to be drawn were a diagonal *line*[3] bottom left to top right, a diagonal *line* top right to bottom left, a *square*, a *circle*, a *figure-8-shape*, and a *spiral*. The drawing had to be performed with the index finger of the tested hand; subjects were promoted to alternate each day between their dominant and non-dominant hand. Each shape

[3] *Line* shapes were not considered in this thesis

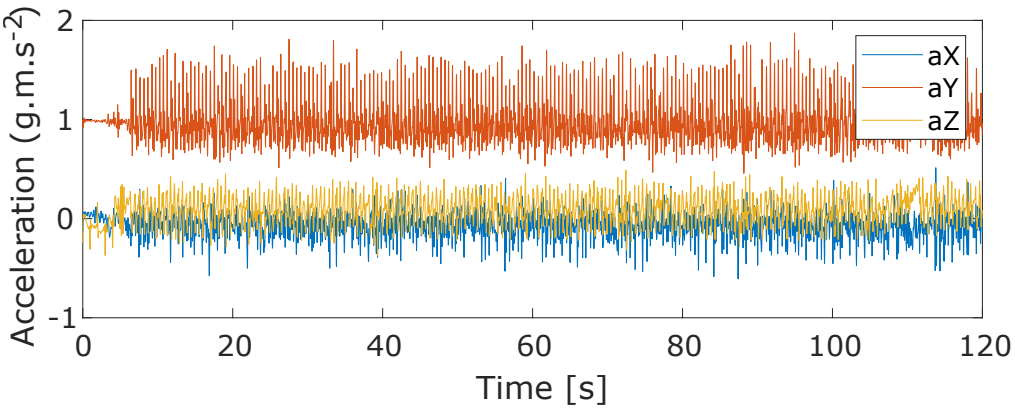


Figure 4.7: Example of Two Minute Walk Test (2MWT) smartphone-based inertial sensor data. Figure depicts a healthy control (HC) during a remotely executed 2MWT as part of FL.

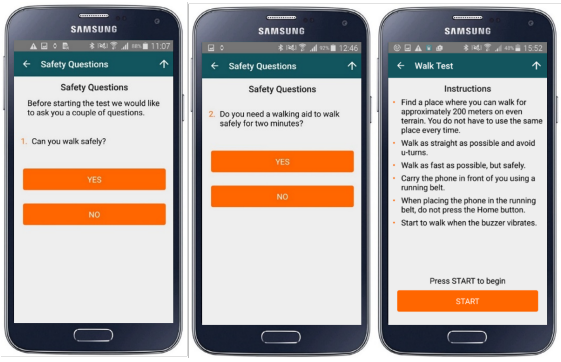


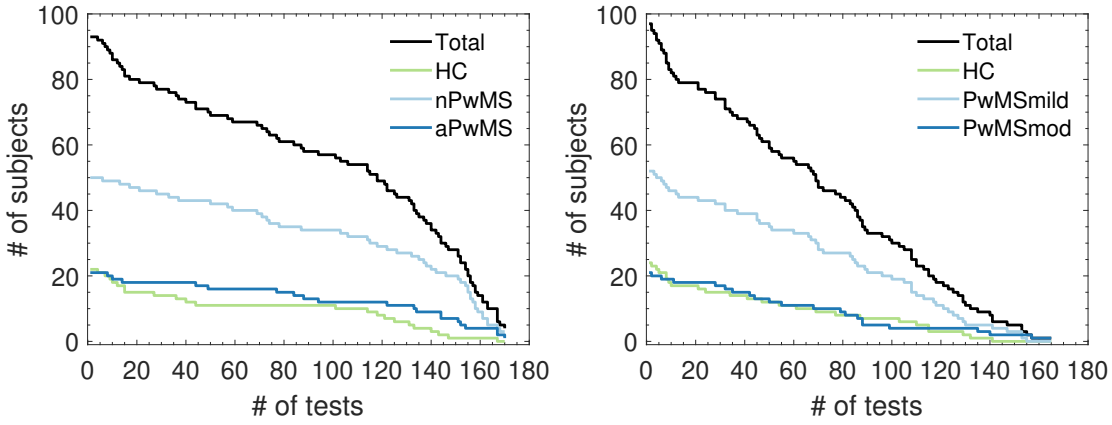
Figure 4.8: Mock up depicting the instructions given to participants to perform the remote smartphone-based Two Minute Walk test (2MWT).

could be attempted up to twice[4]. An example of a healthy volunteer performing the DaS test suite is demonstrated in figure 4.6. The DaS test is presented in more detail, as well as examples of raw shape data, within chapter 5. The number of tests contributed by each subject for the DaS test are depicted in figure 4.9a.

4.3.2 The Two-Minute Walk Test

The *remote* Two-Minute Walk Test (2MWT) instructed participants to walk as fast and as long as they can for two minutes. Participants were asked to walk safely and as straight as they could for at least 200 meters without U-turns. While the instructions given were analogous to those of a *standard* 2MWT, this was not a controlled and clinically assessed 2MWT during site-visits, and therefore the

[4]Only first attempts are considered in this thesis. See chapter 5 for more details.



(a) Draw-a-Shape Test **(b)** Two-Minute Walk Test

Figure 4.9: Survival curves depicting participant adherence rates. Each line depicts the number of **(a)** DaS tests and **(b)** remote Two Minute Walk Test (2MWT) recordings contributed for each subject over the study duration. Adherence is further subdivided into the contributory groups for DaS analysis (HC, nPwMS, aPwMS) and FL 2MWT analysis (HC, PwMSmild, PwMSmod). See the respective DaS, chapter 5, and 2MWT, chapter 6, for an expansion of these tests, subject groups, and subsequent analysis.

outcome of walking distance was not measured. Figure 4.7 depicts an example of smartphone-based inertial sensor data captured from a healthy participant during a remotely executed 2MWT. Figure 4.8 presents a mock up of the instructions given to participants in FL to perform the remote 2MWT. For information on the 2MWT we refer the reader to chapter 6. The number of tests contributed by each subject for the 2MWT are depicted in figure 4.9b.

4.4 Human Activity Recognition Datasets

This section briefly introduces the open-source Human Activity Recognition (HAR) datasets explored as part of the analysis in chapter 7 and chapter 8.

4.4.1 UCI HAR

The "Human Activity Recognition (HAR) Using Smartphones Data Set", denoted UCI (*smartphone*) HAR[5], is a publicly available dataset for HAR [278]. The annotated dataset consists of 10,299 instances of labelled activities (walking, stairs, sitting, standing and laying). Measurements were collected from 30 individuals

[5]The UCI HAR dataset can be found at: `https://archive.ics.uci.edu/ml/datasets/human+activity+recognition+using+smartphones`

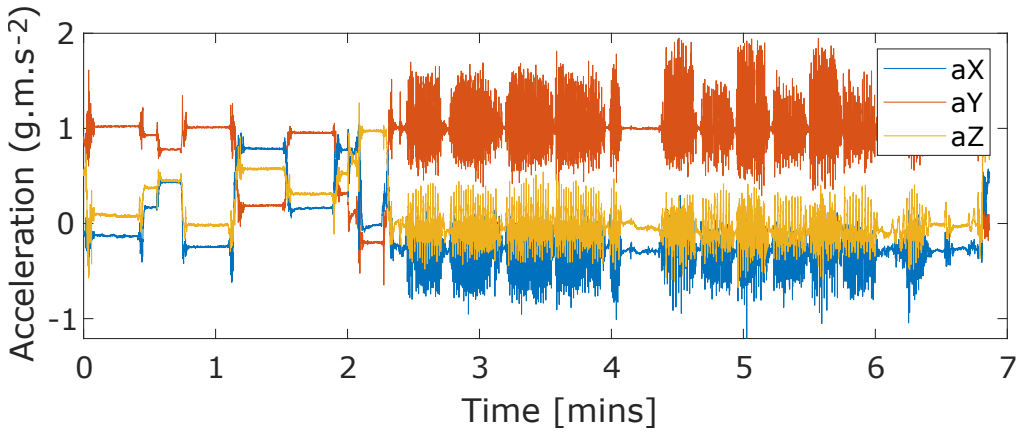


Figure 4.10: Example of smartphone-based inertial sensor data captured from a participant in the UCI HAR dataset.

ranging in age from 19 to 48 years recorded wearing a smartphone (Samsung Galaxy SII, sampled at 50 Hz) affixed to the waist. An example recording from a participant in UCI HAR is depicted in figure 4.10.

4.4.2 WISDM

The "WISDM (Wireless Sensor Data Mining) Smartphone and Smartwatch Activity and Biometrics Dataset"[6] is another publicly available dataset for HAR, among other ADL tasks [279] [280]. Data was collected from 51 subjects, ranging in age from 18-25, each of which had a smartphone (Google Nexus 5/5x or Samsung Galaxy S5 sampled at 20 Hz) in their pocket and a smartwatch on their arm, where they performed 18 different tasks for 3 minutes at a time. Those activities of interest for this study were (walking, jogging, stairs, sitting and standing).

[6]The WISDM dataset can be found at: `https://archive.ics.uci.edu/ml/datasets/WISDM+Smartphone+and+Smartwatch+Activity+and+Biometrics+Dataset+`

5

Smartphone-based Remote Characterisation of Upper Extremity Function using the Draw a Shape Test

Associated publication(s):
Creagh, A. P., Simillion, C., Scotland, A., Lipsmeier, F., Bernasconi, C., Belachew, S., van Beek, J., Baker, M., Gossens, C., Lindemann, M. and de Vos, M. (2020). "Smartphone-based remote assessment of upper extremity function for multiple sclerosis using the Draw a Shape Test". *Physiological Measurement* 41(5): 054002. https://doi.org/10.1088/1361-6579/ab8771

Contents

5.1 Introduction 80
5.2 Background 82
5.3 Methods 83
5.4 Results 91
5.5 Discussion 100
5.6 Conclusion 106

5.1 Introduction

Developing digital health outcomes, *digital biomarkers*, for multiple sclerosis (MS) is a nascent field, where the assessments and metrics evaluated are only in the

early proof-of-concept (PoC) stage of development. To form metrics that may help characterise, predict, or monitor MS disability and symptom manifestation, we must start from first principles. This requires the cultivation of relevant and robust measurements that are able to encapsulate information related to disease and disability. The first steps for these new digitally augmented assessments are to benchmark against established clinical metrics. This litmus test, which aims to reproduce a clinical gold-standard assessment, builds a validity in the measures developed, and allows for further exploration and characterisation of the disease domain. The resulting richer evaluation occurs naturally due to the nature of digital assessments, which are highly granular and can be administered more frequently. Thus, this direct comparison to the established measures not only demonstrates robustness in the digital measurements, but also highlights the greater wealth of information that can be captured in digitally augmenting those traditional physician-lead assessments.

In this chapter we will introduce the “Draw-a-Shape" (DaS) test, a manual dexterity-based digital drawing assessments. This work will first introduce novel DaS-related features that characterise upper-extremity (UE) function and impairment. To benchmark these *digital* measurements, this chapter aims to demonstrate how multivariate modelling of these metrics can reliably predict the 9-Hole Peg Test (9HPT), a clinician-administered UE assessment in people with multiple sclerosis (PwMS).

The primary contributions of this chapter to the field of remote monitoring of multiple sclerosis disease can be considered as:

- Development of a robust and scalable Draw-a-Shape feature pipeline which can be extracted from a smartphone touchscreen;

- Demonstration of how novel features motivated by disease pathology and UE function can be extracted from various shapes (*circle*, *figure-8-shape*, *spiral*, *square*) traced on smartphone touchscreen;

- The characterisation of how examples of these measures relate to UE impairment in PwMS, including the evaluation of various digitised shapes to capture UE PwMS performance;
- Evincement how multivariate modelling of DaS features can reliably predict the 9HPT, an established gold-standard UE assessment in PwMS;
- The establishment of a methodological foundation, critical considerations, and analysis techniques thereof, for evaluating remotely generated UE function derived smartphone PwMS data;
- An appraisal of the limitations associated with in-clinic manual dexterity assessments, such as the 9HPT compared to digitised assessments like Draw-a-Shape.

5.2 Background

5.2.1 The Nine-Hole Peg Test

As eluded to in section 2.2, while various performance tests and patient-reported outcome measures are available [281], the 9-Hole Peg Test (9HPT) is the most frequently used measure of manual dexterity in MS research, clinical trials and clinical practice [23, 91]. To recap, the 9HPT requires participants to repeatedly place and then remove nine pegs into nine holes, one at a time, as quickly as possible [23] and is measured in seconds [s]. The inter-rater and test-retest reliability of the 9HPT is generally high across a range of disability levels, however most studies focus on more disabled PwMS populations [23]. One large scale study with a healthy population for example reports high inter-rater reliability but only moderate test-retest reliability [111]. Regardless, the 9HPT has satisfactory discriminative and ecological validity in PwMS [94], although the coarse nature of this test and its infrequent in-clinic administration has inherent limitations. For instance, it may not be suited for early and sensitive detection of insidious and slowly evolving change of UE function as seen in progressive MS. While the 9HPT measure is restricted to

the stopwatch collection of the overall time to complete a manual dexterity task, it lacks the capacity to discriminate and quantify variable qualitative patterns of alteration of hand or finger motor skills and cannot capture intra-task fluctuations of performances. There is hence a need for more continuous self-administered and refined outcome measures to assess more comprehensively and reliably manual dexterity in MS and other chronic neurological disorders affecting UE function.

5.2.2 Digitally Augmenting Manual Dexterity Assessments

Manual dexterity assessments are easy to implement using smartphones. As section 2.3.3 reviewed, many digital augmentations of UE assessments have been proposed. There is however a lack of clear representation for how digital-based UE assessments relate to (and by proxy, can augment) gold-standard traditional physician-lead assessments.

5.3 Methods

5.3.1 Dataset

The Draw-a-Shape test, as part of the FLOODLIGHT PoC study (FL) instructed PwMS and healthy controls (HC) to trace predefined shapes on a smartphone screen (see chapter 4 for more details on the DaS test). A total of 93 subjects (HC, n=22; PwMS, n=71) contributed both dominant and non-dominant DaS tests for analysis in this study. Subjects were divided into normal (nPwMS) and abnormal (aPwMS) subgroups with respect to their average combined dominant and non-dominant 9HPT times over the entire study [23]. The threshold for abnormal UE function was defined by the average 9HPT times greater than the mean plus 2 standard deviations from normative data on a healthy population, pooled on dominant (9HPT threshold: 17.8 + 2(2.2) [s]) and non-dominant (9HPT threshold: 18.5 + 2(2.3) [s]) tests, where normative values were taken from [282]. Hence, the aPwMS subgroup consisted of PwMS with average 9HPT times >22.7 [s] and the nPwMS subgroup of PwMS with average 9HPT of $\leq$22.7 [s].

Table 5.1: Demographics and characteristics of data subset used in this chapter, for Healthy Controls (HC) and PwMS subgroups, stated as mean ± SD, over the entire study where applicable;

	HC (n=22)	nPwMS (n=50)	aPwMS (n=21)	p[(a)]
# tests per subject	74 ± 59	114 ± 51	107 ± 56	*[1]
Age, year	34 ± 9	40 ± 8	40 ± 8	n.s.[1]
Dominant hand (Right/Left)	(19/3)	(44/6)	(20/1)	n.s.[2]
Male/Female	15/7	14/36	8/13	**[2]
EDSS	NA	2.1 ± 1.26	3.3 ± 1.4	***[1]
9HPT, seconds				
(dominant)	18.3 ± 1.7	19.4 ± 2.1	26.3 ± 5.4	***[1]
(non-dominant)	19.1 ± 1.8	20.1 ± 1.6	27.2 ± 5.0	***[1]
p[(b)]	*[3]	**[3]	n.s.[3]	

nPwMS: PwMS with average 9HPT ≤ 22.7 [s]; aPwMS: PwMS with average 9HPT >22.7 [s]; EDSS, Expanded Disability Status Scale; 9HPT, 9-Hole Peg Test; NA, not applicable.
[(a)] p-value between groups [(b)] p-value between hands
[1] Kruskal-Wallis by ranks tests the null hypothesis that the continuous data in each categorical group (HC, nPwMS, aPwMS) comes from the same distribution;
[2] Chi-squared (χ^2) tests differences between categorical groups (HC, nPwMS, aPwMS);
[3] Wilcoxon signed rank tests 9HPT values between hands used for each group;
* $p<0.05$; ** $p<0.01$; *** $p<0.001$, n.s. - not significant.

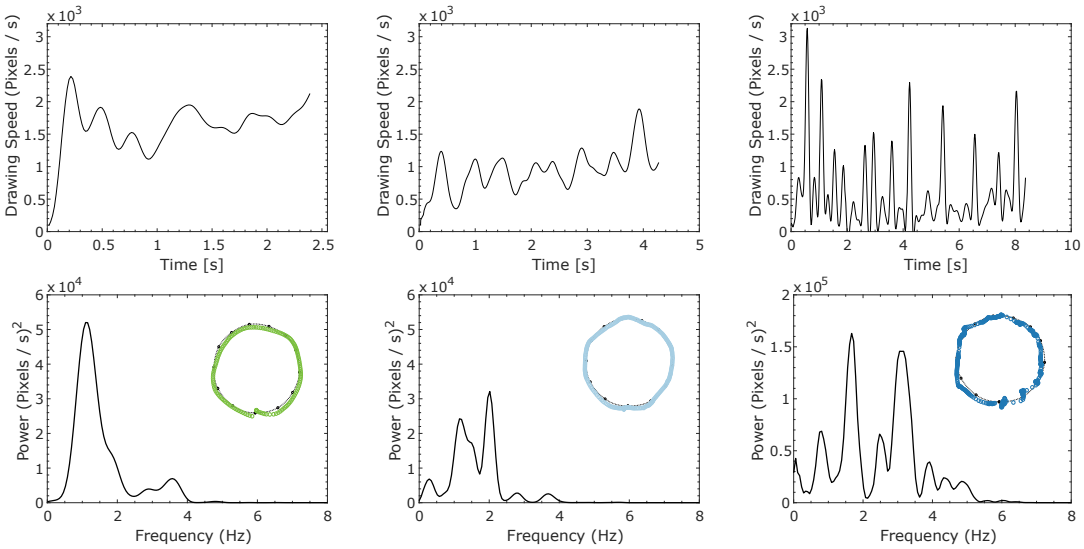
5.3.2 Feature Extraction

5.3.2.1 Raw Data Processing

Raw sensor data was collected from the smartphone touchscreen during the active DaS test and stored as x- and y-screen coordinates with a corresponding timestamp t, (x, y, t). A bespoke MATLAB script extracted attempted and completed shapes from each test, along with the corresponding hand used. All first attempts were used for further feature analysis. All data processing was performed using MATLAB vR2018a (The MathWorks, Natick, MA, USA).

5.3.2.2 Characterization of Manual Dexterity Performance

Multiple features were extracted from each shape capturing temporal, spatial and spatiotemporal aspects involved in the drawing task and potentially reflective of manual dexterity. Furthermore, overall test performance statistics were calculated, such as the time taken to complete all shapes and the number of shapes completed. For a full list of the features extracted from DaS, please see the accompanying appendix A.4.2. A selection of some relevant features are described below and illustrated in figures 5.1-5.4.



(a) HC **(b)** nPwMS **(c)** aPwMS

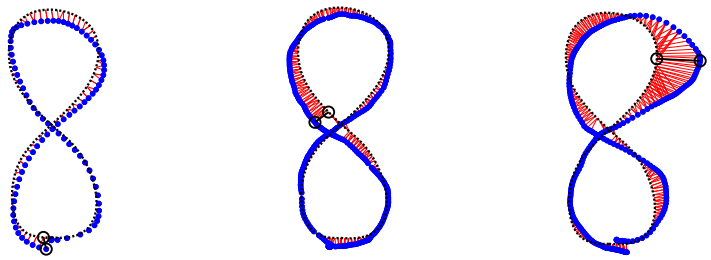
Figure 5.1: Example illustrations of *circle* shape drawn by HC, nPwMS, and aPwMS. Figure depicts **(a)** HC (9HPT 18.4 ± 1.2 [s]); **(b)** nPwMS (9HPT 20.2 ± 2.0); and **(c)** aPwMS (9HPT 25.0 ± 2.1 [s]) subjects. Coloured points depict actual pixel points drawn relative to interpolated reference coordinates (black) for HC (green), nPwMS (light blue) and aPwMS (dark blue). The top row demonstrates examples of drawing speed for duration of time to draw each respective shape. Time series speed signal was first filtered using a low pass filter with a cut off frequency of 8 Hz. The bottom row represents the power spectral density (PSD) estimate of drawing speed which was computed using a Hamming window. Note the time and PSD axis scale between the figures.

5.3.2.3 Temporal Features

Temporal features including drawing velocity, angular and radial velocities and speed distribution measures were computed to assess temporal irregularities, such as delays, smoothness, jerkiness, and rapid finger/hand movement [135, 136, 283]. The dominant frequency and power spectral density was measured for frequencies between 1–7 Hz, which was aimed to capture potential tremulous actions like that of cerebellar intention tremor or to record ataxic movements, both commonly exhibited in PwMS [136, 282]. Examples of drawing speed and power spectral density (PSD) estimate of drawing speed are illustrated in figure 5.1.

5.3.2.4 Spatial Features

Spatial aspects of finger or hand movements were captured by features based on drawing error [134, 136, 283]. A new approach to compute drawing error is presented in this study based on a shape-matching approach known as the Hausdorff distance



(a) HC **(b)** nPwMS **(c)** aPwMs

Figure 5.2: Example illustrations of *figure-8-shapes* drawn by HC, nPwMS, and aPwMS subjects. Figure **(a)** HC (9HPT 18.4 ± 1.2 [s]); **(b)** nPwMS (9HPT 20.2 ± 2.0); and **(c)** aPwMS (9HPT 25.0 ± 2.1 [s]) subjects. Figure depicts actual pixel points drawn (blue) relative to interpolated reference coordinates (black). Hausdorff Distance query points are illustrated with red lines and maximal Hausdorff Distances ($HausD$, as measured in Pixels) are highlighted with black circles; HC (64 Pixels) , nPwMS (90 Pixels), aPwMS (229 Pixels). The total drawing error ($HausDError$) can also be defined as sum of the Hausdorff distances (i.e. the largest minimum distances) between the drawn and reference shape, normalised by the number of touch coordinates drawn. In this example the normalised drawing errors are: HC (31 Pixels), nPwMS (38 Pixels), aPwMS (68 Pixels).

[284, 285]. Let X and Y be two-non empty subsets of a metric space (M, d). The Hausdorff distance $d_H(X, Y)$ is defined as:

$$d_H(X,Y) = \max\left\{\sup_{x\in X}\inf_{y\in Y} d(x,y), \sup_{y\in Y}\inf_{x\in X} d(x,y)\right\} \tag{5.1}$$

where sup is the supremum and inf is the infimum and distance d is computed as the Euclidean ℓ_2 norm (equation 3.42). This metric compares the maximum distance of one set to the nearest point in another set [286], which can be used as a basis to compute the error between the reference way-points (interpolated into a reference shape scaled to the number of pixels drawn) and the subject's drawing attempt. The maximal Hausdorff distance is a measure of the absolute deviation from the reference shape, while the total drawing error can also be defined as sum of the Hausdorff distances (i.e. the largest minimum distances) between the drawn and reference shape, normalised by the number of touch coordinates drawn. An example of Hausdorff distances can be found in figure 5.2.

5.3.2.5 Spatiotemporal Features

Digital drawings are unique in that they encapsulate spatial and temporal performance information simultaneously: each pixel point contains 3-D data relating to the persons hand movement at that time. This information is exploited to create discretized heat maps of touch events (x, y, t). A heat map not only gives a visual representation of performance but can also be used to extract a further important sub-set of features which may be sensitive to motor control or disease fluctuations.

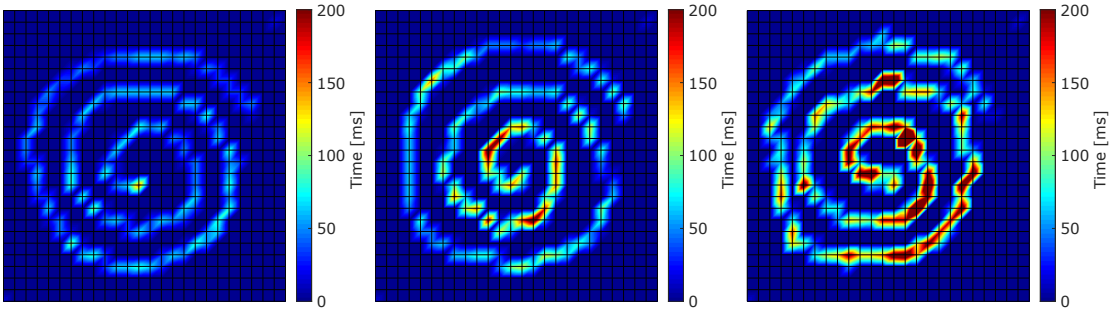
In order to compare intensity maps for image analysis, each shape drawn is scaled to the same coordinates, while the timescales and relating colour intensities are based on global not local pixel counts. Pixels are binned into a coarser grid, both for graphical purposes and as a means to allow comparisons between subjects for each shape drawn. A graded single colour intensity scheme is used to represent pixel densities, which is transposed into an achromatic scale for image feature extraction.

Image features are extracted from both the shape drawings and their transposed heatmaps. An example of a discretised heatmap is illustrated in figure 5.3. Pixel intensities measure the structural composition of such images, while the drawings are also compared with an ideal drawing[1] for similarities using measures such as 2-D image correlation coefficient, or Mutual Information (MI) between the two images [287]. Image entropy of heat map-transposed shape drawings (converted to grayscale) was calculated using:

$$H = -\sum_{k} p_k \log_2(p_k) \tag{5.2}$$

Where k is the number of grey levels and p_k is the probability associated with each grey level k. Entropy is a commonly used measure of disorder in a system and can be used in image analysis for texture mapping [288]. The topography of transposed pixel intensity drawings become a function of finger movements and hence entropy a measure of smooth, non-hesitant drawing.

[1] An ideal drawing is calculated by interpolating reference shape coordinates scaled to the same number of pixel points, irrespective the number of drawn pixels in the subject's attempt.



(a) HC **(b)** nPwMS **(c)** aPwMS

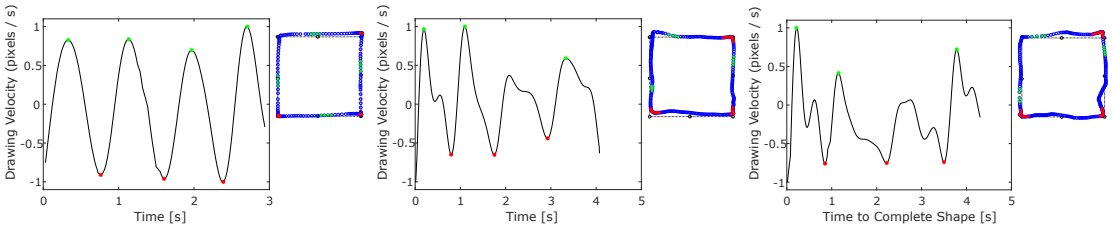
Figure 5.3: Pixel density heat map representation of *spiral* shapes drawn by HC, nPwMS, and aPwMS subjects. Figure depicts: **(a)** HC (9HPT 18.4 ± 1.2 [s]); **(b)** nPwMS (9HPT 20.2 ± 2.0); and **(c)** aPwMS (9HPT 25.0 ± 2.1 [s]) subjects. Screen coordinates are first segmented into 2D bins of fixed width and drawing touch point coordinates are assigned to respective bins. The number of touch coordinates per bin, and hence time, are represented by heat map colour. This builds a spatio-temporal representation of digital *spiral* drawing which encode areas of drawing hesitation or non-movements.

Further features capturing hesitation times and aspects of fine directional changes during drawing were calculated to capture elements of cognitive motor inferences known to affect UE function in MS [75]. Finally, a new measure, celerity, was defined by calculating the ratio of successfully passed waypoints divided by the time taken to complete the shape.

5.3.3 Regression Model to 9HPT

5.3.3.1 Data Selection

This study aims to investigate the prediction of clinical 9HPT times using features relating to UE function computed from the DaS test, where the design matrix $\mathbf{X} \in \mathbb{R}^{N \times P}$ contains the median and standard deviation of each feature per subject over all available test days for dominant and non-dominant handed tests separately. Response variable, $\mathbf{Y} \in \mathbb{R}^{N \times 1}$, were denoted as the average 9HPT time per subject over the entire study (all baseline, week 12, week 24/study completion observations considered) for each respective dominant and non-dominant handed 9HPT separately. It is assumed that drawing performance will generally vary depending on dominance [136, 282], therefore independent models were evaluated based on dominant and non-dominant hands used.



(a) HC **(b)** nPwMS **(c)** aPwMS

Figure 5.4: Example of *square* shapes drawn by HC, nPwMS and aPwMS subjects. Figure depicts: **(a)** HC (9HPT 18.4 ± 1.2 [s]); **(b)** nPwMS (9HPT 20.2 ± 2.0); and **(c)** aPwMS (9HPT 25.0 ± 2.1 [s]) subjects. Normalised drawing velocity is shown with local maximum and minimum speed highlighted in green and red respectively. Corresponding points on shape drawing are also illustrated where local minimum velocity points represent dwell (hesitation) time at corner locations. Note the time axis between figures.

5.3.3.2 Statistical Analysis

Features were assessed for non-normality both by visual inspection and using a Kolmogorov-Smirnov (KS) test (3.4.2.4). Those non-normal features were transformed using box-cox transformations (A.2.6). Pre-processing assessment of the 9HPT response variable displayed a highly-tailed distribution, and as such the 9HPT was also transformed back towards Gaussianity to help fulfil error assumptions of linear prediction [289].

Differences in median clinical metrics (EDSS, 9HPT) and feature values across subject-groups (HC, nPwMS, aPwMS) were tested using a Kruskal-Wallis test (KWt) by ranks (see section 3.4.2.8), and pairwise using a Mann-Whitney U test. Categorical differences in sex were investigated using a Chi-squared (χ^2) test (see section 3.4.2.3). A Brown-Forsythe test (BF) was used to evaluate the null hypothesis that the data in each categorical subject groups (HC, nPwMS, aPwMS) have equal variances, against the alternative that at least two of the data samples do not (see section 3.4.2.9). Wilcoxon signed rank tests (see section 3.4.2.5) were used to investigate differences between dominant and non-dominant handed (non-normal) DaS features and dominant and non-dominant handed 9HPTs, prior to Box-Cox transformation. Handed differences between normal features were evaluated using paired t-tests. Pairwise differences in model residuals and model prediction errors between hands and between models were also assessed

using a Wilcoxon signed rank test. Pearson's correlation (r) and Spearman's rank correlation (ρ) were used to assess the association of features to the 9HPT time univariately, and prior to any applicable box-cox transformations (see section 3.3.2). P-values were corrected using methods described in section 3.4.3 in cases where multiple hypothesis testing was performed.

5.3.3.3 Model Generalisability

To determine the generalisability of our models, stratified 5-fold subject-wise cross-validation (CV) was employed. This consisted of randomly partitioning the dataset into k=5 folds which was stratified with equal proportions of HC, nPwMS and aPwMS where possible. One set was denoted the training set (in-sample), which was further split for into smaller set for parameter selection (validation) using an internal 5-fold CV approach. The remaining data was then denoted testing set (out-of-sample) on which predictions were made. CV was repeated 10 times with new random partitions in order to reduce bias in re-sampling and dataset splitting.

5.3.3.4 Model Evaluation

In order to reveal the (potentially nonlinear) functional relationship between the DaS Test (as represented by the features extracted) and the associated 9HPT, a number of regression models were evaluated based on MAE and RMSE (refer to sections 3.11.4). Least absolute shrinkage and selection operator (LASSO) feature selection was employed (see section 3.6.3 to prevent overfitting and reduce the dimensionality (P) of the DaS features. A top feature ranking table was deduced by interrogating the feature subsets selected by LASSO at each fold and repetition. The relative stability of features selected was assessed by recording the percentage of time that feature is selected at each fold and repetition.

It has been suggested that bias or prediction error can be decreased by performing a separate regression post-LASSO [290]. As such, features were selected using LASSO and those features each presented to linear models where this study investigated the performance of ordinary least squares (OLS), and iteratively re-weighted least squares (IRLS), function (see sections 3.5.1 and 3.5.2 respectively).

It is possible that the DaS features do not combine linearly to predict the 9HPT and therefore non-linear regression was also explored. Support Vector Regression (SVR) with linear and Gaussian radial bias function (RBF) kernel [234] were tested, as well as Random Forest Regression (RFR) [291] (see sections 3.7.3 and 3.7.4 respectively). In the case of SVR, features were selected prior for each CV-fold using LASSO. RF regressors were built over all features and optimised over varying numbers of trees and the number of input variables chosen at each node (see section 3.7.4).

Regression models were evaluated through the coefficient of determination r^2, mean absolute error (MAE), and the root-mean squared error (RMSE) (see section 3.11.4 for further details).

5.4 Results

PwMS subjects in this study were stratified into those with presumably normal (nPwMS) and abnormal (aPwMS) UE function with respect to their pooled average 9HPT times. Table 5.1 represents the demographic information per subject group, HC, nPwMS, aPwMS. The effects between different MS phenotypes such as primary progressive multiple sclerosis (PPMS), secondary progressive multiple sclerosis (SPMS) and relapsing remitting multiple sclerosis (RRMS) were not considered for analysis. The effect of differences in the male-to-female ratio within each subject group (HC, nPwMS and aPwMS), while imbalanced, was also not considered in subsequent analysis.

In the overall population, 9HPT[2] times were found to be significantly different between dominant and non-dominant hands ($p<0.05$). Furthermore, 9HPT times were significantly different between dominant and non-dominant hands for HC ($p<0.05$) and nPwMS ($p<0.01$), but not for aPwMS ($p=0.46$). Two subjects' average 9HPT times (pooled over dominant and non-dominant) were found to be 38.9 and 41.9 [s] respectively, which was greater than the mean plus 4 standard deviations (>38.1 [s]) from the entire FLOODLIGHT population. These subjects

[2]The distribution of the 9HPT in this population subset can be found in previous figure 4.4b.

were considered outliers with respect to the available data and were subsequently removed from final predictive analysis.

5.4.1 Feature Demonstration

A cross-section of relevant features are illustrated in figures [5.1-5.4]. Each figure shows an example from a representative subject from each subject group: HC, nPwMS, aPwMS. Figure 5.2 for instance demonstrates how Hausdorff distance is calculated for the *figure-8-shape*, which has been observed to increase with higher 9HPT times for both dominant (Spearman's ρ: 0.49, $p<0.001$) and non-dominant handed tests (Spearman's ρ: 0.51, $p<0.001$).

As an example from the *circle* shape, drawing speed can be less smooth and more variable in nPwMS and aPwMS than HC subjects, who tend to draw faster and more consistently (figure 5.1). The variability in absolute drawing speed was significantly greater in both PwMS groups (KWt, $p<0.001$) for both hands. The respective frequency distribution of drawing speed also revealed dominant peaks at multiple frequencies for more variable shape drawing.

Pixel density maps can be created based upon the relative sampling stability of the smartphone screen. The longer a finger touch pointer stays in a position the more it will be sampled, and hence a heat map representation can be built from the finger movements both temporally and spatially. Figure 5.3 illustrates *spiral* drawings represented as discretised heatmaps. Areas of hesitation and non-movement are visually apparent and characterised by dense regions of heat intensity. Image entropy encodes this accumulation of hesitation and irregularity of drawing. It was observed that higher *spiral* entropy values significantly correlated to higher 9HPT times (dominant, ρ: 0.40, $p<0.001$; non-dominant, ρ: 0.45, $p<0.001$) (as described in table 5.2). Figure 5.4 demonstrates the calculation of hesitation time and over shoot at the corners of *square* shapes. Values of drawing speed were mapped to the original drawing for visual analysis.

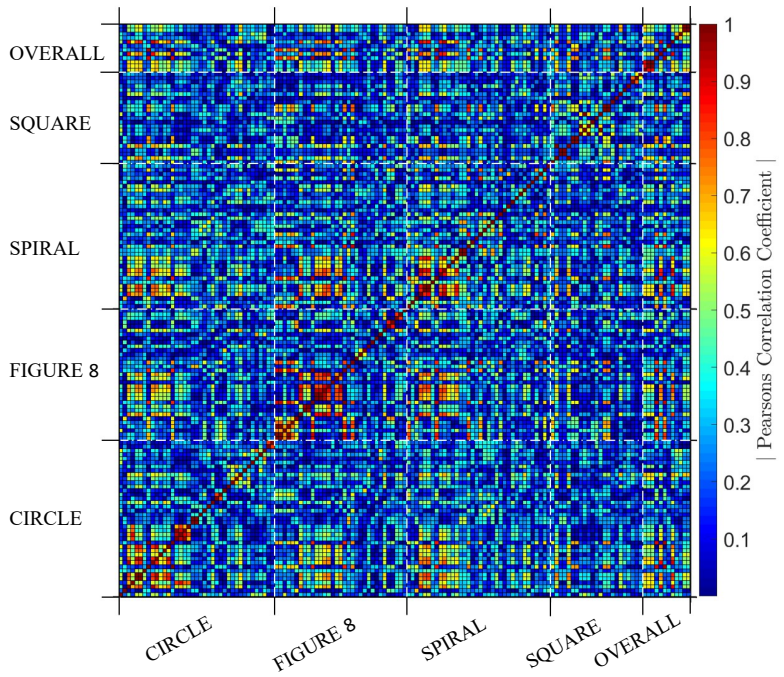


Figure 5.5: Pairwise linear Pearson's (r) correlation matrix. Figure depicts the intra-shape, and inter-shape absolute feature correlation for the top shape features Top features (n=144) were deduced as those features with univariate correlation to the 9HPT [s], $r > 0.25$, $p<0.05$. Feature values are taken as the median feature value per subject (n=93) where the corresponding 9HPT was computed as the average 9HPT per subject over the entire study. It was exhibited that collinearity existed within the shape features, and across shape features, highlighting the need for proficient regularisation. Of the top ranked shape features, *circle* features contribute (n=39), *figure-8-shape* (n=33), *spiral* (n=36), *square* (n=23) and *overall* performance metrics (n=13). Please note, FIGURE 8 refers to the *figure-8-shape* drawn by subjects.

5.4.2 Feature Evaluation

This study found 311 features were significantly correlated to the 9HPT (Pearson's r, $p<0.05$). Applicable hypothesis tests (Wilcoxon signed rank tests or paired t-tests) between these features calculated from dominant and non-dominant handed tests revealed 70% were significantly different between handed tests ($p<0.05$). It was observed that 40% of HC, 73% of nPwMS and 57% of aPwMS subject features differed significantly between hands ($p<0.05$).

Of those significantly correlated features (i.e. n=311), it was observed that n=144 features demonstrated a univariate correlation to the 9HPT [s] greater than 0.25 ($r >0.25$,$p<0.05$) and were considered important features. Of the top ranked shape features, *circle* features contributed (n=39), *figure-8-shape* (n=33), *spiral*

Table 5.2: Comparison of top DaS features between dominant and non-dominant handed models. Features were selected by LASSO across 5-fold CV with 10 repetitions. Each feature was ranked per CV fold by increasing shrinkage regularisation parameter λ, and the percentage (%) of time that feature is chosen in the subset that minimises the CV MSE in the validation set was recorded. For a full list of the features extracted and descriptions see appendix A.4.2.

	Shape	Feature	Description	% Chosen	r	ρ
			Dominant			
1	*spiral*	$ImEntropy(HM)$	Image entropy	100%	0.48***	0.40***
2	*figure-8*	$HausD(X,Y)$	Hausdorff distance	100%	0.46***	0.49***
3	*circle*	$nPeaksNorm(RV)$	# of peaks in radial velocity	82%	0.38***	0.39***
4	*spiral*	$\max(t_{pixel}(x,y))$	Maximum hesitation time	74%	0.43***	0.40***
5	*circle*	$HausD_{middle}(x,y)$	Hausdorff distance	58%	-0.17	-0.30**
6	*figure-8*	$SD(AUC(x))$	Std. deviation in drawing error	68%	0.39***	0.48***
7	*figure-8*	$nPeaksNorm(v)$	# of peaks in velocity	64%	0.43***	0.45***
8	*spiral*	$SD(kurt(RHO))$	Std. deviation in kurtosis in angular velocity	52%	0.31**	0.37***
9	*figure-8*	$SD(nPeaksNorm(RV))$	Std. deviation in # of peaks in radial velocity	42%	-0.17	-0.31**
10	*spiral*	$nPeaks(RHO)$	# of peaks in angular velocity	32%	0.39***	0.37***
			Non-Dominant			
1	*figure-8*	$ImEntropy(HM)$	Image entropy	78%	0.52***	0.47***
2	*figure-8*	$iqr(AUC(x,y))$	Interquartile range of drawing error	96%	0.40***	0.52***
3	*spiral*	$nPeaks(RV)$	# of peaks in radial velocity	84%	0.51***	0.43***
4	*figure-8*	$MI(HM,HM_{ref})$	Spatiotemporal mutual information	70%	-0.50***	-0.58***
5	*spiral*	$ImEntropy(HM)$	Image entropy	66%	0.51***	0.45***
6	*spiral*	$var(RHO)$	Variance in angular velocity	76%	-0.40***	-0.41***
7	*figure-8*	$HausD(X,Y)$	Hausdorff distance	72%	0.49***	0.51***
8	*spiral*	$SD(nPeaks(v))$	Std. deviation in # of peaks in drawing speed	58%	0.51***	0.37***
9	*circle*	$HD_E(HM,x,y)$	Hausdorff distance x Image entropy	52%	0.40***	0.47***
10	*figure-8*	$kurt(RV)$	Kurtosis in radial velocity	36%	0.52***	0.46***

figure-8 referes to the *figure-8-shape* drawn by subjects;
r, Pearson's correlation to 9HPT; ρ, Spearman's correlation to 9HPT;
* p<0.05; ** p<0.01; *** p<0.001.

(n=36), *square* (n=23) and *overall*[3] performance metrics (n=13). The intra-shape, and inter-shape absolute pairwise feature correlation matrix for these features is depicted in figure 5.5. It was exhibited that collinearity existed within the shape features, and across shape features, highlighting the need for proficient reduction and regularisation approaches. LASSO regularisation was consequently employed to

[3]Macro features calculated over all shapes, such as average celerity over all shapes. See appendix A.4.2 for more details.

select rank parsimonious sets of features for predictive modelling of the 9HPT. Table 5.2 describes the top ten features selected by LASSO and relative frequency that were picked for both dominant and non-dominant handed models. Image entropy was the top feature for both handed tests. However, it was computed for *spiral* for dominant and for the *figure-8-shape* for non-dominant handed models, respectively. Among these top ten features, features extracted from the *figure-8-shape* and *spiral* shape features rank most prominently with 4/10 and 5/10 *figure-8-shape* features in dominant and non-dominant handed models, respectively, and 4/10 *spiral* features in both dominant and non-dominant handed models. Interestingly, no *square*-based features are represented in the top ten features for either hand. In contrast, it can be seen that *spiral* image entropy and *figure-8-shape* Hausdorff distance were picked 100% of the time in dominant handed models. *figure-8-shape* image entropy, the top feature for non-dominant models was only picked 78% of the time, yet occupied the top rank. This indicates instances where this feature may not be picked at all, but when it is, it occupies regions of high importance. Nearly all top features for both handed models show moderate-to-strong Pearson's r (linear) and Spearman's ρ (non-linear) correlations with 9HPT (table 5.2).

5.4.3 Model Evaluation

It was observed that non-dominant handed models more accurately predicted 9HPT times (MAE: 2.08 $\pm$ 0.34 [s]) than dominant handed regression models (MAE: 2.32 $\pm$ 0.43 [s]), using simple IRLS across 5 fold CV and 10 repetitions ($p<0.01$). Figure 5.6 compares the out-of-sample test MAE between hands as a function of number of features added to IRLS models across 5 fold CV and 10 repetitions. It was observed that MAE decreases as more features were evaluated in both dominant and non-dominant handed models. Non-dominant handed tests exhibited lowest MAE with 6 features (2.21 $\pm$ 0.04 [s]) compared to dominant handed tests with 16 features (1.93 $\pm$ 0.08 [s]). Presenting the test MAE based on features only chosen from each shape subset (figure 5.7) indicated the shape-wise contribution. For both dominant and non-dominant handed tests, the *figure-8-shape* resulted in the minimum MAE with

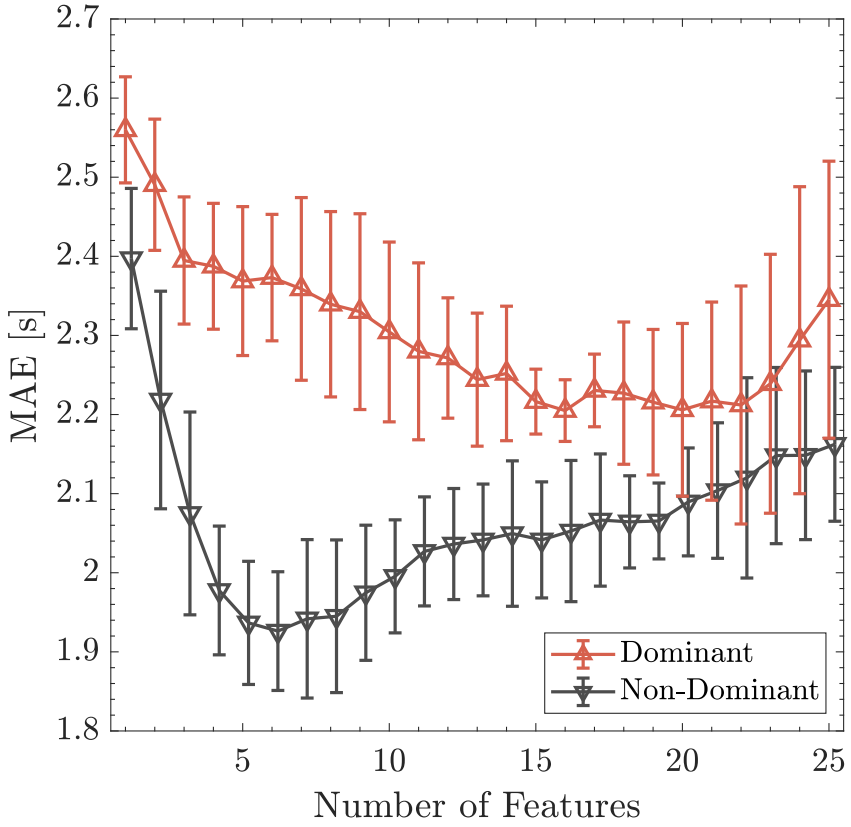
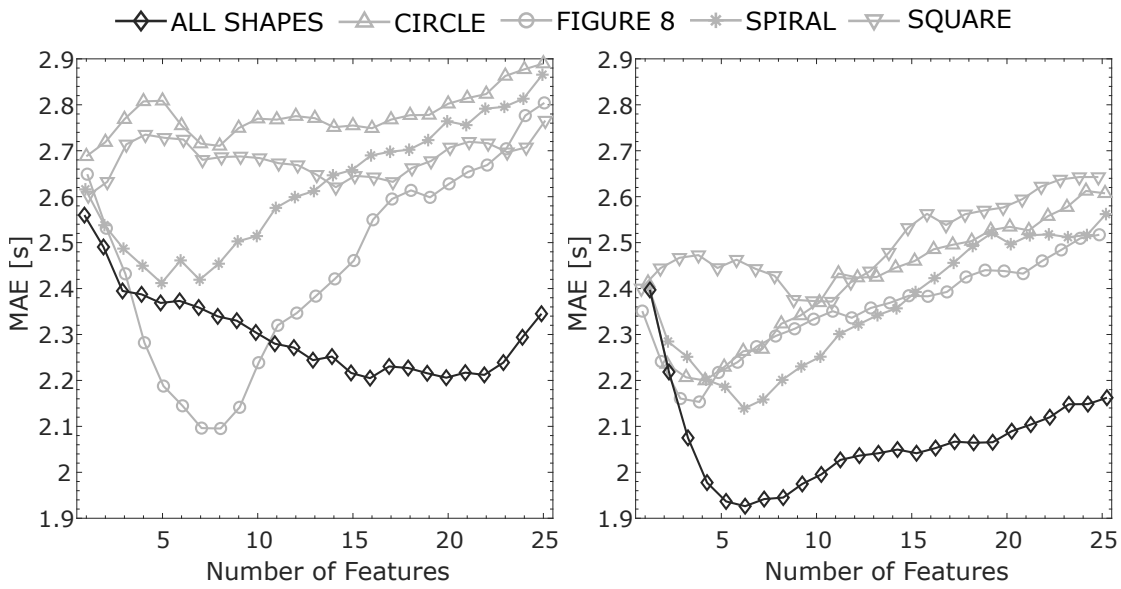


Figure 5.6: Out-of-sample test MAE [s] as features are cumulatively added to an IRLS model. Dominant and non-dominant handed models are built separately for predictions $\hat{Y}$ of average Y 9HPT times [s] for each hand using 5-fold CV with 10 repetitions. Confidence intervals denote one standard deviation (SD) around the quoted mean performance across CV repetitions. Features are pre-selected and ranked within each fold using LASSO feature selection by varying shrinkage parameter λ. The minimum MAE between actual and predicted 9HPTs was obtained with 16 features using the dominant handed model (MAE: 2.21 $\pm$ 0.04 [s]) and with 6 features using the non-dominant handed model (MAE: 1.93 $\pm$ 0.08 [s]).

only a small number of selected features. The *spiral* shape features alone performed similarly well. This was substantiated by the results presented in table 5.2, where *figure-8-shape* and *spiral* shapes dominated the top ranked features. However, the combination of various features extracted across all shapes generally yielded the lowest MAE and importantly, the most robust model (i.e did not over-fit, causing an increase in error like the shape-wise models for a higher number of features).

Scatter plots of the raw 9HPT predictions per subject averaged over all CV-repetitions using IRLS reveal good agreement to the ground truth for dominant (r^2=0.39) and non-dominant (r^2=0.41) tests (figure 5.8). Breakdown of average 9HPT predictions within each subject group demonstrated that HC and nPwMS



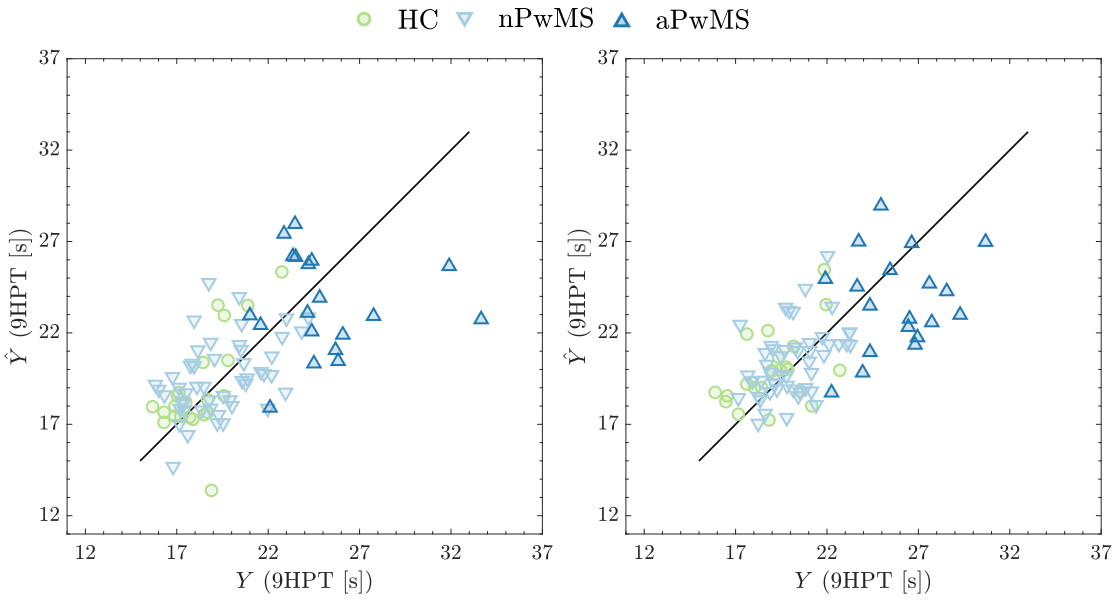
(a) Dominant **(b)** Non-Dominant

Figure 5.7: Comparison of the out-of-sample test MAE [s] across shapes, as features are cumulatively added to an IRLS model. (a) Dominant and (b) Non-dominant handed models are built separately for predictions $\hat{Y}$ of average Y 9HPT times [s] for each hand using 5-fold CV with 10 repetitions. Each line represents features only chosen from each shape subset which are selected and ranked within each fold using LASSO feature selection by varying shrinkage parameter λ. "All shapes" denotes utilising the entire feature space and is analogous of the MAE depicted in figure 5.6.

Table 5.3: Mean absolute error (MAE) in IRLS predictions $\hat{Y}$ of average Y 9HPT times [s]. Results presented per subject group for dominant, non-dominant and pooled dominant and non-dominant handed models. Average MAE is calculated per subject using 5-fold CV over 10 repetitions. Standard deviations (SD) represent the SD per subject-group. Wilcoxon signed rank test between hands used for HC (p=0.83), nPwMS (p=0.09) and aPwMS (p=0.94).

	Dominant	Non-Dominant	Pooled
HC	1.81 ± 1.32	1.93 ± 1.12	1.97 ± 0.98
nPwMS	1.98 ± 1.15	1.62 ± 1.10	1.90 ± 0.89
aPwMS	3.81 ± 2.28	3.51 ± 1.56	3.75 ± 2.00
	2.32 ± 1.66	2.08 ± 1.39	2.31 ± 1.42

had lower MAE compared with aPwMS for both dominant (1.81 ± 1.32, 1.93 ± 1.12 [s]) and non-dominant (1.98 ± 1.15, 1.62 ± 1.10 [s]) handed models (table 5.3). Subject's considered aPwMS were much more difficult to predict for both dominant (3.81 ± 2.28 [s]) and non-dominant (3.51 ± 1.56 [s]) 9HPTs. MAE was higher in non-dominant handed models than dominant for HC subjects, whereas non-dominant handed 9HPTs were predicted more accurately than dominant for nPwMS and aPwMS. While the mean absolute error was not significantly different between



(a) Dominant **(b)** Non-Dominant

Figure 5.8: Scatter plot of average 9-hole peg test (9HPT) times Y versus predicted $\hat{Y}$ 9HPT times per subject. Figure depicts: **(a)** dominant (r^2=0.39, p<0.05) and **(b)** non-dominant (r^2=0.41, p<0.05) handed models using IRLS. Values of $\hat{Y}$ are averaged over all CV repetitions. PwMS were grouped into presumably normal (nPwMS) versus abnormal (aPwMS) hand/arm function based on an upper limit of normal range defined as the average 9HPT time for HCs plus two standard deviations over pooled dominant and non-dominant handed tests (> 22.7 [s]). HC subjects are illustrated using green circles, nPwMS with light blue inverted triangles, and aPwMS are depicted using dark blue triangles. A black line represents perfect predictions.

hands for HC (p=0.83) and aPwMS (p=0.94), a larger trend was exhibited by the nPwMS group (p=0.09). Visual corroboration between (table 5.3 and figure 5.8) revealed that at higher 9HPT values (i.e. aPwMS) the predictions were less accurate by greater magnitudes. Non-linear techniques also exhibited this pattern. Figure 5.9 illustrates the intra- and inter-subject variability of the 9HPT. The within-subject variability of the 9HPT increased with higher 9HPT values (r^2=0.54, p<0.001; r: 0.73, p<0.001; ρ: 0.57, p<0.001). A Brown-Forsythe test for equal variances in Y between subject groups (HC, nPwMS, aPwMS) demonstrated that the between-subject variability in 9HPT also increased with each subject group (p=0.02). In concordance with table 5.3, aPwMS were shown to exhibit greater variability than HC and nPwMS, where higher 9HPT times tended to have much greater within- and between- subject variance, as also depicted in figure 5.9. Training a model all available 9HPTs also demonstrated that the intra-subject variability of Y was higher

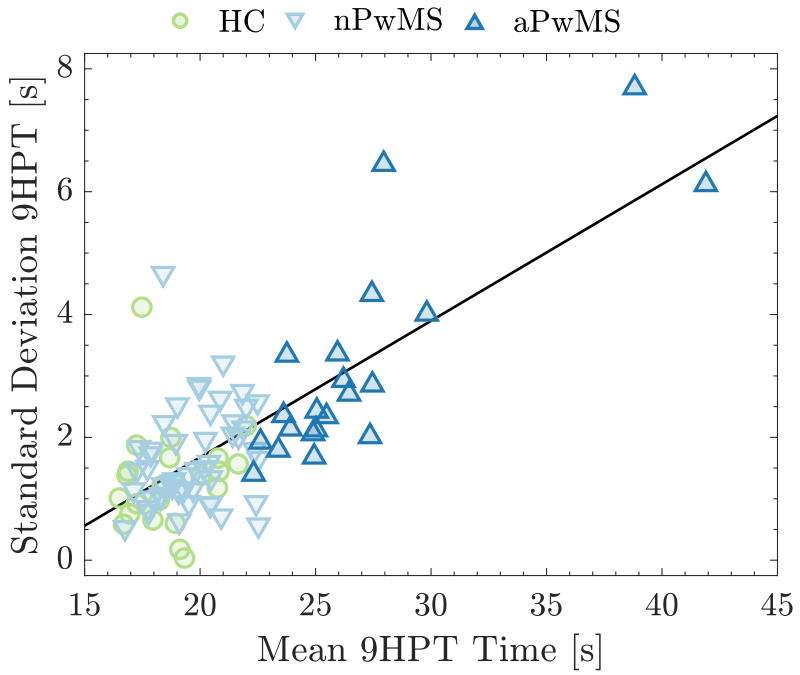


Figure 5.9: Scatter plot of mean 9HPT [s] per subject against the standard deviation (SD). Available 9HPTs are pooled over all visits for both dominant and non-dominant handed tests. A black line represents the least squares line of best fit (r^2=0.54, p<0.001). PwMS were grouped into presumably normal (nPwMS) versus abnormal (aPwMS) hand/arm function based on an upper limit of normal range defined as the average 9HPT time for HCs plus two standard deviations over pooled dominant and non-dominant handed tests (>22.7 [s]). HC subjects are illustrated using green circles, nPwMS with light blue inverted triangles, and aPwMS are depicted using dark blue triangles. The 9HPT is shown to exhibit large heteroscedasticity. It was observed that the within-subject variability of the 9HPT increases with higher 9HPT values (Pearson's r: 0.73, p<0.001; Spearman's ρ: 0.57, p<0.001). A Brown-Forsythe test for equal variances in Y between subject groups (HC, nPwMS, aPwMS) demonstrated that the between-subject variability in 9HPT also increases with each subject group (p=0.02).

for aPwMS participants, than nPwMS or HC (p<0.01). Predicted HC 9HPTs $\hat{Y}$ were less variable than predicted nPwMS and aPwMS NHPT times $\hat{Y}$ (p<0.05), whereas actual HC Y values demonstrated no difference in variance than actual Y nPwMS 9HPT times (see appendix figure B.1). Importantly, the variance in response Y is similarly matched in the variance of predicted $\hat{Y}$ for many HC and nPwMS subjects.

Finally, table 5.4 compared the out-of-sample test error from the 4 respective models built in this study. There were no significant differences observed between any of the model predictions across CV repetitions.

Table 5.4: The out-of-sample test performance metrics of average 9-hole peg test (9HPT) times Y versus predicted $\hat{Y}$ 9HPT times per dominant and non-dominant handed models. Ordinary least squares (OLS) and iteratively re-weighted least squares (IRLS) are compared to Support Vector (SVR) and Random Forest (RFR) regression. Models are built using 5-fold CV with 10 repetitions. The mean MAE and RMSE per CV is presented, where the standard deviation (SD) represents variability across CV repetitions.

	Dominant		Non-Dominant	
	MAE [s]	RMSE [s]	MAE [s]	RMSE [s]
OLS	2.33 ± 0.49	3.03 ± 0.67	2.09 ± 0.36	2.65 ± 0.42
IRLS	2.32 ± 0.43	3.03 ± 0.61	2.08 ± 0.34	2.61 ± 0.39
SVR	2.09 ± 0.42	2.78 ± 0.64	2.01 ± 0.31	2.54 ± 0.38
RFR	**2.09 ± 0.38**	**2.69 ± 0.62**	**1.84 ± 0.26**	**2.34 ± 0.35**

5.5 Discussion

The present study examines UE function in PwMS with mild-to-moderate disability in comparison with HC using DaS, a self-administered digital drawing test captured on a smartphone, and demonstrates how modelling of DaS features from a test can reliably predict the average time of the clinician-administered 9-Hole Peg Test.

It has been proposed that smartphone-based tests developed for repeated assessments in remote settings may offer reliable and objective metrics that could capture a unique window in a subject's disease state and previously unseen or inappropriately estimated characteristics of MS disease [91]. Due to the inherently heterogeneous dissemination in space and time of multiple sclerosis, PwMS experience varying levels of dysfunction or fatigue across different physical domains [52]. As a method to characterise the PwMS population in this study PwMS were divided into those with presumed UE function abnormality (aPwMS), and those with normal UE function (nPwMS), based on average recorded 9HPT times. Abnormal 9HPT times were considered as 9HPT times greater than two standard deviations beyond hand-matched normative data from a healthy population [282]. While applying hard thresholds on clinically administered scales is a blunt stratification method, distinct attributes of each group were apparent and will be discussed with respect to features and predictions.

5.5.1 Feature Discussion

Previous digital upper extremity function assessments have focused mainly on the spiral drawing [137, 138, 142–144, 283] in Parkinson's disease, while those incorporating other types of shapes or drawings have been sparse apropos the information they have extracted [146]. By considering other shapes such as the *circle*, *square*, and *figure-8-shape*, it was hoped to probe all aspects of hand function along with MS-specific pathological impairments such as ataxia, various tremor types, and spasticity [67–69, 292]. Impairments in hand-motor function manifest differently in PwMS as opposed to in Parkinson's disease. This paradigm influenced this work to extract a more exhaustive feature space. Both previously developed and novel features were derived and tested for their clinical validity through multivariate modelling of the 9HPT a typically used UE function test in PwMS. Figures 5.1-5.4 aim to characterise some of the DaS features developed in this study and how the level of UE impairment may influence each shape drawn and resultant feature value. Univariate analysis of these DaS features demonstrated moderate-to-strong Pearson's and Spearman's correlations with the 9HPT (table 5.2), with many coefficients comparable to a range of outcome measures for upper extremity function, collated by Feys *et al.* [23].

Consistent with our study, Feys *et al.* [136] identified handedness as a possibly influential factor on digital drawing performance, although they were unable to test this in a healthy sub-population. Erasmus *et al.* [282] observed a significant difference in drawing error in PwMS subgroups with cerebellar upper limb ataxia, and general worse performance in non-dominant hands across their feature set. Such differences across hands may be more amplified by MS-related impairment and in this study a greater proportion of features differed significantly between hands for nPwMS (70%) and aPwMS (57%) compared with HC (40%). Therefore, digital tests of upper-extremity function that are conditioned on the hand used may be more sensitive to MS disease severity and changes in disease course.

5.5.2 Model Discussion

The DaS testing battery is an information-rich but dimensionally dense test. Similar features can be extracted from 6 different shapes, quickly accumulating the overall volume of features and contributing to redundancy. Many features exhibited collinearity within and between shapes (figure 5.5). Hence, LASSO ℓ_1-regularisation was employed in order to reduce the feature space, minimise redundancy and identify important predictors of 9HPT times. Recording the relative frequency at which features were selected allowed for an interpretation of the feature type and shapes that were most useful to probe aspects of MS disease. Novel features such as Hausdorff distance drawing error or drawing entropy calculated from heat-map transformations appeared as strong predictors of the 9HPT. The *spiral* and *figure-8-shapes* occupied the top feature ranks, especially for non-dominant handed tests. The most accurate reconstruction of 9HPT times was found through the combination of features from various shapes (figure 5.7). Pairwise intra- and inter-shape feature correlation indicated a high redundancy within shape features (figure 5.5). Selecting features across shapes rather than within shapes therefore mitigated against this multi-collinearity to prevent performance problems such as over-fitting, which was typically observed for greater number of features added for shape specific models (figure 5.7).

In addition, overall model performance (table 5.4, figure 5.6) also demonstrated that reconstruction of non-dominant 9HPT was more accurate (MAE: 6 features, 1.93 ± 0.08 [s]), generally with much fewer features than dominant handed 9HPTs (MAE: 16 features, 2.21 ± 0.08 [s]). Comparison of IRLS prediction error between handed models revealed that in HC, 9HPT times are more accurately predicted for dominant handed tests compared with non-dominant handed tests, whereas the improvement in prediction accuracy (i.e. reduced MAE) was strongly driven by the nPwMS and aPwMS groups (table 5.3). It is therefore hypothesised that more complex shapes, tested using a weaker hand, may elicit a wider stratification of PwMS subjects' performance and disease manifestation.

Furthermore, this breakdown of prediction error by subject group demonstrated that HC and nPwMS 9HPT times were accurately predicted by DaS features for both handed tests (table 5.3). High reconstruction accuracy of 9HPT times can be deduced considering that the MAE for HC subjects (1.81 ± 1.32, 1.93 ± 1.12 [s]) was close to the standard deviation of HC 9HPT times (18.3 ± 1.7, 19.1 ± 1.8 [s]) for both dominant and non-dominant handed tests, respectively. Similarly it was shown that MAE for nPwMS (1.98 ± 1.15, 1.62 ± 1.10 [s]) was similar to the variability of their actual 9HPT times (19.4 ± 2.1, 20.1 ± 1.6 [s]).

Higher 9HPT times (those subject observations indicated as aPwMS) were however found to be more erroneously predicted for both handed tests (3.81 ± 2.28, 3.51 ± 1.56 [s]). Visual examination of the distribution of actual 9HPT times (figure 5.8 and figure B.1) demonstrated that greater 9HPT times exhibited higher variance between subject measurements, representative of an inverse Gaussian distribution. Figure 5.9 and B.1 corroborated this heteroscedastic observation, further demonstrating that greater 9HPT times exhibited higher variance within subject measurements ($r^2 = 0.54$, $p<0.001$). As 9HPT times are measured in seconds and are unbounded, hesitations, incorrect movements and the erratic impact of dropping a peg outside of the board—which are more likely in those with greater UE impairment—can compound to greater magnitudes of accumulated 9HPT times. Consequently, higher 9HPT tests become more variable and less stable, both between and within subjects with greater UE impairment. Nonetheless, the MAE for aPwMS (MAE: 3.81 ± 2.28, 3.51 ± 1.56 [s]) was still less than the standard deviation of 9HPT for this group (9HPT: 26.3 ± 5.4, 27.2 ± 5.0 [s]).

Comparison of the out-of-sample test prediction error across CV repetitions demonstrated that RFR and SVR, both non-linear techniques, performed slightly better than OLS or IRLS models. RFR intrinsically uses non-linear feature selection compared to SVR which is dependent on linearly selected features from LASSO, and gave minimum prediction error. However, considering the units of measurements for 9HPT is seconds [s], differences between models of 0.5 seconds can be considered

minimal. As such it can be assumed that a simple linear function can accurately and adequately capture the relationship between 9HPT and DaS features.

5.5.3 Limitations and Considerations

While the results presented in this study demonstrate the utility of using digitally captured DaS features with high concurrent validity as demonstrated by their capacity to predict clinical 9HPT times, there are a number of important limitations which need to be addressed.

First, a limitation of this work is the reliance on estimating a narrow clinical proxy of UE function as a ground truth. As discussed, the 9HPT aims to evaluate UE impairment in PwMS, but given it is measured in seconds it can exhibit a large heteroscedasticity and can be highly variable, especially for longer completion times. It was observed that higher 9HPT times had a higher variance for intra subject measurements over clinical visits (figure 5.9). As such, the 9HPT time should instead not be considered an exact measure of UE impairment, but rather an estimate of a severity range of function that may be impaired. Multiple sclerosis is a heterogeneous disease which can not only manifest differently across people, but symptoms may vary within specific domains, including UE function. Some studies even suggest only moderate test-retest reliability of the 9HPT when examined in a large healthy cohort [111], rather than more disabled PwMS of other studies [23]. Reliance should therefore not be weighted on one test administered infrequently, such as the 9HPT, to effectively capture all aspects of UE function. HC and nPwMS for example were not significantly discriminative of each other based on 9HPT times (dominant: $p=0.11$; non-dominant: $p=0.08$). These are all limitations that should be considered when reporting predictions of any model mapping direct to the 9HPT.

The FL cohort analysed in this study was relatively small (n=93 subjects). Most MS patients were mildly disabled with respect to their overall and motor specific clinical scores (table 5.1), and it was observed that the distribution of the 9HPT was highly tailed and skewed towards shorter 9HPT times. As a result, our models systematically underestimated 9HPT scores for aPwMS groups yet more accurately

predicted shorter 9HPT times, where a greater density of similar observations were available, i.e. the HC and nPwMS subjects. The sparsity in the representation of aPwMS —who additionally are characterised by higher intra- and inter-subject 9HPT variability— limited our ability to learn a more accurate global model on longer 9HPT times. This work may therefore be biased by uneven distributions of UE impairment despite CV stratification.

Another constraint bound by low subject numbers occurs as generalisability problems across cross-validation folds. A low standard deviation in MAE and RMSE regression error ($\pm$ 0.5 [s]) was observed across CV repetitions (table 5.4), demonstrating that results do not change with different permutations of subjects within CVs. Despite this, it was found that feature distributions may not generalise across training, validation and test sets within CV folds, leading to sub-optimal loss minimisation during the training phase. As a result, spurious feature sets and model parameters may be chosen, which can lead to more erroneous 9HPT reconstruction.

This study is longitudinal and data captured can span weeks' worth of testing. A definite limitation is that the temporal aspect of this data is not fully utilised. Instead subject's features are smoothed down as the mean and standard deviation across all their available data. While the standard deviation is a coarse measure of subject variance across the study, more specific time-series modelling of the DaS test may reveal additional detailed insights to the progression and characteristics of PwMS. For example, previous work by Prince *et al.* [187] has shown insights into the longitudinal UE behaviour of patients with Parkinson's disease using a smartphone-based tapping assessment.

Overall, it must be considered that this was a proof-of-concept study with relatively few subjects. While this study helps establish a methodological foundation to construct models that can identify patterns of PwMS UE impairment, further studies —especially with a more heterogeneous and diverse set of subjects— and subsequent analysis will be needed to fully probe the clinical validity of remote smartphone assessments in PwMS.

5.6 Conclusion

This chapter's work —with ongoing further studies— establishes the foundation of how digital sensor-based assessments may enable an out-of-clinic objective augmentation of traditional rater-administered assessments of UE impairment, demonstrating the founding methodologies needed build objective digital biomarkers. Key advantages of digital tests like the DaS test are that they can be administered at high frequency, longitudinally and remotely in free-living environments. More frequent and ecologically valid outcome measures of UE impairment are needed to advance progressive MS research and help make clinical trials more efficient by improving power through sensitive and responsive endpoints. In this respect, the wealth of intra-task UE functional information that encapsulates the DaS feature space administered at higher and potentially daily frequency might be better suited in capturing subtle clinical changes seen in relation to the progressive course of MS than the 9HPT, which is typically administered only every 3 to 6 months.

Specifically, the novel contributions of this chapter can be summarised by:

- We describe the FLOODLIGHT Draw-a-Shape test, a smartphone-based and remotely administered test of upper-extremity function developed for PwMS;
- The analysis from the DaS test expands on the feature space developed by similar studies investigating UE function in other disease areas [143, 144, 293] and contextualises how new and existing features can be used to characterise UE impairment in PwMS;
- Multivariate modelling of these features was shown to reliably predict 9HPT times, a first step in validating UE digital biomarkers for MS;
- While perfect reconstruction of the 9HPT was not possible due to the sparsity of the dataset and the inherent limitations of the 9HPT itself, this work demonstrates how DaS features (and hierarchically, digital biomarkers) may contain a greater wealth of information supplementing beyond discrete 9HPT scores.

MS is a heterogeneous, mutable disease however and subjects may experience symptoms in various domains. The next part of this thesis explores other important functional domains, such as of lower extremity and ambulatory function, leveraging the techniques and insights introduced in this section.

6

Smartphone- and Smartwatch-Based Remote Characterisation of Ambulation during the Two-Minute Walk Test

Associated publication(s):
Creagh, A. P., Simillion, C., Bourke, A., Scotland, A., Lipsmeier, F., Bernasconi, C., van Beek, J., Baker, M., Gossens, C., Lindemann, M. and de Vos, M. (2020). "Smartphone- and Smartwatch-Based Remote Characterisation of Ambulation in Multiple Sclerosis during the Two-Minute Walk Test". *IEEE Journal of Biomedical and Health Informatics*, 25(3), pp.838-849. https://doi.org/10.1109/JBHI.2020.2998187

Contents

6.1 Introduction 108
6.2 Methods 110
6.3 Results 121
6.4 Discussion 131
6.5 Conclusion 137

6.1 Introduction

Ambulatory function has been perceived as one of the most prominent physical impairments in multiple sclerosis (MS) [12]. Previously section 2.1.3 introduced

the clinical foundations motivating the focus on monitoring ambulatory activity in people with multiple sclerosis (PwMS). It was introduced how gait abnormalities [56], postural instability [58], and pronounced gait variability [60] can all be subtle indicators predictive of MS impairment and deterioration. Section 2.3.4 expanded on this, providing more detail on some of the inertial sensor-based methods within the literature that have been successful in characterising gait in PwMS.

The fundamental concept motivating the development of *digital biomarkers* is that administering high-frequency monitoring assessments may be more accurate than conventional outcomes, recorded at periodic visits, to detect subtle progressive sub-clinical changes predictive of disease activity or long-term disability in PwMS [130]. Body worn inertial sensors —especially those embedded in consumer devices such as a smartphone or smartwatches— offer a unique opportunity to unobtrusively and objectively evaluate gait performance out-of-clinic. As reviewed in section 2.3.4, many previous studies have only attempted to asses ambulatory ability using multiple inertial sensors during fixed lengths of controlled walking, in-clinic.

This chapter, ergo, aims to establish a methodological foundation to construct models that can identify patterns of PwMS ambulatory impairment recorded from smartphone and smartwatch inertial sensor data during remote gait assessments. From first principles this chapter will demonstrate how examining the raw sensor data can motivate the creation of specific features, and how established signal-processing metrics can be re-purposed as surrogates to represent aspects of PwMS gait function. Delving into the feature structure, this chapter will explore how ambulatory impairment in PwMS is represented both in a univariate and multivariate domain. Machine learning (ML) techniques, applied to distinguish subgroups of PwMS and HC, can be then used to help identify the patterns and signatures needed to establish a sparse characterisation of PwMS from remote sensor data.

This chapter will finally explore important considerations when analysing remote smartphone and smartwatch PwMS data. Covariates such as subject heterogeneity, adherence, and execution of the assessment procedure will be investigated, as well as a comparison between the methodology and devices used to record data. The

primary contributions of this chapter to the field of remote monitoring of multiple sclerosis disease can be summarised as:

- Development of a robust and scalable gait feature pipeline which can be extracted from smartphone and smartwatch inertial sensor data;
- Characterisation of the raw sensor data and subsequent gait feature space to compare healthy controls (HC) against subgroups of PwMS;
- The successful distinguishing of moderately disabled PwMS from milder PwMS and HC using a sparse representation of the feature structure;
- A comparison of the performance between smartphone and smartwatch devices to capture ambulatory-related PwMS impairments, as well as the performance differences based on discrepancies in test executions (such as smartphone placement in a running belt versus within a pocket);
- The establishment of a methodological foundation, critical considerations, and analysis techniques thereof, for evaluating remotely-generated smartphone and smartwatch PwMS data.

6.2 Methods

6.2.1 Dataset

The PwMS and HC enrolled in this study contributed data which was recorded from a Two-Minute Walk Test (2MWT) performed out-of-clinic and daily over a 24-week period. For more details on the study the reader is referred to chapter 4. A total of 97 subjects contributed both smartphone and smartwatch data and are presented in table 6.1. To allow comparisons between smartphone and smartwatch devices, only test instances where subjects have used both devices during their 2MWT were included in this study. MS is a heterogeneous disease, and in order to differentiate subjects with presumed gait symptoms, subjects were divided into subgroups (mild and moderate) based on their mean EDSS: PwMSmild (n = 52, EDSS [0-3]), and

Table 6.1: Population Demographics for Two-Minute Walk Test subset. Clinical scores taken as the average per subject over the entire study, where the mean ± standard deviation across population are reported; The number of tests are represented as the mean ± standard deviation per subject; EDSS, Expanded Disability Status Scale; T25FW, the Timed 25-Foot Walk; EDSS (amb.) refers to the ambulation sub-score as part of the EDSS; [s], indicates measurement in seconds.

	HC (n=24)	PwMSmild (n=52)[a]	PwMSmod (n=21)[b]	p HC/mild	p HC/mod
Age	35.6 ± 8.9	39.3 ± 8.3	40.5 ± 6.9	0.12[1]	0.07[1]
Sex (M/F)	18/6	16/36	7/14	<0.001[2]	<0.01[2]
EDSS		1.7 ± 0.8	4.2 ± 0.7		
EDSS (amb.)		0.1 ± 0.3	1.9 ± 1.5		
T25FW [s]	5.0 ± 0.9	5.3 ± 0.9	7.9 ± 2.2	0.26[1]	<0.001[1]
smartphone tests	1926	5498	2024		
smartwatch tests	1436	4258	1502		
linked tests	1362	3921	1452		
tests per subject	57.6 ± 47.6	76.4 ± 45.7	70.1 ± 48.2	0.12[1]	0.36[1]
running belt tests	905	2424	1296		
tests per subject	37.7 ± 42.7	46.6 ± 41.8	61.7 ± 48.8	0.41[1]	0.08[1]
pocket tests	457	1497	156		
tests per subject*	30.5 ± 34.4	36.5 ± 37.3	11.1 ± 11.5	0.62[1]	0.18[1]

[a] PwMS with average EDSS [0-3]; [b] PwMS with average EDSS [3.5-5.5];
* HC (n=15), PwMSmild (n=41), PwMSmod (n=14);
[1] Mann-Whitney U-Test between groups; [2] Chi-square (χ^2) test between groups.

PwMSmod (n = 21, EDSS [3.5-5.5]), using a similar threshold to other MS gait studies [154]. As introduced in section 2.2, EDSS is considered a primary outcome for assessing the disease state of PwMS [19, 105]. By definition, gait disorders begin to become prominent in subjects with EDSS≥3.5, and subjects with EDSS<3.5 are mildly impaired [19]. Note: the entire range of subjects' pooled EDSS scores in this study was [0-7]. Differences in clinical characteristics were analysed using the Mann-Whitney U-Test, except categorical differences in sex which were investigated using a Chi-squared (χ^2) test. See methods section 3.4.2.6 and 3.4.2.3 respectively.

6.2.2 Feature Extraction

This section details some of the key methodology for performing feature extraction, from the essential pre-processing techniques, to highlighting and demonstrating some of the useful features analysed in this study. For a further exhaustive list of features extracted in this study please refer to appendix A.4.1. Figure 6.1 schematically illustrates the entire gait processing, model construction and evaluation pipeline.

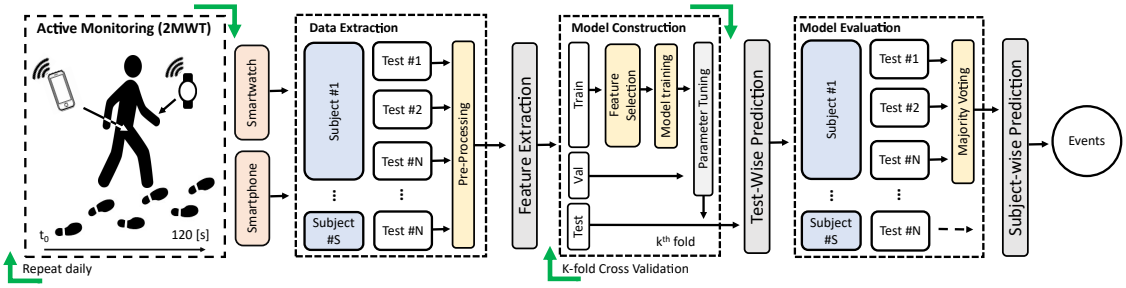
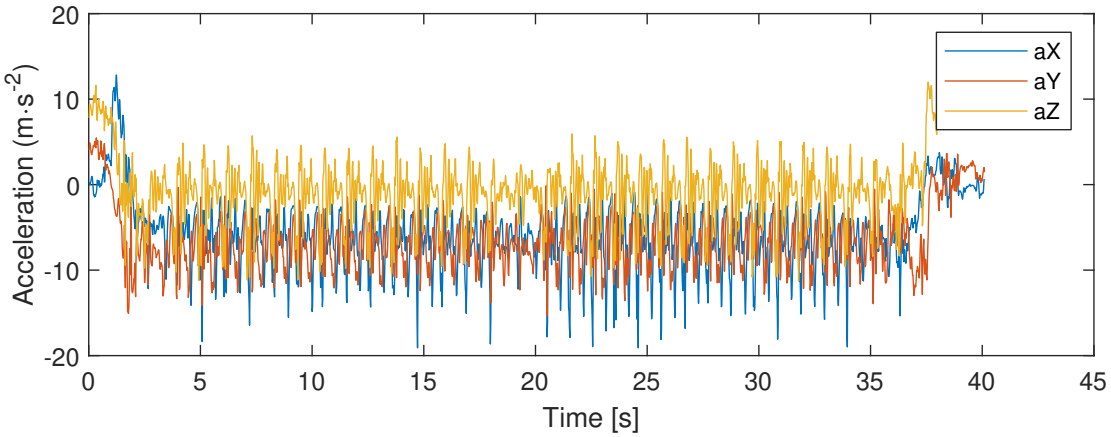


Figure 6.1: Schematic of the 2MWT gait processing pipeline. Users are requested to perform a remote 2MWT daily, for up to 24 weeks, with a smartphone affixed within a waist-worn running belt on the anterior of their waist and smartwatch on their wrist. Sensor data is extracted from both devices independently and signal-based features are computed on each 2WMT per subject. Classification models are then constructed, tuned and evaluated using subject-wise k-fold cross-validation. Individual 2MWT predictions per subject are then majority voted to generate a single prediction per subject, which are used to distinguish subgroups of PwMS and HC as independent binary classification tasks.

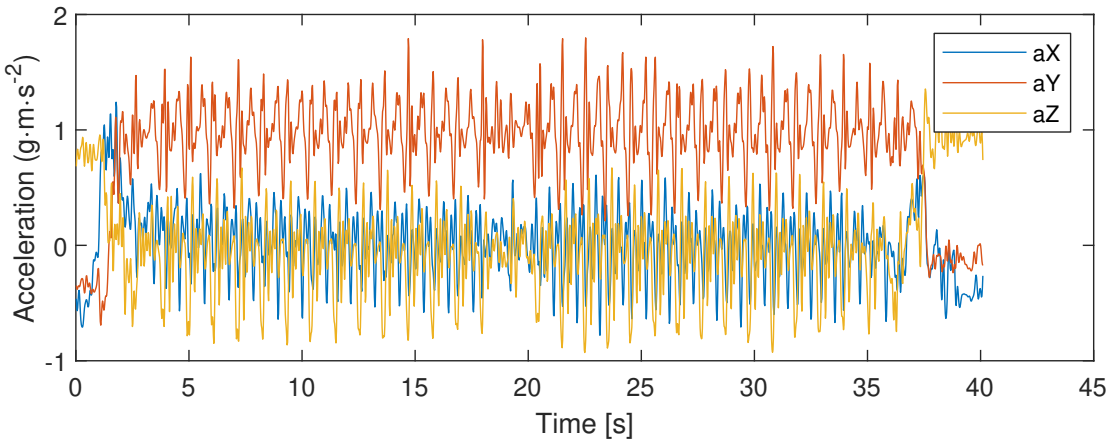
6.2.2.1 Pre-Processing

Figure 6.2 demonstrates an example of the pre-processing pipeline implemented in this study. The raw smartphone-based accelerometer signal captured from a healthy participant during a trial 40 [s] walking bout is presented in figure 6.2a. In this example, the smartphone is affixed to the anterior of the waist in an up-side down orientation to the local smartphone coordinates. Signals were first filtered with a 4^{th} order butterworth filter with a cut-off frequency at 17 Hz to filter non-activity related noise, similar to previous studies [153, 294]. The orientation of the smartphone can be determined by assessing through which axis the mean component of gravity is incident upon during the 2MWT (see previous figure 4.3 for more details). Prior to windowing, the sensor coordinate frame was aligned with the global reference frame through an orientation independent transformation, which is described in appendix section A.2.7. As a result, the anterior-posterior axis ($\mathbf{x}$–) becomes aligned with the direction of motion, which was orthogonal to the vertical ($\mathbf{y}$–) and medial-lateral ($\mathbf{z}$–) axis. This re-orientation with the direction of motion and removal of the gravitational effect is depicted in figure 6.2b.

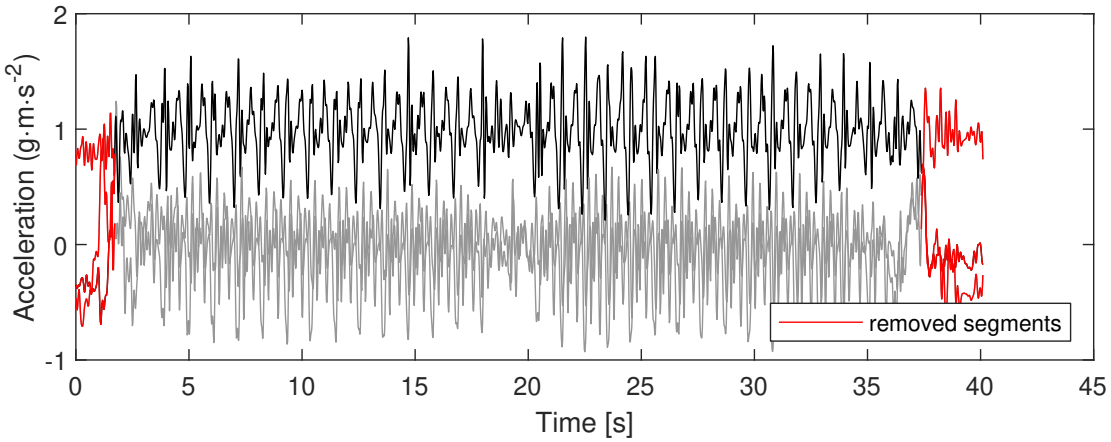
Bouts of non-gait were filtered using the heuristic methods described in [294] and an example of which is depicted in figure 6.2c. The parameters have been adapted from those presented in [294–296]. Features were computed on all accelerometer



(a) Raw Smartphone Sensor Data



(b) Orientation Transformed Smartphone Sensor Data



(c) Gait Bout Detection

Figure 6.2: Inertial sensor data pre-processing steps. Example depicts smartphone sensor data captured from a healthy participant during a 40 second walking trial-run. The smartphone was placed up-side down in a running belt. (a) Represents the raw tri-axial inertial sensor data (aX, aY, aZ) captured during walking; (b) Depicts the transformed sensor data to an orientation invariant coordinate system using three orthogonal versors, whose orientation is independent of that of the smartphone and instead aligned with gravity (aY) and the direction of motion (aX); (c) Illustration of heuristic gait bout detection and segmentation function, where non-gait segments are identified in red.

sensor axes ($\boldsymbol{a}_x, \boldsymbol{a}_y, \boldsymbol{a}_z$) and also on the orientation invariant signal magnitude, for example $\|\boldsymbol{a}\| = (\mathbf{x}^2 + \mathbf{y}^2 + \mathbf{z}^2)^{\frac{1}{2}}$, where $\mathbf{x} = (a_{x_1}, a_{x_2}, ..., a_{x_T})$ and so forth. Subjects' whole 2MWT tests were then windowed into non-overlapping 30 second epochs to help minimise potential signal artefacts, including subject turns during the test. Features were extracted on each epoch and the mean value and standard deviation per 2MWT were taken.

6.2.2.2 Step count and Biomechanical features

The 2MWT was not clinically assessed in this study during site-visits and as such the outcome of walking distance [22, 98] is unavailable. Step counts however have been proposed to approximate the distance travelled [297], and smartphone step count has also been shown to estimate the walking distance in HC over a fixed length during the six-minute walk test (6MWT) [298].

This study implemented a step count to roughly approximate the 2MWT walking distance for comparative purposes. Step counts were compared using two methods; algorithm (a): a step count implemented using methods described by Lee *et al.* [299]; algorithm (b): a step count implemented using methods described by Godfrey *et al.* [300]. Step count represents a simple "biomechanical" feature —a feature that attempts to directly relate or measure a biomechanical form of human locomotion– which can be extracted based on the raw accelerometer signal. Other biomechanical features based on this were extracted such as the step count variability, as well as the step power and step power variability. Specifically, step power was calculated as:

$$\hat{P}_{step} = \frac{1}{s}\int_t^\tau ||\mathbf{a}||\, dx = \frac{1}{2s}\sum_t^\tau \{||a_t|| + ||a_{t+1}||\} \tag{6.1}$$

which calculates the area under the curve of the magnitude of acceleration, $||a||$, relative to the number of steps, s, recorded per epoch of sample length $t = 1, 2, ..., \tau$.

6.2.2.3 Energy Features

This section details some of the energy- and frequency-based features extracted in this study. The continuous wavelet transform (CWT) is a method used to measure the similarity between a signal and an analysing function which can provide a precise

time-frequency representation of a signal [301]. The discretised version of the CWT for a time signal x_n, where $\boldsymbol{x} \in \mathbb{R}^{1\times N}$, with fixed sampling period δ_t, is defined as the convolution of x_n with a scaled and translated mother wavelet $\psi_0(\eta)$ [302]:

$$W_n(s) = \sum_{n'=0}^{N-1} x_{n'} \psi^* \left[\frac{(n'-n)\delta_t}{s} \right] \tag{6.2a}$$

where $(^*)$ denotes the complex conjugate, s is the wavelet scaling factor and n is the localised time index. The subscript $_0$ on ψ has been dropped to indicate that this ψ_0 has been multiplied by $\left(\frac{\delta_t}{s}\right)^{1/2}$, in order to normalise ψ to have unit energy. This ensures that the wavelet transforms $W_n(s)$ at each scale s are directly comparable to each other and to the transforms of other time series; see [302].

It has been shown that the Morlet wavelet effectively captures gait-related spatio-temporal characteristics from acceleration signals obtained from different body locations [303]. A Morlet wavelet consists of a plane wave modulated by a Gaussian:

$$\psi_0(\eta) = \frac{1}{\sqrt[4]{\pi}} e^{i \cdot w_0 \cdot \eta} \cdot e^{\frac{-\eta^2}{2}} \tag{6.2b}$$

where and η is the a non-dimensional time parameter and w_0 is the non-dimensional frequency, here taken to be 6, as per [302], to satisfy the admissibility condition.

The total signal energy at a specific scale can be measured by the scale-dependent energy density spectrum E_s:

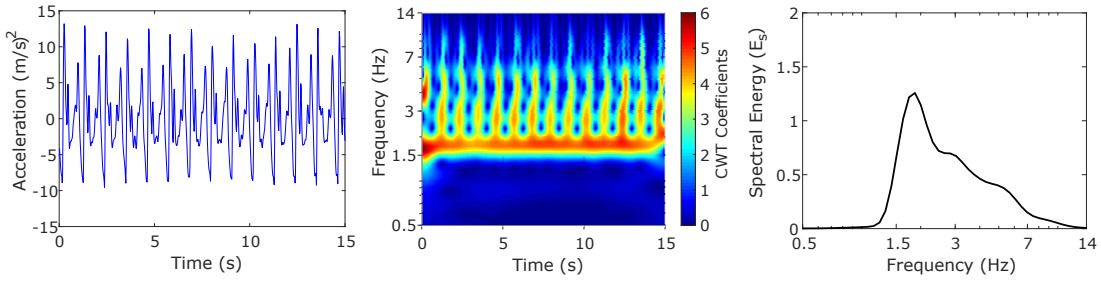
$$E_s = \sum_{n=0}^{N-1} |W_n(s)|^2 \tag{6.3}$$

where $s \in [1, S]$ and $|W_n(s)|^2$ is the 2-D wavelet energy density (scalogram) that measures the total energy distribution of the signal. The frequency f, in Hz, can be approximated from the wavelet scaling factor s such that [304]:

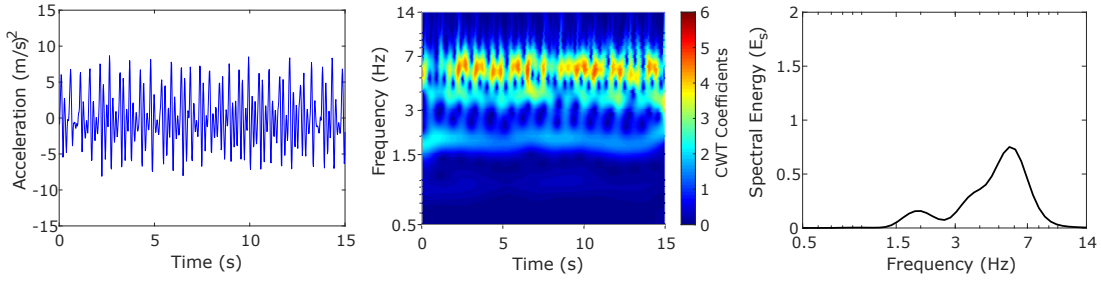
$$f = \frac{f_c}{s} \tag{6.4a}$$

where the center frequency in Hz can be defined by [302]:

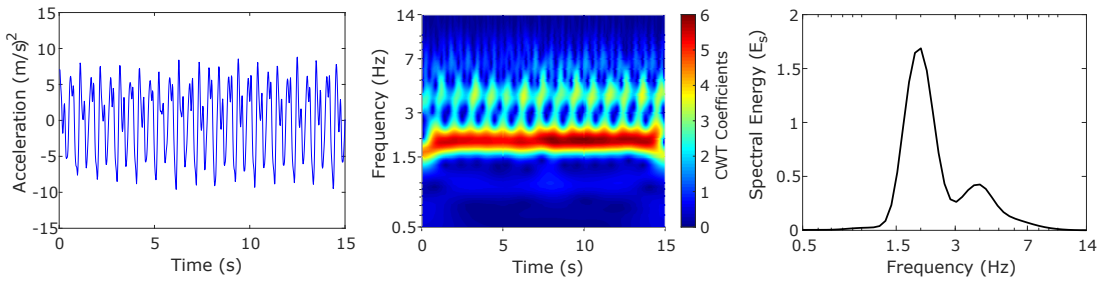
$$f_c = \frac{w_0 + \sqrt{2 + w_0^2}}{4\pi} \tag{6.4b}$$



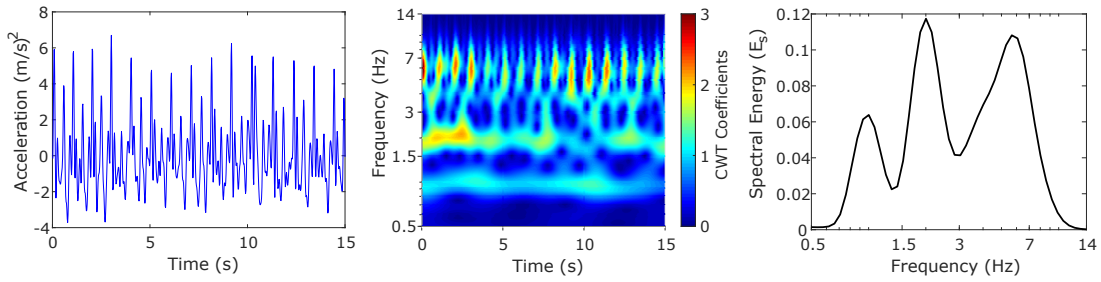
(a) Smartphone - Healthy Control (HC)



(b) Smartphone - PwMSmod

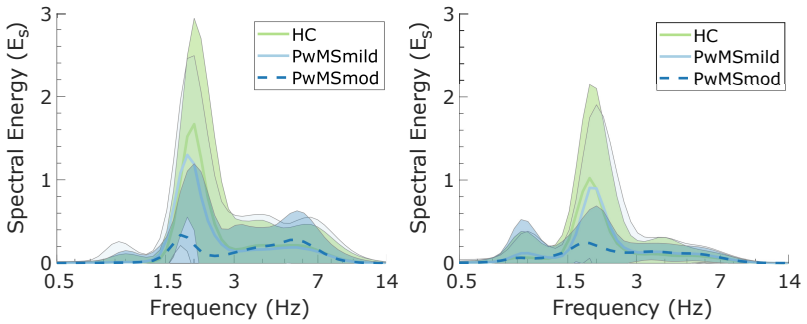


(c) Smartwatch - Healthy Control (HC)



(d) Smartwatch - PwMSmod

Figure 6.3: Continuous Wavelet Transform (CWT) time-frequency representation of sensor data. Typical examples recorded by smartphone device carried in a running belt for representative **(a)** HC and **(b)** PwMSmod subjects, and their respective linked accelerometer data recorded by smartwatch device for the same **(c)** HC and **(d)** PwMSmod subject. The first column represents of raw magnitude acceleration signals $\|\boldsymbol{a}\|$. The second column shows the top view of the CWT scalogram, which is the absolute value of the CWT as a function of time and frequency. The third column corresponds to the scale-dependent (spectral) energy density (E_s) distribution of the CWT coefficients. (HC: T25FW, 3.6 $\pm$ 0.4 [s]); (PwMSmod: EDSS, 4 $\pm$ 0; T25FW, 8.1 $\pm$ 1.3 [s]); Note the axis scales for figure (d).



(a) Smartphone **(b)** Smartwatch

Figure 6.4: The scale-dependent (spectral) energy density E_s distribution of the CWT coefficients. Figure depicts **(a)** smartphone versus **(b)** smartwatch devices. Bold lines and shaded region corresponds to the median and standard deviation in spectral energy amplitude for HC, PwMSmild and PwMSmod groups.

An example of a CWT computed for smartphone and smartwatch example from both a HC and PwMSmod are depicted in figure 6.3, as well the CWT distribution over all HC, PwMSmild and PwMSmod subjects.

A sparse representation of gait signals was also obtained using the Discrete Wavelet Transform (DWT) where the signal was decomposed into a number of different bandwidths expressed by approximation and detail coefficients on which features were computed [305] (see appendix A.3.1). At each decomposition level, or bandwidth, the energy, entropy (H), and the Teager-Kaiser Energy Operator (TKEO) on approximation (cA) and detail (cD) coefficients are calculated. Wavelet Energy (both using CWT and DWT representation) is defined as:

$$E(\mathbf{x}) = \sum_{i=1}^{N} |x_i|^2 \tag{6.5}$$

Wavelet (non-normalised) Shannon Entropy is defined as:

$$H(\mathbf{x}) = -\sum_{i=1}^{N} x_i^2 \log(x_i^2) \tag{6.6}$$

where $\mathbf{x} = cD_j$; $\mathbf{x} = cA_j$ are the detail and approximation coefficients at level $j = 1, 2, ..., L$;

Empirical Mode Decomposition (EMD) has been used previously to characterise the frequency range distributions of gait rhythms from accelerometer signals [306]. Classical EMD decomposes a signal into a small finite number of intrinsic mode

functions (IMFs) using the Hilbert-Huang transform (HHT) to encode instantaneous frequency and amplitude information [307] (see appendix A.3.2 for more details). IMFs offer a data driven approach to analyse non-linear, non-stationary signals. The energy (equation 6.5) of each IMF is computed, where the first IMF represents the "high-frequency (noise)" components with the latter IMFs capturing the relatively "low-frequency (signal)" components of gait rhythm. The relative signal-to-noise ratio (SNR) is then computed as a feature to characterise the ratio of gait-to-higher frequency perturbations in the sensor signal.

6.2.2.4 Statistical and Entropy Features

A number of statistical features were also computed on the sensor signals such as the mean, standard deviation, skewness, kurtosis, zero-crossing rate and auto-correlation coefficients, based on similar features extracted in [186].

Multiscale entropy (MsEn) calculates the sample entropy (SampEn) of a signal at increasingly coarser grains (scales) [308]. MsEn is advantageous to entropy alone in that calculating the entropy of a signal at multiple time scales discriminates long-range correlations in complex systems from completely random signals. Costa *et al.* [308], for example, found that faster and unconstrained walking had more complex dynamics than slower walking, as captured through greater SampEn at different scales. Further to calculating raw MsEn over the first 20 time scales, higher order MsEn-based statistical features were also computed. Similar entropy parameters of embedding dimension $m = 2$, and tolerance $r = 0.2$ were used [308]. Recurrence period density entropy (RPDE) is a method used to characterise the deviations from exact periodicity and stochasticity within a signal [309], proposed here to capture the ability to maintain consistent gait rhythm.

For more details, including an exhaustive list of ambulatory-based features extracted, as well as the parameterisation of those features, the reader is referred to appendix A.4.1.

6.2.3 Exploratory Analysis

Exploratory analysis was first performed in order to characterise the gait features extracted in this study, the underlying structure, as well as univariate associations to gold-standard clinical assessment outcomes.

Univariate feature differences were taken as the median feature value per subject over all available test observations and subsequently investigated using the Mann-Whitney U Test (see 3.4.2.6) for each paired subject group combinations (HC, PwMSmild, and PwMSmod). Associations between mean clinical metrics (EDSS and T25FW) and median feature values per subject were investigated using Spearman's correlation (ρ) (see section 3.3.2). Differences in feature distributions between smartphone locations identified as running belt or pocket were investigated using a U test for subjects who contributed both placements only. In order to reduce our highly dimensional feature space (features derived from 2 devices, 2 sensor sources with 3 axes per sensor) prior to model building all redundant features were first removed based on Spearman's correlation (ρ) to subject group (HC, PwMSmild, PwMSmod) < 0.3, yielding n=93 and n=63 respective smartphone and smartwatch features that were retained. P-values were corrected for multiple hypothesis testing using methods described in section 3.4.3.

Exploratory analysis of data and feature structure was performed using principal component analysis (PCA) (see section 3.3 and A.2.1). Feature reproducibility was investigated using the intraclass correlation coefficient (ICC) metric (see 3.3.3).

6.2.4 Model Construction

A number of machine learning techniques were explored in order to asses the ability to discriminate HC from PwMS sub-groups as binary classification tasks. Model generalisability was determined using 5-fold, subject-wise, stratified, cross-validation (CV). After partitioning the data into training and validation sets, observations were randomly re-sampled to balance class distributions, as subject each contributed unequal quantities of 2MWT observations [266]. CV was repeated 10 times to reduce bias in re-sampling and dataset splitting.

Logistic Regression (LR) was compared to Support Vector Machines (SVM) and a Random Forest classifier (RF) [200]. LASSO regularisation for generalised linear models (*lassoglm*) was employed in order to reduce the dimensions of the extracted feature space into a ranked parsimonious set [200]. In this case *lassoglm* is an extension of LASSO which uses a logit (or inversely a logistic) link function: $\log(\frac{\pi}{1-\pi}) = w_0 + \mathbf{w}^\top \mathbf{x}$, yielding a posterior probability mapping binomial responses (see section 3.5.3). Features were iteratively ranked per CV fold by increasing the shrinkage regularisation parameter λ. These selected features were then cumulatively presented to LR and SVM classifiers. A top feature ranking table was deduced from these by interrogating the feature subsets selected by *lassoglm* at each fold and repetition. The relative stability of features selected was assessed by recording the percentage of time that the feature was selected in the top 5 and top 25 features at each fold and repetition.

Instead of using the the raw *lassoglm* coefficients ($\hat{\boldsymbol{w}}$) for regression problems, it has been suggested that bias or prediction error can be decreased by performing a separate regression post-lasso [290]. Observations were assigned to the class yielding the largest posterior probability, where in the case of the SVM, posterior probabilities were first obtained using methods described by Platt [310]. SVM tuning was performed for each fold via grid-search over internal CV to determine optimal values of the Gaussian radial basis function (RBF) kernel parameter γ and the penalty parameter C (see section 3.7.3.1). The pair that gave the lowest CV misclassification error for each added feature to the classifier were selected [200]. An example of the SVM optimisation strategy is demonstrated in figure 6.5, with a search conducted over the log-scaled range $[10^{-3}, ..., 10^{3}]$ for γ and C. SVM models were not highly sensitive to tuning parameters however.

A selection of RF classifiers were built (using 1500 trees) and trained with a split criterion based on Gini impurity. Classifier performance was examined by varying the number of input variables chosen at each node (denoted as *mtry*). Values of *mtry* used were tested as the square root of the number of features (n=13); double and half this value was also investigated as suggested in [237].

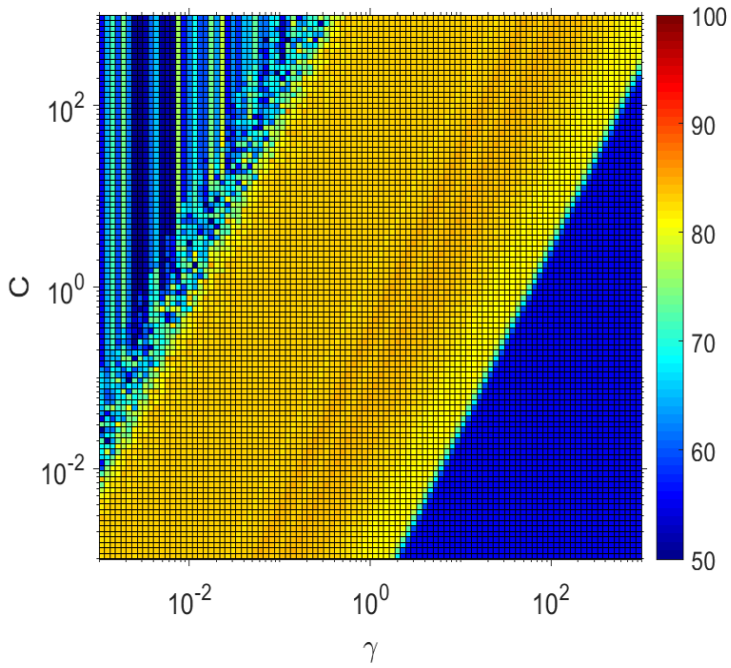


Figure 6.5: SVM optimisation strategy. Figure depicts the CV validation accuracy (%) as a function of kernel scale γ and box constraint C using a Gaussian RBF kernel. Classification outcome is based on HC vs. PwMSmod. It was observed that SVM models were not highly sensitive to tuning parameters.

Classification models were built using individual test observations and metrics based on majority voting of individual test predictions per subject are reported in order to increase prediction robustness. In order to account for the imbalance in the number of subjects within each sub-group, the macro-average of the F_1 score (MF_1) for each class were reported [266] (see section 3.11.3), as well as classification accuracy (acc). Distribution differences in performance results calculated based on feature sets (smartphone, smartwatch, smartphone & smartwatch) and classification models (SVM, LR, RF) built across CV repetitions were tested using a Wilcoxon signed-rank test (see section 3.4.2.5).

6.3 Results

6.3.1 Feature Analysis

Examples of raw sensor signals illustrate visual differences between HC and PwMSmod for both smartphone (figure 6.3a and 6.3b) and smartwatch devices (figure 6.3c and

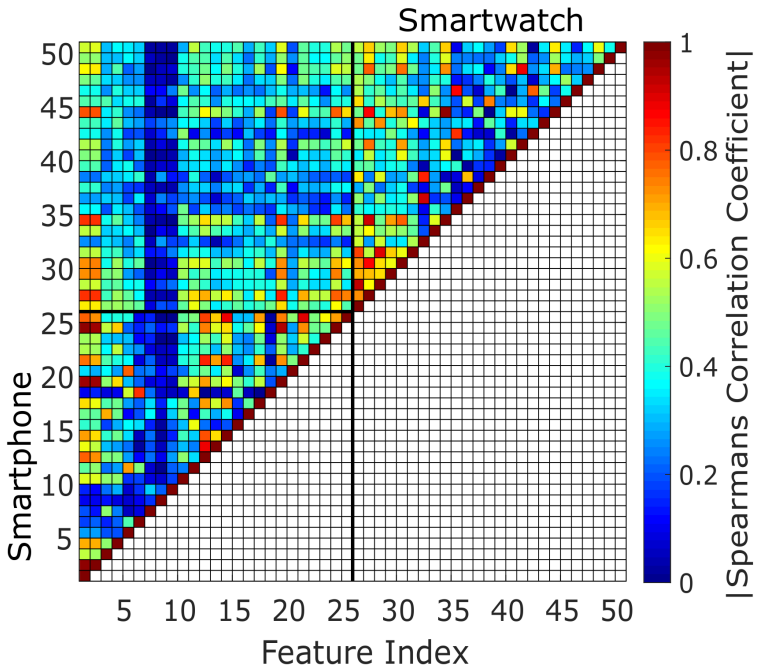


Figure 6.6: Pairwise gait-feature Spearman's ρ correlation matrix. Figure demonstrates the intra- and inter-source correlation for the top 25 ranked smartphone (index 1-25) and smartwatch (index 26-50) features respectively. The top 10 smartwatch features are highly correlated with each other, whereas the top 10 smartphone features exhibit much less inter-correlation. The inter-source correlation is strong in the top 5 features between smartphone and smartwatch devices.

6.3b). The CWT spectral-energy density distribution (figure 6.4) demonstrated that PwMSmod had less power than PwMSmild and HC ($p<0.01$). It was additionally observed that the ratio of total (AUC) spectral energy in the gait domain (0.5–3 Hz) to higher frequency "noise" energy per subject was lower in PwMSmod than in HC or PwMSmild smartphone tests (figure 6.4a, $p<0.001$). This study further focused on subjects who exhibit clinically moderate disease symptoms, including the gait domain (PwMSmod, EDSS [3.5-5.5]). Table 6.2 depicts the top features between HC and PwMSmod as selected by *lassoglm*, the percentage of time chosen in the top 5 and 25 selected features, along with associated statistics and correlation to clinically administered metrics. A number of these top features selected derive from energy and entropy in the frequency bands for gait and movement [151]. Smartphone device features contributed most to the number of top features in the top 15 ranking (11 overall). The top-ranked features demonstrated high stability and consistency with good to excellent ICC values. Furthermore, these features

significantly discriminated HC from PwMSmod using the median feature values per subject. Many top features indicated significant correlation ($p<0.05$) to MS clinical measures in PwMSmod, especially T25FW [21] and EDSS [19].

The intra- and inter-source feature correlation for the overall top 25 smartphone and smartwatch features were investigated in figure 6.6. Many of the top smartwatch features were highly correlated with each other compared to the top 10 smartphone features which exhibited much less inter-feature correlation. The inter-source correlation was also relatively strong in the top 5 features between smartphone and smartwatch devices indicating that these features may capture similar ambulatory information. Visualisation of the feature space of the top 25 smartphone and smartwatch features independently using the first two PCA components —which contain the most variance within the feature space— demonstrated the overall separability of HC, PwMSmild and PwMSmod groups (figure 6.7). Analysis of the mean PCA components per subject (figure 6.7d and figure 6.7h) indicated that PwMSmod were distinctly separable from PwMSmild and HC groups. Furthermore, by visualising the first two PCA components over all observations and labelling each point by participant, many subject specific clusters were apparent. Many of these clusters were homogeneous (indicating all observations were from the same subject, i.e. a low intra-subject variability) and in certain cases distances between the clusters were distinct, indicating the presence of an inter-subject variability and an an underlying heterogeneity within the dataset.

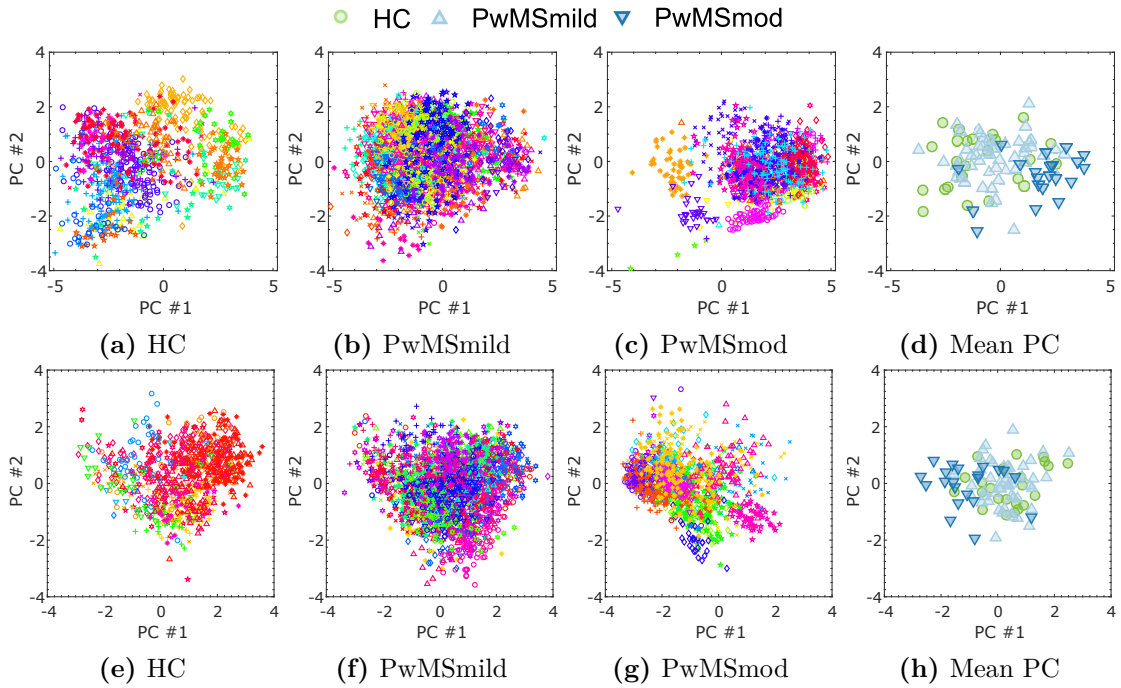
Table 6.2: Comparison of step count and top features between HC and PwMSmod. Features are selected by *lassoglm* across 5-fold CV with 10 repetitions.

	Feature	Description	Top 5, 25 (%)	ICC	p	ρ EDSS	ρ T25FW
	# Steps	Total number of steps counted over entire 2MWT[299].		0.91 (0.81-0.95)	0.08	0.64**	0.52***
1	$H(cD_5(\|\boldsymbol{a}\|))$	Entropy (6.6) of the 5$^{\text{th}}$ DWT coefficient detail quantifies the predictability of the gait signal roughly corresponding to frequency range 3.3-1.5 Hz (faster gait).	78, 80	0.96 (0.93-0.98)	***	0.52*	0.47**
2	$skew(E_s(\|\boldsymbol{a}\|))$	Skewness as a measure of the asymmetry of the scale-dependent energy density distribution of the CWT coefficients E_s calculated from the the acceleration signal. This measures the relative magnitude of how far the distribution deviates from the 'normal', which was used as a proxy for smooth stable gait movements.	72, 74	0.81 (0.61-0.91)	***	0.56**	0.40**
3	$\frac{max_1(E_s(\|\boldsymbol{a}\|))}{max_2(E_s(\|\boldsymbol{a}\|))}$	The ratio of the maximum scale-dependent energy peak density E_s to the next highest peak, corresponding to the frequencies computed using a CWT, over the 2MWT.	38, 70	0.81 (0.61-0.91)	**	0.45*	0.31*
4	$std(MsEn(\|\boldsymbol{a}\|))$	Std. deviation in multiscale entropy over all scales. This quantifies the variation in the predictability of a gait signal over multiple temporal scales, where $MsEn$ characterises dynamic complexity of gait within a signal [308].	30, 100	0.80 (0.59-0.90)	***	0.05	0.19
5	$E(cD_5(\|\boldsymbol{a}\|))$	The energy (6.5) contained in the 5$^{\text{th}}$ DWT coefficient detail roughly corresponding to frequency range 3.3-1.5 Hz. This could be a proxy for (faster) gait power.	24, 86	0.94 (0.88-0.97)	***	0.53*	0.49**
6	$RPDE(\|\boldsymbol{a}\|)$	Recurrence period density entropy characterises the periodicity of a gait signal and the ability to maintain consistent gait rhythm [309].	28, 54	0.87 (0.73-0.93)	**	0.21	0.32*
7	$std(SNR(\|\boldsymbol{a}\|))$	Variability in the SNR between epochs, characterised by IMFs and computed using EMD, with the signal sampled every 0.5 [s]. This is analogous for the variability in the ratio (amount) of gait to higher-frequency perturbations over the 2MWT.	16, 82	0.74 (0.45-0.87)	*	0.22	0.30*

Table 6.2: Comparison of step count and top features between HC and PwMSmod. Features are selected by *lassoglm* across 5-fold CV with 10 repetitions.

	Feature	Description	Top 5, 25 (%)	ICC	p	ρ EDSS	ρ T25FW
8	$std(SNR(\|\boldsymbol{a}\|))$	(see above)	16, 84	0.81 (0.60-0.90)	***	0.20	0.40**
9	$skew(\|\boldsymbol{a}\|)$	Skewness as a measure of the asymmetry of the probability distribution of the acceleration signal values. The relative magnitude of how far a distribution deviates from the normal which used as a proxy for smooth stable ambulatory movements.	4, 92	0.40 (0.20-0.71)	n.s.	0.30	0.06
10	$std(zcr(\boldsymbol{a}_z))$	The std. deviation in the zero-crossing rate calculates the rate of sign-changes along the medial-lateral plane over the 2MWT, which can measure the dynamic sway during ambulation.	10, 76	0.68 (0.32-0.84)	*	0.53*	0.47**
11	$zcr(\|\boldsymbol{a}\|)$	The zero-crossing rate calculates the rate of sign-changes in a signal, roughly capturing the static-to-dynamic transitions within gait.	24, 68	0.82 (0.65-0.92)	*	0.15	0.33*
12	$std(\boldsymbol{a}_z)$	The standard deviation in the acceleration values in the medial-lateral plane for the 2MWT, which can measure the dynamic sway during gait.	4, 95	0.86 (0.65-0.94)	**	0.15	0.10
13	$SNR(\|\boldsymbol{a}\|)$	The mean SNR, where the signal and noise are characterised by IMFs and computed using EMD, with the signal sampled every 0.5 [s]. This is analogous for the ratio (amount) of gait to higher- frequency perturbations over the 2MWT.	10, 76	0.73 (0.46-0.87)	*	0.10	0.32*
14	$\frac{MsEn(\|\boldsymbol{a}\|)_{(:10)}}{MsEn(\|\boldsymbol{a}\|)_{(11:)}}$	The ratio in multi-scale entropy over first 10 to the last 10 scales captures the dynamic complexity of gait versus that of random fluctuations within the gait signal [308].	10, 64	0.76 (0.53-0.87)	**	0.38	0.50**
15	$kurt(\|\boldsymbol{a}\|)$	Kurtosis as a measure of how outlier prone the distribution of the magnitude of acceleration signal values are to quantify gait-related perturbations.	0, 76	0.81 (0.61-0.90)	*	0.55**	0.61***

[1] See supplementary material for a full description of all of the features extracted in this study;
ICC, Intraclass correlation coefficient (95% CI); Other statistics calculated on median feature value per subject (HC, n=24; PwMSmod n=21): p, Mann-Whitney U Test between groups; ρ EDSS, Spearman's correlation to EDSS in PwMSmod; ρ T25FW, Spearman's correlation to Timed-25 ft. walk test;
*p<0.05, **p<0.01, ***p<0.001; n.s., not significant;
- smartphone feature; - smartwatch feature;



(a) HC **(b)** PwMSmild **(c)** PwMSmod **(d)** Mean PC

(e) HC **(f)** PwMSmild **(g)** PwMSmod **(h)** Mean PC

Figure 6.7: Example of PCA performed on top 25 smartphone and smartwatch gait features respectively. Figure depicts: row 1 (a-d) top smartphone features; row 2 (e-h) top smartwatch features. The first two components have been plotted per subject and their respective test observations for HC: smartphone (a) and smartwatch (e); PwMSmild: smartphone (b) and smartwatch (f); PwMSmod smartphone (c) and smartwatch (g). Figures (d) and (h) represent the mean principal component (PC) value per subject, for smartphone and smartwatch features respectively. Intra-subject variability appears low, while inter-subject variability is distinctly high as subjects cluster within themselves.

6.3.1.1 Step Count

The top features selected by *lassoglm* were also compared to total step count (determined using methods described by Lee *et al.* [299]). Step count was significantly correlated with EDSS (ρ: 0.64, p<0.01) and T25FW (ρ: 0.52, p<0.001) for PwMSmod groups only (table 6.2). Total step count did not significantly distinguish HC from PwMSmild (U test, p=0.09) or PwMSmod (U test, p=0.08) groups. However, a significantly lower step count was observed in PwMSmod versus PwMSmild (U test, p<0.01). The Kruskal–Wallis[1] test by ranks (KWt) however demonstrated significant differences between groups for step count, p<0.001; as well as step power std. p<0.01. Step power std. was significantly different for univariate comparisons between HC vs. PwMSmod (U test, p<0.01) and PwMSmild vs.

[1]See section 3.4.2.8 for more details.

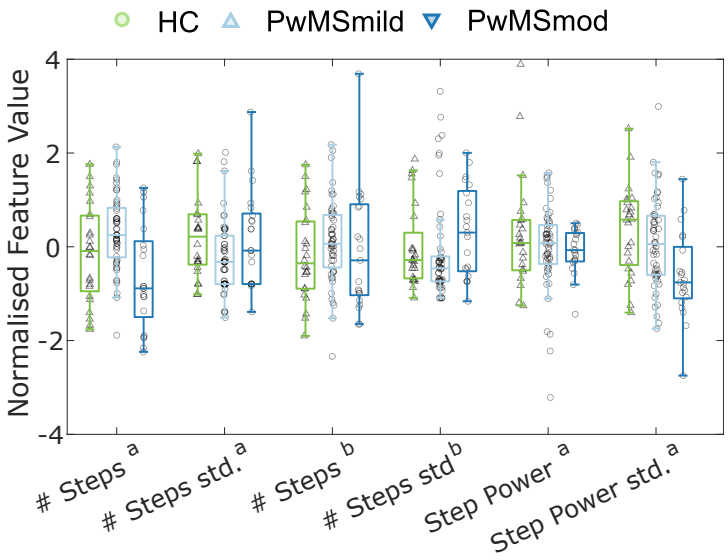


Figure 6.8: Comparison of step count-based biomechanical features extracted from the smartphone-based 2MWT. Feature compared across HC (green), PwMSmild (light blue) and PwMSmod (dark blue) groups. Each data point represents the median feature value per subject over the duration of the study. Features have been standardised for visual comparisons. Superscript[a] denotes step count implemented using methods described by [299], whereas superscript[b] corresponds to step count implemented using methods described by [300]. A Kruskal–Wallis test by ranks demonstrated significant differences between groups for: step count[a], $p < 0.001^{***}$; step power std[a] $p < 0.01^{**}$.

PwMSmod (U test, p<0.01). Distributions of the (median) biomechanical features per subject are depicted in figure 6.8. The agreement between two step counting algorithms (a) [299] and (b) [300] extracted using the smartphone-based 2MWT were compared in figure 6.9. The cadence (steps/min) for algorithm (a) 107.4 ± 11.6 and (b) 106.6 ± 9.7 were not significantly different (paired *t*-tests, p=0.3) and within normative cadence values for HC and PwMS [62, 311, 312]. Disagreement between the two algorithms however occurred generally for slower cadence values (i.e number of steps), particularly for some PwMSmod subjects. Paired *t*-tests demonstrated that the average cadence for HC (a: 106.4 ± 11.7; b: 106.6 ± 9.7), were not significantly different (p=0.2) while the average cadence for PwMSmild (a: 111.0 ± 9.49; b: 107.6 ± 8.5, p<0.001) and PwMSmod (a: 99.39 ± 12.7; 105.77 ± 12.7, p<0.05) were significantly different.

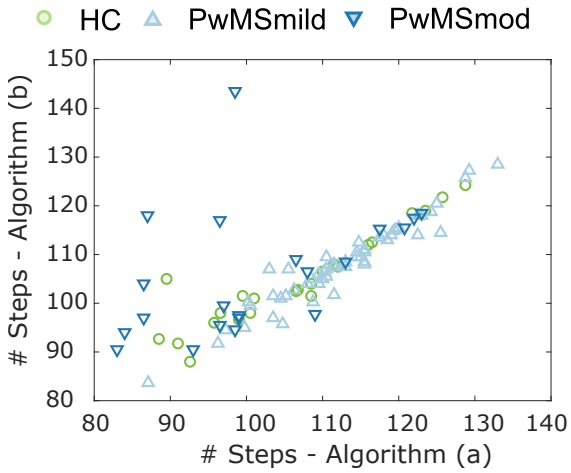


Figure 6.9: The agreement between two step counting algorithms. Step counts extracted upon the smartphone-based 2MWT across HC (green), PwMSmild (light blue) and PwMSmod (dark blue) groups. **(a)** denotes step count implemented using methods described by [299], whereas **(b)** corresponds to step count implemented using methods described by [300]. Each point represents the average step count per 60 [s] walking bout (i.e. cadence). Disagreement between (a) and (b) occurs generally at lower number of steps, particularly for PwMSmod groups.

Table 6.3: Smartphone-based step count classification outcomes by EDSS grouping. Results presented as the mean and standard deviation across CV repetitions (%), using a SVM with a RBF kernel.

Classifier (Step Count†)	Acc.	MF_1
HC vs. PwMSmild	59.1 ± 3.2	54.9 ± 3.4
PwMSmild vs. PwMSmod	68.6 ± 3.7	59.8 ± 3.4
HC vs. PwMSmod	56.7 ± 3.8	60.0 ± 5.1

†Total step count over the whole 2MWT used as a single feature for classification.

6.3.2 Classification Analysis

To gather an understanding of added smartphone and smartwatch feature performance the out-of-sample classification accuracy (HC vs. PwMSmod) was evaluated as the number of features added into an SVM and LR classifier was varied. While subjects were instructed to preferably carry the smartphones in the provided running belt, analysis has found that some participants wore the smartphone on either the running belt or in a pocket during the 24-week testing period (HC, n=15; PwMSmild, n=41; PwMSmod, n=14). Smartphone orientations captured in landscape orientation were deduced to have come from the running belt in the anterior waist location (HC n=905; PwMSmild, n=2424; PwMSmod

n=1296), whereas portrait orientations were labelled as pocket locations (HC, n=457; PwMSmild, n=1497; PwMSmod, n=156). No subject in this analysis contributed only pocket locations.

SVM classification accuracy using running belt tests rather than any location (either pocket or the running belt) (figure 6.10a) yielded improved accuracy; significantly so ($p<0.05$) beyond 3 features thereafter (besides 7-8 features added $p=0.09$ and $p=0.23$ respectively). Classification accuracy plateaued after 15 features are added to the model. Many smartphone features showed significantly different distributions between pocket and running belt locations (n=22, $p<0.05$). Given the smaller number of subjects and highly skewed number of pocket observations contributed per subject, where few subjects contributed the majority of pocket tests, the classification performance of pocket locations alone was unable to be evaluated.

Classification performance was compared using features derived from either smartphone or smartwatch devices or with features derived from both devices (figure 6.10b and table 6.2). Maximum out-of-sample CV subject classification performance was reached using an SVM for smartphone devices with 23 features (Acc. 82.2 ± 2.9, MF_1: 84.3 ± 3.8), compared with 19 features for smartwatch devices (Acc. 71.3 ± 3.6, MF_1. 71.1 ± 3.5). Furthermore, smartphone devices showed significantly better accuracy ($p<0.05$) with at least 8 features added to the classifier.

Additional classification models were also constructed to explore the separability between PwMSmild and PwMSmod groups, and HC and PwMSmild groups separately shown in table 6.4. It was observed that classification of PwMSmild and PwMSmod groups performed similarly: subject classification was maximised (Acc. 82.3 ± 1.9, MF_1: 75.1 ± 2.2) using an SVM classifier with 21 features. Separation between HC and PwMSmild sub-groups however was less visible; where maximum classification accuracy was achieved (Acc. 66.4 ± 4.5, MF_1: 58.6 ± 5.8) with 19 features modelled using an SVM.

Univariate and multivariate modelling of top-ranked signal-based complexity features achieved improved classification performance compared to using total step

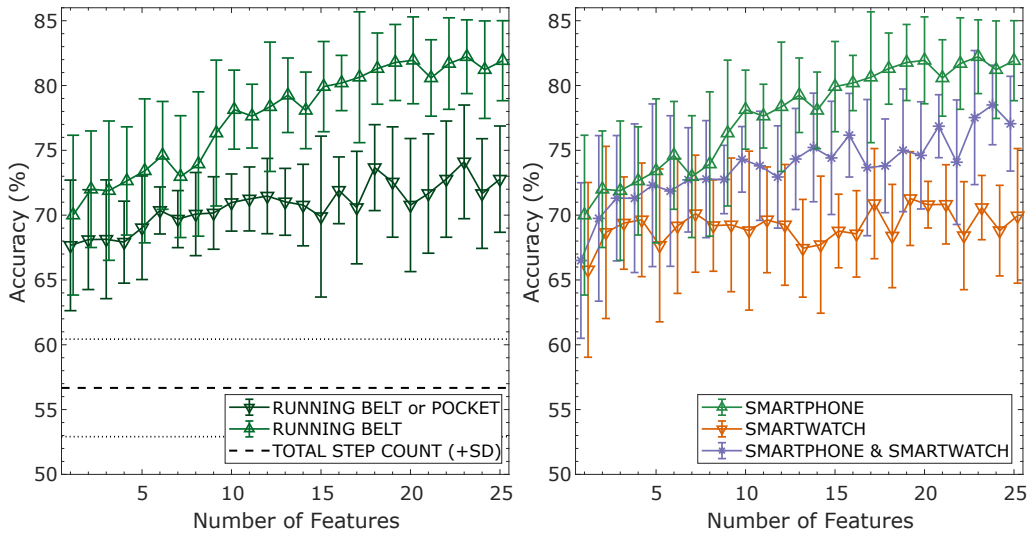


Figure 6.10: Comparison of out-of-sample subject classification performance as features are added to an SVM classifier. Results presented for the HC vs. PwMSmod classification task, using 5-fold CV with 10 repetitions. Figure **(a)** compares classification accuracy between smartphone tests performed using the running belt only versus tests using either the running belt or the pocket. Total step count represents the total step count over the whole 2MWT if used as a feature for classification. Figure **(b)** compares classification performance using features from smartphone, smartwatch and both devices. Confidence intervals denote one standard deviation (SD) around the quoted mean performance. For clarity, only the first 25 steps are presented.

Table 6.4: Smartphone-based subject classification outcomes by EDSS grouping for various classifiers. LR, Logistic Regression; SVM, Support Vector Machines with a RBF kernel; RF, Random Forest. Results presented as the mean and standard deviation across CV repetitions (%).

Task	Classifier	Acc.	MF_1
HC vs. PwMSmild	LR	64.5 ± 5.8	58.3 ± 5.5
	SVM	**66.4 ± 4.5**	**58.6 ± 5.8**
	RF	63.2 ± 4.0	50.4 ± 5.6
PwMSmild vs. PwMSmod	LR	83.7 ± 2.4	77.1 ± 2.7
	SVM	82.3 ± 1.9	75.1 ± 2.2
	RF	**84.0 ± 1.9**	**78.1 ± 2.3**
HC vs. PwMSmod	LR	80.4 ± 5.3	80.0 ± 5.4
	SVM	**82.2 ± 2.9**	**84.3 ± 3.8**
	RF	76.2 ± 3.5	76.5 ± 3.9

count alone as a feature for all classification outcomes based on EDSS grouping (figure 6.10a) and table 6.3.

Classification was compared across different classifiers (LR, SVM and RF) as depicted in table 6.4. SVM models performed best at distinguishing HC from PwMSmild and PwMSmod, whereas the RF was marginally better at separating

PwMSmild vs. PwMSmod. Accuracy was not significantly different between LR, SVM and RF models for all binary classification tasks expect at distinguishing HC from PwMSmod, with both LR and SVM performing significantly better than the RF ($p<0.05$). Maximal subject classification was obtained for all classifiers using ≤ 26 features. It was observed that majority voting improved all classifier's performance.

6.4 Discussion

This chapter examined gait and physical function in PwMS and HC using smartphone and smartwatch sensor data, while subjects performed a remote 2MWT. The aim of this analysis was twofold: (1) can meaningful features be derived from the remotely performed 2MWT that correlate with clinical assessments, and (2) can multivariate modelling of these performance metrics correctly distinguish groups of HC and PwMS with mild and moderate disability.

6.4.1 Feature Evaluation

As introduced in section 6.1, and background section 2.3.4, many previous studies that characterise gait function with inertial sensors in PwMS have done so within controlled in-clinic environments. The "biomechanical" features these studies extracted (and other similar studies but in different disease domains, such as Parkinson's disease introduced in [300, 313]) are highly parameterised and only validated in controlled manners or specific cohorts. From a clinical perspective, the ability to compute interpretable "biomechanical" features directly from smartphone (and smartwatch) inertial sensor data is important to attribute easier clinical meaning and reinforce the current literature, but has practicable limitations applied to remote inertial sensor data. Biomechanical feature methodologies are often highly parameterised [314], bespoke [300] or not flexible enough initially to be adopted to disease-related free-living settings without validation or calibration first [167, 313, 315, 316]. Consequently, applying biomechanical-based algorithms blindly to remote sensor data without a specific prior validation study or without the presence a clinical observer lacks robustness. As such, this work focused on using established

signal-based features as surrogates to represent complex biomechanical features. For instance calculating entropy-based measures to characterise the complexity in a signal as a proxy to capture disease-influenced ambulatory activity rather than modelling actual biomechanical processes.

This study instead only evaluated a simple biomechanical step count and step power features, which indicated of some differences between groups (figure 6.8). Considering that the 2MWT assesses walking distance in PwMS [98], the number of steps taken can be used to indirectly approximate walking distance [297]. In addition, studies have found that cadence (steps/minute) during the 6MWT and daily step count to be significantly different between MS subgroups [61, 62]. However, this simple biomechanical metric, as computed based on Lee *et al.* [299], showed less discriminatory power between groups compared to signal-based complexity features. It must be considered that a recent comparative study calculating step counts in PwMS has also found considerable variability in the precision and accuracy of the algorithms and devices [129]. This mirrored our results as shown in figure 6.9, where two different parameter-reliant step counting algorithms disagreed, significantly for diseased PwMSmild and PwMSmod participants. Interestingly, it was recently demonstrated that CWT-based step detection error was greater during slower walking speeds, for more disabled PwMS [167]. Ideally, step length would be used to calculate distance travelled during the 2MWT. Methods to derive step length however are erroneous or highly complex [317, 318], and their use in this study would again require validation in more controlled (non-remote) settings. Therefore, it was opted to approximate the outcome of distance travelled for the remote 2MWT using a simple step count and its inclusion as a classifying feature is presented here primarily for comparative purposes.

Classification performance, bench-marked against total step count, improved for all classification outcomes based on EDSS group stratification through the usage of these features (table 6.3 and table 6.4).

A number of the top features characterise the energy, frequency and variability of sensor signals recorded in the gait domain (table 6.2). In particular, percentage

energy, multiscale and Shannon's entropy appeared prominently for both devices. Figure 6.4 demonstrated that HC had more power in the gait domains (0.5–3 Hz). Also, the entropy computed in the gait band $H(cD_5(\|\boldsymbol{a}\|))$ was higher in PwMSmild and PwMSmod, i.e. the predictability of the gait signal in PwMS was lower than HC, with PwMS exhibiting lower energy and slower movements in their gait. Mapping clinical meaning to this could attribute PwMS, especially PwMSmod, to be less predictable with increased gait variability. Those with more severe MS could move more erratically as suggested in the overall top features selected from both devices that characterise the signal-to-noise ratio (SNR) and accelerometer skewness, which could capture noise and jerk-like movements.

6.4.2 Classification Evaluation

It was observed that PwMSmod were distinguishable from both HC and PwMSmild (Acc. $82.3 \pm 2.9\%$ and $82.3 \pm 1.9\%$ respectively), whereas PwMSmild were relatively less distinguishable from HC (Acc. $66.4 \pm 4.5\%$). One of the primary means of PwMS disease assessment is the EDSS [19], hence it was used to stratify sub-groups of PwMS in this study. While those subjects with EDSS$\geq$3.5 are considered to have some gait related impairment, analysis of EDSS ambulation sub-scores demonstrated lower levels of ambulatory impact [19]. The heterogeneity of MS as a disease and its effect on symptom manifestation must also be considered when stratifying sub-groups based on clinical assessments and analysis thereafter, especially for those with lower (milder) severity scores [19, 56]. For example, examination of the mean PCA value per subject for smartphone features (figure 6.7d) and for smartwatch features (figure 6.7h) using the first 2 components demonstrated that some PwMSmod can appear like HCs and vice-versa, where PwMSmild bisected regions between the two groups. This is an outcome to be expected from such a model considering that PwMSmild subjects experience very little gait abnormalities, and T25FW times were not significantly different between PwMSmild and HC (p=0.26).

LASSO is a common ML technique with integrated feature selection and regression functionality (in this case LR) which is robust for reducing large numbers of

features [200]. LR, in essence, describes a linear relationship between the predictors with a non-linear mapping to response variables. LR models performed similarly to the SVM and RF classifier, suggesting perhaps a simple linear relationship exists between the combination of gait features. However there may be other non-linear feature selection techniques which may select more optimal features in a more optimal ranking [200]. RF models built in this study, which internally use non-linear feature selection, did however perform significantly worse than LR and SVM models for HC vs. PwMSmod classification tasks ($p<0.05$). A limitation of this work is that each subject's tests were considered independent and identically distributed (i.i.d), where subjects each contributed a varying number of tests. In reality, test observations per subject will be dependent and may be better suited to sequence modelling approaches. For example, there may be alternatives to RF in this application such as Mixed Effect Trees [319], which consider repeated measures (tests) for classification and could help overcome potential model biases related to the varying number of observations per subject.

Comparing the classification accuracy of HC vs. PwMSmod using smartphone and smartwatch devices showed comparative prediction accuracy for a smaller number of features added (<8) to our models (figure 6.10b). However, beyond this smartphone features demonstrated significantly improved performance ($p<0.05$) over smartwatch features. Surprisingly, drawing from both the smartphone and smartwatch feature space did not lead to improved classification performance and maximal accuracy was achieved using smartphone features only (smartphone: Acc. 82.2 ± 2.9%; smartwatch: Acc. 71.3 ± 3.6%). Further investigations revealed a high variability in the type of feature (calculated from smartphone vs. smartwatch) selected at each fold. Interrogating the feature distributions within CV folds highlighted poor feature generalisability between training and testing sets in some cases. Table 6.2 indicated for example the top feature $H(cD_5(\|\boldsymbol{a}\|))$ was picked 80% of the time, but when the feature was picked, it was nearly always in the top 5 features per fold — in other folds the distributions were changed so dramatically they would not be picked. Combining features from both devices amplified these problems

and did not lead to improved results. Finally, the added smartphone classification accuracy beyond 10–15 features was minimal (figure 6.10b), demonstrating that accurate classification can be achieved with a small number of features.

6.4.3 Limitations and Considerations

Although remote monitoring has many advantages such as unobtrusive, high-frequency assessment of disease, a number of confounding factors must be considered when taking measurements in real-world non-laboratory scenarios. Some examples encountered included the differences in subjects' adherence during the study, some subjects not following the prescribed protocols (instances of smartphones in pocket locations or the use of only one device during testing), along with the many degrees of freedom associated with self-generated patient data from real-world testing. While the instructions given (see [31]) were standardised, analogous to that of an in-clinic performed 2MWT [22, 98], the 2MWT in this study was a remotely executed out-of-clinic assessment. The performance of 2MWT can be highly influenced by the testing environment such as the length of the hallways, the number and frequency of subjects' turns, or other factors which cannot be determined remotely.

The number of unique HC (n=24), PwMSmild (n=52) and PwMSmod (n=21) in this study was relatively few. Sampling sufficiently sized data from a more diverse cohort should also be considered in order to build robust and generalizable models. For example, biases may exist related to the mismatch in the male to female ratio between HC and PwMS groups (table 6.1). It was also acknowledged that there was and a high standard deviation in the number of tests contributed by each subject, however no subject group contributed significantly more tests than another (table 6.1). Besides individual subjects' adherence rates, this variability was also partly due to the exclusion criteria imposed on the data used in this analysis, where only linked smartphone and smartwatch tests, and of those, only smartphone tests performed with the running belt were considered for fair comparisons between the devices. As such, performing CV on subject-wise splits with each subject contributing varying numbers of tests can cause the distributions of features to vary between CV partitions.

This helped attribute to the high classification variance between folds and variability in the type and number of features needed for maximal classification. Furthermore, subject heterogeneity can also highly influence model robustness when data is sparse. For example, figure 6.7 illustrated the first two principal components computed from PCA plotted against each other for all HC and PwMS test observations. The top 25 smartphone features (figure 6.7a–6.7d) were used to perform the PCA and hence represented the overall structure of the features which are mostly selected for smartphone classification. It was observed that intra-subject variability appeared low, while inter-subject variability was distinctly high as subjects clustered within themselves. Subject-wise CV is hence confounded not only by low subject numbers and sparse, heterogeneous data, but in this case was also heavily exacerbated by subject's gait feature patterns, which were uniquely associated to the individual. This manifested as problems exhibited in feature selection and generalisability across CV partitions. There has been much discussion within the academic community as to the advantages and disadvantages of subject-wise versus observation-wise CV approaches[263, 264, 320]. Subject-wise CV has been criticised in that it may break assumptions to consistently estimate generalisation error in heterogeneous data and lead to model under-fitting and larger classification error [320]. However, in subject-wise CV it is argued that there is no "leakage" of subject information between training and testing partitions where models can learn individual subjects' characteristics [263]. Further recent analysis by Neto *et al.* [264], using simulated and real-world mobile data, demonstrated how various examples of heterogeneity across subjects can lead to the identification of subjects' characteristics rather than disease in observation-wise CV approaches. As such, this study adopted the conservative subject-wise CV method to eliminate identity-confounding factors, but acknowledges that sub-optimal classification and model under-fitting can occur given the low number of subjects contributing multiple observations in this study.

Finally, a high inter-dependence was observed within the feature space. A number of the top features within a source (device) and features between sources were highly correlated with each other. This indicates that some features may

represent the same information. Figure 6.6 in appendix A shows the inter-source and intra-source feature correlation. This could attribute to predictor redundancy and explain the marginal classification performance beyond 10 features added for all sources (figure 6.10b). However, smartwatch features reached a plateau in added information before smartphone features, suggesting that this device may contain less information and hence more feature redundancy. Every top smartwatch feature (HC vs. PwMSmod) characterises the relative power in gait to non-gait frequency domains derived from CWT and EMD. It should be considered that the location of smartwatch (wrist) sensor to smartphone (running belt) may have a profound effect on the depth of information that can be recorded about gait function. This may be manifested by a larger and more varied number of useful features (as discussed, the top smartwatch features chosen only characterise the relative power in gait to non-gait frequency domains).

Despite applying an orientation transformation during pre-processing, it was found that some smartphone-based features were also location dependent (running belt versus pocket) and classification based on features from only one pose increased the accuracy of our models (figure 6.10a). As such, smartphone-based tests performed with the running belt were considered different to the smartphone in the pocket.

6.5 Conclusion

This chapter demonstrates the benefits of ML and multivariate feature modelling in the identification of the signs of ambulatory function impairment in PwMS from remotely captured smartphone and smartwatch inertial sensor data, where the novel findings can be summarised as:

- A combination of statistical- and signal-based features calculated from both devices performed better than simple biomechanical metrics such as step count, which was used to approximate the standard 2MWT outcome of walking distance;

- This study indicated that sufficient information for accurate MS symptom characterisation may be captured in relatively few features ($\leq$26) obtained from an out-of-clinic 2MWT, using only one device; This contrasts to many previous studies probing the characteristics and separability of PwMS and HC, which have used multiple standalone inertial sensors affixed to the body at various locations during controlled in-clinic assessments [54, 55, 153, 154];

- It was found that PwMSmod, who experience gait-related dysfunction, could be distinguished with a high accuracy from PwMSmild and HC, whom the latter two groups were more difficult to differentiate from each other;

- The work presented here, with on-going future work, helps establish a methodological foundation to construct models that can identify patterns of PwMS ambulatory impairment from remote gait assessments.

Ultimately however, the classification performance presented in this chapter are reliant on the feature structure to sufficiently capture information related to ambulatory function and measures that are specifically capable of characterising ambulatory dysfunction related to MS disease activity. This approach is limited by the effectiveness of our features, which in essence were "hand-crafted" based on our prior hypothesises of what we were attempting to encode. The features introduced in this chapter focused on using established signal-based features, re-purposed as surrogates to capture aspects of hypothesised PwMS ambulation. For instance, creating an entropy-based feature to capture the dynamics of walking, or attempting to characterise perturbed walking by modelling the energy in an accelerometer signal. With such narrow constraints, based solely on prior hypothesis (albeit relatively clinically informed), we risk missing critical information for our task. As such, the next chapter 7 will introduce a more data-driven methodology, whereby a deep network is allowed to learn its own features to perhaps capture a better representation of HC and PwMS disease characteristics from remote smartphone inertial sensor data.

7

Deep Learning for the Remote Characterisation of Ambulation in Multiple Sclerosis using Smartphones

Associated publication(s):
Creagh A.P., Lipsmeier F., Lindemann M. and de Vos, M. (2021). "Interpretable Deep Learning for the Remote Characterisation of Ambulation in Multiple Sclerosis using Smartphones". *Nature Scientific Reports*, 11(1), pp.1-14. https://doi.org/10.1038/s41598-021-92776-x

Contents

7.1 Introduction 139
7.2 Background 141
7.3 Methods 145
7.4 Results 151
7.5 Discussion 157
7.6 Conclusion 160

7.1 Introduction

This chapter builds upon the previous investigations in chapter 6, where it has been shown how smartphones and smartwatches can be used to characterise gait impairments in people with multiple sclerosis (PwMS) from a remotely administered

Two-Minute Walk Test (2MWT). The work presented in chapter 6 demonstrated how inertial sensor-based features can be extracted from consumer devices to develop machine learning models that can distinguish gait-related MS disability from healthy participants. These approaches, however, are essentially constrained approximations of ambulatory function which are based on prior assumptions of what hand-crafted sensor-based features may characterise. For instance, extracting the mean and variance in a sensor signal in an attempt to capture gait variability in PwMS. As such, there may be much greater power in allowing an algorithm to learn it's own features [251, 321], termed *representation learning* [247].

Deep learning, as introduced in section 3.9, is an overarching term given to *representation learning*, where multiple levels of representation are obtained through the combination of a number of stacked (hence *deep*) non-linear model layers. Deep learning models typically describe convolution neural networks (CNN), deep neural networks (DNN), or combined fully-connected deep convolution neural network (DCNN) architectures [202]. While CNNs are omnipresent in image recognition-based tasks, these models can also be applied towards time-series tasks [251]. For example, CNNs have been shown to act as feature extractors capable of learning temporal and spatio-temporal information directly from the raw time-series signals [202, 251]. The features extracted by convolution layers can then be arranged to create a final output (a classification or regression) through fully connected (dense) layers.

The work presented in this chapter compares the performance of CNN extracted features and DCNN models against hand-crafted features proposed in chapter 6 to directly classify healthy controls (HC) and subgroups of PwMS from remote smartphone-based inertial sensor data, captured during the 2MWT.

The primary contributions of this chapter to the field of remote monitoring of multiple sclerosis disease can be summarised as:

- The development of a deep learning framework using DCNNs that can identify characteristics of healthy from PwMS disease, where DCNNs can be applied directly to raw smartphone-based inertial sensor data;

- The empirical comparison of DCNN models versus feature-based approaches introduced in chapter 6 to stratify HC and PwMS subgroups;
- The proposal of transfer learning techniques to adequately train DCNN models in the presence of low subject numbers and heterogeneous data;
- The characterisation of applicable open-source datasets and relevant tasks, which may prove useful to transfer domain knowledge from towards ambulatory-based HC and PwMS disease characterisation, recorded from remotely generated smartphone inertial sensor data.

7.2 Background

7.2.1 Deep Networks Applied to Inertial Sensor Data

Deep networks are a relatively well established methodological approach to solve many medical tasks [5, 322], particularly towards medical imaging analysis [323]. Previously introduced work in section 2.3.6 exemplifies this application, for instance in MRI imaging-based lesion detection in PwMS [192–194]. Recently, however, deep networks are now also being applied towards inertial sensor data for a range of various biomedical related tasks.

7.2.1.1 Human Activity Recognition

Human Activity Recognition (HAR) refers to the recognition of different movement-related and static activities, typically from wearable sensor data. Some early studies have perhaps catalysed the growing interest in the power of deep networks to characterise wearable inertial sensor data [321, 324–326].

Many segmentation algorithms aim to classify between general static versus movement activities such as “sitting” and “standing” from “walking” or “running”. Other, more detailed analysis, have attempted to parse apart more specific gesture based activities, such as those found in at-home scenarios; examples being “opening drawers” or “drinking”, which are often termed as activities of daily living (ADL) [327]. By far the most popular— and most accurate— techniques which have been

applied to HAR-based sensor activities are the use of deep- and recurrent-networks [321, 324, 326, 328]. Combination of CNN and Long-Short-Term Memory (LSTM) layers (architecture termed DeepConvLSTM) applied directly to the raw sensor data have been successful in classifying HAR and ADL [321]. The DeepConvLSTM model typically exhibited around 5%–13% better performance than the next best feature-based approach, at 87–93% accuracy, depending on the task. Other deep activity recognition models have proposed a hybrid approach of DNN and hidden Markov models (DNN-HMM) using spectrogram representations of the raw sensor signal, and have resulted in substantial improvement over traditional feature based approaches and shallower models in WISDM[1] (7% improvement at 93% accuracy), DAPHNET[2] (12% improvement at 92% accuracy) and SKODA[3] datasets (4% improvement at 89% accuracy) [324]. Wang *et al.* [328] offer a comprehensive review on existing works based on deep learning approaches for sensor-based activity recognition, including a detailed collation of sensor modality, architecture, application and datasets in the literature. Importantly, the authors identify that deep networks are the top performing models across a range of HAR- and ADL-based datasets, with generally 83%–96% classification accuracy overall.

7.2.1.2 Subject Identification

A similar field of subject identification (SI), has demonstrated the power of deep network-based representation learning, where wearable inertial sensor data (typically from smartphones and smartwatches) have been used to identify between various people [182, 184] . The architecture presented by Gadaleta *et al.* [184], termed IDNet, demonstrated explicitly how CNN-based features extracted from smartphone inertial sensor-data outperformed hand-crafted features (by between 10%–27% improvement in accuracy) when both were compared using traditional machine learning models (such as SVMs and KNNs[4]). Furthermore, this work demonstrated that the fully connected DCNN model achieved the highest classification accuracy

[1] To detect movement versus static activities.
[2] To detect freezing of gait events in Parkinson's disease.
[3] To detect activities belonging to a car maintenance scenario.
[4] K-Nearest Neighbour Algorithm

comparatively at 99.85%. Iterative studies by other authors, Zou *et al.* [182], achieved similar performance to IDNet on SI tasks with various deep network CNN and LSTM combined architectures, including demonstrating the generalisability of the HAR-based DeepConvLSTM architecture introduced above for SI tasks.

7.2.1.3 Disease-related classification

As previously introduced in section 2.3.6, many studies are beginning to explore disease classification and symptom monitoring with wearable-generated inertial sensor data using deep networks. The bulk of these applications have been within Parkinson's disease (PD) assessment [188–191, 329]. For instance, DCNN models have been shown to distinguish PD from healthy subjects from remotely generated smartphone inertial sensor data during walking (AUROC: 0.87)[5] [188]. Towards MS assessment however, Gong *et al.* [197] have demonstrated that healthy participants could be distinguished from PwMSmild and PwMSmod using DCNNs applied to spectrogram-transformed[6] accelerometry data from multiple sensors during fixed, in-clinic walking. More recent work in MS has also demonstrated that bidirectional LSTM (biLSTM) models applied to wearable accelerometer data successfully identified PwMS at risk of falls (AUROC: 0.88). Importantly, those authors report substantially better performance with biLSTMs compared to wearable sensor-based spatiotemporal gait features (21% improvement) and statistical features (16% improvement).

7.2.2 Deep Transfer Learning

Despite the possibility of significant performance improvements compared to hand-crafted feature-based methods, deep networks require much more training data to make successful, robust and generaliseable decisions [202]. Medical data, and in particular, digital health data can often be sparse. Studies are generally small with the accompanying limitations such as: few participant numbers, heterogeneous cohorts, and as learned in section 6.4.3, missing data and varying patient adherence.

[5] AUROC: Area Under the Receiver Operator Curve.
[6] i.e. the sensor representation is treated as an image.

Transfer learning (TL) is a machine learning technique which aims to overcome these challenges by transferring information learned between related source and target domains (i.e. data or datasets) [202, 330, 331]. While the data may be in different domains, or the distributions may differ from the target and source tasks, TL assumes that the knowledge that is learned in another task and dataset will be useful and related to the new target task[7] [202, 326]. Transfer learning has been successfully implemented in many computer vision tasks [332, 333] and for time series classification [334], including wearable inertial sensor data [335, 336]. Some studies have aimed to pre-train models on larger and similar datasets on the same task, such as EEG sleep-staging [337, 338], then subsequently fine-tune their model to the smaller task dataset. The hypothesis being, hierarchically, initial layers in a neural network learn generalised feature representations, and as the network depth increases, the representations become more task or dataset specific [202, 247]. Kernels[8] in the convolutional layers of CNNs have been shown to act as feature extractors, progressively highlighting more domain-specific features in the upper network layers [202, 339]. Morales *et al.* [327] demonstrated this paradigm using wearable sensors in the HAR domain, whereby they linked similarities in CNN features across source and target HAR datasets using transfer learning. The authors presented that inter-source (between HAR datasets) and inter-modality (between sensor locations) transfer learning was possible. TL significantly improved performance compared to a baseline models ($p < 0.05$) for up to 3 layers transferred. Thereafter, model performance decreased. Recently, Martinez *et al.* [335] trained a network using a publicly available inertial sensor wearable data set for HAR, which was subsequently fine-tuned successfully for a falls-risk classification outcome in elderly participants (AUROC: 0.933), resulting in a 10.5% improvement compared to a model trained without TL. This work demonstrated how inertial sensor features from shallower layers can be learned between dissimilar tasks on related domains (i.e. learning shallow HAR features that are useful for falls-risk outcome, both

[7] A classic example are the similar characteristics that could be transferred from image-based cat identification to dog identification instead.

[8] For more information on CNN kernels the reader is referred to section 3.9.1.

using inertial sensor measurements). This TL approach also outperformed other standard machine learning classifiers, such as logistic regression (LR), support vector machines (SVM), and a random forest (RF) (which are the same models that were also previously investigated in chapter 6).

7.2.3 Leveraging Human Activity Recognition for Transfer Learning

The work presented in this chapter will therefore aim to utilise transfer learning to supplement our model's ability to discriminate between healthy and diseased subjects in the FLOODLIGHT PoC (FL) dataset (see section 4.1). Deep transfer learning was performed by first identifying relevant large (open-) source datasets from which information can be exploited. The similarity between some HAR datasets and FL, the applicability of the HAR domain task (which includes walking bouts), as well as the trove of established HAR deep networks suggests that HAR may be a suitable candidate to transfer domain knowledge. In this study two publicly available datasets are identified, UCI HAR [278] and WISDM [279, 280], which use comparable smartphone and smartwatch devices, and device affixing locations similar to that of FL. More information on UCI HAR and WISDM can be found in the previous section 4.4. Figure 7.1 schematically illustrates the transfer learning approach undertaken in this chapter, where the information learned from a HAR classification task ($\mathcal{T}_S$) and dataset ($\mathcal{D}_S$) are transferred towards a disease recognition task ($\mathcal{T}_T$) within the FL dataset ($\mathcal{T}_T$).

7.3 Methods

This section follows a similar time series classification problem to chapter 6; the raw smartphone sensor data is partitioned into vector sequences where deep learning models are used to classify a given vector sequence, $\boldsymbol{x}$, as having been performed by a HC, PwMSmild or PwMSmod. Three separate classification models were constructed for binary tasks (HC vs. PwMSmild, PwMSmild vs. PwMSmod and HC vs. PwMSmod) to allow for direct comparison of the hand-crafted feature-based

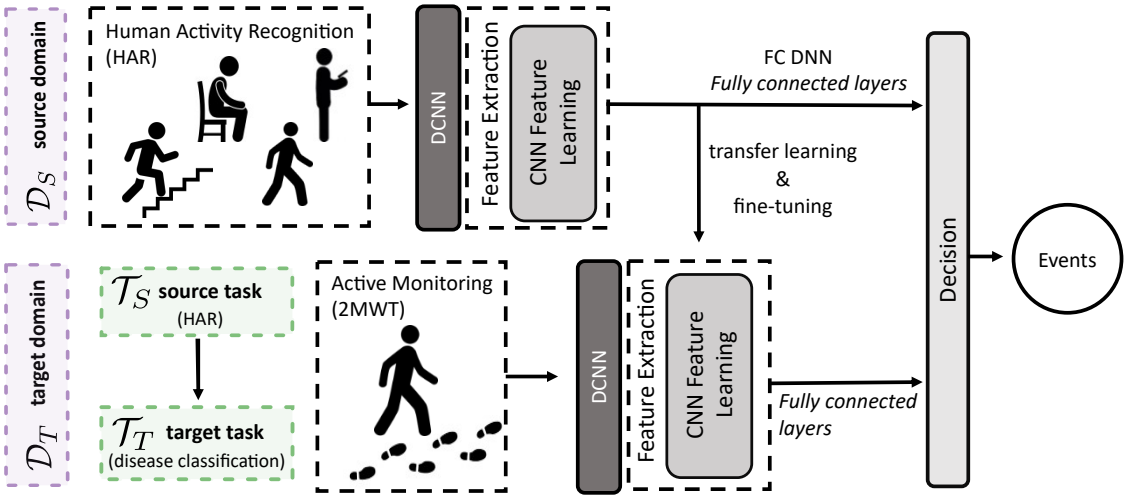


Figure 7.1: Schematic of proposed transfer learning (TL) approach. First, open-source datasets ($\mathcal{D}_S$) were utilised to learn a HAR classification task ($\mathcal{T}_S$) with a DCNN. Learned activity information was then subsequently transferred using the TL framework, where the DCNN model is retrained on the FL datatset ($\mathcal{D}_T$), and parameters are fine-tuned towards the application of a disease recognition task ($\mathcal{T}_T$).

classification outcomes assessed in chapter 6, as well as a unified multi-class model incorporating all three classes simultaneously. The population subset explored for this chapter are the same as in chapter 6, and are depicted previously in table 6.1.

7.3.1 Pre-Processing

Several pre-processing steps were performed to format the raw signals for input into the respective deep networks. To maintain consistency and for comparability using TL approaches, all signals were processed according to the same pipeline as chapter 6. For additional consistency with chapter 6, only subject's 2MWTs identified using the running belt were considered for subsequent analysis in this chapter. All inertial sensor signals were sampled at 50 Hz; in the case of the WISDM dataset, signals were re-sampled to 50 Hz using a shape-preserving piecewise cubic hermite interpolation (PCHIP) (see section A.2.5). Signals were filtered with 4^{th} order Butterworth filter with a cut-off frequency at 17 Hz [153, 294], and as per section 6.2.2.1, the smartphone coordinate axes were aligned with the global reference frame using the PCA-based method described in [184] for IDNet. All signals were detrended (subtracting of the mean) and amplitude normalised with zero mean and unit variance [184]. Sensor signals for each recording were then up-sampled using

fixed-width sliding windows of 2.56 sec and 50% overlap (128 samples/epoch), in accordance with parameters in similar studies [182, 184, 278].

7.3.2 Model Construction

Feature learning was performed by applying a series of one-dimensional kernels on the raw sensor signal with an input (channel 1-4): $\mathbf{X} = (\boldsymbol{a}_x, \boldsymbol{a}_y, \boldsymbol{a}_z, \|\boldsymbol{a}\|)^\top$, where $\boldsymbol{a}$ are acceleration vectors for the **x**-, **y**- and **z**- components containing samples $\boldsymbol{a} = (x_1, x_2, ..., x_T)$ and $\|\boldsymbol{a}\|$ refers to original orientation invariant signal magnitude. The DCNN structure in this study adapted the CNN architecture proposed by [182], which consisted of four CNN blocks (as detailed in figure 7.2): the first block (CNN #1) learned 32 filters with a width of 9 samples and stride length of 1 (9×1); CNN #2 and CNN #3 incorporated 64 filters, with width (3×1); CNN #4 learned 128 filters with a with of 6 (6×1). Varying depth of dense fully connected layers were experimented with (see appendix B.1 information for all model architecture combinations examined)[9]. This architecture also utilised l_2-norm regularisation ($\lambda = 1e^{-3}$) (see section 3.9.3.3) and batch normalisation (BN) layers (see section 3.9.3.2), which have been proposed in similar networks characterising gait PD from smartphone inertial sensors [188, 190]. Max pooling operations (se section 3.9.3.1) were also applied with pool size p=2 and down-scaled by stride factor s=2, as per [182].

7.3.3 Deep Transfer Learning

A model architecture was first trained on source domain $\mathcal{D}_S$ and task $\mathcal{T}_S$, in this instance a HAR classification task on the UCI HAR or WISDM dataset (see section 4.4.1 and section 4.4.2 for more information on these datasets). The parameters and learned weights of source model $f_S(\cdot)$ were then used to initialise and train a new model on domain $\mathcal{D}_T$ and task $\mathcal{T}_T$ by iteratively transferring the source model layers and re-training (fine-tuning) the network towards this new target task $\mathcal{T}_T$ (i.e. in this case the subject group classification between HC, PwMSmild and PwMSmod).

[9]Note: to guarantee determinism between models, all CNN kernel and DNN layer weights were initialised with a He normal initialiser [340] maintaining the same seeding throughout.

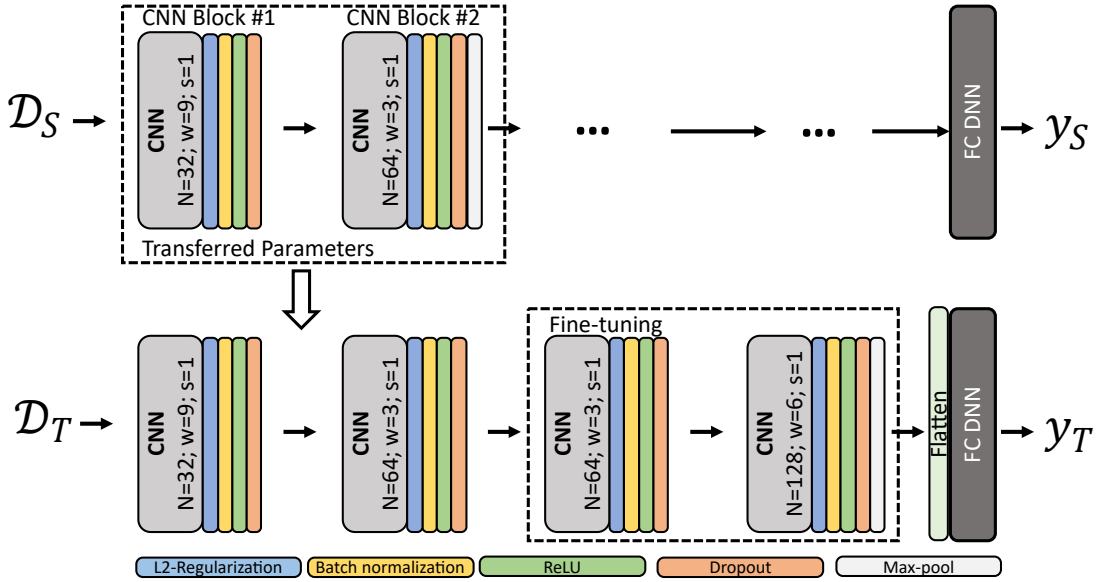


Figure 7.2: Schematic of deep transfer learning approach. $\mathcal{D}_S$ refers to input data from a source domain, in this case a HAR dataset, to learn a task $\mathcal{T}_S$, which is represented by the label space $\mathcal{Y}_S$ (the HAR activity classes). $\mathcal{D}_T$ refers to the target domain, in this case the FLOODLIGHT data, where $\mathcal{Y}_T$ are the disease classification outputs of HC, PwMSmild or PwMSmod for target task $\mathcal{T}_T$. During transfer learning, a model's parameters and learned weights, $f(\cdot)$ of $\mathcal{D}_S$, are then used to initialise and train a model on target domain $\mathcal{D}_T$ and task $\mathcal{T}_T$. Transfer learning is then performed by transferring the source model's layers (where these weights and parameters are "frozen") to subsequently re-train a new model (i.e. fine-tune) using $\mathcal{D}_T$ data for the new target task, $\mathcal{T}_T$. The number of layers transferred can be optimised, where downstream layers in the network are fine-tuned towards this new target task decision $\mathcal{Y}_S$.

Figure 7.2 schematically details the TL approach. Baseline "end-to-end" refers to a DCNN trained and validated exclusively on FL data. "Direct" transfer refers classification that is performed with full HAR trained model and weights; after, only the last fully connected dense layer has been replaced by disease targets and retrained on FL data (all other layers are frozen). "Fixed" transfer refer to a HAR trained architecture, where the convolution blocks and weights are frozen and act as a "fixed feature extractor", however DNN weights thereafter are fine-tuned.

The deep network investigated to build end-to-end and TL models for FL maintain the same parameter and architecture combinations. The work presented below first probes the robustness of these deep network approaches to evaluate HAR tasks in UCI HAR and WISDM dataset, motivating the use of these datasets for TL later.

7.3.4 Model Parametrisation

Models were trained to minimise a cross-entropy loss function (equation 3.76) where the binary target variables $y_k \in 0, 1$ had a 1-of-K coding scheme indicating the class k, depending on task (i.e. the number of classes) analysed. The network outputs were predicted by a softmax function (see 3.73). All DCNN models were trained using *Adam* optimization (see section 3.8) algorithm with ($\beta_1 = 0.9, \beta_2 = 0.999$), the exponential decay rates for the moment estimates of the gradient [202, 246]. This study also experimented with varying initial learning rates ($lr = \{1 \times 10^{-2}, 5 \times 10^{-2}, 1 \times 10^{-3}, \ldots, 5 \times 10^{-5}, 1 \times 10^{-5}\}$) and minibatch size ($bs = \{64, 128, 250, 500\}$). For the fine-tuning process, the use of lower learning rates have shown to work best as the source-learned weights should not deviate too greatly during the fine-tuning optimisation, which could risk losing the transferred information [332, 335]. The architecture and parameters that maximised each classification task can be found in appendix table B.1. All models in this work were developed using using the Keras framework 2.2.4 [341] with a Tensorflow 1.14 (Google Inc., California) back-end [342].

7.3.5 Dataset Construction and Model Evaluation

To determine the generalisability of our models, stratified 5-fold, subject-wise, cross-validation (CV) was employed with the same seeding as in chapter 6. This consisted of randomly partitioning the dataset into k=5 folds which was stratified with equal class proportions where possible. One set was denoted the training set (in-sample), which was further split for into smaller set for validation, using roughly 10% of the training data proportionally. The remaining 20% of the dataset was then denoted testing set (out-of-sample) on which predictions were made.

To help alleviate model biases occurring from the varying number of repeated tests contributed per subject over the duration of the FL study, m number of 2MWTs per subject were randomly selected with replacement to create balanced datasets. Parameterisation of the number of tests per subject was determined using a baseline DCNN prior to building all subsequent models within FL. The classification performance over varying data set sizes was examined by sampling

Table 7.1: Overview of source $\mathcal{D}_S$ and target domain $\mathcal{D}_T$ datasets. Datasets were constructed from the original sensor signal using 2.56 [s] epoch sliding windows with a 50% overlap.

	$\mathcal{D}_S$		$\mathcal{D}_T$
(n)	UCI HAR[1]	WISDM[1]	FL
subjects	30	51	97[a]
tests	61	252	970[b]
samples	10,013	54,781	82,450

See sections: [1] 4.4.1; [2] 4.4.2 for more details.
[a] HC, n=24; PwMSmild, n=52; PwMSmod, n=21; see table 6.1 for more details. [b] Randomly sampling $m = 10$ tests per subject.

$m = \{1, 5, 10, 25, 50\}$ daily tests sampled (with replacement) per subject. It was observed that there was minimal additional classification performance after $m = 10$ 2MWTs sampled per subject across each binary task (see appendix figure B.2 for more details). Class distributions in the training and validation sets were then balanced using the re-sampling approach from section 6.2.4. Imbalances in the HAR training and validation data were also countered using this approach. The total number of observations/epochs for each constructed datasets are depicted in table 7.1. HAR model performance was reported based on the classification of individual epochs into the correct activity class for UCI HAR and WISDM. FL and TL performance was based on the majority prediction over a 2MWT (test-wise), where final subject-wise classification results were computed though majority voting each aggregated individual 2MWT prediction per subject (as per chapter 6).

Multi-class classification metrics were then reported as the median ± std. deviation over one CV and the subject-wise outcome for that CV (in the case of FL), using overall metrics such as the macro accuracy, macro F1-score (MF_1) and Cohen's kappa (k) (see section 3.11 for more details).

The structure of the DCNN features extracted were visually examined using t-SNE[10] (see section 3.3) with respect the each $\mathcal{T}_S$ and $\mathcal{T}_T$.

[10]For clarity of visualisation and in order to speed up convergence and computation times, t-SNE visualisations on FL sampled roughly 50% of the dataset (n=8075 epochs per fold), with equal number of epochs per subject (those figures include: 7.5, 7.6, B.4 B.5)

Table 7.2: HAR classification $\mathcal{T}_S$ performance for each $\mathcal{D}_S$. Results presented as the mean ± standard deviation over one cross-validation run.

$\mathcal{D}_S$	Acc.	κ	MF_1
UCI HAR	90.5 ± 1.8	0.880 ± 0.023	0.893 ± 0.025
WISDM	62.1 ± 3.7	0.525 ± 0.047	0.622 ± 0.038

Actual \ Predicted	laying	sitting	stairs	standing	walking
laying	93.4% (1784)	2.0% (39)	0.05% (1)	4.6% (87)	0% (0)
sitting	10.2% (177)	68.1% (1183)	0.1% (2)	21.6% (376)	0% (0)
stairs	0% (0)	0.04% (1)	98.6% (2745)	0.1% (3)	1.3% (35)
standing	0.05% (1)	7.6% (145)	0.1% (2)	92.3% (1774)	0% (0)
walking	0% (0)	0.2% (4)	4.6% (76)	0% (0)	95.2% (1578)

(a) UCI HAR

Actual \ Predicted	sitting	stairs	standing	walking	jogging
sitting	61.3% (6632)	3.1% (331)	33.5% (3622)	1.2% (128)	1.0% (103)
stairs	3.7% (390)	59.2% (6227)	9.4% (991)	14.8% (1561)	12.8% (1351)
standing	34.2% (3763)	6.6% (727)	57.2% (6293)	1.4% (159)	0.6% (68)
walking	2.2% (254)	30.1% (3452)	3.7% (419)	54.0% (6190)	10.0% (1147)
jogging	2.0% (214)	20.4% (2237)	2.5% (270)	6.1% (665)	69.1% (7587)

(b) WISDM

Figure 7.3: Confusion matrices for $\mathcal{T}_S$ HAR classification results. $\mathcal{D}_S$: (a) UCI HAR dataset and (b) WISDM dataset.

7.4 Results

7.4.1 HAR Evaluation

In this section, $\mathcal{T}_S$ HAR models on the UCI HAR and WISDM $\mathcal{D}_S$ datasets were evaluated, maintaining the same parameter and architecture combinations that were investigated when building end-to-end models for FL. Table 7.2 depicts the cross-validated HAR $\mathcal{T}_S$ classification results for each $\mathcal{D}_S$ dataset. Figure 7.3a and figure 7.3b represent the analogous confusion matrices for each $\mathcal{D}_S$ and $\mathcal{T}_S$.

The latent $\mathcal{T}_S$ DCNN feature space learned on the UCI HAR $\mathcal{D}_S$ is visualised using t-Distributed Stochastic Neighbour Embedding (t-SNE) in figure 7.4.

7.4.2 HAR Evaluation on FLOODLIGHT

Trained HAR models were evaluated directly on FLOODLIGHT sensor data to determine the $\mathcal{T}_S$ predictions $\hat{\mathcal{Y}}_S$. Visualisation of the t-SNE embeddings of UCI HAR DCNN features in FL across 5 CV folds are presented in figure 7.5. For clarity of t-SNE visualisation and in order to speed up convergence times, figure 7.5 depicts roughly 50% of the FL dataset using $n = 40,375$ epochs ($n = 8075$

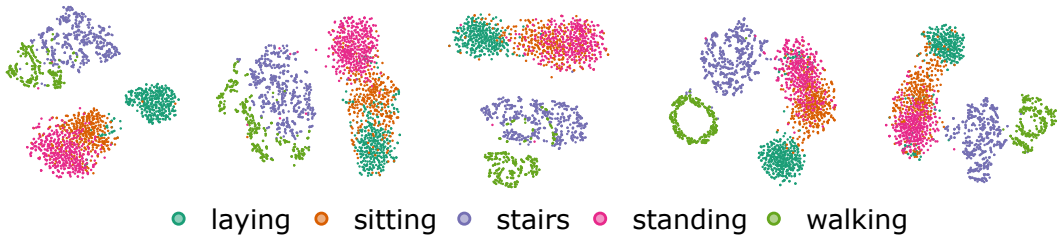
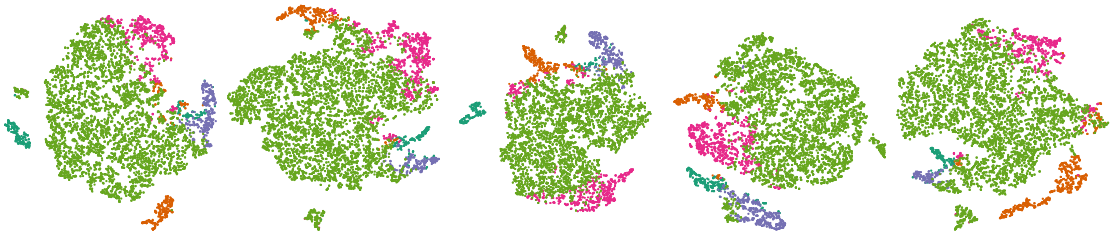


Figure 7.4: Visualisation of UCI HAR DCNN features using *t*-Distributed Stochastic Neighbour Embedding (t-SNE). The UCI HAR ($\mathcal{D}_S$) trained DCNN features are labelled by HAR true activity ($\mathcal{Y}_S$). Each column represents the t-SNE embedding of one fold (test partition) in the UCI HAR dataset (k=5 folds).



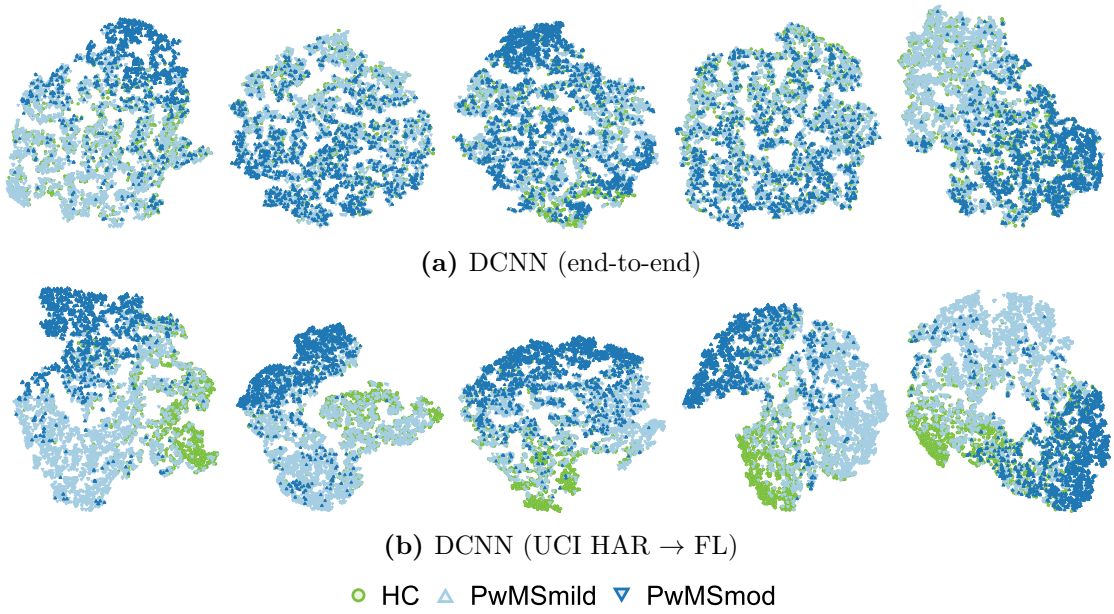
(a) DCNN (UCI HAR → FL, direct): $\mathcal{Y}_T$



(b) DCNN (UCI HAR → FL, direct): $\hat{\mathcal{Y}}_S$

Figure 7.5: Visualisation of UCI HAR trained DCNN features in FLOODLIGHT using *t*-Distributed Stochastic Neighbour Embedding (*t*-SNE). Each column represents one fold (test partition) of the FLOODLIGHT dataset (k=5 folds). Therefore, **(a)** represents the *t*-SNE embeddings of $f_S(\cdot)$ for $\mathcal{D}_T$, i.e. where epochs are labelled by true subject group ($\mathcal{Y}_S$: HC / PwMSmild / PwMSmod) whereas **(b)** depicts the analogous plots coloured by UCI HAR predictions $\hat{\mathcal{Y}}_S$ for $\mathcal{T}_S$ activities.

epochs per fold), with an equal number of observations per subject. HAR DCNN *t*-SNE embeddings in FL are labelled by actual subject group ($\mathcal{Y}_T$: HC, PwMSmild, PwMSmod) in figure 7.5a and labelled by the analogous ($\mathcal{D}_S$, $\mathcal{T}_S$) UCI HAR predictions ($\hat{\mathcal{Y}}_S$) in figure 7.5b. The majority of epochs in FL were predicted as movement activities, "walking" (75.7%) or "stairs" (10.5%). Static activities made up 13.8% of epochs within this sample. A Kruskal–Wallis test by ranks (KWt) (see section 3.4.2.8) demonstrated that the median number of HAR activities



(a) DCNN (end-to-end)

(b) DCNN (UCI HAR → FL)

Figure 7.6: Visualisation of FL trained DCNN features using *t*-Distributed Stochastic Neighbour Embedding (*t*-SNE). Each column represents one fold (test partition) of the FLOODLIGHT dataset (k=5 folds). Figure **(a)** represents a DCNN trained end-to-end on FL; figure **(a)** represents a DCNN model transferred from UCI HAR (UCI HAR → FL). DCNN (UCI HAR → FL) appears to learn a better feature representation capable of stratifying HC, PwMSmild and PwMSmod subject groups.

per subject were not significantly different across HC, PwMSmild and PwMSmod subject groups (appendix table B.2). PwMSmod however tended to have more epochs evaluated as "sitting" than HC and PwMSmild. Appendix table B.2 presents the distribution of predicted UCI HAR activities on the FL 2MWT data for each subject group: HC, PwMSmild and PwMSmod. The analogous HAR CNN only features are depicted in appendix figure B.4.

7.4.3 FLOODLIGHT Evaluation

Similar to figure 7.5, the latent DCNN (UCI HAR → FL)[11] feature structure for a multiclass $\mathcal{T}_T$ was explored using *t*-SNE as depicted in figure 7.6. The analogous CNN (UCI HAR → FL) only features are depicted in appendix figure B.4.

Table 7.3 depicts the classification outcomes for all tasks. Bench-marking against a feature based Support Vector Machines (SVM) classifier introduced in chapter 6, DCNN (end-to-end) model performance was similar in all tasks. As with the results

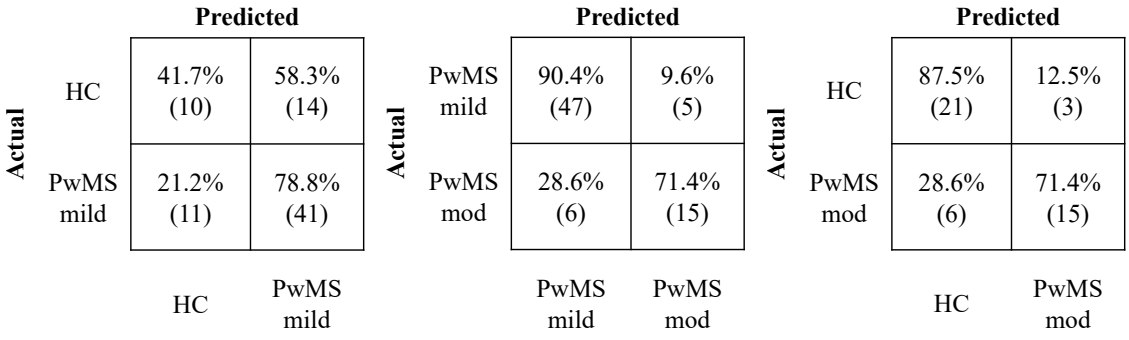
[11] "→" denotes the source HAR dataset $\mathcal{D}_S$ used and transferred to FL $\mathcal{D}_S$ and $\mathcal{T}_S$.

presented in chapter 6, PwMSmod could largely be distinguished from HC and PwMSmild. Incorporating a CNN HAR model as a fixed feature extractor ("fixed"), combined with a newly initialised DNN classifier achieved similar performance as end-to-end DCNNs (table 7.3). HAR DCNN models evaluated directly on FL ("direct") did not distinguish between subject groups. Transfer learning improved disease classification accuracy for all tasks examined relative to feature-based and end-to-end models by between 8%–33%, where TL DCNN based on $\mathcal{D}_S$, UCI HAR and WISDM performed similarly for all target classification tasks $\mathcal{T}_T$ (table 7.3). The associated posterior subject-wise confusion matrices evaluated using SVM, DCNN (end-to-end) and DCNN (UCI HAR $\rightarrow$ FL) for $\mathcal{T}_T$: HC vs.PwMSmild, PwMSmild vs.PwMSmod and HC vs. PwMSmod are depicted in figure 7.7. The parameters of models achieving maximal classification performance within table 7.3 are presented in the accompanying appendix table B.1. Further results expanding on this work can be found in appendix B.2–B.5, including the training regimes for each binary DCNN model (figure B.3), which depicts the validation performance for DCNNs trained end-to-end on FL versus those initialised and fine-tuned using TL.

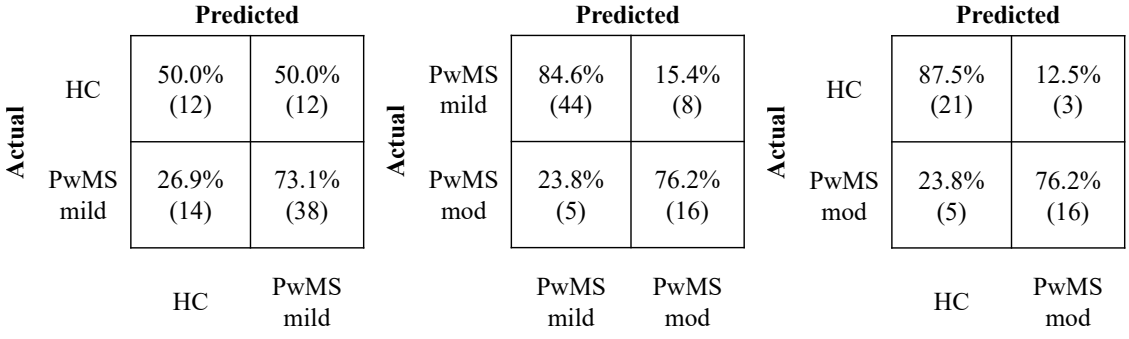
Table 7.3: Comparison of smartphone-based classification results between various models for each task subset ($\mathcal{T}$). Results are presented as: (1) the posterior subject-wise outcome for one cross-validation (CV) run as well as (2) the 2MWT *test-wise* median and interquartile range (IQR) across that CV in brackets. The models that achieved the highest classification performance are highlighted in **bold**. See figure 7.2 for a more detailed description of the TL approach used in this study.

$f(\cdot)$	Acc.	κ	MF_1
		HC vs. PwMSmild	
features + SVM	0.671 (0.576, 0.544–0.696)	0.212 (0.153, 0.088–0.393)	0.605 (0.575, 0.527–0.694)
DCNN (end-to-end)	0.658 (0.601, 0.517–0.641)	0.226 (0.082, 0.037–0.194)	0.613 (0.541, 0.494–0.588)
DCNN (UCI HAR, fixed)	0.632 (0.679, 0.592–0.726)	0.086 (0.106 0.050–0.231)	0.554 (0.539, 0.483–0.626)
DCNN (UCI HAR→FL)	**0.776 (0.741, 0.688–0.767)**	**0.510 (0.435, 0.346–0.481)**	**0.754 (0.716, 0.662–0.737)**
DCNN (WISDM, fixed)	0.533 (0.598, 0.485–0.614)	0.088 (0.138, 0.085–0.200)	0.532 (0.530, 0.480–0.567)
DCNN (WISDM→FL)	0.763 (0.733, 0.698–0.761)	0.486 (0.479, 0.343–0.490)	0.741 (0.727, 0.667–0.743)
		PwMSmild vs. PwMSmod	
features + SVM	0.849 (0.783, 0.706–0.858)	0.627 (0.566, 0.412–0.708)	0.813 (0.778, 0.692–0.853)
DCNN (end-to-end)	0.822 (0.682, 0.617–0.763)	0.583 (0.356, 0.166–0.444)	0.791 (0.675, 0.562–0.721)
DCNN (UCI HAR, fixed)	0.726 (0.700, 0.614–0.747)	0.367 (0.333, 0.123–0.428)	0.683 (0.702, 0.584–0.791)
DCNN (UCI HAR→FL)	0.904 (0.849, 0.839–0.873)	0.776 (0.675, 0.650–0.707)	0.888 (0.837, 0.818–0.871)
DCNN (WISDM, fixed)	0.849 (0.783, 0.725–0.818)	0.592 (0.405, 0.343–0.539)	0.793 (0.696, 0.668–0.796)
DCNN (WISDM→FL)	**0.918 (0.869, 0.833–0.935)**	**0.810 (0.690, 0.630–0.844)**	**0.905 (0.845, 0.812–0.922)**
		HC vs. PwMSmod	
features + SVM	0.800 (0.773, 0.737–0.881)	0.595 (0.546, 0.474–0.763)	0.796 (0.772, 0.737–0.881)
DCNN (end-to-end)	0.822 (0.734, 0.663–0.831)	0.641 (0.462, 0.292–0.657)	0.820 (0.730, 0.618–0.828)
DCNN (UCI HAR, fixed)	0.756 (0.750, 0.481–0.811)	0.507 (0.495, 0.081–0.610)	0.754 (0.745, 0.397–0.805)
DCNN (UCI HAR→FL)	0.889 (0.873, 0.730–0.929)	0.777 (0.743, 0.446–0.847)	0.889 (0.870, 0.723–0.924)
DCNN (WISDM, fixed)	0.844 (0.785, 0.695–0.810)	0.690 (0.559, 0.364–0.606)	0.844 (0.773, 0.674–0.802)
DCNN (WISDM→FL)	**0.911 (0.886, 0.766–0.911)**	**0.821 (0.772, 0.520–0.820)**	**0.911 (0.886, 0.760–0.910)**
		HC vs. PwMSmild vs. PwMSmod	
features + SVM	0.629 (0.551, 0.510–0.577)	0.368 (0.093, 0.020–0.103)	0.580 (0.510, 0.495–0.540)
DCNN (end-to-end)	0.608 (0.503, 0.488–0.516)	0.274 (0.106, 0.081–0.130)	0.523 (0.446, 0.402–0.483)
DCNN (UCI HAR→FL)	**0.814 (0.703, 0.700–0.744)**	**0.673 (0.331, 0.325–0.423)**	**0.796 (0.672, 0.664–0.720)**
DCNN (WISDM→FL)	0.763 (0.690, 0.677–0.737)	0.571 (0.303, 0.274–0.407)	0.725 (0.671, 0.644–0.699)

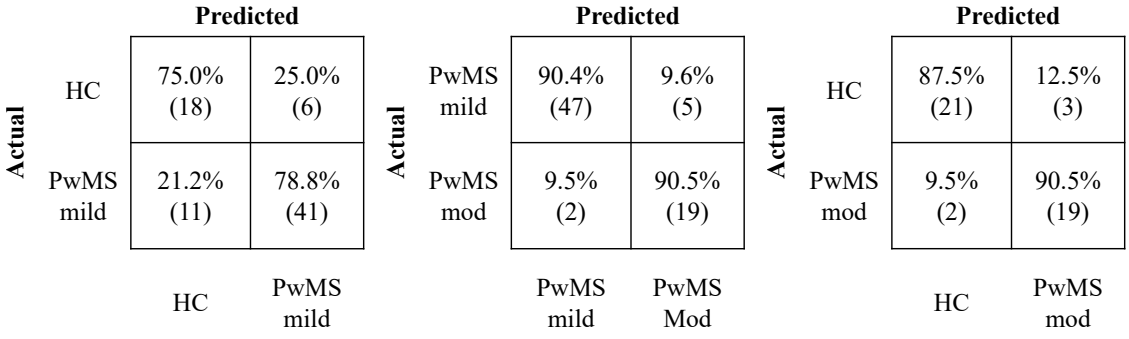
"features + SVM" refers to classification performed using features and a SVM with the pipeline described in chapter 6; "end-to-end", refers to a model trained and validated end-to-end exclusively on $\mathcal{D}_T$ data; "fixed" refers to a $\mathcal{D}_S$ HAR trained architecture, where the convolution blocks and weights are frozen and act as a fixed feature extractor, and DNN weights thereafter are fine-tuned towards FL $\mathcal{S}_T$ and $\mathcal{T}_S$; "→" denotes the source HAR dataset $\mathcal{D}_S$ transferred to $\mathcal{D}_T$.



(a) Features + SVM



(b) DCNN (end-to-end)



(c) DCNN (UCI HAR→FL)

Figure 7.7: Confusion matrix for smartphone-based subject classification outcomes. For, column-wise (L-R) 3 separate models were constructed based on: (1) HC vs. PwMSmild, (2) PwMSmild vs. PwMSmod and (3) HC vs. PwMSmod. Subject-wise classification results are presented using using 5-fold CV for **(a)** hand-crafted features + SVM classifer, **(b)** DCNN model trained end-to-end on FLOODLIGHT and **(c)** DCNN model that has been fine-tuned using transfer learning. For (c), Models are first trained on UCI HAR dataset for a HAR classification task and then fine-tuned on FL for subject-group classification. Subject-wise predictions are computed by majority voting all individual test predictions per subject. Confusion matrices depict the both the raw classification values as well as the percentage (%) recall for that class.

7.5 Discussion

Deep networks may learn a better representation of gait function collected from smartphone-based inertial sensors, than those of traditionally hand-crafted features. This study leveraged DCNNs to extract unconstrained features on raw smartphone accelerometry data captured from remotely performed 2MWTs by HC and PwMS subjects. Rather than relying on hand-crafted features, which are constrained transformations and approximations of the original signal, DCNNs offer data-driven approach to characterise ambulatory related features directly from the raw sensor data.

In this work, higher-level DCNN features learned on open-source HAR datasets were transferred to the FL domain. A DCNN model was first trained as a HAR classification task on two independent open-source HAR datasets. In accordance with other studies [179], excellent discrimination of HAR activities was achieved using deep networks in the UCI HAR database (table 7.2; figure 7.3a). Prediction of WISDM-based HAR activities were less accurate in comparison (table 7.2; figure 7.3b), although much of the relative added confusion in WISDM occurred between similar "movement" activities (such as jogging and walking). Despite this, the prediction of static from movement activities were distinctly separate for both UCI HAR and WISDM. Visualisation of the latent DCNN feature space using t-SNE revealed distinct and interpretable HAR clusters for each fold in UCI HAR (figure 7.4). Analysis of these clusters provided useful and confirmatory evidence about the structure of HAR activities; for instance in many folds, "walking" and "stairs", both dynamic activities, are grouped together yet were bisected by a separating hyperplane. These clusters could be attributed to "'movement" based embeddings, and were always distinct from other "static" classes. "Sitting" and "standing" activities grouped together across fold visualisations, demonstrating a similar underlying structure to each. Laying activities mostly appeared disconnected from other clusters, but were closest to other static groupings of "sitting" and "standing". Importantly, applying a HAR-trained model directly to FL did not identify HC and PwMS subgroups. For example, t-SNE embeddings of the UCI

HAR trained DCNN predictions on the FL $\mathcal{D}_T$ dataset (figure 7.5) indicated that HAR activities were well distributed through each subject group (i.e. that no HAR label indicated healthy, mild or moderate PwMS than any other).

The results presented in this chapter demonstrated that DCNN models can be applied to raw smartphone-based inertial sensor data to successfully distinguish PwMSmod from HC and PwMSmild subjects. The performance of DCNN models applied directly to FL (end-to-end) was similar to that of the feature-based approaches proposed in chapter 6. However, models that were trained end-to-end using only FL data were highly susceptible to over-fitting and struggled to generalise compared to models which had been fine-tuned from a trained HAR network. This paradigm was particularly evident in the poor end-to-end model performance for the more difficult multi-class $\mathcal{T}_T$ task (table 7.3)[12]. As a result, transfer learning improved model robustness and the recognition of subgroups of PwMS from HCs compared to DCNN models trained end-to-end on FL, as well as the feature based methods explored in chapter 6 (by between 8%–15% in binary tasks and up to 33% in the multi-class task). Larger performance gains were particularly evident in the improved recognition of PwMSmild from HC.

The enrichment given by TL models were further explored through comparing the latent t-SNE feature structure against a DCNN (end-to-end) model in figure 7.6. Visually, it appeared that the DCNN (UCI HAR $\rightarrow$ FL) features stratified across the three subject groups, where PwMSmod were distinctly separable from HC and milder subjects in each fold. For DCNN (end-to-end) features however, while PwMSmod tended to cluster in some folds, PwMSmild could not be distinguished from HC.

The improved DCNN model performance through TL could be attributed to a number of rationale. The obvious added benefit of TL is the ability to train a network across more samples overall, through initialising a model's weights with a larger number of samples, and fine-tuning a portion of the network on a separate set of large samples. The TL framework improves generalisation as well as providing regularisation properties by freezing layers; reducing trainable parameters mitigates

[12]The complementary multi-class confusion matrices are given in appendix figure B.6

against over-fitting. More interestingly, both UCI HAR and WISDM comprised of young healthy individuals. In FL there were only n=24 healthy participants (only 16 of which contributed more than 10 unique 2MWT tests, see previous figure 4.9b). As such, initially training on more healthy examples in particular may have allowed the DCNN to more accurately learn a representation of "healthy" walking from inertial sensor data. Models that were initially trained based on $\mathcal{D}_S$ UCI HAR versus WISDM performed similarly for all target classification tasks $\mathcal{T}_T$. Transferring from WISDM tended to marginally perform better at distinguishing PwMSmod however, whereas PwMSmild were better identified from HC when transferring from UCI HAR. Perhaps the activity patterns within each dataset uniquely aided each task. For example, WISDM explicitly learned a unique "jogging" class, which could allow the better representation learning of faster versus slower gait. Moderately disabled PwMS in particular are known to have relatively slower cadence [61], which was also observed within the feature representations (figure 6.8) and previously reported in section 6.4.1. The recognition of walking from various HAR activities was more accurate in the UCI HAR datset (95.2%) than WISDM (54.0%), also perhaps indicating that a more detailed and accurate representation of gait was comparatively learned in $\mathcal{D}_S$ UCI HAR, which subsequently aided the better distinguishment of HC from PwMSmild with subtle gait dysfunction. Other characteristics must also be considered, such as affixing of the phone to the waist in UCI HAR (similar to FL running belt) versus the pocket in WISDM. Regardless, more work is certainly needed to uncover the performance gain and understand explicit reasons for the improvement of TL models applied to remote sensor data. Particularly, further studies will be needed to fully define the attributes of a source domain $\mathcal{D}_S$ and task $\mathcal{T}_S$ which are relevant for the target domain $\mathcal{D}_T$ and tasks $\mathcal{T}_T$, or to determine the optimal $\mathcal{D}_S$ (or combination thereof) among multiple $\mathcal{D}_S$ candidates.

7.5.1 Limitations and Considerations

There are a few limitations to the work presented in this chapter. For example, while TL helped overcome difficulties encountered when attempting to build predictive

models in the presence of low subject numbers and heterogeneous data, the fine-tuning and evaluation of DCNN performance could still be predisposed by these exact limitations. For instance, multiple repeated measurements and differences in the number of unique tests contributed per subject, or even demographic biases, such as the male-to-female ratio mismatches between HC and PwMS, the inclusion of various different MS phenotypes, or that the mild versus moderate sub-groups were bluntly created based on clinician-subjective EDSS scores, are factors that should all be considered.

Furthermore, the DCNN architecture investigated in this study was relatively simple compared to other frameworks [188, 190, 321, 343], future work will investigate the use recurrent layers (such as in LSTMs), which have proven beneficial to characterise the temporal nature of gait recorded within the sensor signal [182, 321]. Later discussions in future work section 9.2.3 will expand on these considerations.

Following the results presented in chapter 6, the work within this chapter only focused on smartphone-based gait assessment using deep networks. While smartphone feature-based classification systematically outperformed smartwatch devices, a more accurate and data-driven representation may also be learned by smartwatch-based deep networks. This could enrich the data captured from smartwatch assessments; DCNNs could improve smartwatch classification alone or the aggregation of information from of the two devices simultaneously.

7.6 Conclusion

The work introduced in this chapter aimed to explore the ability of deep networks to detect impairment in PwMS using remote smartphone inertial sensor data. The novel findings within this chapter can be summarised as:

- Deep networks, specifically DCNNs, investigated in this study can be applied directly to the raw smartphone inertial sensor data. This avoided the need to pre-engineer features and yielded comparable stratification of HC and PwMS subgroups to that of feature-based approaches;

- Remote health data, like FL, can often be too sparse and heterogeneous, due in part to low study participant numbers, to build reliable and generalisable deep models. To overcome this issue, it was demonstrated how common gait-related characteristics can be learned from smartphone-based open-source datasets first using transfer learning;

- Transfer learned DCNN models appeared to characterise a better representation of gait function compared to feature based approaches and DCNNs trained end-to-end of FL, improving the recognition of HC from PwMS subgroups by between 8%–33%;

- Further work is needed to understand the underlying feature structure, along with the optimal source datasets and methods to extract the most appropriate information available to characterise HCs and subgroups of PwMS from remote smartphone inertial sensor data.

Despite the advances deep learning provides to tackling many healthcare related tasks, or specifically the improvements beyond traditional feature-based approaches, deep networks have been criticised as "black-box" algorithms [344]. Deeper models can be highly non-linear and as such there exists an inherent difficulty in interpreting the decisions or patterns that influence a prediction. This lack of transparency remains one of the largest stumbling blocks to the wider acceptance of deep learning for clinical applications [345]. The next chapter will introduce a technique to visualise a DCNNs decisions using relevance heatmaps. By attributing relevance scores to the input inertial sensor data and augmenting this assessment with traditional time-frequency visual representations, this work attempts to visually interpret how disease-related ambulatory activity can be captured by smartphone inertial sensor data in FLOODLIGHT.

8

Developing Interpretable Smartphone-based Sensor Models for Remote Subject Monitoring

Associated publication(s):
Creagh A.P., Lipsmeier F., Lindemann M. and de Vos, M. (2021). "Interpretable Deep Learning for the Remote Characterisation of Ambulation in Multiple Sclerosis using Smartphones". *Nature Scientific Reports*, 11(1), pp.1-14. https://doi.org/10.1038/s41598-021-92776-x.

Creagh A.P., Lipsmeier F., Lindemann M. and de Vos, M. (2022). "Longitudinal Remote Characterisation of Ambulation in Multiple Sclerosis using Smartphones". *IEEE open journal of engineering in medicine and biology*, 3, pp. 202-210. https://doi.org/10.1109/OJEMB.2022.3221306

Contents

8.1 Introduction 163
8.2 Background 165
8.3 Methods 169
8.4 Results 173
8.5 Discussion 186
8.6 Conclusion 195

8.1 Introduction

One of the most significant challenges to the wider acceptance of clinical machine learning, and specifically deep-learning, faces is the inherent difficulty in interpreting the decisions that lead to a prediction [345]. Deep networks can be highly complex and the patterns that these models learn can be so abstract that they are difficult to understand (hence so-called distasteful term "black-box" algorithms [344]). Clear communication between engineers and clinicians will therefore play a vital role in developing the understanding of AI and will help to trust by opening these "black-boxes" [344, 346, 347].

Incorporating expert clinical knowledge through interpretation of network decisions could greatly develop clinicians' fundamental understanding of diseases [346]. For example, breakthroughs in visualising deep networks have been applied to interpreting relevant parts of the brain responsible for the predictions of MS and Alzheimer's disease using Magnetic Resonance Imaging (MRI) [194, 348]. This improved domain knowledge could even be used to reverse engineer features, return to train better models or to develop more sensitive and specific measurements, such as the *digital biomarker* assessments introduced in this work. The clinical utility of explaining deep networks' decisions are illustrated in figure 8.1.

The work presented in this thesis thus far has examined cross-sectional or aggregate disease-related sensor outcomes. As introduced in chapters 6 and 7, considering the time-varying nature of multiple sclerosis (MS) symptoms, the real value of remote monitoring PwMS may ultimately lie in investigating test performance as a function of time. The success of remote subject monitoring in people with multiple sclerosis (PwMS) will be build upon robust, reliable and interpretable models that can accurately represent the longitudinal characteristics of healthy and PwMS symptoms.

The work introduced in this chapter will apply state-of-the-art techniques to visualise and interpret the fully-connected deep convolutional neural network (DCNN) models developed as part of chapter 7. Specifically, the patterns of sensor signal (and therefore the features) that are relevant for distinguishing healthy and

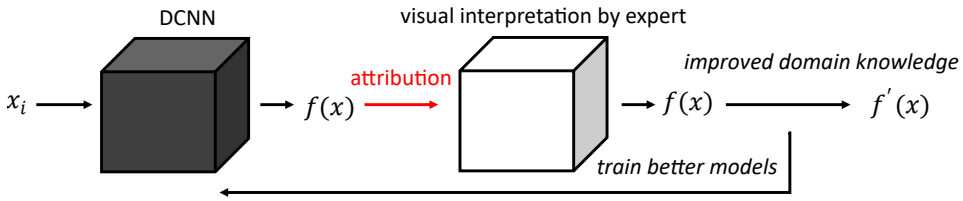


Figure 8.1: Schematic illustrating the clinical utility of interpretable models. "Black-box" algorithms, such as DCNNs, can be interpreted using attribution techniques which aim to map the parts of an input x_i that are responsible for the activations within a network and therefore explain a model's output decision(s). This "opening" of a network's inner workings can allow visual interpretation from clinical experts, expand clinicians' fundamental understanding of these diseases, or to develop more sensitive and specific *digital biomarker* measurements.

disease related gait function will be investigated. Finally, it will be demonstrated how severity representations can characterise the longitudinal performance of healthy controls versus participants with varying manifestations of MS severity.

The primary contributions of this chapter to the field of remote monitoring of multiple sclerosis disease can be summarised as:

- The introduction of a framework to visualise, and therefore understand, the segments of smartphone-based inertial sensor data that influence a model's decision;

- The interpretation of the relevant characteristics of smartphone-based inertial sensor data that are related to healthy gait versus those characteristics which appear to be disease-influenced;

- The development of a continuous, digital-based, estimate of disease severity generated from DCNN predictions, that is capable of stratifying healthy controls (HC) versus PwMS sub-groups;

- The characterisation of the longitudinal performance of typical healthy versus mild and moderate PwMS participants through explainable severity estimate outcomes, particularly demonstrating a sensitivity to monitoring of PwMS symptom manifestations and relapses.

8.2 Background

8.2.1 Explaining Neural Network Decisions

Various methods have emerged in recent years which attempt to communicate how neural network's have made a decision [344, 349]. These can be broadly classed as: (1) *feature visualisation*, i.e. to visualise the features learned within the layers of a network in order to interpret the characteristics of the patterns that have been learned [339, 350]; and (2) *attribution* (sometimes termed saliency maps), which aim to uncover what part of an input is responsible for a network's activations [350].

A common feature visualisation technique is to map the feature space to a low-dimensional embedding (typically in 2-D) as a means to visualise the feature structure, such as with Principal Component Analysis (PCA) or t-distributed Stochastic Neighbour Embedding (t-SNE) (as introduced previously in section 3.3.1, and demonstrated within figures 3.1e and figure 7.4 respectively). Alternative approaches to feature visualisation have aimed to visualise and interpret the individual filters or Kernels[1] within the hidden layers of deep network, in order to infer the pattern each filter, and thus the feature, is capturing [339, 351].

In recent years, the application of attribution methods to explain neural network's decisions have gained considerable momentum. For instance, attention-based methods of attribution have been shown to improve interpretability in time-series models and natural language processing [352, 353]. Attention enables the model to learn deferentially between relevant parts of a sequence and visualising these attention weights allows users the ability to quantify the interdependence and importance of sections of an input sequence [354, 355]. Attention models have been applied successfully in many clinical (time-series) settings, for example to uncover the polysomnography (PSG) factors that influence sleep scoring [356], the vital sign changes guiding early warning systems for real-time patient monitoring [357], and the electrocardiogram (ECG) morphology's pertinent for classifying atrial fibrillation [358].

[1] See section 3.9.1.

Other popular techniques of attribution involve building visual heatmaps, indicating the relevance or contribution across an input example to explain a network's output [359]. One such field of attribution, known as *sensitivity analysis*, typically refers to gradient-based methods which attempt to back-propagate the partial derivatives of the prediction scores across the network with respect to a small change in regions or sections of the input example [359]. The gradient can therefore be utilised to indicate if a change in a region affects the output score for that class, and hence build a heatmap of relevance, such as with Deconvolutional Networks (deconv) [339] and Guided Back-Propagation (guided backprop) [360] which are extensions of this principle.

Activation maximisation methods alternatively attempt to visualise the activation of some specific layer in a network. Perturbation analysis for instance aims to quantify the importance of sections of the input by visualising the probability of the correct class being chosen as a function of perturbing or masking those sections [339, 361]. Class Activation Maps (CAMs) are another method, where the predicted class score is mapped back to the previous convolutional layer to generate heatmaps of relevant sections within an input [362]. Guided Gradient-weighted CAMs (Grad-CAM) are an extension of this, whereby backpropagation from gradient-based approaches are incorporated into the activation mapping [363]. Guided Gradient-weighted CAMs have been successful in aiding the interpretation of automated sleep scoring from electroencephalogram (EEG), electromyogram (EMG), and electrooculogram (EOG) time-series. In this work, the signal regions that Grad-CAM identified as relevant corresponded to clinically known sleep characteristics, for example, sleep spindles. [364].

8.2.2 Layer-wise Relevance Propagation

Layer-wise relevance propagation (LRP) is a backward propagation technique which has gained considerable notoriety as a method to explain and interpret deep networks beyond many existing techniques [256–258, 365]. Backward propagation methods (e.g. LRP, deconv, guided backprop etc.) have been shown to to scale

better for more complex deep neural network (DNN) models than gradient-based methods [366]. These techniques also allow the characterisation and contribution of multiple components of an input by filtering the activations of certain neurons or feature maps. In comparison with non-conserving methods such as guided backprop, LRP also allows the visualisation of parts of the input that contradict the prediction of a class by considering negative activations within a network [256, 257]. Furthermore, LRP has been shown to generate quantitatively more informative heatmaps through decomposing the classification score $f(\boldsymbol{x})$, rather than the gradient (such as sensitivity analysis introduced in section 8.2.1 above), and by conserving the relevance layer-by-layer throughout the back-propagation [367]. Montavon *et al.* [257] and Samek *et al.* [366, 367] provide an insightful and comprehensive comparison between some popular visualisation techniques for deep networks, which the reader is directed towards.

Layer-wise relevance propagation itself has been found clinical utility in interpreting relevant parts of the brain responsible for the predictions of Alzheimer's disease (AD) [348] and multiple sclerosis using MRI images [194]. Attribution through LRP has also been successfully applied to clinical time series data such as EEG trial classification for brain–computer interfacing [368] and, crucially, at identifying gait patterns unique to individual subjects [183]. The latter study, by Horst *et al.*, implemented LRP to attribute segments of a DNNs predictions back to the input signal, which were collected from ground reaction force plates and motion-capture full-body joint angles. This first work reliably demonstrated how LRP could characterise temporal gait patterns, and explained the nuances of particular gait characteristics that distinguished between individual participants in detail (this is essentially a subject identification task, as previously introduced in section 7.2.1.2). For example, LRP identified how one individual had a greater extension of an ankle during a specific phase of the gait cycle. While this work demonstrated how gait patterns can be uniquely associated to an individual, and established this through LRP explanations, an extension of this rationale is that there are gait patterns characteristic of a disease, or diseased sub-population. As such, LRP

decomposition of DCNN predictions may aid the interpretation of MS-related gait impairment from healthy ambulation in this thesis.

Incorporating expert clinical knowledge through visualisations generated from LRP could expand a clinicians' fundamental understanding of ambulatory impairment in PwMS. Since interpretability is considered to be a core component of clinical utility, the subsequent work presented in this chapter proposes to utilise LRP to understand the segments of smartphone-based inertial sensor data that: (1) explain the identification of various human activities in a Human Activity Recognition (HAR) task, specifically to explain the characteristics indicative of healthy walking[2]; and (2) the interpretation of the gait patterns which are relevant for classifying MS diseased versus healthy gait in the FLOODLIGHT PoC study (FL).

8.2.3 Towards Longitudinal Remote Subject Monitoring

Chapter 2 previously established MS as a heterogeneous and mutable disease where subjects may experience symptoms in various domains which hard thresholds on infrequently administered clinical scales may fail to capture. As MS symptoms fluctuate periodically, the real value of remote monitoring PwMS may ultimately lie in investigating test performance as a function of time. Key advantages of objective assessments, like those explored in the FL study, are that they can be administered at high-frequency and longitudinally in out-of clinic environments. As introduced before in sections 2.3.2 and 2.3.6, remote digital outcomes may better capture a patient's habitual disease state or add additional data-driven metrics for clinicians to use when patients return to the clinic. Previously, figure 2.4 conceptualised the proposed utility of digital sensor-based outcomes to enable clinicians to continuously monitor the disease course of PwMS. Remote, continuous, and granular assessments could capture subtle manifestations of habitual MS, as well as the disease course and relapse events that are currently missed by infrequent and less sensitive in-clinic measures. As such, this chapter will aim to develop a severity estimate outcome constructed from the DCNNs introduced in chapter 7. The ambulatory-based

[2]HAR is previously introduced in chapter section 7.2.1.1.

DCNN digital outcome measures will be further appraised in the ability to stratify HC versus PwMS severity longitudinally, and to explore the patterns of subject performance over the repeated test observations they each contributed in this study.

8.3 Methods

8.3.1 Developing Interpretable Smartphone Sensor Models

The work presented in this chapter aimed to decode smartphone sensor data recorded from UCI HAR and FL using attribution-based layer-wise relevance propagation (see section 3.10.1 for more details on the technical framework outlining LRP). In this chapter, trained DCNN models, which are previously described as part of chapter 7, were analysed using LRP to attribute relevance values, R_i, to inertial sensor samples. As schematically shown previously in figure 3.6, LRP decomposes the output $f(\boldsymbol{x})$ of a learned function f, given an input $\boldsymbol{x}$, where relevance values R_i are attributed to individual input samples x_i. In this case, x_i are represented by discrete sensor samples within an testing epoch and therefore R_i is directly embedded in the time domain. The contribution of LRP can also be quantified across the input channels (in this case the sensor axis). Resulting heatmaps can then be produced, where LRP values indicates which parts of the sensor signal were *relevant* for the prediction of that class and the contribution of the individual axis to $f(\boldsymbol{x})$. For more information on the implementation LRP[3] the reader is directed to [256–258, 365]. In this study, individual LRP heatmaps were produced for every epoch per subject in the out-of-sample testing set using the iNNvestigate implementation of LRP [369]

The explanation of DCNN decisions towards individual sensor-epochs was supplemented through the integration of CWT analysis, previously introduced as part of section 6.2.2.3 and within figure 6.3. The top view of the CWT scalogram, which is the absolute value of the CWT as a function of time and frequency, further provides useful insight into the instantaneous frequency composition of the sensor signal. It is therefore proposed that this LRP-CWT framework allows for

[3]Both Keras and PyTorch implementations of the LRP algorithm have been developed and are found at http://heatmapping.org/

the interpretation (at least to some extent) of the DCNN classification decisions within individual out-of-sample sensor epochs. This framework also provides an extra lens to interpret (and validate) the insights gained from DCNN predictions using the aggregation of visualisation from LRP, the raw sensor data and its time-frequency representation, to compare against already existing clinical domain knowledge in PwMS.

8.3.1.1 Constructing Representative Gait Signal Epochs

In order to determine the population-wise gait characteristics pertinent of healthy versus mild and moderate MS disease, an average representative epoch was generated for each subject-group using Dynamic Time Warping Barycenter Averaging (DBA) [370]. First, Dynamic Time Warping (DTW) is a method to measure the similarity (distance) between two sequences in cases where the order of elements in the sequences must be considered [371]. DTW can be used to align the signals such that the similarity between their points is minimised, hence generating a "warped" optimal alignment between sequences. For instance, gait cycle templates have previously been generated for PwMS using DTW [164]. DTW Barycenter Averaging is a global averaging method for an arbitrary set of DTW sequences, which can be used to create a macro-average sequence, in this case, a representative gait epoch. An average gait cycle epoch for each subject-group was constructed by applying DBA to a random selection of correctly classified epochs (n≈2000) with a high posterior probability of that class (Pr.>0.85), with no more than (k< 40) epochs (i.e. $< 50\%$) contributed from a single 2MWT. A previously trained DCNN model was then applied to each representative epoch and relevance scores were attributed using LRP[4].

8.3.2 Estimating Disease Severity

Individual binary models explored in section 7.4.3 can be aggregated into a unified estimate of disease severity. Data from HC, PwMSmild and PwMSmod were

[4]An implementation of DBA can be found at `https://github.com/fpetitjean/DBA/`

combined and models were trained to minimise a multi-class categorical cross-entropy loss function (3.76) for $k \in \{hc, mild, mod\}$. A continuous estimate of severity can then be captured by taking an average of all epoch predictions, (3.73), over a test for a given assessment day, d such that:

$$\hat{y}_d = \frac{1}{N} \sum_{n=1}^{N} \arg\max_k (p(\hat{y}_k = k | \mathbf{x}_n)) \tag{8.1}$$

where N are the number of windowed epochs for a given test date, d, and k lies in an ordinal range of $[0, 1, ...K]$. Therefore $\hat{y}_d$ will be continuous such that $0 \leq \hat{y}_d \leq K$ and can conceptualised as a naïve disease severity score, mapping a severity range from healthy to mild to moderate.

8.3.3 Longitudinal Trend Monitoring of Remote Smartphone-Based Outcomes

Longitudinal trends of specific participants were examined as a time-series by considering the severity scores $\hat{y}$ of repeated 2MWTs over all their available data for the duration of the FL study. Participants in FL were requested to perform a remote Two-Minute Walk Test (2MWT) daily over a 24-week period and were assessed clinically during site-visits at baseline, week 12 and week 24. As discussed in section 6.4.3, some test-dates may be missing, where various participants had differing adherence rates during the study. There are many strategies to overcome sparsity in clinical data [372, 373]; some methods aim to impute feature values and raw data [374], while other techniques aim to account for missingness within the construction of a model itself [190]. As the goal of this work was to perform longitudinal analysis of subject's severity, namely the severity trends over time, missing 2MWT outcomes were imputed by considering $\hat{y}$ as a time-series. Piecewise Linear Interpolation (PLI) (see appendix A.2.4), was used in this study to impute missing test severity observations on a given date. Other simple and efficient interpolation methods also exist such as linear or previous neighbour interpolation, as well as cubic or spline interpolation [372, 373].

A simple trend estimation can be applied to the time sequence of severity scores ($\hat{y}$) across days (d) using a $d-$ centred linear moving average filter (MAF).

$$\hat{z}[i] = \frac{1}{2N+1} \sum_{N}^{j-N} \hat{y}[i+j] \tag{8.2}$$

where $\hat{y}[\cdot]$ is the input sequence (severity scores) and $\hat{z}[\cdot]$ is the output (filtered) sequence (moving severity score estimate) for each d^{th} day; $2N+1$ defines the order of the filter, in this case the number of days d used in the moving average. A 7-day window was implemented in order to capture the trends in $\hat{y}_d$ of various representative participants: a stable HC, a stable PwMSmod and two PwMS subjects that both reported relapses using the FLOODLIGHT application on their smartphone during the study. Finally, the population-wise longitudinal trajectory for HC and each PwMS sub-group was constructed based on the average trend per subject.

8.4 Results

8.4.1 Visualising Remote Sensor Data

As a initial step to implement LRP-based visualisations from remotely captured smartphone sensor data, a DCNN classification model that separated HC from PwMSmod (i.e. $f(\boldsymbol{x}) \in \{-1, 1\}$, where HC $\in \{-1\}$ and PwMSmod $\in \{+1\}$) was investigated in figure 8.2.

Relevance values, R_i, were evaluated independently with respect to probabilistic output per each individual epoch, however it is possible to recover the original signal segment, $\boldsymbol{x}$, from overlapped windowed samples (epochs) to any length, where LRP values in adjacent epochs can be inversely windowed and reconstructed as seen in figure 8.2.

The larger panel plot of figures 8.2 (d-f) represents one epoch within the signal from (a), which corresponds to the shaded grey area highlighted in figure 8.2 (c). Applying the $\alpha_2\beta_1$-rule[5] indicated that positive and negative R_i represented the presence of factors which influenced $f(\boldsymbol{x}) > 0$ (PwMSmod) or inversely as $f(\boldsymbol{x}) < 0$ (HC). As LRP was evaluated independently for each epoch with respect to $f(\boldsymbol{x})$, yet we are aggregating LRP across multiple epochs, to benefit interpretation of LRP examples, the heatmap colours were standardised such that hot hues represent to presence of factors which influence $f(\boldsymbol{x}) > 0$ (PwMSmod), whereas cold hues (inversely) contradict the prediction $f(\boldsymbol{x}) < 0$ (HC), as overlaid upon the accelerometer trace in figures 8.2. Black represented ($R_i \approx 0$) inputs which have little or no influence to the model's decision. Figures 8.2 (a), and (e-f) demonstrate the LRP relevance values overlaid upon the respective (a) and (e) accelerometer trace, as well as the LRP contribution across each channel (1-4, i.e. $\boldsymbol{a}_x, \boldsymbol{a}_y, \boldsymbol{a}_z, \|\boldsymbol{a}\|$) in figure 8.2 (f). A model's final class prediction was derived from maximising the posterior probabilistic output in the last fully connected layer. This output is depicted in 8.2 (c) as the posterior probability of PwMSmod (where the posterior

[5]See section 3.10.1 for more details.

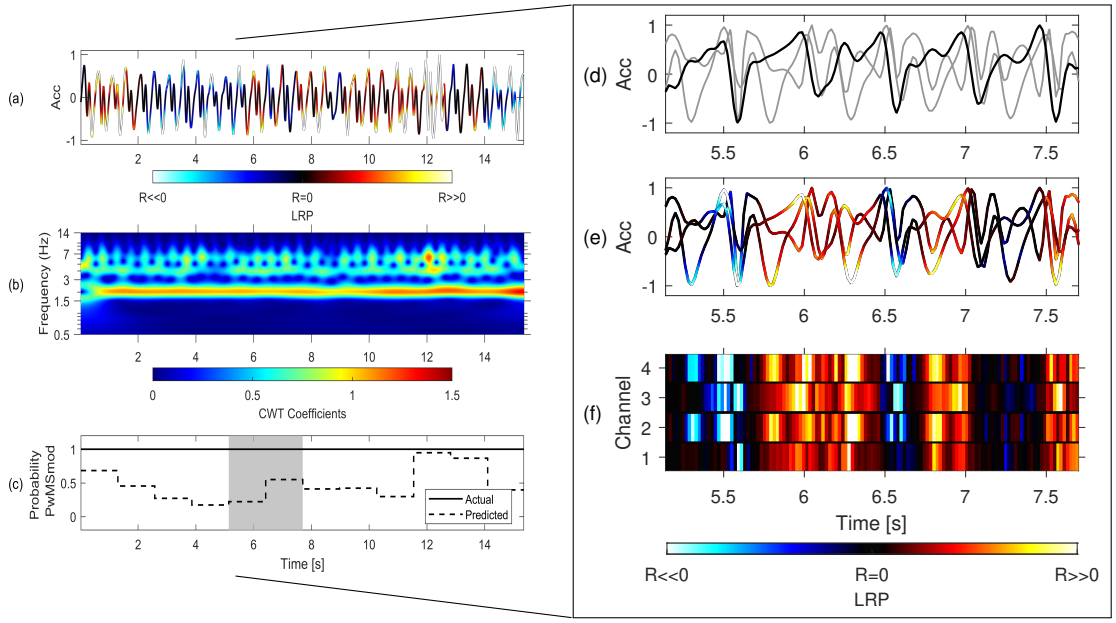


Figure 8.2: Panel plot illustrating the DCNN MS gait-signal explanations using layer-wise relevance propagation (LRP).
Panel **(a)** depicts 15.4 [s] (or 12 epochs with 50% overlap) of the normalised $||\mathbf{a}||$ accelerometer trace ($m \cdot s^{-2}$), from a PwMSmod subject during a 2MWT. The LRP value across overlapping epochs have been superimposed upon the accelerometer trace, which demonstrates the decomposition of the network's output score onto the individual gait samples. Red and hot colours identify input segments denoting positive relevance ($R_i > 0$) indicating $f(\boldsymbol{x}) > 0$ (i.e. PwMSmod). Blue and cold hues are negative relevance values ($R_i < 0$) indicating $f(\boldsymbol{x}) < 0$ (i.e. HC), while black represents ($R_i \approx 0$) inputs which have little or no influence to the DCNNs decision. The second row **(b)** depicts the top view of the CWT scalogram, which is the absolute value of the CWT as a function of time and frequency. Panel **(c)** depicts the probabilistic output hypnogram for a $\mathcal{T}_T$ HC vs. PwMSmod disease classification. The shaded grey area represents an example epoch (n=128 samples, or 2.56 [s]), which is examined in the larger subplot (d-f). Panel **(d)** represents the 3-axis accelerometer trace within that epoch, where the magnitude $||a||$ of the 3-axis signal is highlighted in bold. The LRP values attributed to each axis are superimposed upon the signal in **(e)**, where the individual LRP values per channel (1-4, i.e. $\boldsymbol{a}_x, \boldsymbol{a}_y, \boldsymbol{a}_z, \|\boldsymbol{a}\|$)) are further highlighted in **(f)**[6]. (PwMSmod: EDSS: 3.7 ± 1.0; T25FW: 6.2 ± 0.7)

probability of HC is reflected by the inverse of posterior probability of PwMSmod) and allows the direct comparison of $f(\boldsymbol{x})$ to LRP values in 8.2 (a) (e-f).

Following the same pipleline as introduced in section 6.2.2.3 and figures 6.3, the CWT scalogram representation of instantaneous frequency within the sensor signal are depicted in figures 8.2 (b) and (d).

The $\alpha_2\beta_1$-rule emphasises the weights of positive contributions relative to

[6]Note that the LRP values have been scaled in (a) and (e, f) individually to maximise contrast for visual purposes, and the epoch visualised in (d-f) is the rightmost epoch within the 5.1–7.7 [s] range, where LRP values are computed on epochs independently.

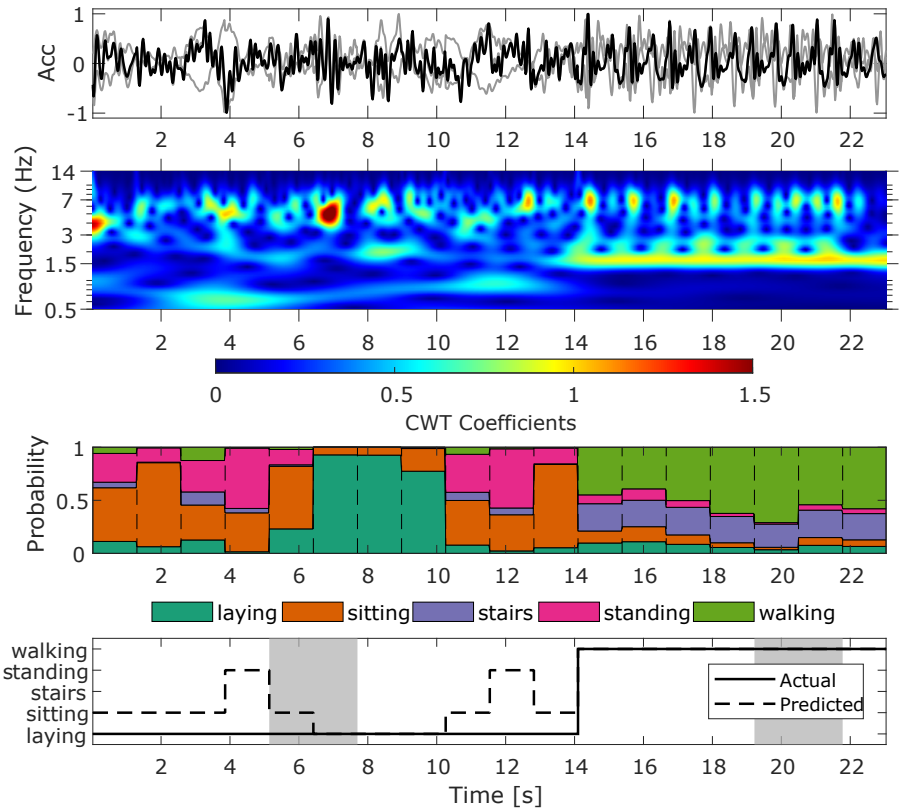
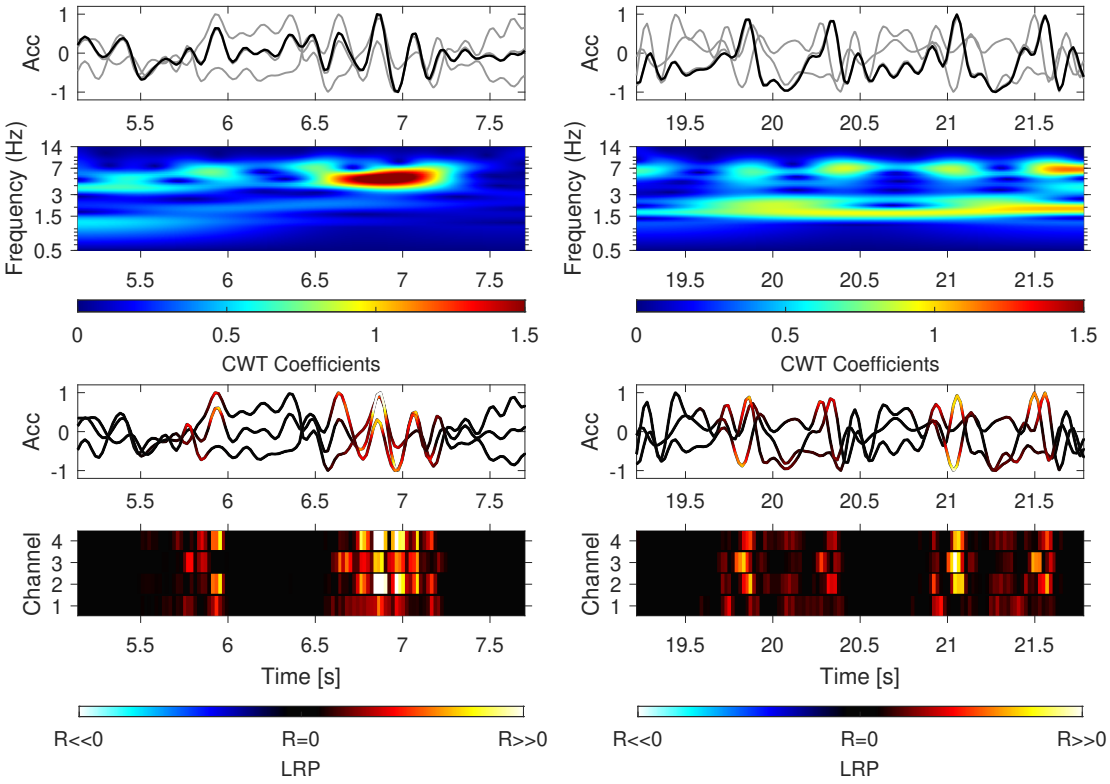


Figure 8.3: Panel plot illustrating example performance of UCI HAR DCNN model.. The top row represents a 3-axis accelerometer trace captured from a smartphone over 24 seconds, which corresponds to 18 epochs of length 128 samples (or 2.56 [s]) with a 50% overlap. The magnitude of the 3-axis signal is highlighted in bold. The second row depicts the top view of the CWT scalogram. The (average) HAR posterior probability for each class $\mathcal{Y}_S$ across overlapped epochs is displayed in the third row. The corresponding output hypnogram for for each class compared to the ground-truth is overlaid in the final row. Shaded grey areas represent individual epochs which are further examined in figure 8.4.

inhibitory contributions that predict $f(\boldsymbol{x})$, while the $\alpha_1\beta_0$-rule allows only positive contributions to be shown in the heatmap. In order to initially generate relatively more simple and explainable LRP heatmaps, the remaining LRP heatmaps introduced in this chapter utilised $\alpha_1\beta_0$-rule to focus on the morphology and characteristics of a sensor signal influencing $f(\boldsymbol{x}) > 0$ with respect to that prediction.

8.4.2 Interpreting HAR decisions

In this section the LRP framework introduced in section 8.4.1 was applied to HAR domain DCNN models, from the UCI HAR database, which have previously been evaluated in section 7.2.1.1. Figure 8.3 details an example 24 second trace including movement and static predicted classes. The a 3-axis accelerometer trace,



(a) "laying" **(b)** "walking"

Figure 8.4: Illustrative examples of static versus movement predicted HAR epochs. **(a)** Represents an example epoch (n=128 samples, or 2.56 [s]) corresponding to the acceleration trace from 5.1 - 7.7 [s] within figure 8.3. **(b)** Represents an example epoch (n=128 samples, or 2.56 [s]) corresponding to the acceleration trace from 19.2 - 21.8 [s] within figure 8.3.
In both **(a)** and **(b)**, the top rows represents the normalised 3-axis acceleration signal, where the magnitude of acceleration signals is highlighted in bold. The second rows depict the top view of the CWT scalogram. The panels underneath represent the decomposition of input relevance values (R_i) using Layer-Wise Relevance Propagation (LRP). The third rows represent the LRP values overlaid upon the accelerometer signal, where the bottom panel represents the LRP activations per channel (i.e. $\boldsymbol{a}_x, \boldsymbol{a}_y, \boldsymbol{a}_z, \|\boldsymbol{a}\|$).

and magnitude $\|\boldsymbol{a}\|$ in bold, are depicted along with the analogous CWT time-frequency representation. The (average) HAR posterior probability for each class $\mathcal{Y}_S$ across overlapped epochs, as well as the corresponding output hypnogram are represented by the 3^{rd} and 4^{th} panel row. Shaded grey areas represent individual static, "laying", and movement, "walking", epochs which are further examined in figure 8.4 through LRP decomposition. Figure 8.5 presents dynamic movement examples of "stairs" compared to "walking".

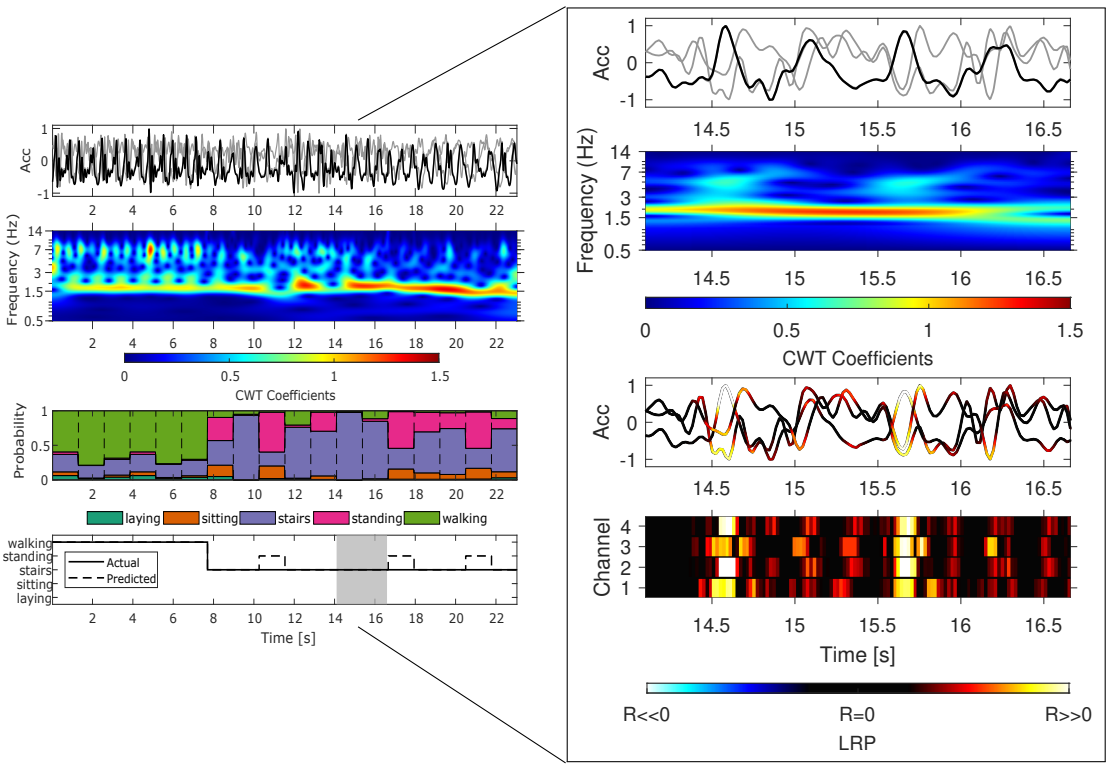


Figure 8.5: Panel plot illustrating movement activity predicted from a UCI HAR DCNN model.
The top row represents a 3-axis accelerometer trace captured from a smartphone over 24 seconds, which corresponds to 18 epochs of length 128 samples (or 2.56 [s]) with a 50% overlap. The magnitude of the 3-axis signal is highlighted in bold. The second row depicts the top view of the CWT scalogram. The (average) HAR posterior probability for each class $\mathcal{Y}_S$ across overlapped epochs is displayed in the third row. The corresponding output hypnogram for for each class compared to the ground-truth is overlaid in the final row. The shaded grey area represent an individual epoch (n=128 samples, or 2.56 [s]) within the acceleration trace from 14.1–16.6 [s], which is examined in the larger subplot through the decomposition of DCNN input relevance values (R_i) using Layer-Wise Relevance Propagation (LRP). Red and hot colours identify input segments denoting positive relevance ($R_i > 0$) influencing that HAR prediction. LRP values are overlaid upon the accelerometer signal, where the bottom panel represents the LRP activations per input channel (i.e. $\boldsymbol{a}_x, \boldsymbol{a}_y, \boldsymbol{a}_z, \|\boldsymbol{a}\|$).

8.4.3 Interpreting MS Remote Sensor Data

The evaluation of classification models aimed to separate HC and PwMSmod subjects from remotely collected smartphone-sensor data were augmented with LRP decomposition, along with the integration of CWT analysis previously introduced in chapter 6. As DCNN classification models are first built for HAR classification tasks and then subsequently fine-tuned for disease recognition in FL (see chapter 7 for more details), further examining this smartphone accelerometer data with

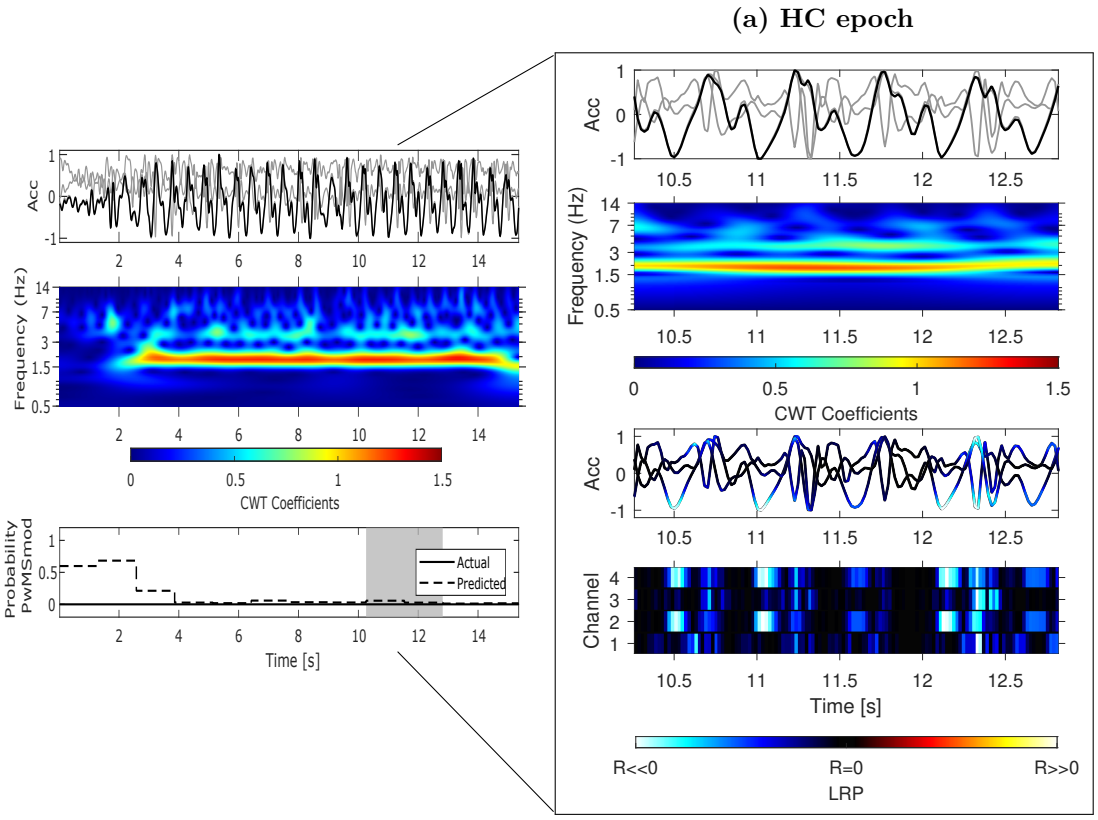


Figure 8.6: (a) Panel plot illustrating example performance of a typical HC subject (true negative) which can be visually interpreted using LRP decomposition and CWT frequency analysis. (HC, T25FW: 4.8 ± 0.35)
The top row represents a 3-axis accelerometer trace captured from a smartphone over 15.4 seconds, which corresponds to 12 epochs of length 128 samples (or 2.56 [s]) with a 50% overlap. The magnitude ($\|\boldsymbol{a}\|$) of the 3-axis signal is highlighted in bold. The second row depicts the top view of the CWT scalogram, which is the absolute value of the CWT as a function of time and frequency. The final row depicts the output hypnogram for a HC vs. PwMSmod disease classification model ($\mathcal{T}_T$). The shaded grey area represents an example epoch (n=128 samples, or 2.56 [s]) within the acceleration trace, which is examined in the larger subplot through the decomposition of DCNN input relevance values (R_i) using Layer-Wise Relevance Propagation (LRP). Red and hot colours identify input segments denoting positive relevance ($R_i > 0$) indicating $f(\boldsymbol{x}) > 0$ (i.e. PwMSmod). Blue and cold hues are negative relevance values ($R_i < 0$) indicating $f(\boldsymbol{x}) < 0$ (i.e. HC), while black represents ($R_i \approx 0$) inputs which have little or no influence to the DCNNs decision. LRP values are overlaid upon the accelerometer signal, where the bottom panel represents the LRP activations per input channel (i.e. $\boldsymbol{a}_x, \boldsymbol{a}_y, \boldsymbol{a}_z, \|\boldsymbol{a}\|$).

respect to the original model's HAR output may prove useful to supplement CWT and LRP analysis.

Following the same visualisation pipeline introduced in section 8.4.1, figure 8.2 compared the example patterns and characteristics captured from depicting the raw sensor signal with LRP-CWT analysis for representative HC, PwMSmild and PwMSmod subjects. Figure 8.6 compared these example patterns and characteristics

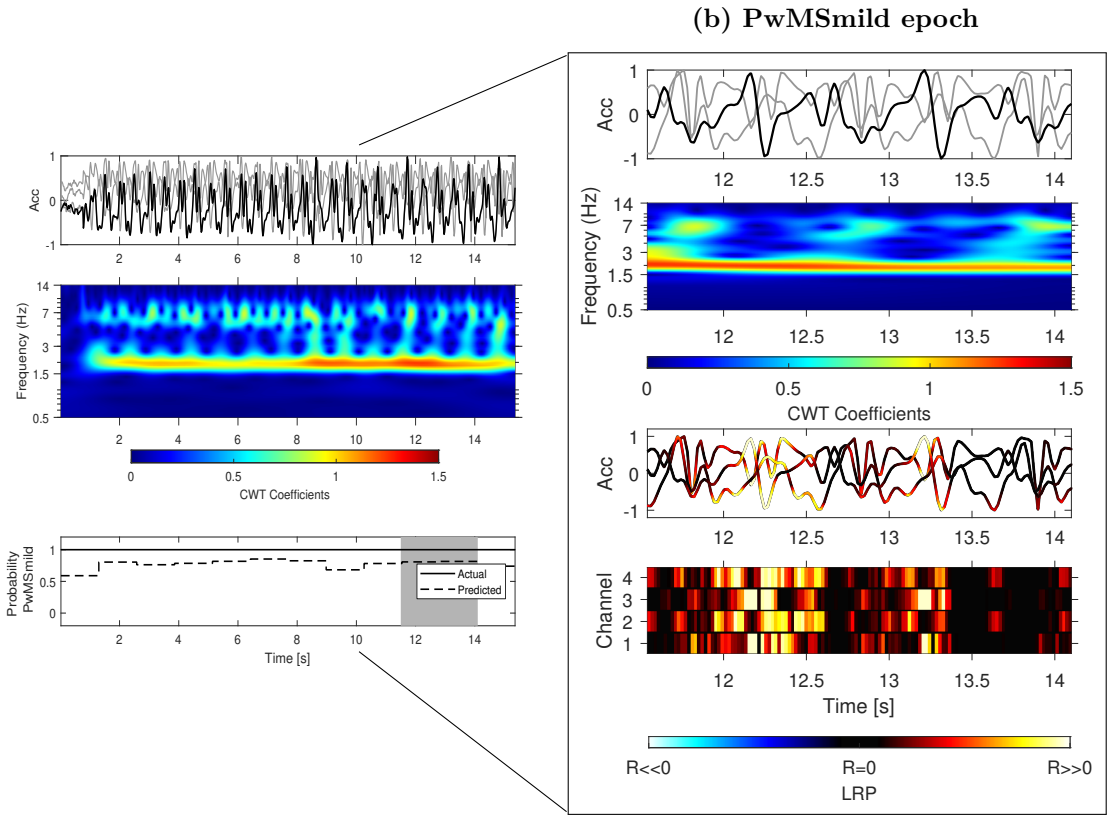


Figure 8.6: Cont.: (b) Panel plot illustrating example performance from a section of correctly classified PwMSmild subject's 2MWT (true positive) which can be visually interpreted using LRP decomposition and CWT frequency analysis. (PwMSmild, EDSS 3.25 ± 0.35; T25FW: 5.5 ± 0.53 [s])

captured from depicting the raw sensor signal, augmented with LRP-CWT analysis for representative examples of correctly classified HC, PwMSmild and PwMSmod subjects respectively. Figure 8.6a first illustrated the performance of a correctly classified HC subject's 2MWT segment, supplemented by the raw sensor data, its CWT representation and the final disease model's probabilistic output $p(\boldsymbol{x})$. In this example, gait signal was apparent from the raw sensor data and supported by strong gait domain energy, E_s, within the CWT representation. This collection of epochs were predicted as predominantly "walking" by a HAR model and corresponded to a confident HC classification with high probability. Focusing on an overlapped epoch example from 10.3–12.8 [s], gait was clearly visible in the time-frequency domain (i.e. large CWT coefficients around 1.5 Hz) and reflected clear steps, as depicted by the magnitude of acceleration. LRP attributed high relevance

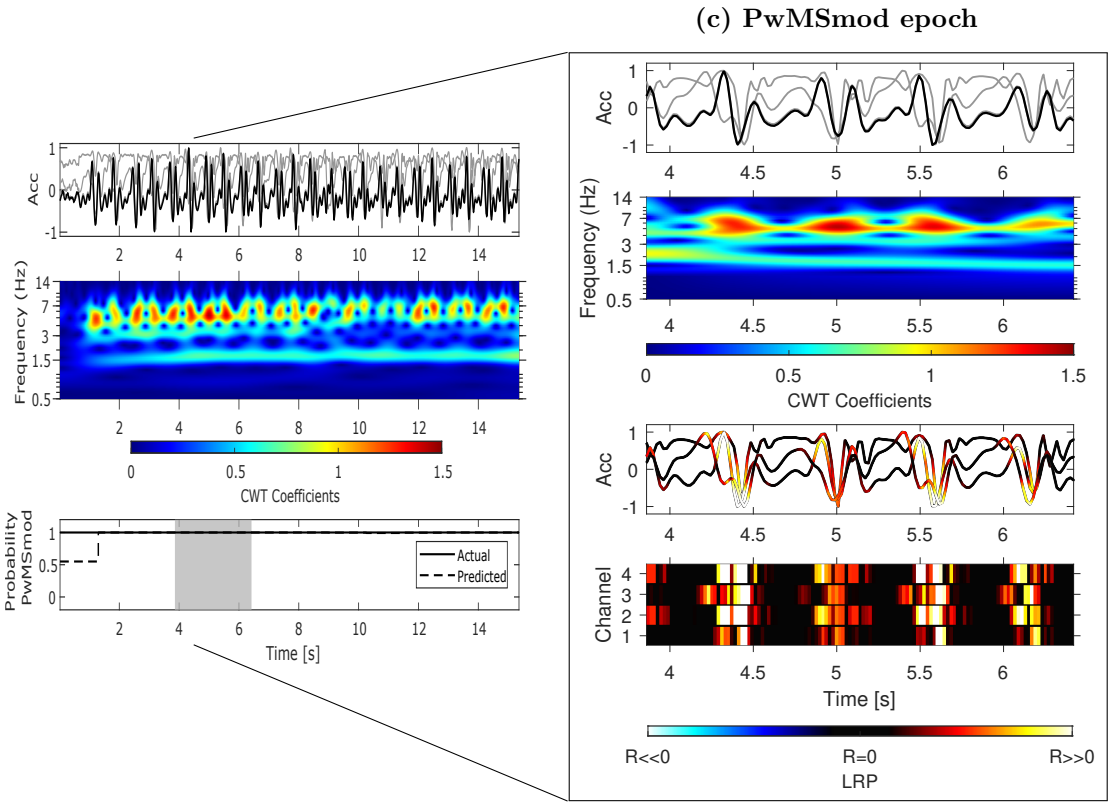
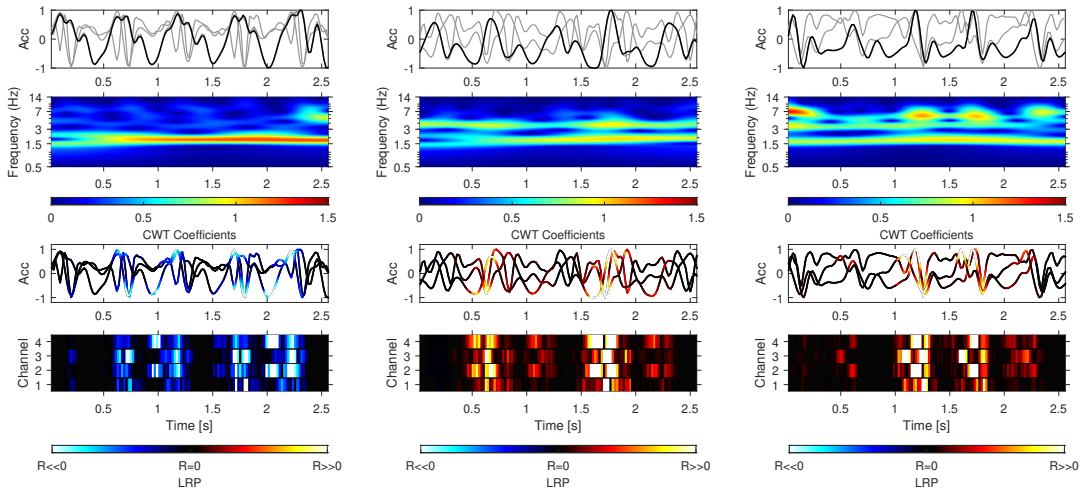


Figure 8.6: Cont.: (c) Panel plot illustrating example performance from a section of a correctly classified PwMSmod subject's 2MWT (true positive) which can be visually interpreted using LRP decomposition and CWT frequency analysis. (PwMSmod, EDSS: 3.8 ± 2.9; T25FW: 6.9 ± 0.5 [s])

scores R_i to these steps in the vertical $\boldsymbol{a}_y$ and orientation invariant signal $||\boldsymbol{a}||$ (i.e. represented by channel 2 and 4).

Figure 8.6b depicted an example 2MWT from a representative correctly classified PwMSmild subject. Similarly to the HC example, gait signal was visible in this example, (i.e. CWT E_s, HAR $f(\boldsymbol{x})$ and clear steps in $||\boldsymbol{a}||$) (Note: HAR posterior probabilities also indicated "walking"). Relevance propagation for an example epoch during 11.5–14 [s] indicated that step occurrences attributed to the "mild" DCNN posterior output. Time-frequency gait signal visualisation through CWT analysis also revealed harmonics occurring at higher frequencies than the gait domain (>3.5 Hz).

Figure 8.6c, in contrast, represented a panel plot illustrating example performance of a typical correctly classified PwMSmod subject's 2MWT segment. In this

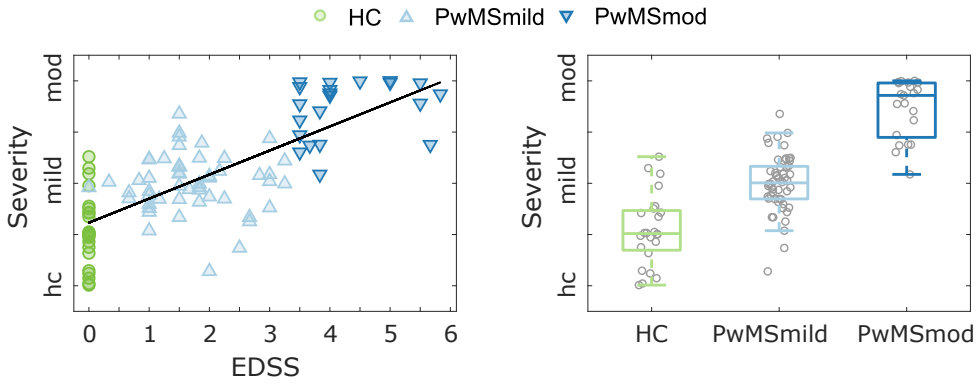


(a) HC (Pr: 0.89) **(b)** PwMSmild (Pr: 0.91) **(c)** PwMSmod (Pr: 0.90)

Figure 8.7: Visualisation of the average gait signal through the LRP-CWT analysis framework. Representative average epochs (n=128 samples, 2.56 [s]) were first created using Dynamic Time Warping Barycenter Averaging (DBA) independently for sets of correctly classified epochs from HC, PwMSmild and PwMSmod subject groups. (Pr; represents the DCNN posterior probability for that class).

example, a concentration of higher frequency E_s disturbances temporally coincided with each step event. These gait-related perturbations were examined in the zoomed sub figure for an example epoch during 3.9–6.4 [s], as highlighted by the shaded shaded grey area during the longer gait example. Relevance decomposition of this epoch attributed all LRP-based relevance R_i to each step and associated high-frequency disturbance (i.e. events influencing $f(\boldsymbol{x}) > 0$ output as PwMSmod). Further examples comparing misclassified HC, PwMSmild and PwMSmod subject epochs are presented in appendix B.7.

Finally, average gait epochs for HC, PwMSmild and PwMSmod groups were created using Dynamic Time Warping Barycenter Averaging. Visualisation of each representative epoch using the LRP-CWT framework was depicted in figure 8.7. The DCNN posterior probabilities for each representative epoch were strongly predictive of the true class (HC, Pr. 0.89; PwMSmild, Pr. 0.91; and PwMSmod, Pr. 0.90).



(a) (b)

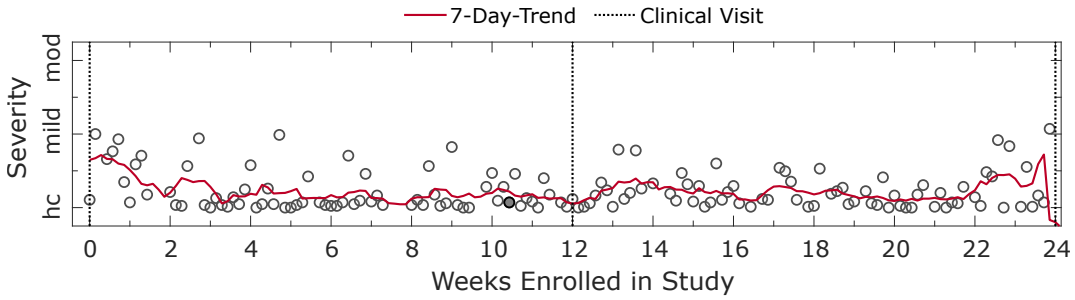
Figure 8.8: The relationship between the continuous disease severity outcome estimate, EDSS and subject group. Figure **(a)** depicts the scatter plot demonstrating the positive correlation ($r : 0.75$; $\rho : 0.71$; $p < 0.001$) between the average severity outcome and EDSS score per subject; whereas **(b)** represents the distribution of the average severity outcomes per each subject group. A DCNN model was constructed based on the average class predictions (HC, PwMSmild, PwMSmod) per subject over all 2MWTs, creating an estimated continuous prediction probability distribution, ranging from healthy to moderate MS. Each point therefore represents the average estimated severity outcome (probability) for that subject. A black line represents the line of best fit between severity and EDSS ($r^2 : 0.56, p < 0.001$).

8.4.4 Digitally Estimated Severity Outcome

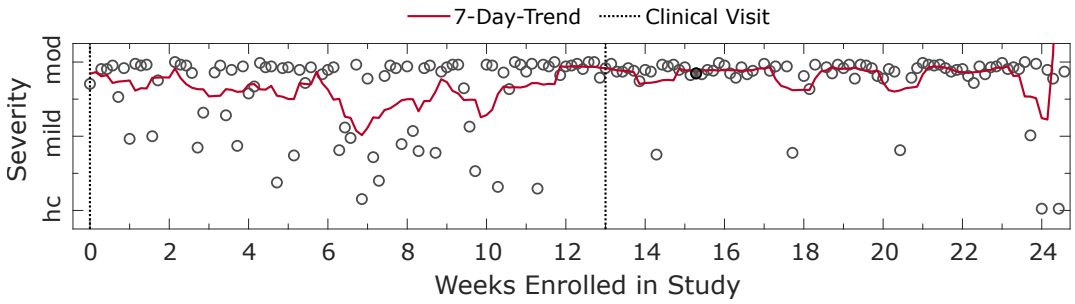
A continuous disease severity outcome was created by averaging class predictions (HC, PwMSmild, PwMSmod) over all 2MWTs within a CV for each subject. Unified disease classification results were presented in table 7.3 as well as the accompanying confusion matrices in figure B.6. A disease severity outcome therefore mapped a posterior probability ranging from healthy, to mild, and to moderate for each subject. The distribution of the average severity per subject was displayed in figure 8.8, and demonstrated the positive relationship between severity outcome and average EDSS, (Pearson's $r : 0.75$; Spearman's $\rho : 0.71$; $p < 0.001$).

8.4.5 Longitudinal Characterisation of Digitally Estimated Severity Outcomes

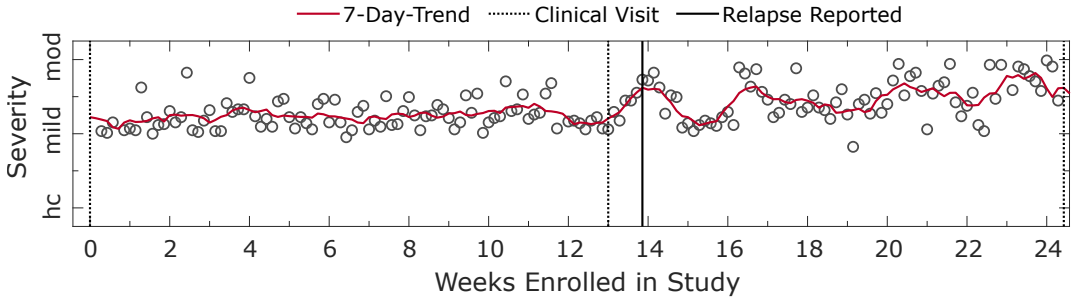
Disease severity outcomes were evaluated for each 2MWT available per subject, allowing longitudinal trends to be monitored by examining severity estimates over the duration of the FL study for individual subjects. Longitudinal examples of the trajectories of daily severity estimates from healthy controls, stable PwMSmod, and



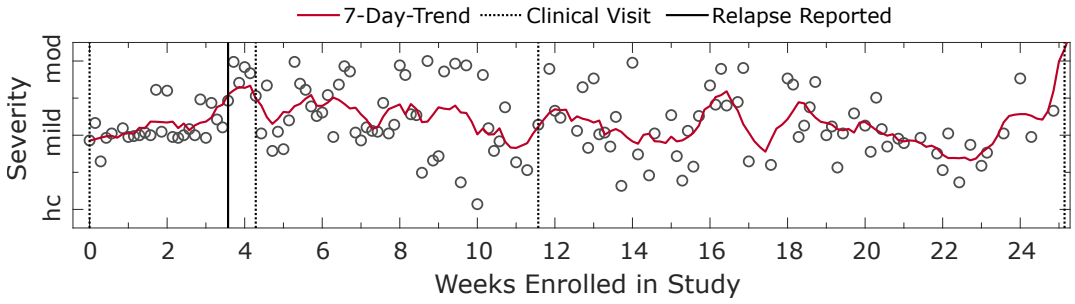
(a) HC Subject



(b) PwMSmod Subject



(c) Relapsing PwMSmod Subject



(d) Relapsing PwMSmild Subject

Figure 8.9: Panel plot illustrating the longitudinal severity estimate outcome for example subjects. Depicted are the disease severity estimate outcome for an example **(a)** HC subject; **(b)** a PwMSmod subject; **(c)** an example of relapsing PwMSmod subject and **(d)** an example of a relapsing PwMSmild subject during the FLOODLIGHT study. Each circle represents the (average) severity outcome estimate for a 2MWT performed on a given date. Shaded red lines depict the and 7-day trend, represented by the d-point centred moving average across days (d). Missing test dates (which are not depicted) were first imputed using piecewise linear interpolation. Dashed black lines represent site-visits where the participant was assessed clinically.

PwMS subjects that experienced relapses were examined in figure 8.9. Individual 2MWT severity outcomes were depicted from the 0^{th} week until study completion in week 24, where dashed black lines represented site-visits where participants were assessed clinically. Red lines depicted the 7-day average trend in severity outcomes.

Figure 8.9a first depicted a HC subject. This participant was examined at baseline (week 0; EDSS 0, T25FW: 5 [s]), midway through the study (week 12; EDSS 0[7]; T25FW: 4.5 [s]) and at the study completion (week 24; EDSS: 0; T25FW: N/A[8] [s]). The test date highlighted in bold for this subject corresponds to the 2MWT previously examined in figures 8.6a. It was observed that this subject was predicted as healthy with a low severity, consistently across the entire study. Many variations in severity outcomes were smoothed out across both 7-day centred window trend.

Figure 8.9b demonstrated a stable PwMSmod subject. This participant was examined at baseline (week 0; EDSS 3.5, T25FW: 5.4 [s]) and midway through the study[9] (week 12; EDSS 4.5; T25FW: 4.9 [s]). The test date highlighted in bold for this subject corresponds to the 2MWT previously examined in figure 8.6c.

Figure 8.9c observed the longitudinal severity outcome trend for a PwMSmod subject who reported a relapse during the FLOODLIGHT study. A black line depicts the date of relapse on-setting during week 13, which was recorded by the participant using the FLOODLIGHT application on their smartphone. Dashed black lines represent site-visits where the participant was assessed clinically. This subject was examined at baseline (week 0; EDSS 3.5, T25FW: 4.9 [s]), midway through the study (week 12; EDSS 3.5; T25FW: 6.6 [s]) and at the study completion (week 24; EDSS: 3.5; T25FW: 10.3 [s]).

Figure 8.9d described the longitudinal severity outcome trend for a PwMSmild subject who reported a relapse during the FLOODLIGHT study. A black line depicts the date of relapse on-setting (Spinal Topography Outbreak) during week 3, which was recorded by the participant using the FLOODLIGHT application on

[7]Note: an EDSS of zero in this case refers to a normal neurological exam, the subject is healthy and has no disability.

[8] N/A denotes scores not assessed at this visit.

[9]Note: clinical assessment scores were not made available for this participant at study completion.

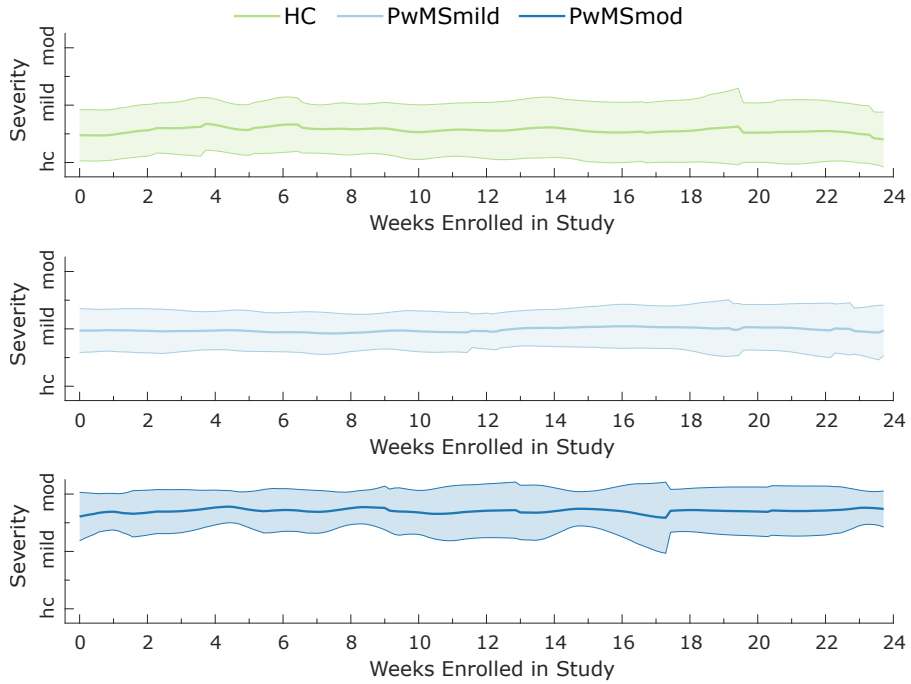


Figure 8.10: The average longitudinal trajectory for HC, PwMSmild and PwMSmod subgroups. Bold lines represent the average trend over all subjects within each subject group, where shaded regions depict the standard deviation around each trend.

their smartphone. Dashed black lines represent site-visits where the participant was assessed clinically. This subject was examined at baseline (week 0; EDSS 1.5, T25FW: 4.9 [s]), week 12 (EDSS 1.5; T25FW: N/A[10]) and at the study completion in week 25 (EDSS: N/A[11]; T25FW: 5.5 [s]). In week 4, 7 days after reporting a relapse, the participant was assessed during an "unscheduled visit" where they exhibited a worsening of MS symptoms, i.e. an increase in EDSS and gait related T25FW (EDSS: 2.5; T25FW, 7.5 [s]). Finally, the population-wise longitudinal trajectories for HC and PwMS groups was presented in figure 8.10, by evaluating the average trend per group over the duration of the study.

[10]See footnote 8
[11]See footnote 8

8.5 Discussion

Recently, breakthroughs in visualising neural networks have paved the way for explanations in deep and complex models, for example heatmaps of "relevant" parts of an input can be built by decomposing the internal neural networks using layer-wise relevance propagation [257, 344, 349]. Visual interpretation of the factors which influence a model's prediction may enable a deeper understanding of how healthy and disease-influenced characteristics are captured from remote smartphone-based inertial sensor data. The work introduced in this chapter aimed to establish a framework to further understand the patterns of healthy and MS disease through multiple viewpoints: visualising the raw inertial sensor signal, its analogous time-frequency CWT representation, as well as augmenting this picture using layer-wise relevance propagation, a deep network visualisation technique. Attribution through LRP has already been successfully applied to visualise gait patterns that are predictive of an individual which were acquired from lab-based ground reaction force plates and infrared camera-based full-body joint angles [183].

In this study, the patterns of "healthy gait" from UCI HAR and within FL, were compared to MS-related gait in FL. Representative examples were initially explained by projecting the LRP-relevance decomposition of a trained DCNN model upon the raw sensor output using heatmaps. This picture was further augmented with CWT time-frequency representations and HAR posterior outputs $p(x)$. Model explanations through this technique thus enabled a holistic interpretation of the activity-recognition or disease-identification outcomes with respect to the raw inertial sensor data and the characterisation of the patterns that were relevant for the prediction of each task.

8.5.1 Interpretation of Activity-based Characteristics

Layer-wise relevance propagation was first applied to a DCNN model trained on the UCI HAR database in order to interpret the network decisions that lead to a HAR prediction. Specifically of interest were the characteristics that were

indicative of walking in this young and "healthy" cohort[12]. Figure 8.3 represented a typical sensor example of static- and activity-based patterns, "laying" and "walking" respectively. Within the sensor trace, walking after 14 seconds was well predicted, where visual analysis of the CWT plot in the same figure demonstrated a strong signal in the gait domain (around 1.5 Hz). Laying, which occurred up until then, was predominately confused for the other similar "static" classes, such as sitting and standing. The shaded grey areas in figure 8.3 represented individual epochs examples for "laying" and "walking" which were correctly classified and further expanded in figure 8.4. Examining these epochs through the LRP lens demonstrated that the higher frequency perturbations in figure 8.4a dominated the R_i score predicting "laying", whereas in figure 8.4b, LRP relevance temporally coincided with the peaks of clear distinct steps (i.e. gait in the 1.5 Hz range).

Comparing these visual aids with the HAR posterior output, $p(\boldsymbol{x})$, enabled a further investigation into the specific differentiation of a "stairs" versus similar "walking" HAR task, which were confused often[13], and visually appeared quite similar in the time-domain (see raw accelerometer trace in figure 8.5). Frequency-based CWT analysis demonstrated that the distinction the DCNN model made between "stairs" and "walking" appeared that the CWT energy E_s in the gait domain (see equation 6.3) was higher than walking, with less higher-frequency-based disturbances. Augmenting CWT with LRP illustrated that clear — and slower— steps influenced $f(\boldsymbol{x})$'s decision as "stairs" (see figure 8.5).

Network decomposition using LRP, in an established task such as HAR, therefore allowed for: (1) a corroboration of our prior beliefs in the characteristics of an accelerometer trace that should influence a classifier's output $f(\boldsymbol{x})$; and (2) validated the HAR model's overall efficacy and robustness. A excellent example of this prior domain knowledge was that these distinct steps (visualised in the time and time-frequency domain) were predictive of "walking" within the sensor signal. As such, returning to the raw data in both the time-frequency domain, supplemented with the decomposition of a model using LRP, enabled a much richer characterisation of

[12]See section 4.4.1 for more demographic information on UCI HAR.
[13]See previously confusion matrix figure 7.3.

the HAR domain task, further cementing the understanding of the sensor patterns that influenced a HAR prediction.

8.5.2 Interpretation of MS-influenced Characteristics

The patterns of healthy gait were first visually inspected in an example HC 2MWT (figure 8.6a) through the LRP-CWT framework presented in section 8.2.1. Comparing these healthy gait templates to PwMSmod examples, including false negatives and false positives, offered a visual interpretation between the differences in the signal characteristics each classifier recognised as important for a prediction. This characterisation of healthy gait could also be supplemented by the findings observed from healthy walking examples in UCI HAR. For instance, it was found that walking in HAR and healthy predicted gait in FL were typically characterised by distinct steps, consistent cadence, and strong gait power (E_s) in the 1.5–3Hz range. Attribution using LRP highlighted step inflections, especially in the vertical $\boldsymbol{a}_y$ and magnitude of acceleration signals $||\boldsymbol{a}||$, as important predictors for both HC ambulation and HAR $p(x)$ of walking. Misclassified PwMS examples also visually resembled that of the HC (appendix figure B.7 (a)–(b)). Similar patterns of the inertial signal were also predictive of HC, for instance LRP also attributed relevance to the $\boldsymbol{a}_y$ and $||\boldsymbol{a}||$ channels during clear step inflections.

The morphology of the sensor data in PwMSmod examples were visually different to correctly classified HC and PwMSmild. In the case of the PwMSmod gait epochs (figure 8.6c) and the false positives (appendix figure B.7 (c)–(d)), these examples exhibited distinct higher frequency "pertubations" in the presence of gait, where further LRP decomposition attributed those disturbances as temporally important for the prediction of PwMSmod in each case. Interestingly, these higher frequency E_s disturbances temporally coincided with each step event from the raw sensor signal. Misclassified examples of HC and PwMSmild as PwMSmod tended to reflect similar properties to the correctly classified example from figure 8.6c, such as as lower E_s gait domain energy (see [2]), evidence of higher frequency perturbations and a visually less well defined step morphology within the raw sensor signal.

As corroboration to the factors influencing these misclassifications, LRP clearly attributed positive relevance (i.e. PwMSmod predictions) to time points of the signal corresponding to higher frequency signal-based activity.

Generating representative correctly classified gait epochs using DTW Barycenter Averaging painted a macro-picture of the average gait patterns for HC, PwMSmild and PwMSmod groups. Visualising these representative epochs using the LRP-CWT framework (figure 8.7), displayed confirmatory patterns observed in previous independently classified 2MWT examples (figure 8.6 (a)–(c)). Importantly, the raw accelerometer signals collected from healthy controls and the DBA-generated average cycles were highly comparable to established characteristics that are representative of healthy walking [**RN1215**, 375]. For example, a DBA representative HC epoch was visually observed to have clear step patterns with discernible initial and final feet contact points, and stronger E_s in the gait domain (0.5–3 Hz). In comparison, milder and moderate MS-predicted average epochs tended to have reduced (gait) signal-to-noise E_s and the presence of higher-frequency perturbations, as described previously.

The uncovering of the inertial sensor characteristics that distinguished MS-related disease from healthy ambulation, through LRP attribution, enabled for clinical interpretations. For example, higher relevance values coinciding with distinct step inflections and higher power (E_s) in the upper-end of the gait domain could represent cadence-based factors associated with faster walking, which are established characteristics that may differentiate healthy versus PwMS ambulation [56, 61, 376]. The attribution of gait disturbances suggesting MS-related impairment, such as in figure 8.6c, could be associated with other accepted indicators, such as gait variability, which have shown to stratify PwMS from HC [56, 64].

More interestingly, hand-crafted features that captured similar characteristics to those visually observed through the CWT-LRP framework appeared previously as top features within table 6.2. Features such as wavelet entropy and energy, capturing predictability and energy in the faster gait domain (1.5–3.3 Hz), separated the same healthy and PwMSmod participants. Other measures, such as signal-to-noise based IMF components, multi-scale entropy and variability of those features also

appeared as distinguishing factors[14]. Validating evidence was further observed in calculated cadence values (steps / min), which were demonstrated to be slower in PwMS than PwMSmild and HC in previous chapter 6, figure 6.8. Similar characteristic gait patterns identified through LRP could also be observed in previous example sensor data in chapter 6, figure 6.3. The HC subject (T25FW, 3.6±0.4 [s]) in figure 6.3a visually demonstrated the same distinct patterns (i.e prominent steps, consistent cadence and strong gait power, E_s, in the 1.5–3Hz range). The PwMSmod participant (EDSS, 4 ± 0; T25FW, 8.1 ± 1.3 [s]), figure 6.3b in comparison, demonstrated typical lower E_s gait domain energy and higher frequency E_s disturbances.

The similarity between hand-crafted features, visual examples, and LRP-explained DCNN features clinically corroborated an interpretation of the factors which may be sensitive to MS-related gait impairment. As discussed before in chapter 7.5, DCNN features outperformed hand-crafted models by up to 15% in binary classification tasks. Returning further to chapter 6, section 6.4.1, the hand-crafted features introduced in this work focused on using established signal-based metrics as surrogates to represent aspects of PwMS gait function. As such, these surrogate features were not engineered to specifically capture complex biomechanical processes in PwMS gait. Data-driven measures may therefore have been more comprehensive, sensitive, or specific to capture the same representation of MS-indicative characteristics, than the approximations from constrained, hand-crafted features.

8.5.3 Longitudinal Trend Monitoring of Remote Smartphone-Based Outcomes

This work demonstrated how a unified classification model introduced in chapter 7.4.3 could (naïvely) estimate disease severity from ordinal classification categories. Severity outcome estimates stratified across HC and PwMS groups, and were highly correlated to actual disease status ($r : 0.75$; $\rho : 0.75$, $p < 0.001$), as measured by the EDSS, which is considered the ground-truth assessment in PwMS [19] (figure

[14]See section 6.4.1 for a more expansive discussion on the clinical interpretation of these features.

8.8). For instance, no misclassifcation of HC as PwMSmod was observed, or vice-versa, indicating that severity estimates could stratify along actual disease severity. More interestingly, those subjects at classification boundaries displayed severities representative of their clinical assessments. For instance, those with EDSS just over 3.5 (i.e. PwMSmod) were miss-classified more as PwMSmild compared to those with EDSS much greater than 3.5, implying that a reflective estimate of disease severity could be captured by transforming a DCNN model into a simple probabilistic outcome per subject.

The longitudinal patterns of healthy controls versus participants with varying manifestations of MS severity could be characterised by examining severity outcomes over the duration of the FL study for individual subjects. For instance, figure 8.9a and figure 8.9b depicted examples of stable trends for HC and PwMSmod subjects respectively. Both participants had "good" and "bad" days, yet the mean severity prediction over each repeated test (overall) reflected the participant's true class grouping. The HC and PwMSmod subjects represented the same example 2MWT examined in figures 8.6a and figures 8.6c respectively using LRP analysis (the specific test date was highlighted in bold). Interestingly, severity was evaluated as mainly "mild" during the first week of testing, where severities thereafter stabilised to "healthy", perhaps suggesting a stabilisation or learning effect where the participant settled into an consistent execution of the 2MWT thereafter.

Evaluating subject's performance longitudinally suggested that severity estimates may be sensitive to capture MS-symptom worsening. An intriguing observation related to the stable PwMSmod participant depicted in figure 8.9b, who was mainly predicted with a severity of PwMSmod, with a relatively consistent 7-day average. Some sequences of tests were predicted as milder however, particularly before the midway clinical visit in week 13. Interestingly, after week 12, this subjects' EDSS rose by +1 to 4.5. A Brown–Forsythe (BF)[15] test demonstrated that this subject had greater variance in their severity outcome before this clinical visit compared to after (BF, $p < 0.01$). Median severity outcomes (Mann-Whitey U test[16]) were

[15]The reader is referred to methods section 3.4.2.9.
[16]The reader is referred to methods section 3.4.2.6

not significantly different between these time-points (U test, p=0.34), however mean severity outcomes were found to be significantly lower before this clinical visit than after (Welch's *t*-test[17]: p<0.05). This same participant's 2MWT sensor data was examined previously in figure 8.6c and corresponded to the specific test date highlighted in bold in figure 8.9b).

During the FL study, two participants experienced relapses which they self-reported using the application on their smartphones. Longitudinal analysis of the trajectories of daily severity scores from these subjects revealed useful insights into the manifestation of relapses expressed in remote inertial sensor data. For instance, both subjects displayed an increased severity outcome up to and around the data of reporting a relapse (figure 8.9c and 8.9d), suggesting that remote severity outcomes were sensitive to remotely capture relapse events.

The relapsing PwMSmod subject presented in figure 8.9c was observed to consistently evaluate as "mild", up until week 13, where this participant reported a relapse on-setting using the FL application on their smartphone. Severity outcomes then increased towards "moderate" during week 13 and peaked at week 14, around the suspected relapse date reported at the end of week 13. Thereafter, severity outcomes stabilised to "mild" before becoming more variable and "moderate" until the end of the study. Considering the relapse reporting date as a threshold, it was found that severity outcomes were significantly "milder" before relapse (where severity outcomes evaluate this subject as PwMSmild) than after relapse on-setting (where severity outcomes evaluate this subject as PwMSmod) when testing between mean (Welch's *t*-test, $p< 0.001$) and between median (U test, $p< 0.001$) severity outcomes. A BF test also signified that severity outcome variability was higher after relapse on-setting than before ($p< 0.01$). This subject was misclassified as PwMSmild using all available 2MWTs, but interestingly was narrowly labelled a PwMSmod and not a PwMSmild subject using their available EDSS scores (EDSS, 3.5 ± 0).

[17]The reader is referred to methods section 3.4.2.2.

In the case of relapsing PwMSmild participant (figure 8.9d), a severity score increase after reporting a relapse was corroborated with a worsening of clinically assessed symptoms from baseline (week 0; EDSS 1.5, T25FW: 4.9 [s]) to the unscheduled clinical visit, which was prompted by the relapse (EDSS: 2.5; T25FW, 7.5 [s]). Examination of severity outcomes leading up to week 3 demonstrated consistent "mild" trends using 7-day moving averages. Interestingly, after the date of onset of self-reported relapse, severity scores rose towards "moderate", indicating that MS symptom manifestation had worsened. Longer term analysis demonstrated that there was a significantly higher variability in predicted severity outcomes after relapse date than before (BF, $p< 0.001$). This subject was further assessed during week 12, where their EDSS returned to as it was reported at baseline (EDSS 1.5; T25FW: N/A). Severity outcomes also returned to consistently "mild" towards the end of the study from weeks 18 onwards, where median (U test, $p= 0.24$) and mean (Welch's t-test, $p= 0.13$) severity outcomes where not significantly different before- and after-relapse. This subject was predicted as PwMSmild over their entire 2MWT outcome measures.

The overall longitudinal population characteristics for healthy participants, mild and moderate subgroups of PwMS were investigated in figure 8.10. The average trend per population, which included incorrectly classified participants, remained stable and consistent regarding each group over the duration of the study. As such, average population-wise estimated severity outcomes both stratified and distinguished between each group longitudinally.

8.5.4 Limitations and Considerations

There are a number of limitations which should be considered in the appraisal of the cross-sectional and longitudinal presentations of this work.

Firstly, despite the clinical utility LRP heatmaps could hold in visualising and interpreting neural network decisions, there are several limitations which must be considered. Heatmap interpretations were only qualitatively evaluated based on visual assessment in this study, albeit motivated by a clinical hypothesis. For

instance, LRP relevance values coinciding with distinct step inflections and higher power (E_s) in the upper-end of the gait domain could represent cadence-based factors associated with faster walking, which are established characteristics that may differentiate healthy versus PwMS ambulation [61, 376]. Other objective methods to evaluate heatmap representations have been proposed which involve perturbing the model's inputs [366]. Verifying that LRP has attributed meaningful relevance is inherently difficult however due to the remote nature of the 2MWTs performed by participants in this study. There ergo exists an inherent lack of ground-truth where we are importantly blinded by *how* the participant performed the actual gait assessment. Further studies should aim to evaluate HC and PwMS gait function in more controlled settings, such as under visual observation or using in-clinic gait measurement systems. Only then, against a ground-truth, can the underlying attributions of LRP be fully qualitatively and quantitatively evaluated.

Besides LRP there are many other methods to interpret neural networks (as discussed in section 8.2.1). More robust analysis should also aim to compare various other techniques, especially similar attribution methods, to evaluate smartphone-based remote sensor models. This future work could be used to further verify the predictive patterns uncovered with one method (e.g. the confirmatory hypothesis of another attribution method also picking up on the same pattern as LRP highlighted).

As initial steps, this thesis focused on establishing clear and concise interpretations of smartphone-based inertial sensor models to characterise patterns of gait and disease-influenced ambulation by first focusing on only the positive contributions towards class predictions (i.e. LRP $\alpha_1\beta_0$-rules). With this baseline, further work (and especially in more controlled settings) should aim to apply LRP for more complex tasks to develop more full-bodied explanations, such as visualising contributions which contradict the prediction of a class (e.g. using LRP $\alpha_2\beta_1$-rules). Attribution methods to disentangle the DCNN decisions, and to uncover this gain in identification performance, could greatly aid clinician's understanding of the sensor factors that characterise milder MS, especially sub-clinical disease manifestations.

The severity outcomes explored in this work were naïve scores; although outcomes captured a trend of increased impairment with higher severity (as modelled by EDSS, figure 8.8), they should not be considered an exact measure of MS, nor a surrogate clinical outcome to permit any "clinical diagnosis".

Furthermore, as with the limitations outlined in chapter 7 section 7.5, the DCNN models used in this study to create longitudinal severity outcomes did not truly utilise the time-series nature of repeated 2MWT measurements in the FL data. Each repeated test was treated as independent, and as such, trajectories did not incorporate any temporal information across a population or within a subject (e.g. whether the previous day's $d-1$ test can affect the outcome at d). It would be assumed that this is critically missing temporal information which could help build more reliable and accurate longitudinal models, and should be considered as a key next step for future work.

Ultimately, the manifestation of gait impairment in PwMS is a complex and heterogeneous process. This study does not aim to introduce a global model for how gait dysfunction is represented in PwMS, but rather introduces a promising framework to visualise how sensor data can be represented to clinically understand the decisions that lead to a prediction for individual inertial sensor examples. This framework can then be used to generate added insight into how PwMS impairment can be characterised remotely from smartphone-based sensors, and how PwMS can be monitored longitudinally. By aggregating these discrete clinical interpretations, and framing the findings in the context of a macro- and longitudinal performance, we can then start to overlay some of the important understandings to build a picture for how the characteristics of inertial sensor data that may represent healthy versus PwMS disease.

8.6 Conclusion

Deep networks have often been called "black-box" algorithms as they can be highly non-linear that it can be incredibly difficult to interpret the function that influenced a prediction. This chapter therefore aimed to establish a framework to

explain and build transparent smartphone-based remote sensor models leveraging multiple viewpoints such as: (1) visualising the raw inertial sensor signal; (2) its analogous time-frequency CWT representation; (3) the $\mathcal{T}_S$ HAR output, (4) as well as augmenting this picture using LRP, a deep network visualisation technique. With this framework, the patterns captured from smartphone-based inertial sensor data that are distinctive of those who are healthy versus PwMS can begin to be established.

Incorporating expert clinical knowledge through interpretation of network decisions, such as the LRP decomposition proposed here, could greatly develop clinicians' fundamental understanding of diseases. This improved domain knowledge could be used to reverse engineer features, develop more robust models or to help refine more sensitive and specific measurements.

Administering digital assessments remotely and at a high-frequency may be a powerful tool to characterise patterns of MS ambulatory impairment. Next, this chapter explored the longitudinal characteristics of repeated and continuous model severity outcomes for HC versus mild and moderate PwMS. It was observed that longitudinal trends of PwMS were discernible from healthy participants and that the signs of MS relapse could be captured from inertial sensor outcomes. Incorporating the visualisation framework introduced in this work at cross-sectional time-points throughout the FL study for individual participants, allowed for a richer and contextual understanding of HC and PwMS performance. This application both framed the overall (macro-) and longitudinal patterns representative of various participants and groups, but also offered insightful context to the (micro-) patterns influencing the raw inertial sensor data; hence forming overall characterisation of disease-influenced versus healthy activity which has been recorded from out-of-clinic, smartphone-based, inertial sensor data. This study, with on-going future work, therefore further demonstrates the clinical utility of objective, and more transparent, out-of-clinic assessment outcomes to develop digital biomarkers for monitoring PwMS.

9
Conclusions and Future Work

This thesis aimed to explore the use of digital health technologies, specifically smartphones and smartwatches, to characterise and monitor disability in people with multiple sclerosis (PwMS), remotely and out-of-clinic. The wider characterisation of objective, quantifiable, and repeatable assessments for those with neurodegenerative diseases using digital technologies has been referred to as *digital biomarkers*. The significant advantages of digital biomarker outcomes are the ability to collect objective, rich and sensitive health-related measurements at high frequency and remotely.

Digital biomarkers is a emerging field however, where the assessments and metrics evaluated are only in the early proof-of-concept (PoC) stage of development. This thesis therefore first aimed to explore how digital sensor data from smartphones and smartwatches can represent upper-extremity and ambulatory dysfunction in PwMS using data acquired from the FLOODLIGHT PoC study (FL). This work demonstrated how measurements from touch-screen and inertial sensors could be developed to accurately replicate gold-standard clinician-administered assessments or distinguish between healthy controls (HC) and sub-groups of PwMS. Building upon these foundations —which also established a much deeper understanding of the covariates and considerations for developing digital biomarkers captured from remote settings— this thesis focused on leveraging state-of-the-art machine learning

(ML) techniques to further characterise HC function versus PwMS dysfunction using smartphone sensor data.

In this concluding chapter, a summary of results from each previous chapter will be provided, highlighting their respective novelties and contributions to the field of digital biomarkers. The findings of this thesis will also be discussed with respect to their intrinsic limitations, as well as the possible directions of future work aimed to further advance the development of remote digital biomarker assessments for PwMS.

9.1 Summary of Contributions

The vast majority of previous attempts to characterise multiple sclerosis (MS) using wearable digital health data have been applied in controlled clinical settings; these technologies however offer the unique opportunity to monitor the signs and symptoms of MS objectively, subtlety and remotely. Nowadays, smartphones and smartwatches are ubiquitous in everyday life, where these tools offer the potential to measure and create digital outcomes to assess PwMS. This thesis therefore focused on exploring the use of smartphone and smartwatches to characterise the signs and symptoms of MS function out-of-clinic. Starting from first principles, this thesis introduced a methodology to develop measures (termed "features") which can capture the physical manifestation of healthy versus of MS dysfunction. The initial steps for these new digitally augmented assessments were to benchmark against established clinical metrics.

The first results chapter demonstrated how this litmus test can be performed, whereby the ability of digitally captured smartphone touch-screen features to represent MS hand-motor function (HMF), and the ability to stratify across MS severity levels, was investigated. This was achieved through univariate and multivariate modelling of features developed from the smartphone-based remotely executed Draw-a-Shape test (DaS) against the clinically administered Nine Hole Peg Test (9HPT). The 9HPT is the gold-standard assessment for upper-extremity function (UE) in PwMS and thus represented a ground-truth for which to assess the ability of DaS features to accurately characterise UE performance.

The work presented there initially introduced how novel features motivated by disease pathology and UE function could be extracted from various shapes *(circle, figure-8-shape, spiral, square)* traced on smartphone touchscreen. This introduction to the development of digital biomarkers demonstrated how these digital measures relate to UE impairment in PwMS, including the evaluation of various digitised shapes to capture UE PwMS performance. This, to the authors knowledge, is the first study demonstrating how digitally recorded drawing data can predict the 9HPT, the gold-standard upper-extremity function assessment in MS. It was found that a simple linear combination of features from various shapes resulted in a suitably accurate reconstruction of the 9HPT, which were modelled separately for dominant ($r^2 = 0.41$, $p < 0.05$; MAE: 2.08 ± 0.34 [s]) and non-dominant handed tests ($r^2 = 0.39$, $p < 0.05$; MAE: 2.32 ± 0.43 [s]). The fact that non-dominant handed tests were more accurately predicted than dominant tests indicated that perhaps more complex shapes, tested using a weaker hand, elicited a wider stratification of PwMS subjects' performance and disease manifestation. Moreover, it was observed that subjects with higher 9HPT times (indicating greater disability) had more variable 9HPT times, between- and within-subject. This variability resulted in less accurate prediction of longer 9HPT times, however it was noted that the mean absolute error (MAE) in predicted 9HPTs was comparable to the variability of actual 9HPT times within HC, and PwMS groups, respectively. Importantly, this work ignites the exploration of digital biomarkers by demonstrating how DaS features may contain a greater wealth of information supplementing beyond discrete 9HPT scores. Chapter 5 represented an important first step in validating UE digital biomarkers for MS, particularly as few previous studies have advanced beyond classical confirmatory tests to implement machine learning tasks for disease classification and severity regression.

Another essential hallmark of digital biomarkers is the ability to perform disease classification, or to distinguish between different severity categories of disease (such as in chapter 6). Classification, compared to specifically modelling a clinical metric, as introduced previously in chapter 5, largely circumvents the overt limitations of

coarse, discrete, and sometimes subjective clinical assessments. Directly modelling the presence of disease might allow more disease-specific features rather than clinical assessment-specific features to be developed.

While chapter 5 laid the framework to perform remote digital assessments in PwMS, it was considered that MS is a heterogeneous and mutable disease where subjects may experience symptoms in various domains. Ambulatory function is one of the most prominent and disabling physical impairments in multiple sclerosis. As such, chapter 6 explored the ability of inertial sensor data captured from smartphones and smartwatches to distinguish (i.e. classify) healthy controls from mild and moderately impaired PwMS. This work established how examining the raw sensor data could motivate the creation of specific features, and how established signal-processing metrics could be re-purposed as surrogates to represent aspects of PwMS gait function. A sparse representation of the gait feature structure was established using LASSO feature selection which explored how inertial sensor features could characterise ambulatory impairment both in a univariate and multivariate domain. Classifications of subject types were compared using data obtained from smartphone, smartwatch and the fusion of features from both devices. Models built on smartphone features alone achieved the highest classification performance, indicating that accurate and remote measurement of the ambulatory characteristics of HC and PwMS could be achieved with only one device. It was observed however that smartphone-based performance was affected by inconsistent placement location (running belt versus pocket). Results clearly demonstrated that PwMSmod, who experience gait-related dysfunction, could be distinguished from HC subjects (Acc. 82.2 $\pm$ 2.9%, MF_1 84.3 $\pm$ 3.8), and PwMSmild (Acc. 82.3 $\pm$ 1.9%, MF_1 75.1 $\pm$ 2.2) using an SVM classifier with a Radial Basis Function (RBF). PwMSmild were shown to exhibit HC-like behaviour and were thus less distinguishable from HC (Acc. 66.4 $\pm$ 4.5%, MF_1 58.6 $\pm$ 5.8). Importantly, chapter 6 introduced the important considerations, and analysis techniques thereof, when evaluating remote smartphone and smartwatch PwMS data. Subject heterogeneity, test adherence, and execution of the assessment procedure were shown to act as salient covariates.

Although chapter 6 helped establish a methodological framework to construct models that could identify patterns of PwMS ambulatory impairment from remote 2MWT gait assessments, ultimately, the classification performance presented there was reliant on the feature structure to sufficiently encapsulate information related to PwMs ambulatory function. Chapter 7 instead proposed that we are limited by the effectiveness of our features, which have been "hand-crafted" to encode what we a-priori believe to be important characteristics of ambulatory dysfunction related to MS disease activity. However, as introduced in chapter 6, these hand-crafted gait features were generally established signal-processing metrics re-purposed as surrogates to represent aspects of PwMS gait. With such narrow constraints, based solely on prior hypothesis (albeit relatively clinically informed), risked missing critical PwMS gait information. Chapter 7 therefore attempted to introduce a more data-driven methodology to gait feature extraction, whereby a deep network was allowed to learn its own features to perhaps capture a better representation of HC and PwMS disease characteristics from remote smartphone-based inertial sensor data. This deeper analysis focused on the gait domain, demonstrating how gait representations learned from fully-connected deep convolutional neural networks (DCNNs) could outperform feature-based approaches. To overcome the typical limitations associated with remotely generated health data, such as low subject numbers and heterogeneous data, transfer learning (TL) from similar large- and open-source datasets was proposed. This work leveraged the information first learned from Human Activity Recognition (HAR) tasks from wearable smartphone sensor data. It was demonstrated that fine-tuning TL DCNN HAR models towards the disease classification tasks introduced in chapter 7 outperformed feature based methods, as well as out-performing DCNN models trained exclusively on FL, by between 8%–33%.

Despite the improvement of TL DCNNs demonstrated over traditional feature-based methodologies, and demonstrate within the wider sphere of tackling many healthcare related tasks, a deep network's lack of transparency remains one of the larger stumbling blocks to the wider acceptance of deep learning for clinical applications. Chapter 8 aimed to open these "black-box" algorithms by introducing

a technique to visualise a DCNNs decisions using relevance heatmaps, known as layer-wise relevance propagation (LRP). First, LRP relevance scores were attributed to HAR classification examples in order to establish the patterns DCNN models identified as gait-related activities, such as walking and stair climbing. Next, healthy gait examples from FL were examined by visualising the raw sensor data and its instantaneous time-frequency representation, through the continuous wavelet transform (CWT), with respect to a HAR posterior output and LRP attribution. Combining the knowledge ascertained from HAR analysis and HC analysis in FL, a template of healthy gait could be established by which to compare PwMS examples. The subsequent investigations in chapter 8, though examining multiple instances of PwMS and HC 2MWTs using the CWT-LRP framework of visualising the data from multiple viewpoints, including decomposing DCNN predictions using LRP, helped establish and characterise the gait patterns that are distinctive of those who are healthy versus PwMS. Interpretations suggested that cadence-based measures, gait speed, and temporal perturbations were distinct patterns which distinguished healthy controls and PwMS captured from smartphone-based inertial sensor data.

Finally, chapter 8, section 8.4.4 explored how surrogate continuous model severity outcomes stratifying HC versus mild and moderate PwMS could be constructed from the raw DCNN outputs. By holistically examining DCNN predictions as a function of time (section 8.4.5), it was demonstrated how longitudinal trends of PwMS were discernible from healthy participants. These surrogate severity scores even demonstrated a sensitivity to the signs of MS relapse, as evinced by two participant examples. Incorporating the CWT-LRP visualisation framework introduced earlier in the chapter for collections of epochs within individual 2MWTs of interest, enhanced an understanding towards the patterns that may indicate healthy versus MS-influenced symptoms. This aggregation approach integrated the longitudinal (macro-) patterns and a deeper understanding ascertained from the (micro-) patterns influencing the model's predictions (and hence the patterns characterising the presence of ambulatory-related disease). This final contribution encapsulated the key clinical usefulness of objective out-of-clinic assessment outcomes

that can be assessed at high-frequency and ultimately represented to key benefits motivating the development of digital biomarkers for monitoring PwMS.

9.2 Limitations Leading to Future Work

The findings, and shortcomings, presented in this thesis motivate several potential avenues of both technical and clinical future work.

9.2.1 Study Design and Considerations

The specific limitations borne in each chapter provide the obvious first routes for future work. The common thread limiting the work throughout this thesis were ultimately linked to data constraints and study design within FL. Underlying biases manifested due to low subject numbers, heterogeneous data and varying participant adherence; multiple repeated measurements and differences in the number of unique tests contributed per subject, or even demographic biases, such as the milder-skewed distribution in disease severity of PwMS subjects, the inclusion of various different MS phenotypes, and male-to-female ratio mismatches between HC and PwMS should all be considered. For instance, these issues were expressed in chapters 5 and 6 as biases within CV, such as unexpected and dramatic changes in the distribution of features between different folds, or the selection of spurious feature sets or even inconsistent model parameterisation. Moreover, most MS patients were mildly disabled with respect to their EDSS or motor specific 9HPT and Timed-25 foot Walk (T25FW) clinical scores. As a result, classification models were only as powerful as the total number of subjects in the minority class, or that regression models were observed to systematically underestimate 9HPTs for longer (more severe) actual 9HPTs. Learning more accurate global models were therefore biased by the diversity, representation, and size of the training data available.

In reality, the NCT02952911 FLOODLIGHT PoC study was only intended as a small proof-of-concept investigation to assess the feasibility to remotely monitor PwMS subjects, yet has provided many meaningful insights which can be implemented in future studies. Follow-up trials with larger, more diverse cohorts

are already being undertaken, such as FLOODLIGHT OPEN, a crowd sourced dataset where the general public can contribute their own data [377].

In chapter 7 transfer learning from larger open-source datasets and fine-tuning a model for a smaller dataset, like FL, was proposed as a useful technique to circumvent training data-related constraints. More work is needed however to fully characterise the benefits of both deep networks and transfer learning applied to sparse remote sensor data. Further work, for example, will be needed to understand the underlying feature structure, along with the most applicable source datasets and methods to extract the most appropriate information available to characterise HCs and subgroups of PwMS from remote smartphone inertial sensor data. Despite the usefulness which TL could accomplish, fine-tuning a model to a dataset with biases will nonetheless encode those same biases. For example, the mismatch in the sex of HC (mostly male) versus PwMS (mostly female) is an obvious inherent bias in the FL dataset. The design of future studies will learn from the limitations of those in FL.

The richest learning's from FL will include the confounding factors encountered when taking measurements in real-world non-laboratory scenarios. Some examples encountered included the differences in subjects' adherence during the study, some subjects not following the prescribed protocols (during the 2MWT there were instances of smartphones in pocket locations or the use of only one device during testing), along with the many degrees of freedom associated with self-generated patient data from real-world testing. The results from this study emphasised the importance of a standardised approach to remote sensor monitoring and advocated the use of a consistent sensor location such as a running belt for future studies. Advantages of running belts are that they offer a fixed and standard placement location to capture gait characteristics and avoid the need for participants to have pockets or another means to carry their device.

Adherence to the prescribed protocols was an issue observed in this study. It is acknowledged that instructing participants to regularly carry their smartphone and smartwatch for a daily 2MWT, and to affix their smartphone using a running belt is both obtrusive and inconvenient. A missing link not explored in chapter 7

was the characterisation and performance of smartphone versus smartwatch-based classification of HC and PwMS subgroups using DCNN models, and becomes a vital next step for the work that was explored later in sections 7 and 8.4.5. This proposed analysis might inform other future investigations where perhaps the use of a smartphone within a pocket, or even a single smartwatch during passively collected free-living gait may be a better option for the unobtrusive, long-term monitoring of PwMS subjects.

Before these suggested future avenues of exploration are opened, more work is needed first to understand in greater detail: (1) the effect of placement location of smartphone devices; and (2) to investigate the differences in information captured by smartphone and smartwatch devices for quantifying gait dysfunction in PwMS, particularly in more controlled settings. These further studies should also aim to compare the outcome measures investigated in this work to clinically administered 2MWTs and in-clinic gait measurement systems. This suggested analysis would allow the further evaluation, understanding and improvement of the most optimal protocols designed to explore how sensor data can represent PwMS impairment remotely and out-of-clinic.

As eluded to in chapter 8, section 8.5.2, verifying that LRP has attributed meaningful relevance to various gait samples is inherently difficult as we are essentially blinded by the remote nature of the 2MWTs performed by participants in this study. Examining gait performance in more controlled settings such as under visual observation or using in-clinic gait measurement systems would therefore allow the LRP attributions to be clinically evaluated.

9.2.2 The Migration to Unobtrusive and Free-Living Ambulatory Monitoring of PwMS

Recently, it has been hypothesised that ambulatory behaviour in uncontrolled (free-living) environments may lead to more novel insights on the accumulation and distribution of gait function in PwMS [378]. While the 2MWT is a common walking outcome in PwMS, thought to measure fatigability in addition to overall

walking disability, it is a constrained and scripted testing protocol which may not represent an ecologically valid picture of the underlying disease [174]. For instance, previous work investigating free-living gait assessments have been shown to distinguish HC from PD patients [6, 7]. Similar studies examining free-living gait suggest it emphasises ongoing environmental and cognitive challenges that impair gait performance in Parkinson' disease (PD) [379]. In PwMS, free-living gait assessments are relatively less well established, although it has been suggested that free-living gait measures from short walking bouts may distinguish disability levels in more disabled PwMS groups (EDSS>5) better than longer walking bouts or in-clinic laboratory gait (as determined by pairwise significance tests for various biomechanical features) [167]. Other studies suggest that holistic activity indicators such as a daily step count, daily energy expenditure and walking intensity metrics may capture important aspects of MS ambulatory and physical function [62, 173, 174, 380]. As such, future work stemming from this thesis should aim to investigate the passive monitoring data collected as part of FL. Proposed work should first explore the fidelity of passively collected gait, the differences between passive and active gait explored as well as the characterisation of the features, metrics and parameters developed in this thesis applied to habitual gait.

9.2.3 Improving Severity Outcome Estimates

The severity outcome measures introduced in this work begin to establish how robust and objective markers of MS disease can be composed from remotely generated smartphone data. It should be considered that the DCNN architecture investigated in this study was relatively less complex compared to other frameworks [188, 190, 321, 343] (see also section 7.2.1). Proposed DCNN models did not incorporate recurrent layers for instance, which may be beneficial to characterise the temporality of gait recorded within the sensor signal [182, 321]. Long Short-Term Memory (LSTM) layers are particularly capable of characterising temporal information within a sequence or epoch and learning long term dependencies such as the accumulation of evidence related to a pattern or feature over that sequence [202, 381]. Architectures

based on LSTMs have been successfully applied to time-series data for example in speech recognition [382], in the modelling of Electronic Health Records (EHR) clinical measurements [383] and they have become important additions integrated into convolutional neural network (CNN) based HAR [321] and subject identification architectures [182]. Other models such as temporal convolutional neural networks (TCNN) may also yield promising improvements for sequence modelling (as in this gait task) and have been shown to outperform LSTMs with considerably less computational requirements [384].

A clear gap in the analysis introduced in this thesis is that non-feature based approaches, such as the DCNN, were not applied towards the Draw-a-Shape data. CNNs are a staple methodology in computer vision image recognition tasks [385, 386], many of which would appear similar to the hand drawing of shapes in DaS. The seminal example of which is the classification of handwritten digits in the MSNIT database by LeCun *et al.* (1998) using LeNet-5, a CNN architecture [387]. Work has already begun on applying CNNs towards disease assessment in Parkinson' disease using digitally drawn spirals and other tasks [388–391]. The use of LRP to decompose a DaS trained DCNN would arguably lead to a more intuitive relevance heatmap than the gait application investigated in this thesis. In particular, LRP has shown to produce highly interpretable heatmaps for handwritten digits in the MSNIT database [256].

The background to MS disease in section 2.1.2, presented the pathophysiological characterisation of MS as a heterogeneous and mutable disease which affects subjects across various domains. While this thesis focused on the physical impairments of ambulatory function and upper-extremity function, these domains were treated as separate and independent. In reality, the complex demyelination process of MS may be expressed as symptoms in multiple domains simultaneously, which could be inextricably linked. This thesis never investigated the aggregation of 2MWT gait and DaS performance metrics to classify or predict disease status. Furthermore, those severity scores and subsequent longitudinal analysis in this thesis only considered the ambulatory domain and as such my miss important manifestation of MS associated

with other domains. Ensuing work should aim to incorporate UE function measures captured from the DaS test to create a unified physical function severity outcome estimate. Extensions of this could aggregate information obtained from the other domains and tests within FL, such as the cognitive assessments.

The models presented throughout this thesis are ultimately hindered however by an intrinsic reliability on clinical labels. In chapter 5, DaS features were constructed to predict the clinician-administered 9HPT; in chapters 6–8 classification models were based on clinician-rated EDSS scores. While these imperfect measures are rife with rater-biases, coarseness- and variability-based limitations, the supervised models in this work were subjugated to these issues. For instance, in classifying PwMSmild versus PwMSmod sub-groups, that were bluntly split by subjective EDSS thresholds (albeit clinically motivated). Perhaps should we best let a model uncover the phenotypes within PwMS directly from the data? There exist unsupervised methodologies, such as Deep Embedded Clustering (DEC) [392] which can be extended to combine Deep-Convolutional Autoencoders (DCAE) [393] with a clustering loss to learn representations aimed to produce explicitly separable clusters in the embedded space. These clusters could be explored to learn data-driven PwMS phenotypes, or reveal unseen patterns of MS through the deep-latent embedding.

9.2.4 Longitudinal Remote Monitoring of PwMS

Finally, as MS symptoms fluctuate periodically, the real value of remote monitoring PwMS may ultimately lie in investigating test performance as a function of time. To sufficiently capture the time-varying nature of MS ambulatory impairments it will require both robust outcome measures and the design and standardisation of feasible and unobtrusive protocols for objective assessments —that can be delivered remotely and administered frequently— such as those introduced in this thesis.

The DCNN severity outcome estimates developed as part of section 8.4.5 naïvely treated each 2MWT as independent. As such, the rich time-series nature of those repeated 2MWTs per subject were never truly exploited. Trajectories or long-term monitoring did not incorporate any temporal information or trends across

a population or within a subject. For instance, the repeated FL assessments, and therefore digital biomarkers that were extracted, could be applied to models that exploit this aggregation of temporal information directly, for example the methodology behind generalised additive models (GAM) [394]. Encoding this critical information could help construct more reliable, adaptive, and predictive models to estimate MS disease severity and disease course using remotely collected smartphone and smartwatch data. The long-term monitoring of PwMS performance could open up the space for true personalisation: the clustering of disease trajectories or similar patients, estimating the likelihood of disease progression, quantifying response to different treatments as a population or an individual, as well catching the mutable patterns of MS disease that are only visible out-of-clinic and as a function of time.

As a research community we are only beginning to develop *digital biomarkers*, where the goal of personalised healthcare in the digital age has only started to be conceptualised. What is clear however, is that the enrichment of information that can be generated at high-frequency over time, from remotely administered smartphone- and smartwatch-based outcomes, will allow us to learn and characterise this chronic disease more than we ever have before. This thesis aims to become a small cog to help contribute towards an effective understanding of multiple sclerosis outside of the clinic, and ultimately, I hope will help towards finding more effective treatments for this terrible disease.

Appendices

A

Methods: Additional Distributions, Algorithms and Features

A.1 Additional Distributions

A.1.1 The *z*-distribution

The $z-$distribution represents the standard Normal distribution, $\mathcal{N}(0,1)$, (i.e. unit variance) over the z-score (z) [204]. For a random variable $\boldsymbol{x}$ with mean $\mu = \bar{\boldsymbol{x}}$ and standard deviation σ, the z-score of a value x is

$$z = \frac{x - \mu}{\sigma} \tag{A.1}$$

The general form of its probability density function is defined as:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{A.2}$$

The $p-$value (p) is determined to define the cut-off (critical value) of z which represents that the probability is less than a pre-defined significance threshold. Typically, this threshold is defined at the 95% level, where $p = 0.05$. For $p < 0.05$, critical z values for a two-tailed test are ± 1.96.

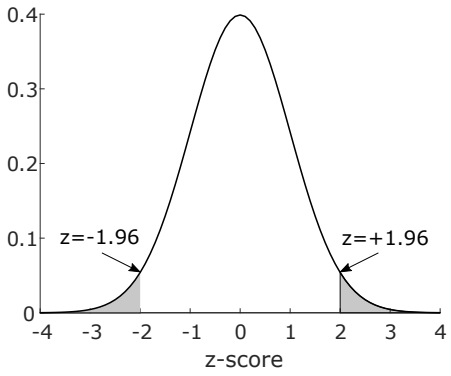


Figure A.1: The standard z-distribution. To assess if a test is statistically significant, the z-score is compared to a critical value for a given $p-$value. For $p < 0.05$, critical z-scores are $z < -1.96$ and $z > 1.96$ (as shown).

A.1.2 The Student's t-distribution

The Student's t-distribution is defined as a function $f(t)$, over a t-statistic (x) [204, 213]:

$$f(x|d) = \frac{\Gamma\left(\frac{d+1}{2}\right)}{\sqrt{d\pi} \cdot \Gamma(\frac{d}{2})} \cdot \left(1 + \frac{x^2}{d}\right)^{-(\frac{d+1}{2})} \tag{A.3}$$

where d is the degrees of freedom and $\Gamma(\cdot)$ is the Gamma function. The result $f(x)$ is therefore the probability of observing a value of x from the Student's t-distribution with d degrees of freedom. The evaluation of the t-distribution parameters by various d choices is depicted in figure A.2a. Figure A.2b further depicts the critical values of t that achieves $p < 0.05$ for a two-tailed test for $d = 100$.

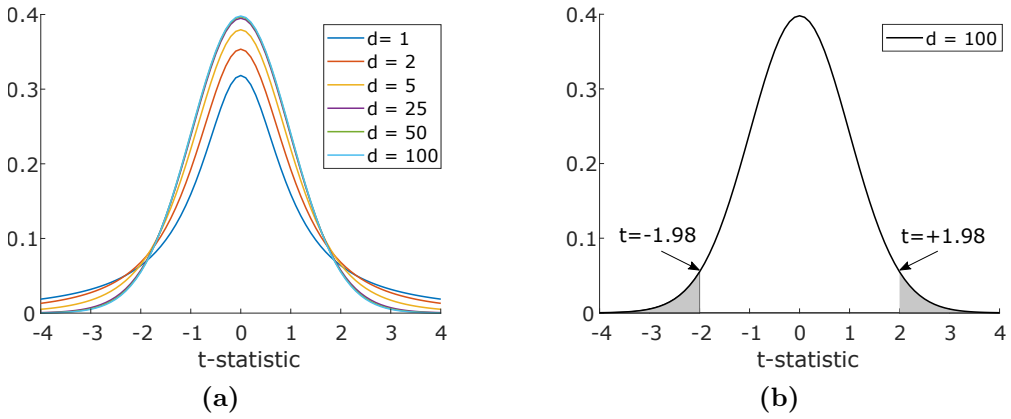


Figure A.2: The $t-$distribution. Figure **(a)** The $t-$distribution for various values of d. To assess if a test is statistically significant, the $t-$statistic is compared to critical values for a given significance level. (b) The $t-$distribution with $d = 100$. The critical $t-$statistics for $p < 0.05$ are $t < -1.98$ and $t > 1.98$ (as shown).

A.1.3 The Chi-Square distribution χ^2

The chi-square χ^2 distribution is defined as a function $f(x)$, over a χ^2-statistic (x) [204]:

$$f(x|d) = \frac{x^{\frac{(d-1)}{2}} e^{\frac{-x}{2}}}{2^{\frac{d}{2}}\Gamma(\frac{d}{2})} \tag{A.4}$$

where d is the degrees of freedom and $\Gamma(\cdot)$ is the Gamma function. The result $f(x)$ is therefore the probability of observing a value of x from the χ^2-distribution with d degrees of freedom. The evaluation of the χ^2-distribution parameters by various d choices is depicted in figure A.3a. Figure A.3b further depicts the critical values of t that achieves a significance $p < 0.05$ for $d = 4$.

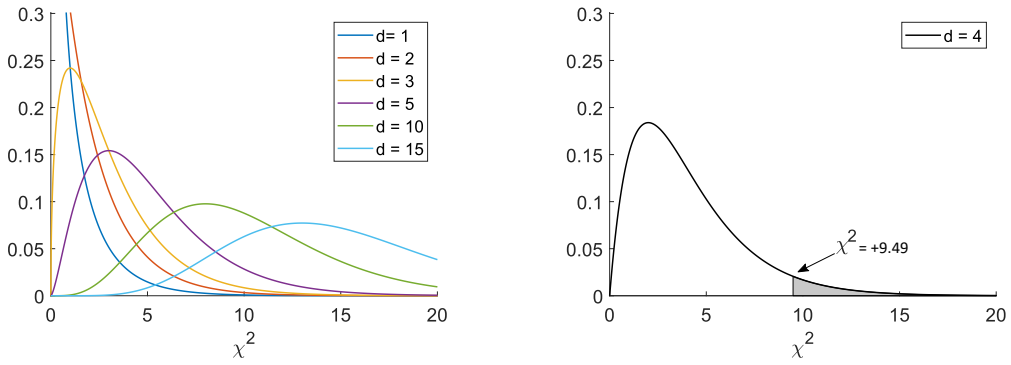


Figure A.3: The χ^2-distribution. Figure **(a)** The χ^2-distribution for various values of d. To assess if a test is statistically significant, the $t-$statistic is compared to critical values for a given significance level. **(b)** The χ^2-distribution with $d = 4$. The critical χ^2-statistic for $p < 0.05$ is $t > +9.49$(as shown).

A.1.4 The F-distribution

The F-distribution is defined as a function $f(x|d_1, d_2)$, over a F-statistic (x) [204]:

$$f(x|d_1, d_2) = \frac{\Gamma(\frac{d1+d2}{2})}{\Gamma(\frac{d1}{2})\Gamma(\frac{d2}{2})} \cdot \left(\frac{d1}{d2}\right)^{\frac{d1}{2}} \cdot \frac{\left(x^{\frac{d1-2}{2}}\right)}{\left(1+\frac{d1}{d2}x\right)^{\frac{d1+d2}{2}}} \tag{A.5}$$

where $d1$ and $d2$ are the degrees of freedom and $\Gamma(\cdot)$ is the Gamma function. The F-distribution is dependent on $d1$ and $d2$; it is therefore non-symmetrical, positive-definite and suitable for only one-tailed tests, as can be seen in figure A.4a. The critical value of F that achieves a significance of $p < 0.05$. for $d1 = d2 = 100$ is depicted in figure A.4b.

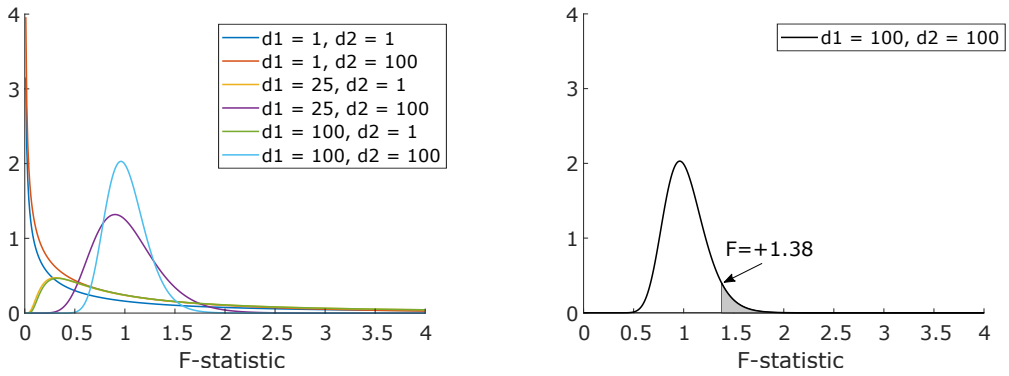


Figure A.4: The $F-$distribution. Figure **(a)** The $F-$distribution for various values of $d1$ and $d2$. To assess if a test is statistically significant, the $F-$statistic is compared to a critical value for a given significance level. **(b)** The $F-$distribution with $d1 = 100$ and $d2 = 100$. The critical $F-$statistic for $p < 0.05$ is $F > 1.38$ (as shown).

A.2 Algorithms

A.2.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) can be formally defined as the orthogonal projection of the data, $\mathbf{X} \in \mathbb{R}^{N \times P}$, onto a lower dimensional space, the *principal component* space, such that the variance of the projected data is maximised [230] [204]. The directions of the latent *principal components* (PCs) are orthogonal, correlated, and ordered such that the first PC is aligned in the direction that maximises the variance of the projected data. The goal of PCA is to project data onto a space having dimensionality $Q < P$, while maximising the variance of the projected data, especially useful when $N \ll P$. Firstly, the covariance of $\mathbf{X} \in \mathbb{R}^{N \times P}$ is denoted as $\mathbf{\Sigma} \in \mathbb{R}^{P \times P}$, and can be defined by:

$$\mathbf{\Sigma} = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top \tag{A.6}$$

where $\mathbf{x}_i \in \mathbb{R}^P$ corresponds to the i^{th} observation in $\mathbf{X}$ and $\bar{\mathbf{x}} \in \mathbb{R}^P$ is the sample set mean. The sample covariance can also be written in the linear algebra form (omitting the unbiased estimator term $1/(1-N)$):

$$\mathbf{\Sigma} = \mathbf{X}^\top \mathbf{X} \tag{A.7}$$

Next, the *eigenvalues* and *eigenvectors* are computed from $\mathbf{\Sigma}$ using *eigen decomposition* of $\mathbf{\Sigma}$ such that:

$$\mathbf{\Sigma} = \mathbf{V}\mathbf{D}\mathbf{V}^\top \tag{A.8}$$

where $\mathbf{D}$ is a diagonal matrix of *eigenvalues* ordered in decreasing value, and $\mathbf{V}^{P\times P}$ is matrix of the corresponding *eigenvectors* as orthonormal columns.

The Singular Value Decomposition (SVD) is a fast method to way of obtaining the eigenvectors [200]. The SVD of the centred matrix[1] $\mathbf{X} \in \mathbb{R}^{N\times P}$ is:

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top \tag{A.9}$$

where $\mathbf{U}^{N\times N}$ matrix, $\mathbf{V}^{P\times P}$ matrix and are orthonormal. The diagonal entries of (a non-negative and decreasing) $\mathbf{\Lambda}^{N\times P}$ matrix are known as the singular values of $\mathbf{X}$. SVD is actually applied to the covariance matrix of $\mathbf{X}$ (in the form of equation A.7) yielding:

$$\begin{aligned}\mathbf{\Sigma} = \mathbf{X}^\top\mathbf{X} &= \mathbf{V}\mathbf{\Lambda}^\top\mathbf{U}^\top\mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top \\ &= \mathbf{V}(\mathbf{\Lambda}^\top\mathbf{\Lambda})\mathbf{V}^\top\end{aligned} \tag{A.10}$$

This is the same form of the *eigen decompositon* in equation A.8 of $\mathbf{X}^\top\mathbf{X}$ (and of $\mathbf{\Sigma}$, up to a factor N), and we need to only find $\mathbf{U}$ and $\mathbf{\Lambda}$, where $D = \mathbf{\Lambda}^\top\mathbf{\Lambda}$ are a diagonal matrix of eigenvalues ordered in decreasing value and $\mathbf{U}^{P\times P}$ is a matrix of the corresponding *eigenvectors* as orthonormal columns. This can be achieved by using the much more computationally efficient *eigen decompositon* of $\mathbf{X}\mathbf{X}^\top$ when $N \ll P$:

$$\mathbf{X}\mathbf{X}^\top = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top\mathbf{V}\mathbf{\Lambda}\mathbf{U}^\top = \mathbf{U}(\mathbf{\Lambda}\mathbf{\Lambda}^\top)\mathbf{U}^\top \tag{A.11}$$

Lastly, we can denote denote the $\mathbf{Z}$ as the scores matrix with the projections onto the PC, or the transformed latent variables, with P principal component directions:

$$\mathbf{Z} = \mathbf{X}\mathbf{V} = \mathbf{U}\mathbf{\Lambda} \tag{A.12}$$

The PCs are ordered by the proportion of variance each explain and as such the first $Q < P$ chosen latent components can be kept to approximate the data in a lower dimension.

[1] i.e. standardised with zero mean and unit variance (standard normal distribution, see A.1.1).

A.2.2 *t*-distributed Stochastic Neighbour Embedding

$t-$distributed Stochastic Neighbour Embedding embeds high-dimensional points in low dimensions with the aim of preserving the similarities between points. Nearby points in the high-dimensional space, $\mathbf{X} \in \mathbb{R}^{N \times P}$ correspond to nearby embedded low-dimensional points $\mathbf{Y} \in \mathbb{R}^{N \times Q}$, $Q \ll P$, and vice-versa. The general steps to performing $t-$SNE can be summarised as [201]:

1. Data preparation, including the standardising $\mathbf{X}$ to have zero mean and unit variance, and in cases of $N \ll P$, summarising $\mathbf{X}$ in a lower-dimension using PCA first.

2. Calculate the pairwise distances $d(\mathbf{x}_i, \mathbf{x}_j)$ between high-dimensional points in $\mathbf{X}$. Any distance metric can be used, such as the Euclidean distance, $d(\mathbf{x_i}, \mathbf{x_j})$=$||\mathbf{x}_i - \mathbf{x}_j||$ (see equation 3.42 for more details).

3. Calculate the standard deviation σ_i for each i^{th} row, and re-define this in terms of a predefined *perplexity*[2]:

$$perplexity(P_i) = 2^{H(P_i)} \tag{A.13}$$

   where $H(P_i)$ is the Shannon entropy of P_i:

$$H(P_i) = -\sum_j p(j|i) \log_2(p_{j|i}) \tag{A.14}$$

   and P_i represents the conditional probability distribution over all other data points given data points $\mathbf{x}_i$. The *perplexity* can be interpreted as a smooth measure of the effective number of neighbours. SNE optimises over σ_i to produce a P_i matching the predefined perplexity.

4. $t-$SNE computes the probabilities $p_{j|i}$ that are proportional to the similarity of $\mathbf{x}_i$ and $\mathbf{x}_j$. First, define the conditional probability of $j|i$

$$p_{j|i} = \frac{\exp\left(-||\mathbf{x}_i - \mathbf{x}_j||^2/(2\sigma_i^2)\right)}{\sum_{k \neq i} \exp\left(||\mathbf{x}_i - \mathbf{x}_k||^2/(2\sigma_i^2)\right)} \tag{A.15}$$

[2]Any σ_i induces a probability distribution, P_i, over all of the other datapoints, which has an corresponding and proportional entropy H.

where $i \neq j$; $p_{i|i} = 0$; σ_i is the variance of the Gaussian that is centred on datapoint $\mathbf{x}_i$.

5. The similarity matrix is calculated using the joint probability p_{ij} which then be defined by symmetrizing the conditional probabilities:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N} \tag{A.16}$$

6. Initially embed $\mathbf{X}$ in a low-dimensional space $t-$SNE aims to minimise the Kullback-Leibler (KL) divergence between the model Gaussian distribution of the points in $\mathbf{X}$ and a Student's $t-$distribution (with one degree of freedom) of points $\mathbf{Y}$ in the low-dimensional map. With this distribution, the joint probabilities q_{ij} are defined as:

$$q_{ij} = \frac{(1 + |||\mathbf{y}_i - \mathbf{y}_j||^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + ||\mathbf{y}_k - \mathbf{y}_l||^2)^{-1}} \tag{A.17}$$

where $q_{ii} = 0$. The KL divergence between the joint distribution P and Q is then:

$$KL(P||Q) = \sum_j \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{A.18}$$

using p_{ij} and q_{ij} from equations A.16 and A.17 respectively.

7. The Kullback-Leibler divergence becomes the loss function, which is minimised using a modified gradient descent (see section 3.8 for more details) with respect to $\mathbf{Y}$ to iteratively update the low-dimensional points:

$$\frac{\partial KL(P||Q)}{\partial \mathbf{y}_i} = 4 \sum_{j \neq i} Z(p_i - q_{ij}) q_{ij} (\mathbf{y}_i - \mathbf{y}_j) \tag{A.19}$$

where the normalisation term is defined as:

$$Z = \sum_k \sum_{l \neq k} \left(1 + ||\mathbf{y}_k - \mathbf{y}_l||^2\right)^{-1} \tag{A.20}$$

A.2.3 The SVM Dual Representation

This section outlines the formulation of the Support Vector Machines (SVM) dual representation which is useful for efficiently constructing and training an SVM (see section 3.7.1 for more details). First, consider the data: for the i^{th} observation row, $\mathbf{x}_i \in \mathbf{X}$, where $\mathbf{X} \in \mathbb{R}^{N \times P}$; and the response variable $\mathbf{y}_i \in \mathbf{y}$, where $\mathbf{y} \in \mathbb{R}^{N \times 1}$; and where P are the number of features and N are the number of observations. The SVM ***primal*** problem for estimating the parameters ($\mathbf{w}$, b) of a binary SVM can be specified as the following constrained quadratic program (see previous equation 3.50 for the construction of this objective function):

$$\begin{aligned} &\underset{\mathbf{w},b,\xi_\mathbf{i}}{\arg\min} \left\{ \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N} \xi_i \right\} \\ &\text{subject to}: \quad \xi_i \geq 0,\ y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i,\ \forall i \end{aligned} \tag{A.21}$$

where C is the regularisation cost parameter defining how "soft" the margin is yet allowing the maximal margin (i.e. how many violations (ξ_i) are permitted, where $\sum_i \xi_i \leq C$; points violate their margin by $M\xi_i$, whereas points on the correct side of the margin represent $\xi_i = 0$). For each sample, the primal requires two constraints:

$$\begin{aligned} \xi_i &\geq 0 \\ \xi_i &\geq 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b) \end{aligned}$$

$\xi_i \geq 0$, and $\xi_i \geq 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)$. We can multiply each constraint by $\alpha_i \geq 0$ and $\mu_i \geq 0$ respectively, and add them up to obtain:

$$(\alpha_i + \mu_i)\xi_i \geq \alpha_i(1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)) \tag{A.22}$$

letting $\alpha_i + \mu_i = C$ we can then express the lower bound of the objective function (A.21) as:

$$C\xi_i \geq \alpha_i(1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)) \tag{A.23}$$

$$\frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N} \xi_i \geq \frac{1}{2}||\mathbf{w}||^2 + \sum_{i=1}^{N} \alpha_i(1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)) \tag{A.24}$$

To minimise the lower bound of the objective function (i.e the right-hand side of $\geq$) we can differentiate with respect to $\partial/\partial\mathbf{w} = 0$ and $\partial/\partial b = 0$, resulting in:

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \tag{A.25}$$

$$\sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i = \mathbf{w} \tag{A.26}$$

Substituting the value of $\mathbf{w}$ (A.26) back into the lower bound of (A.24) provides us with the dual objective. This should be maximised to obtain the largest lower bound on the primal problem under the constraints that α_i and μ_i have to satisfy, therefore:

$$\begin{aligned} &\arg\max_{\boldsymbol{\alpha}} \left\{ \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{j,k=1}^{N} \alpha_j \alpha_k y_j y_k \mathbf{x}_j^\top \mathbf{x}_k \right\} \\ &\text{subject to :} \quad \sum_{i=1}^{N} \alpha_i y_i = 0, \\ &\qquad\qquad \alpha_i \geq 0,\ \mu_i \geq 0,\ \forall i, \\ &\qquad\qquad \alpha_i + \mu_i = C,\ \forall i, \end{aligned} \tag{A.27}$$

The constrain bound by μ_i is redundant and can be dropped, yielding a solution for the SVM ***(dual)*** Lagrangian form as:

$$\begin{aligned} &\arg\max_{\boldsymbol{\alpha}} \left\{ \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{j,k=1}^{N} \alpha_j \alpha_k y_j y_k \mathbf{x}_j^\top \mathbf{x}_k \right\} \\ &\text{subject to :} \quad \sum_{i=1}^{N} \alpha_i y_i = 0,\ 0 \leq \alpha_i \leq C,\ \forall i \end{aligned} \tag{A.28}$$

Finally, substituting (A.26) into the *primal* representation of the linear SVM (see equation 3.47), allows the SVM to be represented in its dual form:

$$\textbf{Primal:} \qquad f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b \tag{A.29}$$

$$\textbf{Dual:} \qquad f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i y_i (\mathbf{x}_i^\top \mathbf{x}) + b \tag{A.30}$$

A.2.4 Piecewise Linear Interpolation

Piecewise Linear Interpolation (PLI), was used in this study to impute missing test severity observations [395]. First, for any point z, given a set of $\mathbf{x}[1:N]$ and $\mathbf{y}[1:N]$ pairs on the interval partition $[\alpha, \beta]$, with a monotonic vector $\mathbf{x}$, i.e.

$x_k < x_{k+1}$, the index k is such that $x_k \leq z < x_{k+1}$ and $y_k = f(x_k)$ and $k = 1 : N$. The PLI is therefore built upon the local linear interpolants ($L_k(z)$) of f at the points $x = x_k$ and $x = x_{k+1}$, on each subinterval $[x_k, x_{k+1}]$:

$$L_k(z) = a_k + b_k(z - x_k) \tag{A.31}$$

where for $k = 1 : N - 1$, the coefficients are defined by:

$$a_i = y_k \qquad\qquad b_k = \frac{y_{k+1} - y_k}{x_{i+1} - x_i} \tag{A.32}$$

The concatenation of linear interpolants between each pair of data points results in a continuous curve, with a discontinuous derivative (C^0):

$$L(z) = \begin{cases} L_1(z), & \text{if } x_1 \leq z < x_2 \\ L_2(z), & \text{if } x_2 \leq z < x_3 \\ & \vdots \\ L_{n-1}(z), & \text{if } x_{n-1} \geq z \leq x_n \end{cases} \tag{A.33}$$

L can therefore be evaluated at a point $z \in [\alpha, \beta]$, one just needs to to determine the subinterval that contains z.

A.2.5 Shape-Preserving Piecewise Cubic Hermite Interpolation

Shape-Preserving Piecewise Cubic Hermite Interpolation(PCHIP) [396] interpolates using a piecewise cubic polynomial $P(x)$. PCHIP divides the interval into smaller sub-intervals $x_k \leq x \leq x_{k+1}$, and constructs different low degree polynomial $P(x)$ approximations (with specified derivatives at the interpolation points) on the sub-intervals. $P(x)$ approximates f as $P(x_i) = f(x_i)$ and the first derivative $\frac{dP}{dx}$ is continuous C^1. PCHIP computes the derivatives by imposing local monotonicity in each interval $x_k x_k + 1)$, which preserves the shape of the data.

A.2.6 The Box-Cox Transformation

The Box-Cox transformation is a monotonic power transform which can be used to shift non-normally distributed data such that it follows an approximately normal

distribution [397, 398]. The Box–Cox transformation of $\mathbf{x} \in \mathbb{R}^{N\times 1}$ can be defined by

$$\mathbf{x}^{(\lambda)} = \begin{cases} \dfrac{\mathbf{x}^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln \mathbf{x} & \text{if } \lambda = 0, \end{cases} \tag{A.34}$$

The optimal λ value can be found by maximising the log-likelihood as described in section 3.5.3.

A.2.7 Orientation Independent Transformation

In order to mitigate the effects of smartphone placements in varying orientations, which can occur within a pocket or a running belt during a gait assessment, an orientation independent transformation can be applied in order to align the smartphone coordinate system with gravity and the direction of motion. As such, all smartphone gait assessments are projected to a global coordinate system that remains fixed, regardless of the original placement of the device. Following an orientation independent transformation described by Gadaleta *et al.* [184], which is similar to previous studies [399–401], Principal Component Analysis (PCA) is used to determine the direction with the largest variance, assumed to be parallel to the direction of motion (based on [399]). Orthogonal versors[3] are determined, ($\hat{\boldsymbol{e}}_{\boldsymbol{x}}$, $\hat{\boldsymbol{e}}_{\boldsymbol{y}}$, $\hat{\boldsymbol{e}}_{\boldsymbol{z}}$) which are aligned with the direction of gravity; such that $\hat{\boldsymbol{e}}_{\boldsymbol{x}}$ are aligned with the anterior-posterior axis (direct of motion); $\hat{\boldsymbol{e}}_{\boldsymbol{y}}$ are orthogonal to the vertical axis; $\hat{\boldsymbol{e}}_{\boldsymbol{z}}$ aligns with the medial-lateral axis.

First, the accelerometer sensor axes is described by ($\boldsymbol{a}_x, \boldsymbol{a}_y, \boldsymbol{a}_z$), with an acceleration matrix $\mathbf{A} = [\boldsymbol{a}_x, \boldsymbol{a}_y, \boldsymbol{a}_z]^{\top}$. The average component of gravity is determined; letting the mean vector direction of gravity be represented by the average acceleration for each axis: $\vec{\mathbf{g}} = [\overline{\boldsymbol{a}}_x, \overline{\boldsymbol{a}}_y, \overline{\boldsymbol{a}}_z]^{\top}$. The gravitational aligned versor $\hat{\boldsymbol{e}}_{\boldsymbol{y}}$, in the vertical direction, is therefore defined by:

$$\hat{\boldsymbol{e}}_{\boldsymbol{y}} = \frac{\vec{\mathbf{g}}}{||\vec{\mathbf{g}}||} \tag{A.35}$$

[3] In this context versors are unit vectors to represent the axes of a Cartesian coordinate system.

The projected acceleration $\boldsymbol{a}_{\hat{\boldsymbol{e}}_{\boldsymbol{y}}}$ across the vertical axis versor $\hat{\boldsymbol{e}}_{\boldsymbol{y}}$ is therefore described by:

$$\boldsymbol{a}_{\hat{\boldsymbol{e}}_{\boldsymbol{y}}} = \mathbf{A} \cdot \hat{\boldsymbol{e}}_{\boldsymbol{y}} \tag{A.36}$$

The gravity-independent acceleration matrix $\mathbf{A}' = [\boldsymbol{a}'_x, \boldsymbol{a}'_y, \boldsymbol{a}'_z]^\top$ can then be corrected by:

$$\mathbf{A}' = \mathbf{A} - \hat{\boldsymbol{e}}_{\boldsymbol{y}} \cdot \boldsymbol{a}^\top_{\hat{\boldsymbol{e}}_{\boldsymbol{y}}} \tag{A.37}$$

Next, Principal Component Analysis (PCA) is applied to $\mathbf{A}'$ in order to determine the acceleration and directions of variance that are independent from gravitational effects. The matrix $\mathbf{A}'$ is first centred, where the eigenvalues and eigenvectors of the corresponding covariance matrix $\boldsymbol{\Sigma}$ of $\mathbf{A}'$ are evaluated using PCA (as described previously in section A.2.1). The eigenvector $\vec{\boldsymbol{\nu}}$ associated with the maximum eigenvalue (i.e. the first principal component) therefore identifies the direction of maximum variance. This direction is the anterior-posterior axis (i.e. the direction of motion, as found in [399]) and can be defined through the versor $\hat{\boldsymbol{e}}_{\boldsymbol{x}}$:

$$\hat{\boldsymbol{e}}_{\boldsymbol{x}} = \frac{\vec{\boldsymbol{\nu}}}{||\vec{\boldsymbol{\nu}}||} \tag{A.38}$$

The medial-lateral axis versor $\hat{\boldsymbol{e}}_{\boldsymbol{z}}$ is finally obtained through a cross product of:

$$\hat{\boldsymbol{e}}_{\boldsymbol{z}} = \hat{\boldsymbol{e}}_{\boldsymbol{y}} \times \hat{\boldsymbol{e}}_{\boldsymbol{x}} \tag{A.39}$$

where $\hat{\boldsymbol{e}}_{\boldsymbol{x}}$ and $\hat{\boldsymbol{e}}_{\boldsymbol{y}}$ are orthogonal. Again, the orientation transformed accelerometer sensor data vectors are obtained by projecting along versors $\hat{\boldsymbol{e}}_{\boldsymbol{x}}$ and $\hat{\boldsymbol{e}}_{\boldsymbol{z}}$ respectively:

$$\boldsymbol{a}_{\hat{\boldsymbol{e}}_{\boldsymbol{x}}} = \mathbf{A} \cdot \hat{\boldsymbol{e}}_{\boldsymbol{x}} \tag{A.40}$$

$$\boldsymbol{a}_{\hat{\boldsymbol{e}}_{\boldsymbol{z}}} = \mathbf{A} \cdot \hat{\boldsymbol{e}}_{\boldsymbol{z}} \tag{A.41}$$

A.3 Signal Decompositions

A.3.1 The Discrete Wavelet Transform (DWT)

The Discrete Wavelet Transform (DWT) generates a sparse representation of a signal as a series of wavelet 'basis functions', which are spatially localised (non-zero

over a finite period), and derived from dilated, translated, and scaled versions of a "*mother wavelet*". Considering a signal $\mathbf{x} \in \mathbb{R}^{1\times N}$ (of any given length, N) that is convolved with a filter $\mathbf{g} \in \mathbb{R}^{1\times M}$ of length $M < N$, the resulting signal $\mathbf{y}[n]$ is:

$$\mathbf{y}[n] = (\mathbf{x} * \mathbf{g})[n] = \sum_{k=-L}^{L} \mathbf{x}[k]\mathbf{g}[n-k] \tag{A.42}$$

The DWT provides a decomposition to analyse a signal across various frequency bands with different resolutions using coarse approximation coefficients and detail information detail coefficients. A DWT can obtain a multi-resolution representation of $\mathbf{x}$ through a series of cascading filter banks (as shown in figure A.5). At each level L, a series of low-pass $\mathbf{g}[\cdot]$ and high-pass $\mathbf{h}[\cdot]$ filters are applied, evenly bisecting the frequency into its constituent low- and high-frequency components at that level using a convolution between the signal and filter as per equation A.42. After each filter operation, the resulting signal is down-sampled by a factor of 2 (dyadic decimation), where $c\boldsymbol{D}_j$ and $c\boldsymbol{A}_j$ represent the respective detail and approximation coefficients at level $j = 1, 2, ..., L$. The result is the decomposition of $\mathbf{x}$ into multiple frequency bands, where after each level, the frequency resolution is doubled while the time resolution is halved. Therefore the decomposed signal y at the j^{th} level, of length 2^N, roughly corresponds to a signal in the pass-band $\left[\frac{\pi}{2^j}, \frac{\pi}{2^{j-1}}\right]$, and coefficients in the range $\left[2^{N-j}, 2^{N-j+1}\right]$. Many different "*mother wavelets*" exist (such as Daubechies, Symlets and Coiflets), with each having unique combinations of low-pass and high-pass filters coefficients. Debauchies-4 (db4) wavelets have

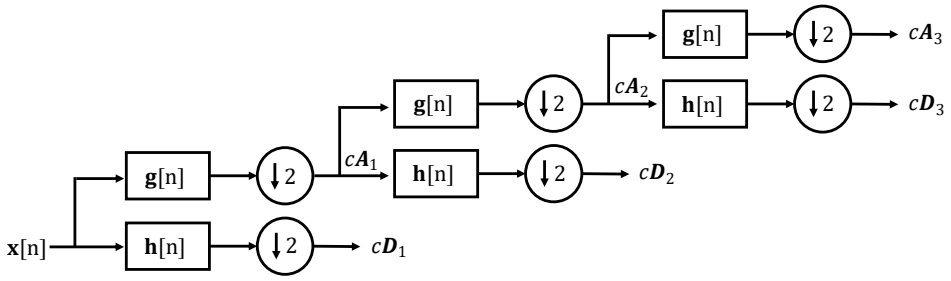
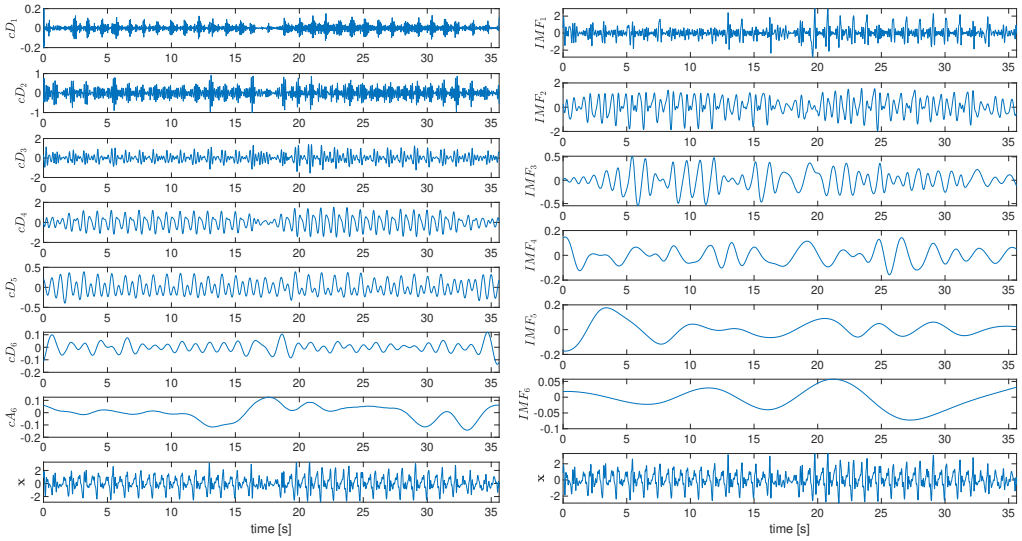


Figure A.5: The Discrete Wavelet Transform (DWT) filterbank. Diagram depicting a cascading DWT filter bank consisting of a series of low-pass $\mathbf{g}[\cdot]$ and high-pass $\mathbf{h}[\cdot]$ filters applied to a signal $\mathbf{x} \in \mathbb{R}^{1\times N}$ for 3 levels. After each filter operation, the resulting signal is down-sampled by a factor of 2 (dyadic decimation), where $c\boldsymbol{D}_j$ and $c\boldsymbol{A}_j$ represent the respective detail and approximation coefficients at level $j = 1, 2, ..., L$.



(a) DWT decomposition **(b)** EMD decomposition

Figure A.6: The multi-resolution decomposition of a gait-signal. Comparison of the multi-resolution decomposition of a sample gait signal $\mathbf{x}$ using **(a)** DWT detail $c\boldsymbol{D}_j$ and approximation $c\boldsymbol{A}_j$ coefficients versus **(b)** EMD represented by a series of IMFs.

successfully distinguished gait-related events from gait-freezing events in Parkinson' disease using similar wearable sensor-based accelerometry data and was hence used in this study [402]. The db4 form of the orthogonal low-pass ($\mathbf{g}[\cdot]$) and high-pass ($\mathbf{h}[\cdot]$) filters are defined as:

$$\mathbf{g}[\cdot] = \left[\frac{1+\sqrt{3}}{4\sqrt{2}}, \frac{3+\sqrt{3}}{4\sqrt{2}}, \frac{3-\sqrt{3}}{4\sqrt{2}} \frac{1-\sqrt{3}}{4\sqrt{2}}\right] \tag{A.43}$$

$$\mathbf{h}[\cdot] = \left[\frac{1-\sqrt{3}}{4\sqrt{2}}, \frac{-3+\sqrt{3}}{4\sqrt{2}}, \frac{3+\sqrt{3}}{4\sqrt{2}} \frac{-1-\sqrt{3}}{4\sqrt{2}}\right] \tag{A.44}$$

Figure A.6a illustrates a gait sample, $\mathbf{x} \in \mathbb{R}^{1\times N}$, decomposed using DWT for $j = 6$ levels using a db4 wavelet. A multi-resolution decomposition is obtained, where each j^{th} level detail coefficients $c\boldsymbol{D}_j$ contain gait signal information between descending frequency bands.

A.3.2 Empirical Mode Decomposition (EMD)

Empirical mode decomposition (EMD) decomposes a signal $\mathbf{x}(\mathbf{t})$ into intrinsic mode functions (IMFs) which correspond the different frequency bands[4]. The

[4]For a more expansive description of the EMD procedure and IMFs we refer the reader to [307, 403, 404].

concept underpinning EMD is the "*shifting*" process, in which a signal $\mathbf{x}(\mathbf{t})$ is considered as a fast oscillation superimposed on a slower one. Each oscillatory IMF component consists of different frequency bands in descending order that are extracted recursively [307, 403, 404]:

1. Find the local minima and maxima of $\mathbf{x}(\mathbf{t})$. Construct lower and upper envelopes $s_{-}(t)$ and $s_{+}(t)$ of $\mathbf{x}(t)$. Obtain the residual: $\boldsymbol{h}(t) = \mathbf{x}(t) - \bar{\mathbf{x}}(t)$, where $\bar{\mathbf{x}}(\mathbf{t})$ is the mean.

2. Denote $\boldsymbol{r}_0(t) = \mathbf{x}(\mathbf{t})$ and let $i = 0$. Before sifting, check $\boldsymbol{r}_i(t)$ against the conditions:

   (a) Find the sum of local extrema of $\boldsymbol{h}_i(t)$ and zero-crossings zcr (see equation A.55), which must differ by at most one.

   (b) At any point, the mean between the $s_{-}(t)$ and $s_{+}(t)$ envelopes of $\boldsymbol{h}_i(t)$ are zero.

3. Discontinue the procedure if an source stopping condition is met (for instance based on number of extremia or an energy ratio [403]).

4. If $h_i(t)$ satisfies the requirements of step 2, it becomes IMF_1. Otherwise, steps 1–2 are recursively repeated, where $\mathbf{x}(\mathbf{t}) = \boldsymbol{h}_i(t)$, and recursive component $\boldsymbol{h}_{ik}$ is extracted and assessed against the criteria in step 2. The most common stoppage criteria for shifting steps is the *Sift Relative Tolerance*, where the sum of the differences, $\Sigma\Delta_k$ is defined fro the k^{th} sifted component by:

$$\Sigma\Delta_k = \sum_{i=1}^{N} \frac{|\boldsymbol{h}_{1(k-1)i} - \boldsymbol{h}_{1ki}|^2}{\boldsymbol{h}_{1(k-1)i}^2} \tag{A.45}$$

   where $\boldsymbol{h}_{1ki}$ denotes the i^th sample in $\boldsymbol{h}_{1k}$. Sifting terminates if $\Sigma\Delta_k > d$, where typically a threshold $d = 0.3$ is employed, and IMF_1 is set to $\boldsymbol{h}_{1k}$.

5. A new residual signal $\boldsymbol{r_i}$ is defined by $r_i = \mathbf{x}(t) - IMF_1$, with $i = i + 1$, and the processes from step 1 are repeated to find the next IMF_2 and so forth.

As a result of the decomposition procedure, the first IMFs contain the highest frequency information in descending order of IMF to the lowest. This paradigm is illustrated for a gait sample decomposed into IMFs using EMD in figure A.6b.

A.4 Feature Extraction

This section mathematically describes the features that were extracted from the Draw a Shape Test relating to chapter 5; and the gait features that were extracted on the 2MWT relating to chapter 6.

A.4.1 Chapter 6: Gait Feature Space

Features were computed on accelerometer sensor axes ($\boldsymbol{a}_x, \boldsymbol{a}_y, \boldsymbol{a}_z$) and also on the orientation invariant signal magnitude, for example $\|\boldsymbol{a}\| = (\mathbf{x}^2 + \mathbf{y}^2 + \mathbf{z}^2)^{\frac{1}{2}}$, where $\mathbf{x} = (a_{x_1}, a_{x_2}, ..., a_{x_T})$ and so forth. Table A.1 details the gait-related features extracted and the corresponding signal used to calculate each.

Table A.1: Description of the features extracted from the Two-Minute Walk Test

Feature	Description	Parameterisation
$steps(\|\boldsymbol{a}\|)$	The number of steps counted, as per [299].	See [299].
$mean(\boldsymbol{a}_x)$ $mean(\boldsymbol{a}_y)$ $mean(\boldsymbol{a}_y)$ $mean(\|\boldsymbol{a}\|)$	The mean value of vector $\mathbf{x} \in \mathbb{R}^{1\times N}$: $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ (A.46)	
$var(\boldsymbol{a}_x)$ $var(\boldsymbol{a}_y)$ $var(\boldsymbol{a}_z)$ $var(\|\boldsymbol{a}\|)$	The variance which characterises the spread of a set of (random) numbers from their average value. For a random variable vector $\boldsymbol{x} \in \mathbb{R}^{1\times N}$, the variance is defined as: $\sigma = \frac{1}{N-1}\sum_{i=1}^{N} \|x_i - \overline{x}\|^2$ (A.47) where $\overline{x}$ is the mean of $\boldsymbol{x}$ (A.46).	
$std(\boldsymbol{a}_x)$ $std(\boldsymbol{a}_y)$ $std(\boldsymbol{a}_z)$ $std(\|\boldsymbol{a}\|)$	For $\boldsymbol{x} \in \mathbb{R}^{1\times N}$, the standard deviation is defined as: $std = \sqrt{\sigma^2}$ (A.48)	
$p_{25}(\|\boldsymbol{a}\|)$	The 25^{th} percentile value of $\boldsymbol{x} \in \mathbb{R}^{1\times N}$.	
$p_{75}(\|\boldsymbol{a}\|)$	The 75^{th} percentile value of $\boldsymbol{x} \in \mathbb{R}^{1\times N}$.	
$iqr(\|\boldsymbol{a}\|)$	The interquartile range is the difference between 75^{th} and 25^{th} percentiles of $\boldsymbol{x} \in \mathbb{R}^{1\times N}$.	
$med(\|\boldsymbol{a}\|)$	The median value of $\boldsymbol{x} \in \mathbb{R}^{1\times N}$.	
$mode(\|\boldsymbol{a}\|)$	The mode of $\boldsymbol{x} \in \mathbb{R}^{1\times N}$.	
$range(\|\boldsymbol{a}\|)$	The range, $\|\max(x) - \min(x)\|$, of $\boldsymbol{x} \in \mathbb{R}^{1\times N}$.	

Continued on Next Page...

Table A.1: Description of the features extracted from the Two-Minute Walk Test

Feature	Description	Parameterisation
$skew(\boldsymbol{a}_x)$ $skew(\boldsymbol{a}_y)$ $skew(\boldsymbol{a}_y)$ $skew(\|\boldsymbol{a}\|)$	Skewness as a measure of the asymmetry of the probability distribution of the signal values $\boldsymbol{x} \in \mathbb{R}^{1\times N}$. The relative magnitude of how far a distribution deviates from the normal which used as a proxy for smooth stable ambulatory movements. Positive values indicate that the mass of the distribution is predominantly skewed to the left, while the inverse is true for negative values. Skewness is defined as [405]: $$skew = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}{\left[\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2\right]^{3/2}} \quad (A.49)$$ where x_i denotes each i^{th} sample in $\boldsymbol{x}$ and $\overline{x}$ denotes the mean of the samples in $\boldsymbol{x}$.	
$kurt(\boldsymbol{a}_x)$ $kurt(\boldsymbol{a}_y)$ $kurt(\boldsymbol{a}_y)$ $kurt(\|\boldsymbol{a}\|)$	Kurtosis measures the level of 'taildness' in the probability distribution of the magnitude of $\boldsymbol{x} \in \mathbb{R}^{1\times N}$ signal values, compared to the normal distribution, and is used as a measure of how outlier prone a distribution is. Values > 3 indicate tails that exponentially decay at a slower rate to the normal distribution (more outliers present) and vice-versa. Kurtosis is defined by [405]: $$kurt = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^4}{\left[\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2\right]^{2}} \quad (A.50)$$ where x_i denotes each i^{th} sample in $\boldsymbol{x}$ and $\overline{x}$ denotes the mean of the samples in $\boldsymbol{x}$.	
$H(\boldsymbol{a}_x)$ $H(\boldsymbol{a}_y)$ $H(\boldsymbol{a}_z)$ $H(\|\boldsymbol{a}\|)$	For $\boldsymbol{x} \in \mathbb{R}^{1\times N}$, the "Shannon" (non-normalised) Entropy (H) is defined as [406]: $$H(x) = -\sum_{i=1}^{N} x_i^2 \log(x_i^2) \quad (A.51)$$	
$ApEn(\|\boldsymbol{a}\|)$	Approximate entropy characterises the irregularity and subtle fluctuations in a time series $\boldsymbol{x} \in \mathbb{R}^{1\times N}$. The sequence is first divided into m length consecutive segments (templates), which are recursively compared to all others. The total number of similar templates are determined $C_i^{(m)}$ where pairs i of templates are defined as similar by calculating the Chebyshev distance, the maximum absolute difference between each pair of respective template values $< r$, a threshold value. ApEn is then defined as: $$ApEn = \Phi^m - \Phi^{m+1}, \quad (A.52a)$$ $$\Phi^m(r) = \frac{\sum_{i=1}^{N-m+1} \ln C_i^m(r)}{N-m+1} \quad (A.52b)$$ $ApEn \to 0$ denotes similar signal behaviour after $m = m+1$, implying greater regularity within $\boldsymbol{x}$. See Pincus (1991) for more detailed description of $ApEn$ [407].	See [65].

Continued on Next Page...

Table A.1: Description of the features extracted from the Two-Minute Walk Test

Feature	Description	Parameterisation
$H(\boldsymbol{a}_y\|\boldsymbol{a}_z)$ $H(\boldsymbol{a}_x\|\boldsymbol{a}_z)$ $H(\boldsymbol{a}_y\|\boldsymbol{a}_x)$	The conditional entropy $H(Y\|X)$ of the medio-lateral and vertical, anterior-posterior and vertical, and medio-lateral and anterior-posterior axis. $H(Y\|X)$ quantifies the amount of information needed to describe the outcome of a random variable Y given that the value of another random variable X is known. The conditional entropy of Y given X is defined as [406]: $$H(Y\|X) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)} \quad \text{(A.53)}$$ where $\mathcal{X}$ and $\mathcal{Y}$ denote the support sets of X and Y.	
$RPDE(\|\boldsymbol{a}\|)$	Recurrence period density entropy ($RPDE$) is a method used to characterise the deviations from exact periodicity and stochasticity within a signal [309], and was proposed here to capture the ability to maintain consistent gait rhythm. The time series is first embedded in phase space such that: $\mathbf{x}_n = [x_n, x_{n+\tau}, x_{n+2\tau}, \ldots, x_{n+(M-1)\tau}]$ for each value of x_n, where M is the embedding dimension, and τ is the embedding delay. Then, around each point $\mathbf{x}_n$ in the embedded phase space, a recurrence neighbourhood of radius ε is formed [408]. Each recurrence of the time series through this neighbourhood is recorded, and the time difference T between successive returns is recorded in a histogram. The normalised entropy of this recurrence period density function $P(T)$ is then computed as: $$H_{norm} = -(\ln T_{\max})^{-1}\sum_{t=1}^{T_{\max}} P(t)\ln P(t) \quad \text{(A.54)}$$ is the RPDE value, where $T_{\max}$ is the largest recurrence value. For more information we refer readers to [309].	
$\max(jerk(\|\boldsymbol{a}\|))$ $\min(jerk(\|\boldsymbol{a}\|)$ $mean(jerk(\|\boldsymbol{a}\|))$ $std(jerk(\|\boldsymbol{a}\|))$ $skew(jerk(\|\boldsymbol{a}\|))$ $kurt(jerk(\|\boldsymbol{a}\|))$	The maximum, minimum, mean, variability, skewness and kurtosis of the jerk values. Jerk is the third derivative, $\frac{d^3x}{dt^3}$ of $\boldsymbol{x} \in \mathbb{R}^{1\times N}$ and helps characterises sharp sensor movements.	
$zcr(\boldsymbol{a}_x)$ $zcr(\boldsymbol{a}_y)$ $zcr(\boldsymbol{a}_z)$ $zcr(\|\boldsymbol{a}\|)$	The zero-crossing rate calculates the rate of sign-changes in a signal, roughly capturing the static-to-dynamic transitions within gait. Assuming a zero-mean signal $\boldsymbol{x} \in \mathbb{R}^{1\times N}$, zcr can be defined as [409]: $$zcr = \frac{1}{N-1}\sum_{t=1}^{N-1} \mathbb{1}_{\mathbb{R}_{<0}}(x_t, x_{t-1}) \quad \text{(A.55)}$$ where $\mathbb{1}_{\mathbb{R}_{<0}}$ is an indicator function.	

Continued on Next Page...

Table A.1: Description of the features extracted from the Two-Minute Walk Test

Feature	Description	Parameterisation
$\max(\widehat{P}(f))$	The maximum power spectral density (PSD) of the sequence $\boldsymbol{x} \in \mathbb{R}^{1\times N}$, in this case given as $\|\boldsymbol{a}\|$. The PSD estimate of $\boldsymbol{x} \in \mathbb{R}^{1\times N}$ is performed using the modified periodogram, calculated using Welch's overlapped segment averaging estimator, with a Hamming window function h_n. The modified periodogram PSD is defined by: $$\widehat{P}(f) = \frac{\Delta t}{N}\left|\sum_{n=0}^{N-1} h_n x_n e^{-j2\pi f n}\right|^2 \quad \text{(A.56)}$$ where the periodogram is the Fourier transform of the biased estimate of the autocorrelation sequence for the n^{th} sample x_n, sampled at f_s per unit time. Δt is the sampling interval and $-\frac{1}{2\Delta t} < f \leq \frac{1}{2\Delta t}$	
$\arg\max_f(\widehat{P}(f))$	The dominant frequency, f (Hz), in the sequence $\boldsymbol{x} \in \mathbb{R}^{1\times N}$, in this case given as $\|\boldsymbol{a}\|$.	
$\max(\widehat{P}(f_g))$	The maximum power spectral density (PSD) of the sequence $\boldsymbol{x} \in \mathbb{R}^{1\times N}$ within the gait domain such that $f_g \in [0.5 - 3.5]$ Hz. $\boldsymbol{x}$ is given as $\|\boldsymbol{a}\|$.	
$\arg\max_{f_g}(\widehat{P}(f_g))$	The dominant frequency, f_g (Hz), in the sequence $\boldsymbol{x} \in \mathbb{R}^{1\times N}$, within the gait domain such that $f_g \in [0.5 - 3.5]$ Hz.	
$DFA(\|\boldsymbol{a}\|)$	Detrended Fluctuation Analysis is a method for determining the statistical rate of decay of the autocorrelation in $\boldsymbol{x} \in \mathbb{R}^{1\times N}$ [410] and has been used to compare alterations in gait dynamics in healthy and diseased gait [411–413].	Parameters given by [412].
$\max(r(\|\boldsymbol{a}\|))$	The maximum autocorrelation within the time-series $\boldsymbol{x} \in \mathbb{R}^{1\times N}$ [414]. The autocorrelation function measures the correlation between x_t and x_{t+k}, where $k = 0, \ldots, K$. The autocorrelation for lag k is: $$r_k = \frac{c_k}{c_0} \quad \text{(A.57a)}$$ $$c_k = \frac{1}{T}\sum_{t=1}^{T-k}(x_t - \bar{\boldsymbol{x}})(x_{t+k} - \bar{\boldsymbol{x}}) \quad \text{(A.57b)}$$ where $\bar{\boldsymbol{x}}$ is the mean of $\boldsymbol{x}$; c_0 is the sample variance (A.47) of the time-series.	
$r_1(\|\boldsymbol{a}\|))$	The autocorrelation coefficient, r at time lag 1 within the time-series [171, 414].	
$MI(\boldsymbol{a}_z, \boldsymbol{a}_y)$	The mutual information [415] between the anterio-posterior and medio-lateral sensor axis.	
$\rho(\boldsymbol{a}_x, \boldsymbol{a}_y)$	Spearman's correlation between the anterio-posterior and vertical sensor axis.	
$\rho(\boldsymbol{a}_y, \boldsymbol{a}_z)$	Spearman's correlation between the vertical and medio-lateral sensor axis.	
$\rho(\boldsymbol{a}_x, \boldsymbol{a}_z)$	Spearman's correlation between the anterio-posterior and medio-lateral sensor axis.	

Continued on Next Page...

Table A.1: Description of the features extracted from the Two-Minute Walk Test

Feature	Description	Parameterisation
$\max(\widehat{f_h}(\|\boldsymbol{a}\|))$	The maximum Kernel density estimation (KDE) value [161, 162]. For any values of $\boldsymbol{x} \in \mathbb{R}^{1\times N}$, the KDE is given by: $$\widehat{f_h}(\boldsymbol{x}) = \frac{1}{hN}\sum_{i=1}^{N} K\left(\frac{\boldsymbol{x} - x_i}{h}\right) \quad \text{(A.58)}$$ where $[x_1, x_2, \ldots, x_N]$ are random samples from an unknown distribution, N is the sample size, $K(\cdot)$ is the kernel smoothing function, and h is the bandwidth.	
$\arg\max_{\|\boldsymbol{a}\|}(\widehat{f_h}(\|\boldsymbol{a}\|))$	The value of x that maximises the KDE (A.58).	
$\overline{cD_j}(\|\boldsymbol{a}\|)$ $std(cD_j(\|\boldsymbol{a}\|))$ $\max(cD_j(\|\boldsymbol{a}\|))$ $\min(cD_j(\|\boldsymbol{a}\|))$	The average, standard deviation, maximum and minimum values of the j^{th} DWT detail cD_j coefficients. A sparse representation of gait signals were obtained using the Discrete Wavelet Transform (DWT), where the signal was decomposed into a number of different bandwidths expressed by approximation cA_j and detail cD_j coefficients at level $j = 1, 2, 3, \ldots L$. See section A.3.1 for more information on the DWT.	Daubechies-4 (db4) [305]
$E(cA_j(\|\boldsymbol{a}\|))$ $E(cD_j(\|\boldsymbol{a}\|))$	Energy E of the j^{th} DWT approximation cA_j and detail cD_j coefficient. The energy is defined as: $$E(y) = \sum_{i=1}^{N} \|y_i\|^2 \quad \text{(A.59)}$$ where $y = cD_j$; $y = cA_j$; are the detail and approximation coefficients at level $j = 1, 2, 3, \ldots L$.	
$H(cA_j(\|\boldsymbol{a}\|))$ $H(cD_j(\|\boldsymbol{a}\|))$	Entropy H of the j^{th} DWT approximation cA_j and detail cD_j coefficients (see equation A.51).	
$\overline{\Phi}(cA_j(\|\boldsymbol{a}\|))$ $\overline{\Phi}(cD_j(\|\boldsymbol{a}\|))$ $std(\Phi(cA_j(\|\boldsymbol{a}\|)))$ $std(\Phi(cD_j(\|\boldsymbol{a}\|)))$ $\max(\Phi(cA_j(\|\boldsymbol{a}\|)))$ $\max(\Phi(cD_j(\|\boldsymbol{a}\|)))$	The average, standard deviation and maximum value of the Teager-Kaiser Energy Operator (TKEO, Φ) of the j^{th} DWT approximation cA_j and detail cD_j coefficients. TKEO is calculated directly in the time-domain, however it incorporates both amplitude- and frequency-specific information by measuring local differences between adjacent signal samples and is defined as [416]: $$\Phi(x_i) = x_i^2 - x_{i+1} \cdot x_{i-1} \quad \text{(A.60)}$$ where i denotes the index of the input vector.	
E_{pk}	The maximum scale-dependent energy density E_s (equation 6.3) of the Continuous Wavelet Transform (CWT), $E(CWT)$ (see equation 6.2a for more details). $$E_{pk} = \max([E_1, \ldots, E_S]) \quad \text{(A.61)}$$ where S are the number of scales.	
$\arg\max_{f}(E(CWT))$	The frequency (f_{max}), in Hz, which maximises $E(CWT)$ over all scales s (6.3). See equations 6.4 for more details on the approximation of f.	
$\sum_S E_s(CWT)$	The total scale dependent energy E over all scales s (6.3).	
$\overline{E}_s$	The average E value over all scales s (6.3).	

Continued on Next Page...

Table A.1: Description of the features extracted from the Two-Minute Walk Test

Feature	Description	Parameterisation
$std(E(CWT))$	The standard deviation in the values of E over all scales s (6.3).	
$skew(E(CWT))$	Skewness (A.49) as a measure of the asymmetry of the probability distribution of the scale dependent energy distribution E over all scales s (6.3).	
$kurt(E(CWT))$	The kurtosis in the values of E over all scales s (6.3).	
$AUC(E(CWT))$	The area under the curve (AUC) approximate integral of E over all scales s (6.3), with with respect to the frequencies f, using the trapezoidal method as defined by: $$AUC(E(CWT)) = \int_f E_s dS \approx \frac{\Delta f}{2N} \sum_{s=1}^{S} (E_{s-1} + E_s) \quad \text{(A.62)}$$ where s represents the scale factor, S are the number of scales and the spacing between each point is equal to the scalar value $\frac{\Delta f}{N}$.	
$prom(E_{pk})$	The prominence of the maximum scale-dependent energy density peak E_{pk} (A.61). Prominence measures how much a peak stands out due to its intrinsic height and its location relative to other peaks.	
$\frac{E_{pk_1}}{E_{pk_2}}$	The ratio of the maximum scale-dependent energy peak density to next highest peak. $$\frac{E_{pk_1}}{E_{pk_2}} = \frac{max_1(E(CWT))}{max_2(E(CWT))} \quad \text{(A.63)}$$	
$width(E_{pk})$	The width E_w of the maximum scale-dependent energy density E_{pk}. The width of the peak is estimated as the distance between the points where the descending edges intercept a horizontal reference line beneath the peak at a vertical distance equal to half the peak prominence ($prom$).	
$\frac{E_{w_1}}{E_{w_2}}$	The ratio of the width of the maximum scale-dependent energy peak density to next highest peak. $$\frac{E_{w_1}}{E_{w_2}} = \frac{width\,(max_1(E(CWT)))}{width\,(max_2(E(CWT)))} \quad \text{(A.64)}$$	
$\frac{f_{max_1}}{f_{max_2}}$	The ratio of the frequency that maximises the scale-dependent energy peak density, f_{max}, to the frequency that maximises the next highest peak.	

Continued on Next Page...

Table A.1: Description of the features extracted from the Two-Minute Walk Test

Feature	Description	Parameterisation
$MsEn_{\tau}(\|\boldsymbol{a}\|)$	Multiscale entropy ($MsEn$) calculates the sample entropy ($SampEn$) of a signal at increasingly coarser grains (scales). For more details on $SampEn$ we refer the reader to [417]. $MsEn$ characterises dynamic complexity of gait within a signal [308, 418]. Signals are first segmented by taking the mean of data points in non-overlapping windows of increasing length, τ. Given a time series, $\boldsymbol{x} \in \mathbb{R}^{1\times N}$, we construct consecutive coarse-grained time series by averaging a successively increasing number of data points in non-overlapping windows. Each element of the coarse-grained time series $y_k^{(\tau)}$ is computed as: $$y_j^{(\tau)} = \frac{1}{\tau}\sum_{i=(j-1)\tau+1}^{j\tau} x_i \quad \text{(A.65)}$$ where τ represented the scale factor and $1 \leq j \leq N/\tau$. The length of each coarse-grained time series is N/τ. $MsEn_{\tau}(x)$ therefore denotes the sample entropy ($SampEn$) [419] calculated for each coarse-grained time series, $y^{(\tau)}$, defined by τ. Lower values of $SampEn$ indicate more self-similarity in the time series.	As per [308]: Maximum scale factor $\tau = 20$; Embedding dimension $m = 2$; Tolerance $r = 0.2$;
$\overline{MsEn_{(:)}}(\|\boldsymbol{a}\|)$	Average multiscale mntropy ($MsEn$) over all timescales.	
$std(MsEn_{(:)}(\|\boldsymbol{a}\|))$	Standard deviation in multiscale entropy ($MsEn$) over all scales.	
$AUC(MsEn_{(:)}(\|\boldsymbol{a}\|))$	The area under the curve (AUC) of $MsEn$ over all timescales. AUC is computed using numerical integration via the trapezoidal method such that $$AUC = \frac{b-a}{2N}\sum_{\tau=1}^{S}(MsEn_{(\tau-1)} + MsEn_{(\tau)}) \quad \text{(A.66)}$$ where τ represents the scale factor, S are the number of scales and the spacing between each point is equal to the scalar value $\frac{b-a}{N}$, in this case 1.	
$p_{25}(MsEn_{(:)}(\|\boldsymbol{a}\|))$	25^{th} percentile of $MsEn$ values over all scales.	
$p_{75}(MsEn_{(:)}(\|\boldsymbol{a}\|))$	75^{th} percentile of $MsEn$ values over all scales.	
$p_{25}(...$ $AUC(MsEn_{(:)}(\|\boldsymbol{a}\|)))$	25^{th} percentile in AUC values computed from $MsEn$ over all scales (A.66).	
$p_{75}(...$ $AUC(MsEn_{(:)}(\|\boldsymbol{a}\|)))$	75^{th} percentile in AUC values computed from $MsEn$ over all scales (A.66).	
$\frac{MsEn(\|\boldsymbol{a}\|)_{(:10)}}{MsEn(\|\boldsymbol{a}\|)_{(11:)}}$	The ratio in $MsEn$ over first 10 to last 10 scales captures the dynamic complexity of gait versus that of random fluctuations in a signal [308].	
$\frac{std(MsEn(\|\boldsymbol{a}\|)_{(:10)})}{std(MsEn(\|\boldsymbol{a}\|)_{(11:)})}$	The ratio in the standard deviation in $MsEn$ over first 10 to last 10 scales captures the variance in the dynamic complexity of gait versus the variance in that of random fluctuations in a signal [308].	

Continued on Next Page...

Table A.1: Description of the features extracted from the Two-Minute Walk Test

Feature	Description	Parameterisation
$\sum \left\lvert \frac{E(IMF_{(n:)}(\lVert \boldsymbol{a} \rVert))}{E(IMF_{(:n)}(\lVert \boldsymbol{a} \rVert))} \right\rvert$ $\sum \left\lvert \frac{\Phi(IMF_{(n:)}(\lVert \boldsymbol{a} \rVert))}{\Phi(IMF_{(:n)}(\lVert \boldsymbol{a} \rVert))} \right\rvert$ $\sum \left\lvert \frac{H(IMF_{(n:)}(\lVert \boldsymbol{a} \rVert))}{H(IMF_{(:n)}(\lVert \boldsymbol{a} \rVert))} \right\rvert$	EMD decomposes a signal into a small finite number of intrinsic mode functions (IMFs). See section A.3.2 for more details. The first i IMFs represents the "high-frequency (noise)" components with the latter IMFs capturing the relatively "low-frequency (signal)" components of gait rhythm. The sum of the signal-to-noise (SNR) energy, E (A.59), entropy, H (A.51), and TKEO, Φ (A.60), are then computed using IMFs. This is analogous for the variability in the ratio (amount) of gait to higher- frequency perturbations over the signal.	Computed with the signal sampled every: $\Delta t = \{0.01, 0.5, 1\}$ seconds [s] and $n = 1$.

A.4.2 Chapter 5: The Draw a Shape Feature Space

Raw sensor data is collected from the smartphone touchscreen during the active Draw a Shape Test and stored as x- and y-screen coordinates with a corresponding timestamp t: (x, y, t), where N is the total number of touch data points. Features were extracted from various shapes such as the (*circle*, *figure-8-shape*, *spiral*, *square*) traced on smartphone touchscreen. Drawing velocity (v) is calculated using [139]:

$$v = \sum_{i=1}^{N-1} \frac{\sqrt{(x_{i+1} - x_i)^2 - (y_{i+1} - y_i)^2}}{t_{i+1} - t_i} \tag{A.67}$$

Radial velocity (RV) is also calculated, using the following equation [139]:

$$RV = \sum_{i=1}^{N-1} \frac{r_{i+1} - r_i}{t_{i+1} - t_i} \tag{A.68}$$

where r is the radius, defined by: $\sqrt{(x^2 + y^2)}$.

Angular velocity (RHOV) is also calculated, using the following equation [139]:

$$RV = \sum_{i=1}^{N-1} \frac{\theta_{i+1} - \theta_i}{t_{i+1} - t_i} \tag{A.69}$$

where $\theta = \text{atan2}(\frac{y}{x})$ and atan2 is the four-quadrant inverse tangent, $\tan^{-1}(\frac{y}{x})$. All features calculated on v were also computed using the RV and $RHOV$.

Touch screen x- and y- coordinates can be converted to a grayscale achromatic intensity image IM that contain values in the range 0 (black) to 1 (white). Coordinates were first scaled to the original resolution of the smartphone used. IM_{ref} refers to the reference shape x- and y-coordinates interpolated to fixed length L and converted to a grayscale intensity image. Discretised heat maps of touch events, which are described in the methods section 5.3, are denoted by HM. Those heat maps constructed with reference coordinates are referred to as HM_{ref}. The table A.2 describes all features extracted from the Draw-a-Shape Test. Features which were extracted on all shapes are denoted using, $\langle$*shape*$\rangle$ reference. Those features that are shape dependent reference the specific shape that the feature was extracted on. $\langle$ *overall* $\rangle$ refers to any feature computed over all shapes, where the value was taken as the mean or sum over all shapes, which are detailed and referenced accordingly.

Table A.2: Description of features extracted from the Draw-a-Shape Test

Shape	Feature	Description
⟨ *shape* ⟩	acc	The total number of successfully passed waypoints divided by the number of waypoints per shape.
⟨ *shape* ⟩	$apen(v)$	The approximate entropy of v. Time-series x is first interpolated to a fixed length L and using the parameters defined in [139, 143].
⟨ *shape* ⟩	$asym(x, y)$	Shape asymmetry in the x- and y- plane [139].
⟨ *shape* ⟩	$AUC(x, y)$	Drawing error as represented by area-under-the-curve (AUC) integration using the trapezoidal method with unit spacing over the drawn-touch points with respect to the touch coordinates. Using similar methodology to [134].
⟨ *shape* ⟩	$iqr(AUC(x, y))$	The interquartile range of the AUC.
⟨ *circle, spiral* ⟩	$AUC_{CS}(x, y)$	The mean AUC error in the circular shapes.
⟨ *line, square* ⟩	$AUC_{LS}(x, y)$	The mean AUC error in the line & square shapes.
⟨ *shape* ⟩	$celerity(x, y, t)$	Celerity, defined as the ratio of successfully passed waypoints divided by the time taken to complete the shape.
⟨ *circle, spiral* ⟩	$celerity_{CS}(x, y, t)$	The mean celerity in the circular shapes.
⟨ line, *square* ⟩	$celerity_{LS}(x, y, t)$	The mean celerity in the line and square shapes.
⟨ *figure-8-shape* ⟩	$centre_{offshoot}(x, y)$	Euclidean distance by which the touch coordinates misses the centre of the shape.
⟨ *overall* ⟩	$\Sigma shape_1$	The number of shapes successfully completed on the first attempt, divided by the total number of attempts.
⟨ *overall* ⟩	$\Sigma shape_2$	The number of shapes successfully completed on the second attempt, divided by the total number of attempts.
⟨ *overall* ⟩	$\Sigma shape$	The number of shapes successfully completed on the first attempt and on the second attempt, divided by the total number of attempts.
⟨ *shape* ⟩	$corr(IM, IM_{ref})$	The 2-D image correlation coefficient between shape drawn and reference shape image.
⟨ *shape* ⟩	$cov(v)$	The coefficient of variation in v.
⟨ *shape* ⟩	$var(v)$	The variance in v.
⟨ *shape* ⟩	$drawPathRMSE(x, y)$	The root-mean-squared errror (RMSE) between the square root of the x- and y- touch points squared and the square root of the x- and y- reference points squared, similar to [139].
⟨ *shape* ⟩	$dwell\text{-}to\text{-}move(x, y)$	The ratio of total dwell time to total time taken to complete the SQUARE shape. See dwellTime feature.
⟨ *square* ⟩	$dwellTime(x, y)$	Dwell time is used as a measure to assess hesitation time ability to change direction at the corners of square shapes. Velocity (v) is first calculated as described previously, where the local maxima in v represent local max drawing speed and local minima represent the stoppages. Dwell time is calculated as the time distance (width) at half the corresponding minima height.
⟨ *square* ⟩	$dwellTime_m(x, y)$	Mean dwell time. See dwellTime feature.
⟨ *square* ⟩	$dwellTime_{sd}(x, y))$	The standard deviation in the dwell time. See $dwellTime$ feature.
⟨ *spiral* ⟩	$GOF(x, y)$	RMSE value calculated linear regression between ideal (reference) and drawn touch coordinates. See [138] for more details.
⟨ *shape* ⟩	$HausD(x, y)$	The maximum Hausdorff distance between drawn and reference shape which is assessed as an approximate for maximum drawing error[284, 285].
⟨ *shape* ⟩	$HausDError(x, y)$	The sum of the Hausdorff distances (largest minimum distance to closest iterative points) between drawn and reference shape, normalised by the number of touchpoints.
⟨ *shape* ⟩	$HausD_{25}(x, y)$	The sum of the Hausdorff distances (largest minimum distance to closest iterative points) between drawn and reference shape, normalised by the number of touchpoints - assessed as an approximate for drawing error at start of drawing.

Continued on Next Page...

Table A.2: Description of features extracted from the Draw-a-Shape Test

Shape	Feature	Description
⟨ *shape* ⟩	$HausD_{75}(x, y)$	The sum of the Hausdorff distances (largest minimum distance to closest iterative points) between drawn and reference shape, normalised by the number of touchpoints - assessed as an approximate for drawing error at end of drawing.
⟨ *shape* ⟩	$hausD_t(x, y)$	The Hausdorff distance normalised by the time taken to draw the shape.
⟨ *shape* ⟩	$hausDmiddle(x, y)$	Hausdorff distance using touch coordinates from the middle of drawing (i.e. 15-85% range).
⟨ *shape* ⟩	$hausDmiddle_t(x, y)$	Hausdorff distance using touch coordinates from the middle of drawing (i.e. 15-85% range) weighted by the time taken to complete the shape.
⟨ *shape* ⟩	$HD_E(IM, x, y)$	Image entropy of IM multiplied (*) by the maximum Hausdorff Distance as a measures of penalised drawing irregularity.
⟨ *shape* ⟩	$cov(hf(v))$	The variance of the high-frequency wavelet coefficients of v. The high frequency components are first extracted by a 1-level Discrete Wavelet Transform (DWT) using Daubechies 10 wavelet function. This is proposed as a feature for detecting the variation in short and fast movement changes during the spiral drawing task [139].
⟨ *shape* ⟩	$hf_m(v)$	The mean high-frequency wavelet coefficients of x. The high frequency components are first extracted by a 1-level Discrete Wavelet Transform (DWT) using Daubechies 10 wavelet function, similar to [139].
⟨ *shape* ⟩	$max(hf(v))$	The maximum high-frequency wavelet coefficients of v. The high frequency components are first extracted by a 1-level Discrete Wavelet Transform (DWT) using Daubechies 10 wavelet function.
⟨ *shape* ⟩	$hf_{sd}(v)$	The standard deviation in the high-frequency wavelet coefficients of v. The high frequency components are first extracted by a 1-level Discrete Wavelet Transform (DWT) using Daubechies 10 wavelet function, similar to [139].
⟨ *shape* ⟩	$ImEntropy(IM)$	Image entropy of shape drawn [288].
⟨ *shape* ⟩	$\frac{ImEntropy(IM)}{ImEntropy(IM_{ref})}$	Image entropy of shape drawn with respect to reference shape [288].
⟨ *shape* ⟩	$imRatio(IM, IM_{ref})$	Mean pixel intensity ratio between shape drawing transposed as pixel density heat map and reference image.
⟨ *shape* ⟩	$iqr(HausD)$	Interquartile range of Hausdorff distance computed between drawn touch points and reference coordinates. See hausD.
⟨ *shape* ⟩	$isdRatio(IM, IM_{ref})$	Standard deviation in the pixel intensity ratio between shape drawing and reference shape.
⟨ *shape* ⟩	$kurt(v)$	Kurtosis as a measure of the taildness of the probability distribution of v values, and is used as a measure of how outlier prone a distribution is.
⟨ *shape* ⟩	$max(v)$	The maximum velocity, v.
⟨ *shape* ⟩	$\max(t_{pixel}(x, y))$	Maximum hesitation time [s], defined as the maximum time a finger remains in one (x, y) coordinate regions, which are defined by the binning of touch events into discretized heat maps.
⟨ *shape* ⟩	$mean(v)$	The mean of v [137, 420]
⟨ *overall* ⟩	$celerity_m$	The mean celerity over all shapes, defined as the ratio between the number of successfully completed way-points and the time taken to complete the shape.
⟨ *shape* ⟩	$MI(IM, IMref)$	The mutual information between the shape drawing touch coordinates and reference coordinate [287, 421].
⟨ *shape* ⟩	$mIm(IM)$	Mean pixel intensity for shape drawing image.
⟨ *shape* ⟩	$mPeak(v)$	The mean number of peaks in v [139].
⟨ *shape* ⟩	$mPI(IM)$	Mean pixel intensity for shape drawing image.
⟨ *shape* ⟩	$nMSE$	The 2-D mean square error between shape drawing touch coordinates and reference coordinate [422].
⟨ *shape* ⟩	$nPeaks(v)$	The number of peaks in v [139].

Continued on Next Page...

Table A.2: Description of features extracted from the Draw-a-Shape Test

Shape	Feature	Description
⟨ *shape* ⟩	$nPeaksNorm(v)$	The number of peaks in v, normalised by the number of samples in v. This has been proposed as a measure of fine drawing control [139].
⟨ *shape* ⟩	$nRMSE$	The 2-D root-mean square error between shape drawing touch coordinates and reference coordinate [422].
⟨ *shape* ⟩	$numHits$	Number of successfully passed waypoints.
⟨ *circle, spiral* ⟩	$numHits_{CS}$	The total number of waypoint hits in the circular shapes.
⟨ *line, square* ⟩	$numHits_{LS}$	The total number of hits in the line & square shapes.
⟨ *square* ⟩	$overShoot(x, y)$	The total Euclidian distance error by which a subject overshoots or undershoots (misses) a corner.
⟨ *shape* ⟩	$overShoot_m(x, y)$	The mean Euclidian distance error by which a subject overshoots or undershoots (misses) over all 4 corners.
⟨ *shape* ⟩	$overShoot_{sd}(x, y)$	The variance in the Euclidian distance error by which a subject overshoots or undershoots (misses) a corner over all over all 4 corners, as measured by the standard deviation.
⟨ *shape* ⟩	$pDchisq(v)$	The Chi-square distance between the two histograms of heatmap transposed pixel intensity counts which can be used as a rough similarity index [423, 424].
⟨ *shape* ⟩	$pSNR(IM)$	The peak signal-to-noise ratio between in shape drawn with the ideal (reference) touch coordinates [422].
⟨ *shape* ⟩	$\max(PSD(v))$	Maximum Power Spectral Density (PSD) estimate of v, computed using a Welsh periodogram with a Hamming window. Similar to [282].
⟨ *shape* ⟩	$\arg\max_f(PSD(v))$	The dominant frequency refers to the frequency which maximises the Power Spectral Density (PSD) estimate of v.
⟨ *shape* ⟩	$pRMS(v)$	Root-mean squared value of v.
⟨ *spiral* ⟩	$R_{lfit}(x, y)$	r^2 (adjusted) value calculated linear regression between ideal (reference) and drawn touch coordinates. See [138] for more details.
⟨ *overall* ⟩	shp_c	The number of the correctly completed shapes over a test instance, regardless of attempt number.
⟨ *overall* ⟩	shp_{fail}	The number of failed shape drawing attempts over a test instance.
⟨ *shape* ⟩	$skew(v)$	The skewness of distribution of v.
⟨ *shape* ⟩	$sp(v)$	Spearman's rank correlation (R_{sp}) coefficient of the level of variation with time of v. This is computed as the change of variation over time in number of peaks during a test trial, calculated using a sliding window, which compared standard deviations over 30 overlapping windows. For more details we refer the reader to the original description in [139]
⟨ *shape* ⟩	$SSI(IM, IM_{ref})$	Structural Similarity Index, as described in [425].
⟨ *shape* ⟩	$std(v)$	The standard deviation of v [420].
⟨ *shape* ⟩	$time$	Total time taken to complete the shape.
⟨ *shape* ⟩	$TOTSYMM(x, y)$	Total summation of the shape asymmetry in the x-plane and y-plane [139].

B
Supplementary Results

B.1 Chapter 5: The Distribution of All Clinical 9HPT Predictions

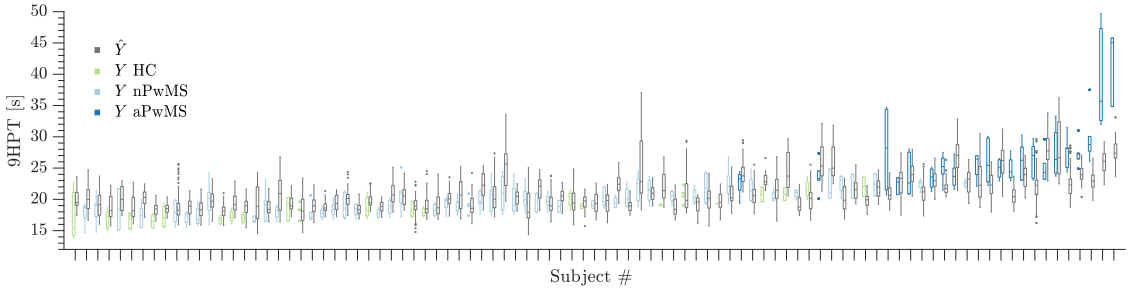


Figure B.1: Distribution of actual Y and predicted $\hat{Y}$ NHPT times [s] per subject. Dominant and non-dominant handed 9HPT times from all 3 clinical visits are pooled. PwMS were grouped into presumably normal (nPwMS) versus abnormal (aPwMS) hand/arm function based on an upper limit of normal range, defined as the average NHPT times for HCs plus two standard deviations over pooled dominant and non-dominant handed tests. IRLS models were trained with the median feature value for a 14-day window around each respective clinical test date. Predicted HC 9HPTs $\hat{Y}$ were less variable than predicted nPwMS and aPwMS NHPT times $\hat{Y}$ (p <0.05), whereas actual HC Y values demonstrated no difference in variance than actual Y nPwMS 9HPT times. The variance in response Y is similarly matched in the variance of predicted $\hat{Y}$ for HC and some nPwMS subjects.

B.2 Chapter 7: DCNN Model Parameterisation

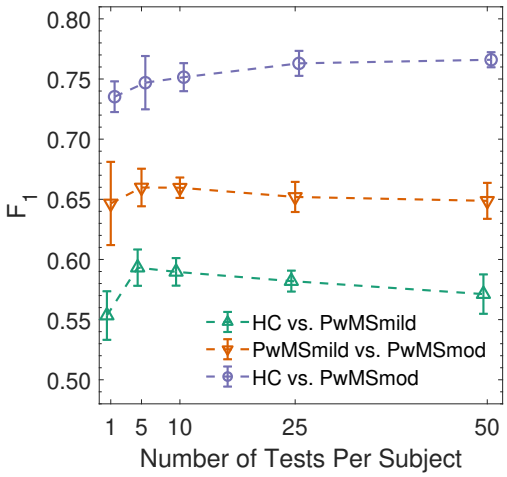


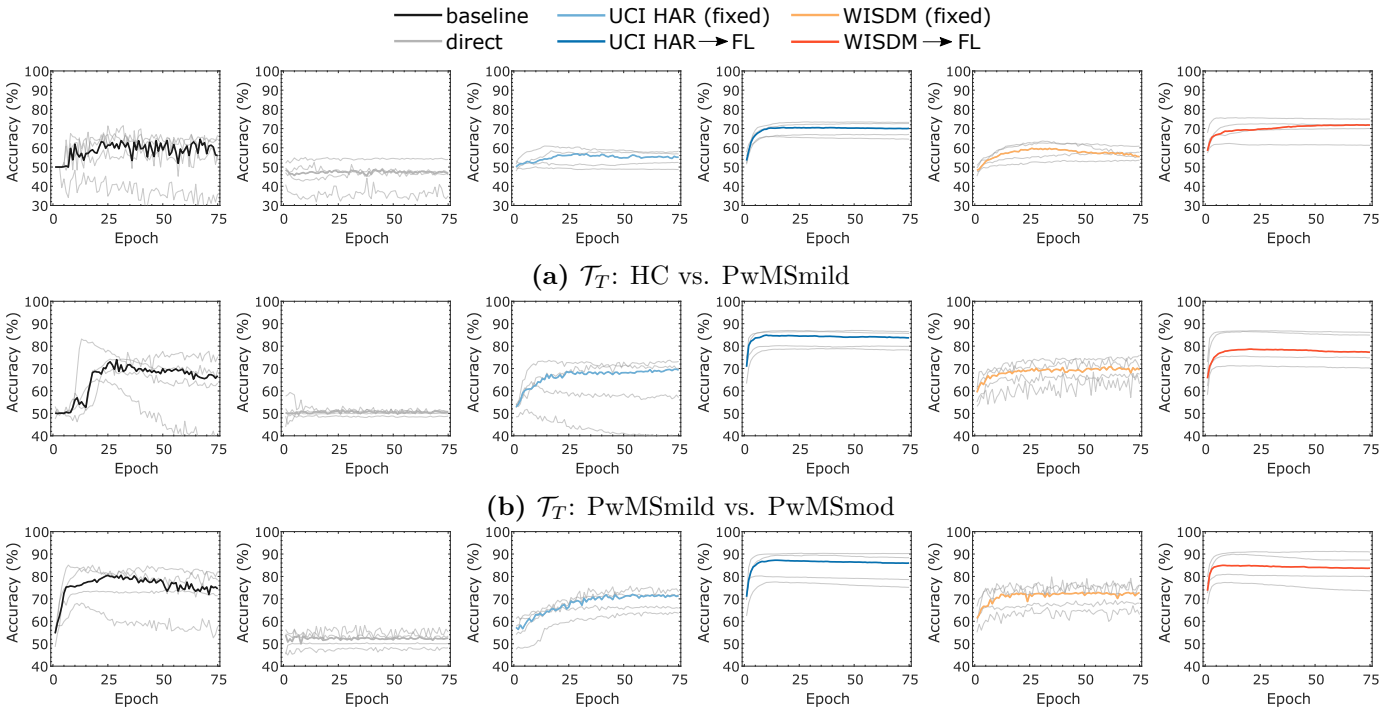
Figure B.2: Classification performance over varying number of daily tests sampled per subject. Number of daily tests sampled per subject that are used for training and validation. Results are presented as the mean and standard deviation across 10 CV repetitions to minimise biases in test sampling. DCNNs are trained end-to-end on FL.

Table B.1: Parameters that maximised each target classification task, $\mathcal{T}_T$.

$\mathcal{T}_T$	$\mathcal{D}_S$	FC DNN[1]	Initial Learning Rate	Mini-Batch Size
HAR	UCI HAR	$\{64, 32, 5\}$	1×10^{-3}	128
	WISDM	$\{5\}$	1×10^{-4}	128
HC vs. PwMSmild	(end-to-end)	$\{2\}$	1×10^{-3}	64
	(UCI HAR, fixed)	$\{2\}$	1×10^{-3}	250
	(WISDM, fixed)	$\{64, 32, 2\}$	5×10^{-3}	250
	(UCI HAR→FL)	$\{2\}$	1×10^{-5}	250
	(WISDM→FL)	$\{32, 2\}$	1×10^{-5}	250
PwMSmild vs. PwMSmod	(end-to-end)	$\{32, 2\}$	1×10^{-3}	64
	(UCI HAR, fixed)	$\{2\}$	1×10^{-2}	250
	(WISDM, fixed)	$\{32, 2\}$	1×10^{-3}	250
	(UCI HAR→FL)	$\{64, 32, 2\}$	1×10^{-4}	250
	(WISDM→FL)	$\{64, 32, 2\}$	5×10^{-5}	250
HC vs. PwMSmod	(end-to-end)	$\{32, 2\}$	1×10^{-2}	250
	(UCI HAR, fixed)	$\{2\}$	1×10^{-2}	128
	(WISDM, fixed)	$\{2\}$	1×10^{-3}	250
	(UCI HAR→FL)	$\{32, 2\}$	5×10^{-5}	250
	(WISDM→FL)	$\{64, 32, 2\}$	5×10^{-5}	250
HC vs. PwMSmild vs. PwMSmod	(end-to-end)	$\{3\}$	1×10^{-4}	64
	(UCI HAR→FL)	$\{3\}$	1×10^{-5}	250
	(WISDM→FL)	$\{3\}$	1×10^{-5}	250

[1] Neural network layers and the number of hidden layers at each node, including final softmax output.

B.3 Chapter 7: DCNN Transfer Learning



(a) $\mathcal{T}_T$: HC vs. PwMSmild

(b) $\mathcal{T}_T$: PwMSmild vs. PwMSmod

(c) $\mathcal{T}_T$: HC vs. PwMSmod

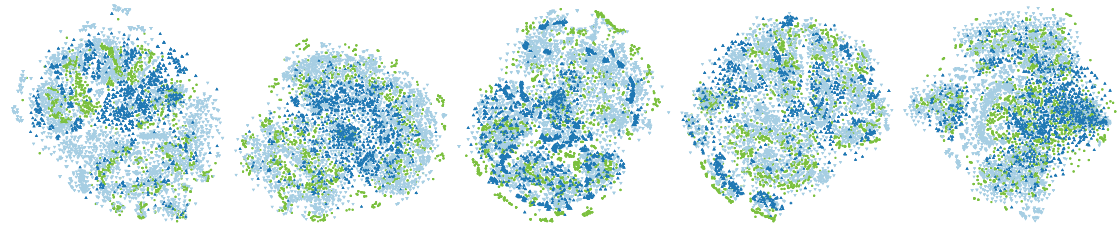
Figure B.3: Comparison of DCNN transfer learning approaches versus a DCNN trained end-to-end on FLOODLIGHT (FL). $\mathcal{T}_T$: **(a)** HC vs.PwMSmild, **(b)** PwMSmild vs.PwMSmod and **(c)** HC vs.PwMSmod. Figure represents the median out-of-sample validation set accuracy against training iteration (epoch) over each (k=5) CV fold, depicted in grey. A black line, "baseline", refers to a DCNN trained and validated end-to-end exclusively on FL data.

B.4 Chapter 7: Activity Predictions in FL

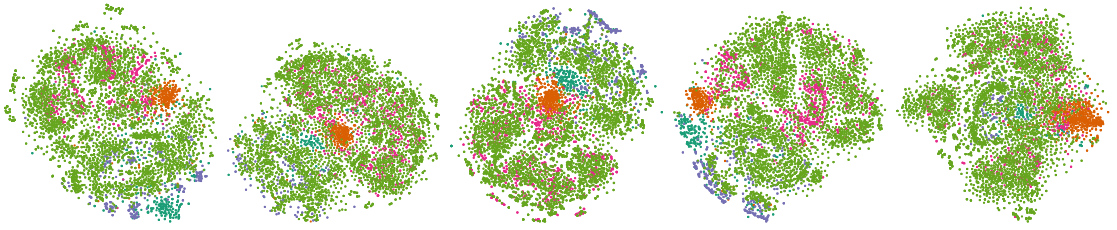
Table B.2: Breakdown in the distribution of predicted HAR activities in the FL 2MWT data. The mean ± standard deviation in the number of tests identified as each HAR activity per subject for each group; otherwise percentage (%) breakdown of each activity class along with count values are reported.

$\hat{\mathcal{Y}}$	Total	HC	PwMSmild	PwMSmod	p[1]
laying	3.4% (2817)	3.7% (754) 34.3 ± 30.8	3.5% (1530) 33.3 ± 34.5	3.0% (533) 25.4 ± 22.7	0.65
sitting	5.8% (4798)	6.3% (1275) 53.1 ± 47.8	4.3% (1891) 37.8 ± 38.3	9.1% (1632) 77.7 ± 68.5	0.05
stairs	10.5% (8630)	9.8% (1996) 90.7 ± 100.7	10.4% (4615) 96.2 ± 119.6	11.3% (2019) 96.1 ± 122.5	0.95
standing	4.6% (3799)	9.2% (1874) 104.1 ± 134.0	3.5% (1552) 39.8 ± 39.1	2.1% (373) 28.7 ± 45.3	0.13
walking	75.7% (62406)	71.1% (14501) 604.2 ± 160.5	78.3% (34612) 665.6 ± 143.2	74.5% (13293) 633.0 ± 134.0	0.12

[1] A Kruskal-Wallis test by ranks (see section 3.4.2.8);



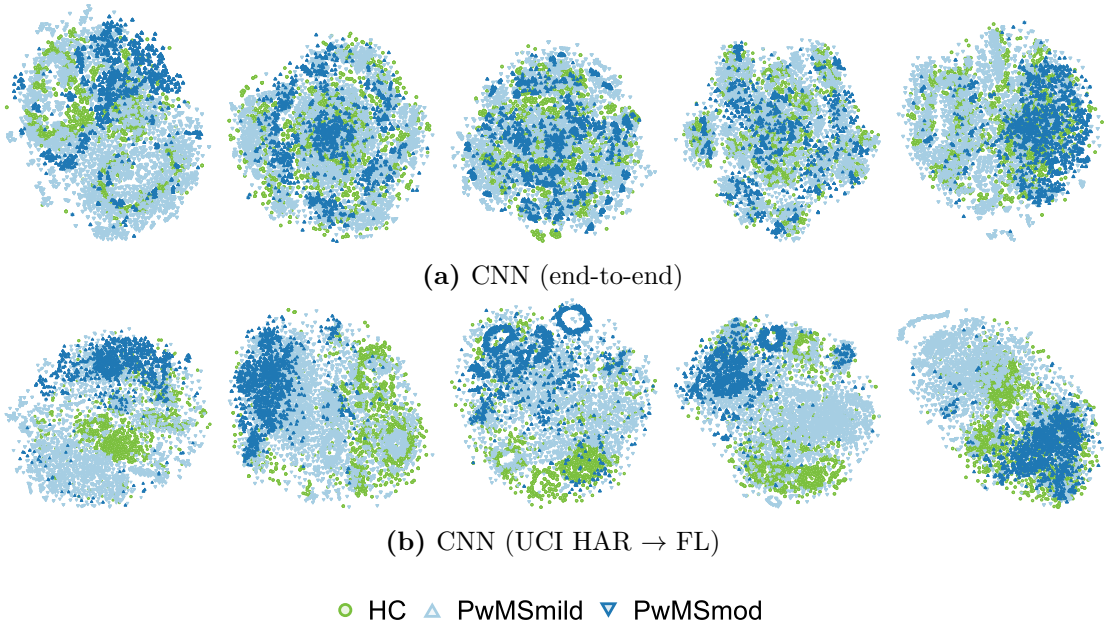
(a) CNN (UCI HAR → FL, direct): $\mathcal{Y}_T$



(b) CNN (UCI HAR → FL, direct): $\hat{\mathcal{Y}}_S$

Figure B.4: Visualisation of UCI HAR trained CNN features (direct) extracted on FLOODLIGHT using t-Distributed Stochastic Neighbour Embedding (t-SNE). Each column represents one fold (test partition) of the FLOODLIGHT dataset (k=5 folds). Therefore, **(a)** represents the t-SNE embeddings of $f_S(\cdot)$ for $\mathcal{D}_T$, i.e. where epochs are labelled by true subject group ($\mathcal{Y}_S$: HC / PwMSmild / PwMSmod) whereas **(b)** depicts the analogous plots coloured by UCI HAR predictions $\hat{\mathcal{Y}}_S$ for $\mathcal{T}_S$ of activities.

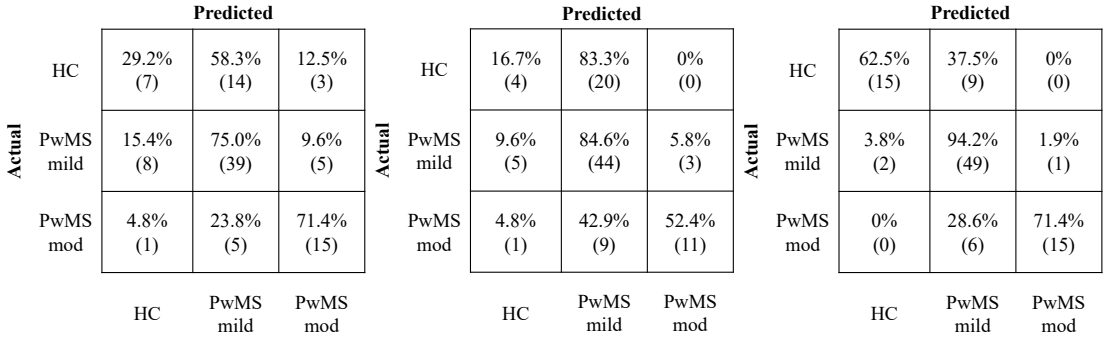
B.5 Chapter 7 & 8: Disease Representation in FL



(a) CNN (end-to-end)

(b) CNN (UCI HAR → FL)

Figure B.5: Visualisation of the CNN features trained on FLOODLIGHT (FL) using *t*-Distributed Stochastic Neighbour Embedding (t-SNE). Each column represents one fold (test partition) of the FL dataset (k=5).



(a) features + SVM **(b)** DCNN (end-to-end) **(c)** DCNN (UCI HAR → FL)

Figure B.6: Confusion matrix for subject classification outcomes for unified multi-class models constructed based on HC vs. PwMSmild vs. PwMSmod. Subject-wise classification results are presented using using 5-fold CV for **(a)** hand-crafted features + SVM classifer, **(b)** DCNN model trained end-to-end on FLOODLIGHT and **(c)** DCNN model that has been fine-tuned using transfer learning. Subject-wise predictions are computed by majority voting all individual test predictions per subject. Confusion matrices depict the both the raw classification values as well as the percentage (%) recall for that class.

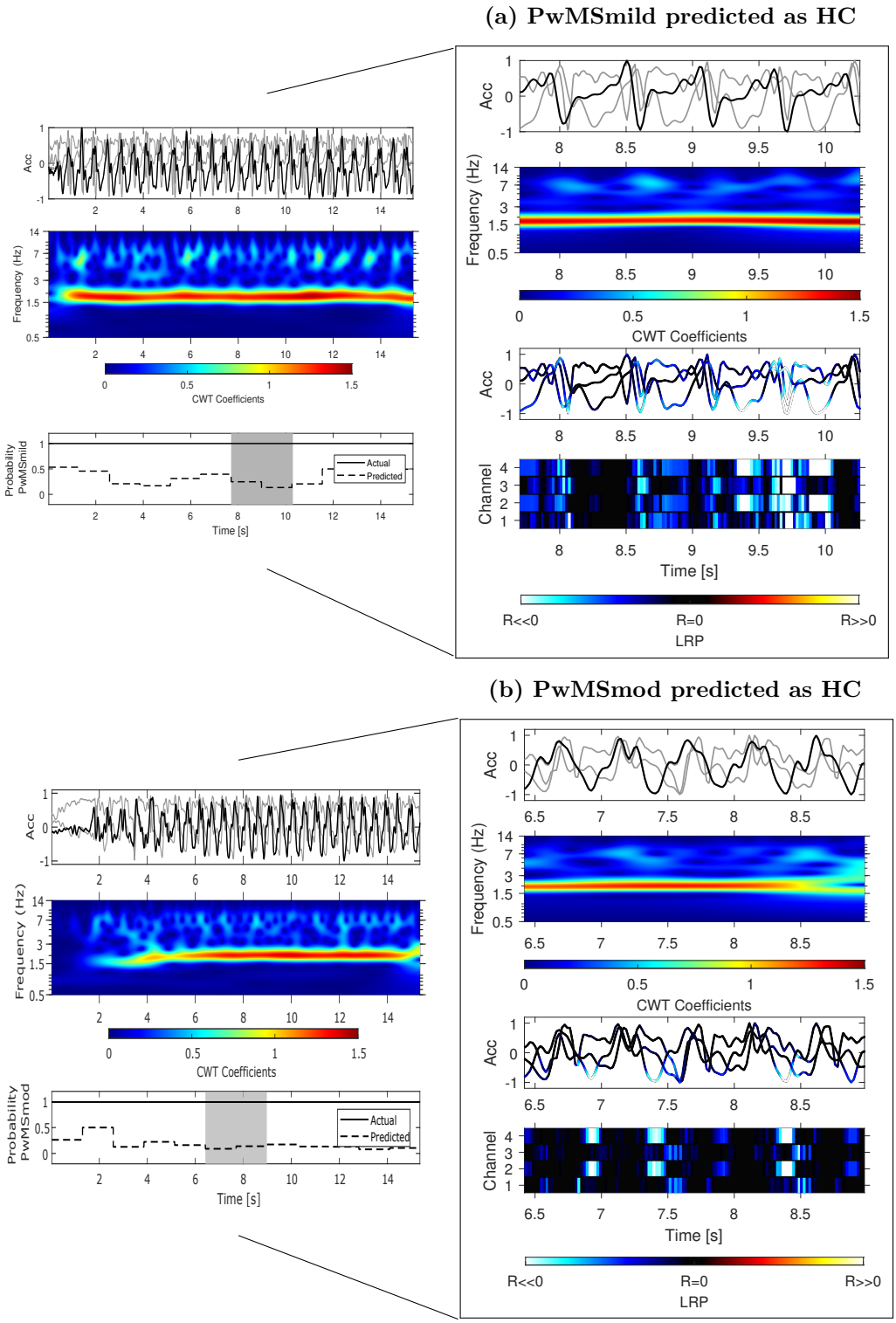


Figure B.7: **Panel plot illustrating example performance from a section of misclassified (a) PwMSmild as HC; (b) PwMSmod as HC 2MWT (false negatives) which can be visually interpreted using LRP decomposition and CWT frequency analysis.**

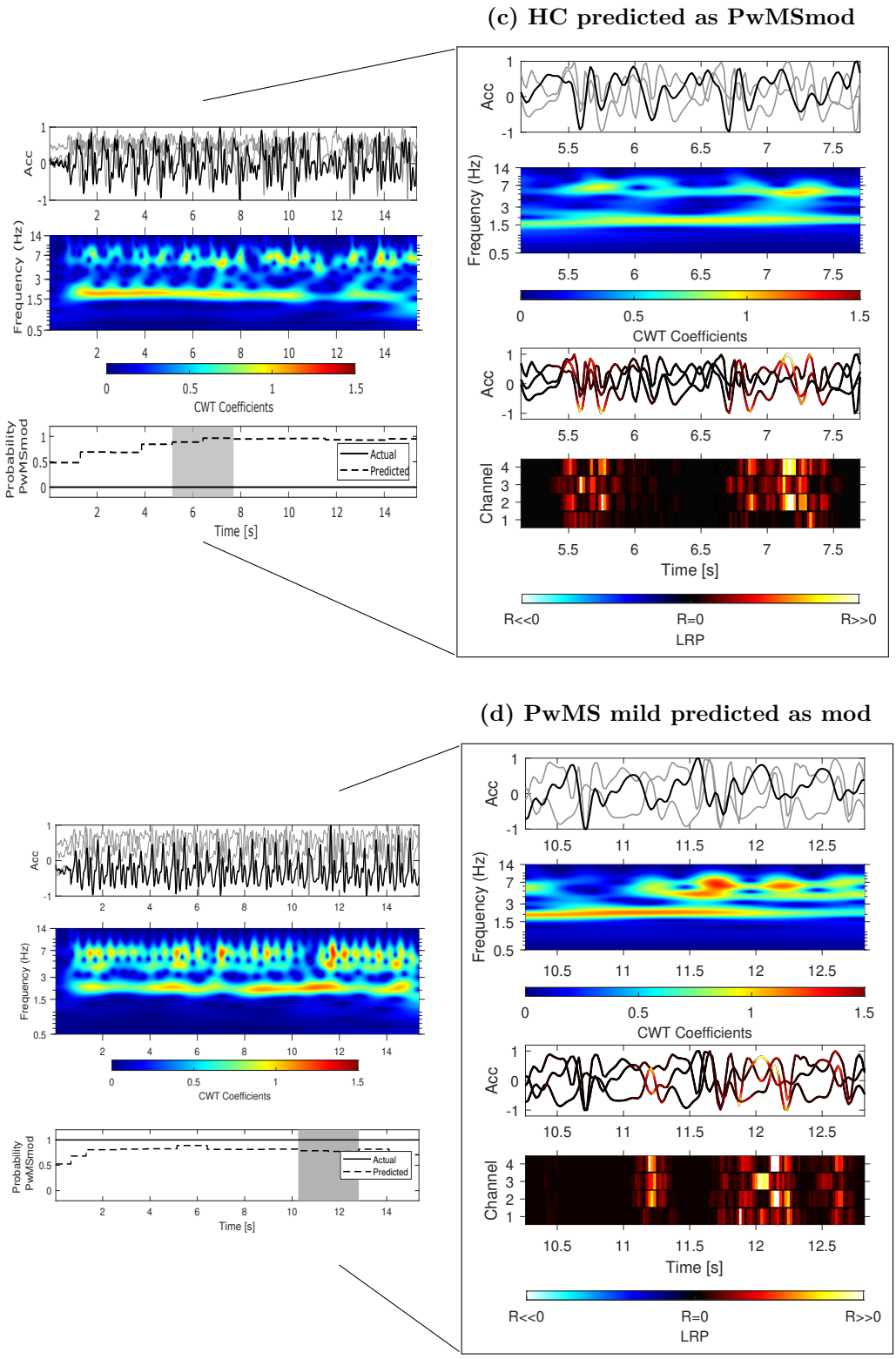


Figure B.7: **Cont.: Panel plot illustrating example performance from a section of misclassified (c) HC as PwMSmod; (d) PwMSmild as PwMSmod 2MWT (false positives) which can be visually interpreted using LRP decomposition and CWT frequency analysis.**

References

[1] A. P. Creagh, C. Simillion, A. Scotland, F. Lipsmeier, C. Bernasconi, S. Belachew, J. van Beek, M. Baker, C. Gossens, M. Lindemann, and M. De Vos. "Smartphone-based remote assessment of upper extremity function for multiple sclerosis using the Draw a Shape Test". In: *Physiological Measurement* 41.5 (2020), p. 054002. URL: http://dx.doi.org/10.1088/1361-6579/ab8771.

[2] A. P. Creagh, C. Simillion, A. Bourke, A. Scotland, F. Lipsmeier, C. Bernasconi, J. van Beek, M. Baker, C. Gossens, M. Lindemann, and M. De Vos. "Smartphone- and Smartwatch-Based Remote Characterisation of Ambulation in Multiple Sclerosis during the Two-Minute Walk Test". In: *IEEE Journal of Biomedical and Health Informatics* (2020). URL: https://doi.org/10.1109/JBHI.2020.2998187.

[3] Andrew P Creagh, Florian Lipsmeier, Michael Lindemann, and Maarten De Vos. "Interpretable deep learning for the remote characterisation of ambulation in multiple sclerosis using smartphones". In: *Scientific Reports* 11.1 (2021), pp. 1–14.

[4] Andrew P Creagh, Frank Dondelinger, Florian Lipsmeier, Michael Lindemann, and Maarten De Vos. "Longitudinal trend monitoring of multiple sclerosis ambulation using smartphones". In: *IEEE open journal of engineering in medicine and biology* 3 (2022), pp. 202–210.

[5] Andrew P. Creagh. "AI and the Digital Healthcare Revolution Uniting Disciplines". In: *Phenotype* 34 (2020), pp. 14–16.

[6] Wei-Yi Cheng, Florian Lipsmeier, Andrew Creagh, Alf Scotland, Timothy Kilchenmann, Liping Jin, Jens Schjodt-Eriksen, Detlef Wolf, Yan-Ping Zhang-Schaerer, Ignacio Fernandez Garcia, et al. "Large-Scale Continuous Mobility Monitoring of Parkinson's Disease Patients Using Smartphones". In: Wireless Mobile Communication and Healthcare. Springer International Publishing, 2017, pp. 12–19.

[7] Wei-Yi Cheng, Florian Lipsmeier, Alf Scotland, Andrew Creagh, Timothy Kilchenmann, et al. "Smartphone-based continuous mobility monitoring of Parkinsons disease patients reveals impacts of ambulatory bout length on gait features". In: *IEEE Life Sciences Conference (LSC)*. IEEE, 2017, pp. 166–169.

[8] Marvin M Goldenberg. "Multiple sclerosis review". In: *Pharmacy and Therapeutics* 37.3 (2012), p. 175.

[9] Nils Koch-Henriksen and Per Soelberg Sørensen. "The changing demographic pattern of multiple sclerosis epidemiology". In: *The Lancet Neurology* 9.5 (2010), pp. 520–532.

[10] Howard L Weiner. "Multiple sclerosis is an inflammatory T-cell–mediated autoimmune disease". In: *Archives of neurology* 61.10 (2004), pp. 1613–1615.

[11] Laura Comber, Rose Galvin, and Susan Coote. "Gait deficits in people with multiple sclerosis: A systematic review and meta-analysis". In: *Gait & Posture* 51 (2017), pp. 25–35.

[12] C Heesen, J Böhm, C Reich, J Kasper, M Goebel, and SM Gold. "Patient perception of bodily functions in multiple sclerosis: gait and visual function are the most valuable". In: *Multiple Sclerosis Journal* 14.7 (2008), pp. 988–991.

[13] Janet L Poole, Trisha Nakamoto, Tina McNulty, Janeen R Montoya, Deedra Weill, Kathy Dieruf, and Betty Skipper. "Dexterity, visual perception, and activities of daily living in persons with multiple sclerosis". In: *Occupational Therapy in Health Care* 24.2 (2010), pp. 159–170.

[14] S Anand Trip, Patricio G Schlottmann, Stephen J Jones, Daniel R Altmann, Garway-Heath, et al. "Retinal nerve fiber layer axonal loss and visual dysfunction in optic neuritis". In: *Annals of neurology* 58.3 (2005), pp. 383–391.

[15] Kristina Bacher Svendsen, Troels Staehelin Jensen, Hans Jacob Hansen, and Flemming Winther Bach. "Sensory function and quality of life in patients with multiple sclerosis and pain". In: *Pain* 114.3 (2005), pp. 473–481.

[16] Nancy D Chiaravalloti and John DeLuca. "Cognitive impairment in multiple sclerosis". In: *The Lancet Neurology* 7.12 (2008), pp. 1139–1151.

[17] Lawrence Steinman. "Multiple sclerosis: a two-stage disease". In: *Nature immunology* 2.9 (2001), pp. 762–764.

[18] Antonio Scalfari, Volker Knappertz, Gary Cutter, Douglas S Goodin, Raymond Ashton, and George C Ebers. "Mortality in patients with multiple sclerosis". In: *Neurology* 81.2 (2013), pp. 184–192.

[19] John F Kurtzke. "Rating neurologic impairment in multiple sclerosis an expanded disability status scale (EDSS)". In: *Neurology* 33.11 (1983), pp. 1444–1444.

[20] RA Rudick, G Cutter, and S Reingold. "The multiple sclerosis functional composite: a new clinical outcome measure for multiple sclerosis trials". In: *Multiple Sclerosis Journal* 8.5 (2002), pp. 359–365.

[21] R. W. Motl, J. A. Cohen, R. Benedict, G. Phillips, N. LaRocca, L. D. Hudson, and R. Rudick. "Validity of the timed 25-foot walk as an ambulatory performance outcome measure for multiple sclerosis". In: *Mult Scler* 23.5 (2017), pp. 704–710.

[22] D. A. Scalzitti, K. J. Harwood, J. R. Maring, S. J. Leach, E. A. Ruckert, and E. Costello. "Validation of the 2-Minute Walk Test with the 6-Minute Walk Test and Other Functional Measures in Persons with Multiple Sclerosis". In: *Int J MS Care* 20.4 (2018), pp. 158–163.

[23] Peter Feys, Ilse Lamers, Gordon Francis, Ralph Benedict, Glenn Phillips, et al. "The Nine-Hole Peg Test as a manual dexterity performance measure for multiple sclerosis". In: *Multiple Sclerosis Journal* 23.5 (2017), pp. 711–720.

[24] V. Khurana, H. Sharma, N. Afroz, A. Callan, and J. Medin. "Patient-reported outcomes in multiple sclerosis: a systematic comparison of available measures". In: *Eur J Neurol* 24.9 (2017), pp. 1099–1107.

[25] Caspar EP van Munster and Bernard MJ Uitdehaag. "Outcome measures in clinical trials for multiple sclerosis". In: *CNS drugs* 31.3 (2017), pp. 217–236.

[26] EW Willoughby and DW Paty. "Scales for rating impairment in multiple sclerosis A critique". In: *Neurology* 38.11 (1988), pp. 1793–1793.

[27] Frederik Barkhof, Massimo Filippi, David H Miller, Philip Scheltens, Adriana Campi, Chris H Polman, Giancarlo Comi, Herman J Ader, Nick Losseff, and Jacob Valk. "Comparison of MRI criteria at first presentation to predict conversion to clinically definite multiple sclerosis". In: *Brain: a journal of neurology* 120.11 (1997), pp. 2059–2069.

[28] Massimo Filippi, Wolfgang Brück, Declan Chard, Franz Fazekas, Jeroen JG Geurts, Christian Enzinger, Simon Hametner, Tanja Kuhlmann, Paolo Preziosa, and Àlex Rovira. "Association between pathological and MRI findings in multiple sclerosis". In: *The Lancet Neurology* 18.2 (2019), pp. 198–210.

[29] Amy M Lavery, Leonard H Verhey, and Amy T Waldman. "Outcome measures in relapsing-remitting multiple sclerosis: capturing disability and disease progression in clinical trials". In: *Multiple sclerosis international* 2014 (2014).

[30] Thomas R Fleming and David L DeMets. *Surrogate end points in clinical trials: are we being misled?* 1996.

[31] Luciana Midaglia, Patricia Mulero, Xavier Montalban, Jennifer Graves, Stephen L Hauser, Laura Julian, Michael Baker, Jan Schadrack, Christian Gossens, and Alf Scotland. "Adherence and Satisfaction of Smartphone-and Smartwatch-Based Remote Active Testing and Passive Monitoring in People With Multiple Sclerosis: Nonrandomized Interventional Feasibility Study". In: *Journal of Medical Internet Research* 21.8 (2019), e14863.

[32] Gavin Giovannoni, Helmut Butzkueven, Suhayl Dhib-Jalbut, Jeremy Hobart, Gisela Kobelt, George Pepper, Maria Pia Sormani, Christoph Thalheim, Anthony Traboulsee, and Timothy Vollmer. "Brain health: time matters in multiple sclerosis". In: *Multiple sclerosis and related disorders* 9 (2016), S5–S48.

[33] Giancarlo Comi, Massimo Filippi, Frederik Barkhof, Luca Durelli, Gilles Edan, Oscar Fernández, Hans-Peter Hartung, Pierrette Seeldrayers, Per Soelberg Sørensen, and Marco Rovaris. "Effect of early interferon treatment on conversion to definite multiple sclerosis: a randomised study". In: *The Lancet* 357.9268 (2001), pp. 1576–1582.

[34] Kirsten I. Taylor, Hannah Staunton, Florian Lipsmeier, David Nobbs, and Michael Lindemann. "Outcome measures based on digital health technology sensor data: data- and patient-centric approaches". In: *npj Digital Medicine* 3.1 (2020), p. 97.

[35] E Ray Dorsey. *Digital Biomarkers: Bridging computer science and biomedicine.* Web Page. [Accessed: 24-07-2020]. 2020. URL: `https://www.karger.com/Journal/Details/271954`.

[36] Lmar M Babrak, Joseph Menetski, Michael Rebhan, Giovanni Nisato, Marc Zinggeler, Noe Brasier, Katja Baerenfaller, Thomas Brenzikofer, Laurenz Baltzer, and Christian Vogler. "Traditional and Digital Biomarkers: Two Worlds Apart?" In: *Digital Biomarkers* 3.2 (2019), pp. 92–102.

[37] E Ray Dorsey, Spyros Papapetropoulos, Mulin Xiong, and Karl Kieburtz. "The first frontier: digital biomarkers for neurodegenerative disorders". In: *Digital Biomarkers* 1.1 (2017), pp. 6–13.

[38] Olle Olerup and Jan Hillert. "HLA class II-associated genetic susceptibility in multiple sclerosis: a critical evaluation". In: *Tissue antigens* 38.2 (1991), pp. 1–15.

[39] Pentti J Tienari, A Sajantila, L Peltonen, J Wikström, and J Palo. "Genetic susceptibility to multiple sclerosis linked to myelin basic protein gene". In: *The Lancet* 340.8826 (1992), pp. 987–991.

[40] Stephen Sawcer, Garrett Hellenthal, Matti Pirinen, Chris CA Spencer, Nikolaos A Patsopoulos, et al. "Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis". In: *Nature* 476.7359 (2011), p. 214.

[41] Alberto Ascherio, Kassandra L Munger, Evelyne T Lennette, Donna Spiegelman, Miguel A Hernán, Michael J Olek, Susan E Hankinson, and David J Hunter. "Epstein-Barr virus antibodies and risk of multiple sclerosis: a prospective study". In: *Jama* 286.24 (2001), pp. 3083–3088.

[42] U Roelcke, L Kappos, J Lechner-Scott, H Brunnschweiler, S Huber, W Ammann, A Plohmann, S Dellas, RP Maguire, and J Missimer. "Reduced glucose metabolism in the frontal cortex and basal ganglia of multiple sclerosis patients with fatigue: A 18F-fluorodeoxyglucose positron emission tomography study". In: *Neurology* 48.6 (1997), pp. 1566–1571.

[43] George C Ebers. "Environmental factors and multiple sclerosis". In: *The Lancet Neurology* 7.3 (2008), pp. 268–277.

[44] National Library of Medicine (US). *Multiple Sclerosis.* Web Page. [Accessed: 09-09-2020]. Bethesda (MD): The Library, 2005. URL: `https://ghr.nlm.nih.gov/condition/multiple-sclerosis`.

[45] Brian G Weinshenker. "Epidemiology of multiple sclerosis". In: *Neurologic clinics* 14.2 (1996), pp. 291–308.

[46] Joost Smolders, Jan Damoiseaux, Paul Menheere, and Raymond Hupperts. "Vitamin D as an immune modulator in multiple sclerosis, a review". In: *Journal of neuroimmunology* 194.1-2 (2008), pp. 7–17.

[47] Kassandra L Munger, Lynn I Levin, Bruce W Hollis, Noel S Howard, and Alberto Ascherio. "Serum 25-hydroxyvitamin D levels and risk of multiple sclerosis". In: *Jama* 296.23 (2006), pp. 2832–2838.

[48] Fred D Lublin and Stephen C Reingold. "Defining the clinical course of multiple sclerosis: results of an international survey". In: *Neurology* 46.4 (1996), pp. 907–911.

[49] Christian Confavreux, Sandra Vukusic, Thibault Moreau, and Patrice Adeleine. "Relapses and progression of disability in multiple sclerosis". In: *New England Journal of Medicine* 343.20 (2000), pp. 1430–1438.

[50] Elizabeth Fisher, Jar-Chi Lee, Kunio Nakamura, and Richard A Rudick. "Gray matter atrophy in multiple sclerosis: a longitudinal study". In: *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society* 64.3 (2008), pp. 255–265.

[51] Anders Österberg and Jörgen Boivie. "Central pain in multiple sclerosis–sensory abnormalities". In: *European Journal of Pain* 14.1 (2010), pp. 104–110.

[52] A. Compston and A. Coles. "Multiple sclerosis". In: *Lancet* 359.9313 (2002), pp. 1221–31.

[53] Nicholas G LaRocca. "Impact of walking impairment in multiple sclerosis". In: *The Patient: Patient-Centered Outcomes Research* 4.3 (2011), pp. 189–201.

[54] Jessie M Huisinga, Martina Mancini, Rebecca J St George, and Fay B Horak. "Accelerometry reveals differences in gait variability between patients with multiple sclerosis and healthy controls". In: *Annals of biomedical engineering* 41.8 (2013), pp. 1670–1679.

[55] RI Spain, RJ St George, A Salarian, M Mancini, JM Wagner, FB Horak, and Dennis Bourdette. "Body-worn motion sensors detect balance and gait deficits in people with multiple sclerosis who have normal walking speed". In: *Gait & posture* 35.4 (2012), pp. 573–578.

[56] Jacob J Sosnoff, Brian M Sandroff, and Robert W Motl. "Quantifying gait abnormalities in persons with multiple sclerosis with minimal disability". In: *Gait & posture* 36.1 (2012), pp. 154–156.

[57] Dragana Frzovic, Meg E Morris, and Lindsay Vowels. "Clinical tests of standing balance: performance of persons with multiple sclerosis". In: *Archives of physical medicine and rehabilitation* 81.2 (2000), pp. 215–221.

[58] Clarissa L Martin, BA Phillips, TJ Kilpatrick, Helmut Butzkueven, Niall Tubridy, Elizabeth McDonald, and MP Galea. "Gait and balance impairment in early multiple sclerosis in the absence of clinical disability". In: *Multiple Sclerosis Journal* 12.5 (2006), pp. 620–628.

[59] Kevin John Kelleher, William Spence, Stephan Solomonidis, and Dimitrios Apatsidis. "The characterisation of gait patterns of people with multiple sclerosis". In: *Disability and Rehabilitation* 32.15 (2010), pp. 1242–1250.

[60] SJ Crenshaw, TD Royer, JG Richards, and DJ Hudson. "Gait variability in people with multiple sclerosis". In: *Multiple Sclerosis Journal* 12.5 (2006), pp. 613–619.

[61] Robert W Motl, Yoojin Suh, Swathi Balantrapu, Brian M Sandroff, Jacob J Sosnoff, John Pula, Myla D Goldman, and Bo Fernhall. "Evidence for the different physiological significance of the 6-and 2-minute walk tests in multiple sclerosis". In: *BMC neurology* 12.1 (2012), p. 6.

[62] An Neven, Annelien Vanderstraeten, Davy Janssens, Geert Wets, and Peter Feys. "Understanding walking activity in multiple sclerosis: step count, walking intensity and uninterrupted walking activity duration related to degree of disability". In: *Neurological Sciences* 37.9 (2016), pp. 1483–1490.

[63] S. Shema-Shiratzky, E. Gazit, R. Sun, K. Regev, A. Karni, J. J. Sosnoff, T. Herman, A. Mirelman, and Jeffrey M. Hausdorff. "Deterioration of specific aspects of gait during the instrumented 6-min walk test among people with multiple sclerosis". In: *Journal of Neurology* 266.12 (2019), pp. 3022–3030.

[64] Jacob J Sosnoff, Madeline Weikert, Deirdre Dlugonski, Douglas C Smith, and Robert W Motl. "Quantifying gait impairment in multiple sclerosis using GAITRite technology". In: *Gait and posture* 34.1 (2011), pp. 145–147.

[65] Jeffrey P Kaipust, Jessie M Huisinga, Mary Filipi, and Nicholas Stergiou. "Gait variability measures reveal differences between multiple sclerosis patients and healthy controls". In: *Motor Control* 16.2 (2012), pp. 229–244.

[66] Lisa Holper, Michaela Coenen, Andrea Weise, Gerold Stucki, Alarcos Cieza, and Jürg Kesselring. "Characterization of functioning in multiple sclerosis using the ICF". In: *Journal of neurology* 257.1 (2010), p. 103.

[67] Sean J Pittock, Robyn L McClelland, William T Mayr, Moses Rodriguez, and Joseph Y Matsumoto. "Prevalence of tremor in multiple sclerosis and associated disability in the Olmsted County population". In: *Movement disorders* 19.12 (2004), pp. 1482–1485.

[68] Sundus Husni Alusi, J Worthington, S Glickman, and PG Bain. "A study of tremor in multiple sclerosis". In: *Brain* 124.4 (2001), pp. 720–730.

[69] SH Alusi, J Worthington, S Glickman, LJ Findley, and PG Bain. "Evaluation of three different ways of assessing tremor in multiple sclerosis". In: *Journal of Neurology, Neurosurgery & Psychiatry* 68.6 (2000), pp. 756–760.

[70] JE Freal, GH Kraft, and JK Coryell. "Symptomatic fatigue in multiple sclerosis". In: *Archives of physical medicine and rehabilitation* 65.3 (1984), pp. 135–138.

[71] Lauren B Krupp, Luis A Alvarez, Nicholas G LaRocca, and Labe C Scheinberg. "Fatigue in multiple sclerosis". In: *Archives of neurology* 45.4 (1988), pp. 435–437.

[72] IK Penner, C Raselli, M Stöcklin, K Opwis, L Kappos, and P Calabrese. "The Fatigue Scale for Motor and Cognitive Functions (FSMC): validation of a new instrument to assess multiple sclerosis-related fatigue". In: *Multiple Sclerosis Journal* 15.12 (2009), pp. 1509–1517.

[73] Nuray Yozbatıran, Ferdi Baskurt, Zeliha Baskurt, Serkan Ozakbas, and Egemen Idiman. "Motor assessment of upper extremity function and its relation with fatigue, cognitive function and quality of life in multiple sclerosis patients". In: *Journal of the neurological sciences* 246.1-2 (2006), pp. 117–122.

[74] WI McDonald and D Barnes. "The ocular manifestations of multiple sclerosis. 1. Abnormalities of the afferent visual system". In: *Journal of neurology, neurosurgery, and psychiatry* 55.9 (1992), p. 747.

[75] YC Learmonth, LA Pilutti, and Robert W Motl. "Generalised cognitive motor interference in multiple sclerosis". In: *Gait & Posture* 42.1 (2015), pp. 96–100.

[76] Julie A Bobholz and Stephen M Rao. "Cognitive dysfunction in multiple sclerosis: a review of recent developments". In: *Current opinion in neurology* 16.3 (2003), pp. 283–288.

[77] Stephen M Rao, Gary J Leo, Linda Bernardin, and Frederick Unverzagt. "Cognitive dysfunction in multiple sclerosis. I. Frequency, patterns, and prediction". In: *Neurology* 41.5 (1991), pp. 685–691.

[78] Anne Katrine Andreasen, PE Spliid, Henning Andersen, and J Jakobsen. "Fatigue and processing speed are related in multiple sclerosis". In: *European journal of Neurology* 17.2 (2010), pp. 212–218.

[79] Massimo Filippi, Maria A Rocca, Olga Ciccarelli, Nicola De Stefano, Nikos Evangelou, Ludwig Kappos, Alex Rovira, Jaume Sastre-Garriga, Mar Tintorè, and Jette L Frederiksen. "MRI criteria for the diagnosis of multiple sclerosis: MAGNIMS consensus guidelines". In: *The Lancet Neurology* 15.3 (2016), pp. 292–303.

[80] Peter A Brex, Olga Ciccarelli, Jonathon I O'Riordan, Michael Sailer, Alan J Thompson, and David H Miller. "A longitudinal study of abnormalities on MRI and disability from multiple sclerosis". In: *New England Journal of Medicine* 346.3 (2002), pp. 158–164.

[81] Martin Daumer, Anneke Neuhaus, S Morrissey, Rogier Hintzen, and George Cornell Ebers. "MRI as an outcome in multiple sclerosis clinical trials". In: *Neurology* 72.8 (2009), pp. 705–711.

[82] Y Ge. "Multiple sclerosis: the role of MR imaging". In: *American Journal of Neuroradiology* 27.6 (2006), pp. 1165–1176.

[83] Ludwig Kappos, David Li, Peter A Calabresi, Paul O'Connor, Amit Bar-Or, Frederik Barkhof, Ming Yin, David Leppert, Robert Glanzman, and Jeroen Tinbergen. "Ocrelizumab in relapsing-remitting multiple sclerosis: a phase 2, randomised, placebo-controlled, multicentre trial". In: *The Lancet* 378.9805 (2011), pp. 1779–1787.

[84] Xavier Montalban, Stephen L Hauser, Ludwig Kappos, Douglas L Arnold, Amit Bar-Or, Giancarlo Comi, Jerome De Seze, Gavin Giovannoni, Hans-Peter Hartung, and Bernhard Hemmer. "Ocrelizumab versus placebo in primary progressive multiple sclerosis". In: *New England Journal of Medicine* 376.3 (2017), pp. 209–220.

[85] Stephen L Hauser, Amit Bar-Or, Giancarlo Comi, Gavin Giovannoni, Hans-Peter Hartung, Bernhard Hemmer, Fred Lublin, Xavier Montalban, Kottil W Rammohan, and Krzysztof Selmaj. "Ocrelizumab versus interferon beta-1a in relapsing multiple sclerosis". In: *New England Journal of Medicine* 376.3 (2017), pp. 221–234.

[86] Vincenzo Parisi, Gianluca Manni, Maria Spadaro, Gaspare Colacino, Rita Restuccia, Stefano Marchi, Massimo G Bucci, and Francesco Pierelli. "Correlation between morphological and functional retinal impairment in multiple sclerosis patients". In: *Investigative ophthalmology & visual science* 40.11 (1999), pp. 2520–2527.

[87] Lauren S Talman, Esther R Bisker, David J Sackel, David A Long Jr, Kristin M Galetta, John N Ratchford, Deacon J Lile, Sheena K Farrell, Michael J Loguidice, and Gina Remington. "Longitudinal study of vision and retinal nerve fiber layer thickness in multiple sclerosis". In: *Annals of neurology* 67.6 (2010), pp. 749–760.

[88] Christina V Warner, Stephanie B Syc, Aleksandra M Stankiewicz, Girish Hiremath, Sheena K Farrell, Ciprian M Crainiceanu, Amy Conger, Teresa C Frohman, Esther R Bisker, and Laura J Balcer. "The impact of utilizing different optical coherence tomography devices for clinical purposes and in multiple sclerosis trials". In: *PLoS One* 6.8 (2011), e22947.

[89] Jennifer B Fisher, Dina A Jacobs, Clyde E Markowitz, Steven L Galetta, Nicholas J Volpe, M Ligia Nano-Schiavi, Monika L Baier, Elliot M Frohman, Heather Winslow, and Teresa C Frohman. "Relation of visual function to retinal nerve fiber layer thickness in multiple sclerosis". In: *Ophthalmology* 113.2 (2006), pp. 324–332.

[90] MAA Van Walderveen, F Barkhof, OR Hommes, CH Polman, H Tobi, STFM Frequin, and J Valk. "Correlating MRI and clinical disease activity in multiple sclerosis: relevance of hypointense lesions on short-TR/short-TE (T1-weighted) spin-echo images". In: *Neurology* 45.9 (1995), pp. 1684–1690.

[91] G. R. Cutter, M. L. Baier, R. A. Rudick, D. L. Cookfair, J. S. Fischer, J. Petkau, et al. "Development of a multiple sclerosis functional composite as a clinical trial outcome measure". In: *Brain* 122 (Pt 5) (1999), pp. 871–82.

[92] Gammon M Earhart, Jim T Cavanaugh, Terry Ellis, Matt P Ford, K Bo Foreman, and Lee Dibble. "The 9-hole PEG test of upper extremity function: average values, test-retest reliability, and factors contributing to performance in people with Parkinson disease". In: *Journal of Neurologic Physical Therapy* 35.4 (2011), pp. 157–163.

[93] Donald E Goodkin, Doris Hertsguard, and Judy Seminary. "Upper extremity function in multiple sclerosis: improving assessment sensitivity with box-and-block and nine-hole peg tests". In: *Archives of physical medicine and rehabilitation* 69.10 (1988), pp. 850–854.

[94] JJ Kragt, F AH van der Linden, JM Nielsen, B MJ Uitdehaag, and CH Polman. "Clinical impact of 20% worsening on Timed 25-foot Walk and 9-hole Peg Test in multiple sclerosis". In: *Multiple Sclerosis Journal* 12.5 (2006), pp. 594–598.

[95] D. Cadavid, J. A. Cohen, M. S. Freedman, M. D. Goldman, H. P. Hartung, E. Havrdova, D. Jeffery, R. Kapoor, A. Miller, F. Sellebjerg, D. Kinch, S. Lee, S. Shang, and D. Mikol. "The EDSS-Plus, an improved endpoint for disability progression in secondary progressive multiple sclerosis". In: *Mult Scler* 23.1 (2017), pp. 94–105.

[96] Cristina Forn, Antonio Belenguer, M Antonia Parcet-Ibars, and César Avila. "Information-processing speed is the primary deficit underlying the poor performance of multiple sclerosis patients in the Paced Auditory Serial Addition Test (PASAT)". In: *Journal of clinical and experimental neuropsychology* 30.7 (2008), pp. 789–796.

[97] J. P. Stellmann, A. Neuhaus, N. Götze, S. Briken, C. Lederer, M. Schimpl, C. Heesen, and M. Daumer. "Ecological Validity of Walking Capacity Tests in Multiple Sclerosis". In: *PLOS ONE* 10.4 (2015), e0123822.

[98] R. W. Bohannon, Y. C. Wang, and R. C. Gershon. "Two-minute walk test performance by adults 18 to 85 years: normative values, reliability, and responsiveness". In: *Arch Phys Med Rehabil* 96.3 (2015), pp. 472–7.

[99] Domien Gijbels, BO Eijnde, and Peter Feys. "Comparison of the 2-and 6-minute walk test in multiple sclerosis". In: *Multiple Sclerosis Journal* 17.10 (2011), pp. 1269–1272.

[100] Andreas Buch Møller, Bo Martin Bibby, Anders Guldhammer Skjerbæk, Ellen Jensen, Henrik Sørensen, Egon Stenager, and Ulrik Dalgas. "Validity and variability of the 5-repetition sit-to-stand test in patients with multiple sclerosis". In: *Disability and rehabilitation* 34.26 (2012), pp. 2251–2258.

[101] Davide Cattaneo, Alberto Regola, and Matteo Meotti. "Validity of six balance disorders scales in persons with multiple sclerosis". In: *Disability and rehabilitation* 28.12 (2006), pp. 789–795.

[102] JH Noseworthy, MK Vandervoort, CJ Wong, and GC Ebers. "Interrater variability with the Expanded Disability Status Scale (EDSS) and Functional Systems (FS) in a multiple sclerosis clinical trial". In: *Neurology* 40.6 (1990), pp. 971–971.

[103] MH Verdier-Taillefer, M Zuber, O Lyon-Caen, M Clanet, O Gout, C Louis, and A Alpérovitch. "Observer disagreement in rating neurologic impairment in multiple sclerosis: facts and consequences". In: *European neurology* 31.2 (1991), pp. 117–119.

[104] Donald E Goodkin, D Cookfair, K Wende, D Bourdette, P Pullicino, B Scherokman, and R Whitham. "Inter-and intrarater scoring agreement using grades 1.0 to 3.5 of the Kurtzke Expanded Disability Status Scale (EDSS)". In: *Neurology* 42.4 (1992), pp. 859–859.

[105] Sandra Meyer-Moock, You-Shan Feng, Mathias Maeurer, Franz-Werner Dippel, and Thomas Kohlmann. "Systematic literature review and validity evaluation of the Expanded Disability Status Scale (EDSS) and the Multiple Sclerosis Functional Composite (MSFC) in patients with multiple sclerosis". In: *BMC neurology* 14.1 (2014), p. 58.

[106] MJ Hohol, EJ Orav, and HL Weiner. "Disease steps in multiple sclerosis: a longitudinal study comparing disease steps and EDSS to evaluate disease progression". In: *Multiple Sclerosis Journal* 5.5 (1999), pp. 349–354.

[107] LVAE Bosma, JJ Kragt, L Brieva, Z Khaleeli, X Montalban, CH Polman, AJ Thompson, M Tintore, and BMJ Uitdehaag. "Progression on the Multiple Sclerosis Functional Composite in multiple sclerosis: what is the optimal cut-off for the three components?" In: *Multiple Sclerosis Journal* 16.7 (2010), pp. 862–867.

[108] JJ Kragt, JM Nielsen, F AH van der Linden, B MJ Uitdehaag, and CH Polman. "How similar are commonly combined criteria for EDSS progression in multiple sclerosis?" In: *Multiple Sclerosis Journal* 12.6 (2006), pp. 782–786.

[109] RA Rudick, CH Polman, JA Cohen, MK Walton, et al. "Assessing disability progression with the multiple sclerosis functional composite". In: *Multiple Sclerosis Journal* 15.8 (2009), pp. 984–997.

[110] SR Schwid, AD Goodman, MP McDermott, CF Bever, and SD Cook. "Quantitative functional measures in MS: what is a reliable change?" In: *Neurology* 58.8 (2002), pp. 1294–1296.

[111] Kimatha Oxford Grice, Kimberly A Vogel, Viet Le, Ana Mitchell, Sonia Muniz, and Mary Ann Vollmer. "Adult norms for a commercially available Nine Hole Peg Test for finger dexterity". In: *American Journal of Occupational Therapy* 57.5 (2003), pp. 570–573.

[112] O.R. Pearson, M.E. Busse, R.W.M. van Deursen, and C.M. Wiles. "Quantification of walking mobility in neurological disorders". In: *QJM: An International Journal of Medicine* 97.8 (2004), pp. 463–475.

[113] Richard W Bohannon, A Williams Andrews, and Michael W Thomas. "Walking speed: reference values and correlates for older adults". In: *Journal of Orthopaedic & Sports Physical Therapy* 24.2 (1996), pp. 86–90.

[114] Myla D Goldman, Robert W Motl, John Scagnelli, John H Pula, Jacob J Sosnoff, and Diego Cadavid. "Clinically meaningful performance benchmarks in MS: timed 25-foot walk and the real world". In: *Neurology* 81.21 (2013), pp. 1856–1863.

[115] Joshua T Cohen. "Walking speed and economic outcomes for walking-impaired patients with multiple sclerosis". In: *Expert review of pharmacoeconomics & outcomes research* 10.5 (2010), pp. 595–603.

[116] Francois Bethoux and Susan Bennett. "Evaluating walking in patients with multiple sclerosis: which assessment tools are useful in clinical practice?" In: *International journal of MS care* 13.1 (2011), pp. 4–14.

[117] Kyle Strimbu and Jorge A. Tavel. "What are Biomarkers?" In: *Current opinion in HIV and AIDS* 5.6 (2010), pp. 463–466.

[118] Victor G De Gruttola, Pamela Clax, David L DeMets, Gregory J Downing, Susan S Ellenberg, Lawrence Friedman, Mitchell H Gail, Ross Prentice, Janet Wittes, and Scott L Zeger. "Considerations in the evaluation of surrogate endpoints in clinical trials: summary of a National Institutes of Health workshop". In: *Controlled clinical trials* 22.5 (2001), pp. 485–502.

[119] Richard Frank and Richard Hargreaves. "Clinical biomarkers in drug discovery and development". In: *Nature reviews. Drug discovery* 2.7 (2003), p. 566.

[120] Christopher J Weir and Rosalind J Walley. "Statistical evaluation of biomarkers as surrogate endpoints: a literature review". In: *Statistics in medicine* 25.2 (2006), pp. 183–203.

[121] C. Malmeström, S. Haghighi, L. Rosengren, O. Andersen, and J. Lycke. "Neurofilament light protein and glial fibrillary acidic protein as biological markers in MS". In: *Neurology* 61.12 (2003), pp. 1720–5.

[122] Giampaolo Brichetto, Ludovico Pedullà, Jessica Podda, and Andrea Tacchino. "Beyond center-based testing: Understanding and improving functioning with wearable technology in MS". In: *Multiple Sclerosis Journal* 25.10 (2019), pp. 1402–1411.

[123] Florian Lipsmeier, Kirsten I Taylor, Timothy Kilchenmann, Detlef Wolf, Alf Scotland, et al. "Evaluation of smartphone-based testing to generate exploratory outcome measures in a phase 1 Parkinson's disease clinical trial". In: *Movement Disorders* 33.8 (2018), pp. 1287–1297.

[124] Florian Lipsmeier, Cedric Simillion, Atieh Bamdadian, et al. "Preliminary Reliability and Validity of a Novel Digital Biomarker Smartphone Application to Assess Cognitive and Motor Symptoms in Huntington's Disease (HD)(P1. 8-042)". In: AAN Enterprises, 2019.

[125] Lampros C Kourtis, Oliver B Regele, Justin M Wright, and Graham B Jones. "Digital biomarkers for Alzheimer's disease: the mobile/wearable devices opportunity". In: *NPJ digital medicine* 2.1 (2019), pp. 1–9.

[126] Sven Meister, Wolfgang Deiters, and Stefan Becker. "Digital health and digital biomarkers – enabling value chains on health data". In: *Current Directions in Biomedical Engineering* 2.1 (2016), p. 577.

[127] John Torous, Jorge Rodriguez, and Adam Powell. "The New Digital Divide For Digital Biomarkers". In: *Digital Biomarkers* (2017).

[128] Mike Baker, Johan van Beek, and Christian Gossens. *Digital health: Smartphone-based monitoring of multiple sclerosis using Floodlight*. 2019.

[129] J. M. Balto, D. L. Kinnett-Hopkins, and R. W. Motl. "Accuracy and precision of smartphone applications and commercially available motion sensors in multiple sclerosis". In: *Mult Scler J Exp Transl Clin* 2 (2016), p. 2055217316634754.

[130] R. Bove, C. C. White, G. Giovannoni, B. Glanz, V. Golubchikov, J. Hujol, C. Jennings, D. Langdon, M. Lee, A. Legedza, J. Paskavitz, S. Prasad, J. Richert, A. Robbins, S. Roberts, H. Weiner, R. Ramachandran, M. Botfield, and P. L. De Jager. "Evaluating more naturalistic outcome measures: A 1-year smartphone study in multiple sclerosis". In: *Neurol Neuroimmunol Neuroinflamm* 2.6 (2015), e162.

[131] Elisabeth Maillart, Pierre Labauge, Mikael Cohen, Adil Maarouf, Sandra Vukusic, Cécile Donzé, Philippe Gallien, Jérôme De Sèze, Bertrand Bourre, and Thibault Moreau. "MSCopilot, a new Multiple Sclerosis self-assessment digital solution: results of a comparative study versus standard tests". In: *European journal of neurology* (2019).

[132] Alexandra K. Boukhvalova, Emily Kowalczyk, Thomas Harris, Peter Kosa, Alison Wichman, Mary A. Sandford, Atif Memon, and Bibiana Bielekova. "Identifying and Quantifying Neurological Disability via Smartphone". In: *Frontiers in Neurology* 9.740 (2018).

[133] Daphne Kos, Jynke Raeymaekers, A Van Remoortel, MB D'hooghe, G Nagels, Miguel D'Haeseleer, E Peeters, T Dams, and T Peeters. "Electronic visual analogue scales for pain, fatigue, anxiety and quality of life in people with multiple sclerosis using smartphone and tablet: a reliability and feasibility study". In: *Clinical Rehabilitation* 31.9 (2017), pp. 1215–1225.

[134] Somayeh Aghanavesi, Dag Nyholm, Marina Senek, Filip Bergquist, and Mevludin Memedi. "A smartphone-based system to quantify dexterity in Parkinson's disease patients". In: *Informatics in Medicine Unlocked* 9 (2017), pp. 11–17.

[135] K. Banaszkiewicz, M. Rudzinska, S. Bukowczan, A. Izworski, and A. Szczudlik. "Spiral drawing time as a measure of bradykinesia". In: *Neurol Neurochir Pol* 43.1 (2009), pp. 16–21.

[136] P. Feys, W. Helsen, A. Prinsmel, S. Ilsbroukx, S. Wang, and X. Liu. "Digitised spirography as an evaluation tool for intention tremor in multiple sclerosis". In: *J Neurosci Methods* 160.2 (2007), pp. 309–16.

[137] Min Wang, Bei Wang, Junzhong Zou, and Masatoshi Nakamura. "A new quantitative evaluation method of spiral drawing for patients with Parkinson's disease based on a polar coordinate system with varying origin". In: *Physica A: Statistical Mechanics and its Applications* 391.18 (2012), pp. 4377–4388.

[138] Mitchell Grant Longstaff and Richard A Heath. "Spiral drawing performance as an indicator of fine motor function in people with multiple sclerosis". In: *Human movement science* 25.4 (2006), pp. 474–491.

[139] R. Rudick, J. Antel, C. Confavreux, G. Cutter, G. Ellison, J. Fischer, F. Lublin, A. Miller, J. Petkau, S. Rao, S. Reingold, K. Syndulko, A. Thompson, J. Wallenberg, B. Weinshenker, and E. Willoughby. "Clinical outcomes assessment in multiple sclerosis". In: *Ann Neurol* 40.3 (1996), pp. 469–79.

[140] Michael P. Caligiuri, Hans-Leo Teulings, J. Vincent Filoteo, David Song, and James B. Lohr. "Quantitative measurement of handwriting in the assessment of drug-induced parkinsonism". In: *Human Movement Science* 25.4–5 (2006), pp. 510–522.

[141] Xuguang Liu, Camille B. Carroll, Shou-Yan Wang, John Zajicek, and Peter G. Bain. "Quantifying drug-induced dyskinesias in the arms using digitised spiral-drawing tasks". In: *Journal of Neuroscience Methods* 144.1 (2005), pp. 47–52.

[142] Manuela Galli, Sara L Vimercati, Elena Manetti, Veronica Cimolin, Giorgio Albertini, and Maria F De Pandis. "Spiral analysis in subjects with Parkinson's disease before and after levodopa treatment: a new protocol with stereophotogrammetric systems". In: *Journal of applied biomaterials & functional materials* 12.2 (2014).

[143] M. Memedi, S. Aghanavesi, and J. Westin. "A method for measuring Parkinson's disease related temporal irregularity in spiral drawings". In: *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. 2016, pp. 410–413.

[144] Aleksander Sadikov, Vida Groznik, Martin Možina, Jure Žabkar, Dag Nyholm, Mevludin Memedi, and Dejan Georgiev. "Feasibility of spirography features for objective assessment of motor function in Parkinson's disease". In: *Artificial Intelligence in Medicine* (2017).

[145] Aghanavesi Somayeh, Mevludin Memedi, Mark Dougherty, Dag Nyholm, and Jerker Westin. "Measuring temporal irregularity in spiral drawings of patients with Parkinson's disease". In: *21st International Congress of Parkinson's Disease and Movement Disorders, Vancouver, BC, Canada*. Vol. 32. 2017.

[146] Andrea Vianello, Luca Chittaro, Stefano Burigat, and Riccardo Budai. "MotorBrain: A mobile app for the assessment of users' motor performance in neurology". In: *computer methods and programs in biomedicine* 143 (2017), pp. 35–47.

[147] Judith J. W. van Beek, Erwin E. H. van Wegen, Cleo D. Bol, Marc B. Rietberg, Christian P. Kamm, and Tim Vanbellingen. "Tablet App Based Dexterity Training in Multiple Sclerosis (TAD-MS): Research Protocol of a Randomized Controlled Trial". In: *Frontiers in Neurology* 10.61 (2019).

[148] L Eduardo Cofré Lizama, Fary Khan, Peter VS Lee, and Mary P Galea. "The use of laboratory gait analysis for understanding gait deterioration in people with multiple sclerosis". In: *Multiple Sclerosis Journal* 22.14 (2016), pp. 1768–1776.

[149] Charlotte Elsworth-Edelsten, Alice Bonnefoy-Mazure, Magali Laidet, Stephane Armand, Frederic Assal, Patrice Lalive, and Gilles Allali. "Upper limb movement analysis during gait in multiple sclerosis patients". In: *Human movement science* 54 (2017), pp. 248–252.

[150] Maddalena Sparaco, Luigi Lavorgna, Renata Conforti, Gioacchino Tedeschi, and Simona Bonavita. "The role of wearable devices in multiple sclerosis". In: *Multiple sclerosis international* 2018 (2018).

[151] Delaram Jarchi, James Pope, Tracey KM Lee, Larisa Tamjidi, Amirhosein Mirzaei, and Saeid Sanei. "A Review on Accelerometry Based Gait Analysis and Emerging Clinical Applications". In: *IEEE Reviews in Biomedical Engineering* (2018).

[152] ACRMDOG Godfrey, Richard Conway, David Meagher, and Gearoid Ó'Laighin. "Direct measurement of human movement by accelerometry". In: *Medical engineering & physics* 30.10 (2008), pp. 1364–1386.

[153] Barry R Greene, Stephanie Rutledge, Iain McGurgan, Christopher McGuigan, Karen O'Connell, Brian Caulfield, and Niall Tubridy. "Assessment and classification of early-stage multiple sclerosis with inertial sensors: comparison against clinical measures of disease state". In: *IEEE journal of biomedical and health informatics* 19.4 (2015), pp. 1356–1361.

[154] Rebecca I Spain, Martina Mancini, Fay B Horak, and Dennis Bourdette. "Body-worn sensors capture variability, but not decline, of gait and balance measures in multiple sclerosis over 18 months". In: *Gait & posture* 39.3 (2014), pp. 958–964.

[155] Michael Psarakis, David A Greene, Michael H Cole, Stephen R Lord, Phu Hoang, and Matthew Brodie. "Wearable technology reveals gait compensations, unstable walking patterns and fatigue in people with multiple sclerosis". In: *Physiological measurement* 39.7 (2018), p. 075004.

[156] Lorenza Angelini, William Hodgkinson, Craig Smith, Jessie Moorman Dodd, Basil Sharrack, Claudia Mazzà, and David Paling. "Wearable sensors can reliably quantify gait alterations associated with disability in people with progressive multiple sclerosis in a clinical setting". In: *Journal of Neurology* (2020).

[157] Christopher Buckley, Brook Galna, Lynn Rochester, and Claudia Mazzà. "Upper body accelerations as a biomarker of gait impairment in the early stages of Parkinson's disease". In: *Gait & posture* 71 (2019), pp. 289–295.

[158] Andrew Creagh, Isabelle Killane, Fiachra Maguire, Orna Donoghue, Rose Anne Kenny, and Richard Reilly. "Examining the links between domains of gait and tests of global cognition: a factor analysis approach". In: *Age and Ageing* 45.suppl$_2$ (2016), pp. ii13–ii56.

[159] Sue Lord, Brook Galna, Joe Verghese, Shirley Coleman, David Burn, and Lynn Rochester. "Independent domains of gait in older adults and associated motor and nonmotor attributes: validation of a factor analysis approach". In: *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences* 68.7 (2012), pp. 820–827.

[160] Rosie Morris, Aodhán Hickey, Silvia Del Din, Alan Godfrey, Sue Lord, and Lynn Rochester. "A model of free-living gait: A factor analysis in Parkinson's disease". In: *Gait & Posture* 52 (2017), pp. 68–71.

[161] S. R. Dandu, M. M. Engelhard, A. Qureshi, J. Gong, J. C. Lach, M. Brandt-Pearce, and M. D. Goldman. "Understanding the Physiological Significance of Four Inertial Gait Features in Multiple Sclerosis". In: *IEEE Journal of Biomedical and Health Informatics* 22.1 (2018), pp. 40–46.

[162] Asma Qureshi, Maite Brandt-Pearce, Matthew M Engelhard, and Myla D Goldman. "Relationship between kernel density function estimates of gait time series and clinical data". In: *IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 2017, pp. 329–332.

[163] Sriram Raju Dandu, Matthew M Engelhard, Myla D Goldman, and John Lach. "Determining physiological significance of inertial gait features in multiple sclerosis". In: *IEEE 13th International Conference on: Wearable and Implantable Body Sensor Networks (BSN)*. IEEE, 2016, pp. 266–271.

[164] Matthew M. Engelhard, Sriram Raju Dandu, Stephen D. Patek, John C. Lach, and Myla D. Goldman. "Quantifying Six-Minute Walk Induced Gait Deterioration with Inertial Sensors in Multiple Sclerosis Subjects". In: *Gait & posture* 49 (2016), pp. 340–345.

[165] Ruopeng Sun, Katherine L. Hsieh, and Jacob J. Sosnoff. "Fall Risk Prediction in Multiple Sclerosis Using Postural Sway Measures: A Machine Learning Approach". In: *Scientific Reports* 9.1 (2019), p. 16154.

[166] Vrutangkumar V. Shah, James McNames, Martina Mancini, Patricia Carlson-Kuhta, Rebecca I. Spain, et al. "Quantity and quality of gait and turning in people with multiple sclerosis, Parkinson's disease and matched controls during daily living". In: *Journal of Neurology* 267.4 (2020), pp. 1188–1196.

[167] Fabio A Storm, KPS Nair, Alison J Clarke, Jill M Van der Meulen, and Claudia Mazza. "Free-living and laboratory gait characteristics in patients with multiple sclerosis". In: *PloS one* 13.5 (2018).

[168] Akara Supratak, Gourab Datta, Arie R. Gafson, Richard Nicholas, Yike Guo, and Paul M. Matthews. "Remote Monitoring in the Home Validates Clinical Gait Measures for Multiple Sclerosis". In: *Frontiers in Neurology* 9.561 (2018).

[169] M. J. Bradshaw, S. Farrow, R. W. Motl, and T. Chitnis. "Wearable biosensors to monitor disability in multiple sclerosis". In: *Neurol Clin Pract* 7.4 (2017), pp. 354–362.

[170] A. Zhan, S. Mohan, C. Tarolli, and et al. "Using smartphones and machine learning to quantify parkinson disease severity: The mobile parkinson disease score". In: *JAMA Neurology* (2018).

[171] Andong Zhan, Max A Little, Denzil A Harris, Solomon O Abiola, E Dorsey, Suchi Saria, and Andreas Terzis. "High Frequency Remote Monitoring of Parkinson's Disease via Smartphone: Platform Overview and Medication Response Detection". In: *arXiv preprint arXiv:1601.00960* (2016).

[172] Robert J Ellis, Yee Sien Ng, Shenggao Zhu, Dawn M Tan, Boyd Anderson, Gottfried Schlaug, and Ye Wang. "A validated smartphone-based assessment of gait and gait variability in Parkinson's disease". In: *PLoS one* 10.10 (2015), e0141694.

[173] VJ Block, Antoine Lizee, E Crabtree-Hartman, CJ Bevan, Jennifer S Graves, Riley Bove, AJ Green, Bardia Nourbakhsh, Matthew Tremblay, and P-A Gourraud. "Continuous daily assessment of multiple sclerosis disability using remote step count monitoring". In: *Journal of neurology* 264.2 (2017), pp. 316–326.

[174] P. DasMahapatra, E. Chiauzzi, R. Bhalerao, and J. Rhodes. "Free-Living Physical Activity Monitoring in Adult US Patients with Multiple Sclerosis Using a Consumer Wearable Device". In: *Digital Biomarkers* 2.1 (2018), pp. 47–63.

[175] Seungeun Chung, Jiyoun Lim, Kyoung Ju Noh, Gague Kim, and Hyuntae Jeong. "Sensor Data Acquisition and Multimodal Sensor Fusion for Human Activity Recognition Using Deep Learning". In: *Sensors* 19.7 (2019), p. 1716.

[176] Mohammed Mehedi Hassan, Md Zia Uddin, Amr Mohamed, and Ahmad Almogren. "A robust human activity recognition system using smartphone sensors and deep learning". In: *Future Generation Computer Systems* 81 (2018), pp. 307–313.

[177] Xiangyu Jiang, Yonggang Lu, Zhenyu Lu, and Huiyu Zhou. "Smartphone-Based Human Activity Recognition Using CNN in Frequency Domain". In: Web and Big Data. Springer International Publishing, 2018, pp. 101–110.

[178] Wesllen Sousa Lima, Eduardo Souto, Khalil El-Khatib, Roozbeh Jalali, and Joao Gama. "Human activity recognition using inertial sensors in a smartphone: An overview". In: *Sensors* 19.14 (2019), p. 3213.

[179] Kun Xia, Jianguang Huang, and Hanyu Wang. "LSTM-CNN Architecture for Human Activity Recognition". In: *IEEE Access* 8 (2020), pp. 56855–56866.

[180] Anantha Narayanan, Farzanah Desai, Tom Stewart, Scott Duncan, and Lisa Mackay. "Application of Raw Accelerometer Data and Machine-Learning Techniques to Characterize Human Movement Behavior: A Systematic Scoping Review". In: *Journal of Physical Activity and Health* 17.3 (2020), pp. 360–383.

[181] Matthew Willetts, Sven Hollowell, Louis Aslett, Chris Holmes, and Aiden Doherty. "Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 UK Biobank participants". In: *Scientific reports* 8.1 (2018), pp. 1–10.

[182] Qin Zou, Yanling Wang, Qian Wang, Yi Zhao, and Qingquan Li. "Deep Learning-Based Gait Recognition Using Smartphones in the Wild". In: *IEEE Transactions on Information Forensics and Security* 15 (2020), pp. 3197–3212.

[183] Fabian Horst, Sebastian Lapuschkin, Wojciech Samek, Klaus-Robert Müller, and Wolfgang I. Schöllhorn. "Explaining the unique nature of individual gait patterns with deep learning". In: *Scientific Reports* 9.1 (2019), p. 2391.

[184] Matteo Gadaleta and Michele Rossi. "Idnet: Smartphone-based gait recognition with convolutional neural networks". In: *Pattern Recognition* 74 (2018), pp. 25–37.

[185] S Arora, V Venkataraman, A Zhan, S Donohue, KM Biglan, ER Dorsey, and MA Little. "Detecting and monitoring the symptoms of Parkinson's disease using smartphones: a pilot study". In: *Parkinsonism & related disorders* 21.6 (2015), pp. 650–653.

[186] Siddharth Arora, Vinayak Venkataraman, Sean Donohue, Kevin M Biglan, Earl R Dorsey, and Max A Little. "High accuracy discrimination of Parkinson's disease participants from healthy controls using smartphones". In: *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 3641–3644.

[187] John Prince, Siddharth Arora, and Maarten de Vos. "Big data in Parkinson's disease: using smartphones to remotely detect longitudinal disease phenotypes". In: *Physiological measurement* 39.4 (2018), p. 044005.

[188] Hanrui Zhang, Kaiwen Deng, Hongyang Li, Roger L. Albin, and Yuanfang Guan. "Deep Learning Identifies Digital Biomarkers for Self-Reported Parkinson's Disease". In: *Patterns* (2020), p. 100042.

[189] Patrick Schwab and Walter Karlen. "PhoneMD: Learning to diagnose Parkinson's disease from smartphone data". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 2019, pp. 1118–1125.

[190] John Prince, Fernando Andreotti, and Maarten De Vos. "Multi-Source Ensemble Learning for the Remote Prediction of Parkinson's Disease in the Presence of Source-Wise Missing Data". In: *IEEE Transactions on Biomedical Engineering* 66.5 (2018), pp. 1402–1411.

[191] Muriel Lang, Franz MJ Pfister, Jakob Fröhner, Kian Abedinpour, Daniel Pichler, Urban Fietzek, Terry Um, et al. "A Multi-Layer Gaussian Process for Motor Symptom Estimation in People With Parkinson's Disease". In: *IEEE Transactions on Biomedical Engineering* 66.11 (2019), pp. 3038–3049.

[192] Sergi Valverde, Mariano Cabezas, Eloy Roura, Sandra González-Villà, Deborah Pareto, Joan C Vilanova, Lluís Ramió-Torrentà, Àlex Rovira, Arnau Oliver, and Xavier Lladó. "Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach". In: *NeuroImage* 155 (2017), pp. 159–168.

[193] Tom Brosch, Youngjin Yoo, Lisa YW Tang, David KB Li, Anthony Traboulsee, and Roger Tam. "Deep convolutional encoder networks for multiple sclerosis lesion segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 3–11.

[194] Fabian Eitel, Emily Soehler, Judith Bellmann-Strobl, Alexander U Brandt, Klemens Ruprecht, René M Giess, Joseph Kuchling, Susanna Asseyer, Martin Weygandt, and John-Dylan Haynes. "Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation". In: *NeuroImage: Clinical* 24 (2019), p. 102003.

[195] Yudong Zhang, Siyuan Lu, Xingxing Zhou, Ming Yang, Lenan Wu, Bin Liu, Preetha Phillips, and Shuihua Wang. "Comparison of machine learning methods for stationary wavelet entropy-based multiple sclerosis detection: decision tree, k-nearest neighbors, and support vector machine". In: *Simulation* 92.9 (2016), pp. 861–871.

[196] Yijun Zhao, Brian C. Healy, Dalia Rotstein, Charles R. G. Guttmann, Rohit Bakshi, Howard L. Weiner, Carla E. Brodley, and Tanuja Chitnis. "Exploration of machine learning techniques in predicting multiple sclerosis disease course". In: *PLOS ONE* 12.4 (2017), e0174866.

[197] Gong, Goldman, and Lach. "Deepmotion: a deep convolutional neural network on inertial body sensors for gait assessment in multiple sclerosis". In: *IEEE Wireless Health (WH)*. 2016, pp. 1–8.

[198] Brett M Meyer, Lindsey J Tulipani, Reed D Gurchiek, Dakota A Allen, Lukas Adamowicz, Dale Larie, Andrew J Solomon, Nick Cheney, and Ryan McGinnis. "Wearables and Deep Learning Classify Fall Risk from Gait in Multiple Sclerosis". In: *IEEE journal of biomedical and health informatics* (2020).

[199] Svante Wold, Kim Esbensen, and Paul Geladi. "Principal component analysis". In: *Chemometrics and intelligent laboratory systems* 2.1-3 (1987), pp. 37–52.

[200] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[201] Laurens van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE". In: *Journal of Machine Learning Research* 9.Nov (2008), pp. 2579–2605.

[202] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[203] Karl Pearson. "VII. Note on regression and inheritance in the case of two parents". In: *proceedings of the royal society of London* 58.347-352 (1895), pp. 240–242.

[204] Wayne W Daniel and Chad L Cross. *Biostatistics: a foundation for analysis in the health sciences*. Wiley, 2018.

[205] Charles Spearman. "The proof and measurement of association between two things". In: (1961).

[206] Patrick E Shrout and Joseph L Fleiss. "Intraclass correlations: uses in assessing rater reliability". In: *Psychological bulletin* 86.2 (1979), p. 420.

[207] Joseph P Weir. "Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM". In: *The Journal of Strength & Conditioning Research* 19.1 (2005), pp. 231–240.

[208] Terry K Koo and Mae Y Li. "A guideline of selecting and reporting intraclass correlation coefficients for reliability research". In: *Journal of chiropractic medicine* 15.2 (2016), pp. 155–163.

[209] Kenneth O McGraw and Seok P Wong. "Forming inferences about some intraclass correlation coefficients". In: *Psychological methods* 1.1 (1996), p. 30.

[210] Myles Hollander, Douglas A Wolfe, and Eric Chicken. *Nonparametric statistical methods*. Vol. 751. John Wiley & Sons, 2013.

[211] Jean Dickinson Gibbons and Subhabrata Chakraborti. *Nonparametric Statistical Inference: Revised and Expanded*. CRC press, 2014.

[212] Egon Sharpe Pearson, William Sealy Gosset, RL Plackett, and George Alfred Barnard. *Student: a statistical biography of William Sealy Gosset*. Oxford University Press, USA, 1990.

[213] (William Sealy Gosset) Student. "The probable error of a mean". In: *Biometrika* (1908), pp. 1–25.

[214] Bernard L Welch. "The generalization of student's' problem when several different population variances are involved". In: *Biometrika* 34.1/2 (1947), pp. 28–35.

[215] Frank J Massey Jr. "The Kolmogorov-Smirnov test for goodness of fit". In: *Journal of the American statistical Association* 46.253 (1951), pp. 68–78.

[216] Henry B Mann and Donald R Whitney. "On a test of whether one of two random variables is stochastically larger than the other". In: *The annals of mathematical statistics* (1947), pp. 50–60.

[217] Olive Jean Dunn and Virginia A Clark. *Applied statistics: analysis of variance and regression*. Report. John Wiley & Sons, 1987.

[218] William H Kruskal and W Allen Wallis. "Use of ranks in one-criterion variance analysis". In: *Journal of the American statistical Association* 47.260 (1952), pp. 583–621.

[219] Morton B Brown and Alan B Forsythe. "Robust tests for the equality of variances". In: *Journal of the American Statistical Association* 69.346 (1974), pp. 364–367.

[220] Juliet Popper Shaffer. "Multiple hypothesis testing". In: *Annual review of psychology* 46.1 (1995), pp. 561–584.

[221] John D Storey. "A direct approach to false discovery rates". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.3 (2002), pp. 479–498.

[222] Mikel Aickin and Helen Gensler. "Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods". In: *American journal of public health* 86.5 (1996), pp. 726–728.

[223] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[224] Calyampudi Radhakrishna Rao. *Linear statistical inference and its applications*. Vol. 2. Wiley New York, 1973.

[225] Julian J Faraway. *Linear models with R*. CRC press, 2014.

[226] Paul W Holland and Roy E Welsch. "Robust regression using iteratively reweighted least-squares". In: *Communications in Statistics-theory and Methods* 6.9 (1977), pp. 813–827.

[227] William Dumouchel and Fanny O'Brien. "Integrating a robust option into a multiple regression computing environment". In: *Computer science and statistics: Proceedings of the 21st symposium on the interface*. American Statistical Association Alexandria, VA, 1989, pp. 297–302.

[228] Peter McCullagh. *Generalized linear models*. Routledge, 2018.

[229] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*. Vol. 398. John Wiley & Sons, 2013.

[230] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[231] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

[232] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), pp. 267–288.

[233] Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[234] Bernhard Schölkopf, Alexander J Smola, and Francis Bach. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[235] Sebastian Raschka. "Model evaluation, model selection, and algorithm selection in machine learning". In: *arXiv preprint arXiv:1811.12808* (2018).

[236] Jason Weston, Sayan Mukherjee, Olivier Chapelle, Massimiliano Pontil, Tomaso Poggio, and Vladimir Vapnik. "Feature selection for SVMs". In: *Advances in neural information processing systems*. 2001, pp. 668–674.

[237] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[238] Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. "Conditional variable importance for random forests". In: *BMC bioinformatics* 9.1 (2008), p. 307.

[239] Robin Genuer, Jean-Michel Poggi, and Christine Tuleau-Malot. "Variable selection using random forests". In: *Pattern recognition letters* 31.14 (2010), pp. 2225–2236.

[240] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[241] Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization". In: *Advances in neural information processing systems*. 2014, pp. 2933–2941.

[242] Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. "Pegasos: Primal estimated sub-gradient solver for svm". In: *Mathematical programming* 127.1 (2011), pp. 3–30.

[243] Léon Bottou. "Large-scale machine learning with stochastic gradient descent". In: *Proceedings of COMPSTAT*. Springer, 2010, pp. 177–186.

[244] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. "On the importance of initialization and momentum in deep learning". In: *International conference on machine learning*. 2013, pp. 1139–1147.

[245] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*. Vol. 87. Springer Science & Business Media, 2013.

[246] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[247] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015), pp. 436–444.

[248] Christopher M Bishop. *Neural networks for pattern recognition*. Oxford university press, 1995.

[249] Sepp Hochreiter. "The vanishing gradient problem during learning recurrent neural nets and problem solutions". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6.02 (1998), pp. 107–116.

[250] 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.

[251] Zhicheng Cui, Wenlin Chen, and Yixin Chen. "Multi-scale convolutional neural networks for time series classification". In: *arXiv preprint arXiv:1603.06995* (2016).

[252] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *arXiv preprint arXiv:1502.03167* (2015).

[253] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting". In: *The journal of machine learning research* 15.1 (2014), pp. 1929–1958.

[254] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. "Towards better understanding of gradient-based attribution methods for deep neural networks". In: *arXiv preprint arXiv:1711.06104* (2017).

[255] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. "Layer-wise relevance propagation: an overview". In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 2019, pp. 193–209.

[256] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation". In: *PloS one* 10.7 (2015), e0130140.

[257] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. "Methods for interpreting and understanding deep neural networks". In: *Digital Signal Processing* 73 (2018), pp. 1–15.

[258] Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J Anders, and Klaus-Robert Müller. "Toward Interpretable Machine Learning: Transparent Deep Neural Networks and Beyond". In: *arXiv preprint arXiv:2003.07631* (2020).

[259] Maximilian Kohlbrenner, Alexander Bauer, Shinichi Nakajima, Alexander Binder, Wojciech Samek, and Sebastian Lapuschkin. "Towards best practice in explaining neural network decisions with LRP". In: *arXiv preprint arXiv:1910.09840* (2019).

[260] Sebastian Lapuschkin. "Opening the machine learning black box with layer-wise relevance propagation". In: (2019).

[261] Ron Kohavi. "A study of cross-validation and bootstrap for accuracy estimation and model selection". In: *Ijcai*. Vol. 14. Montreal, Canada, 1995, pp. 1137–1145.

[262] Sylvain Arlot and Alain Celisse. "A survey of cross-validation procedures for model selection". In: *Statistics surveys* 4 (2010), pp. 40–79.

[263] Sohrab Saeb, Luca Lonini, Arun Jayaraman, David C Mohr, and Konrad P Kording. "The need to approximate the use-case in clinical machine learning". In: *Gigascience* 6.5 (2017), gix019.

[264] Elias Chaibub Neto, Abhishek Pratap, Thanneer M. Perumal, et al. "Detecting the impact of subject characteristics on machine learning-based diagnostic applications". In: *npj Digital Medicine* 2.1 (2019), p. 99.

[265] Elias Chaibub Neto, Abhishek Pratap, Thanneer M Perumal, et al. "Learning Disease vs Participant Signatures: a permutation test approach to detect identity confounding in machine learning diagnostic applications". In: *arXiv preprint arXiv:1712.03120* (2017).

[266] Haibo He and Edwardo A Garcia. "Learning from imbalanced data". In: *IEEE Transactions on knowledge and data engineering* 21.9 (2009), pp. 1263–1284.

[267] Haibo He and Yunqian Ma. *Imbalanced learning: foundations, algorithms, and applications*. John Wiley & Sons, 2013.

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

[269] Robert C Holte, Liane Acker, and Bruce W Porter. "Concept Learning and the Problem of Small Disjuncts". In: *IJCAI*. Vol. 89. Citeseer, 1989, pp. 813–818.

[270] Christopher D Manning, Hinrich Schütze, and Prabhakar Raghavan. "Introduction to information retrieval". In: (2008).

[271] Jacob Cohen. "A coefficient of agreement for nominal scales". In: *Educational and psychological measurement* 20.1 (1960), pp. 37–46.

[272] Rob J Hyndman and Anne B Koehler. "Another look at measures of forecast accuracy". In: *International journal of forecasting* 22.4 (2006), pp. 679–688.

[273] P. Mulero, L. Midaglia, X. Montalban, J. Graves, S. Hauser, L. Julian, M. Baker, J. Schadrack, C. Gossens, F. McDougall, A. Scotland, Y. Zhang Schaerer, A. Creagh, S. Belachew, and M. Lindemann. "A Prospective Pilot Study (FLOODLIGHT) to Evaluate the Feasibility of Conducting Remote Patient Monitoring With the Use of Digital Technology in Patients With Multiple Sclerosis". In: *EAN Congress*. 2017.

[274] Kirsten Van Kessel, Duncan R Babbage, Nicholas Reay, Warren M Miner-Williams, and Paula Kersten. "Mobile technology use by people experiencing multiple sclerosis fatigue: survey methodology". In: *JMIR mHealth and uHealth* 5.2 (2017), e6.

[275] Sashank Prasad, Ravi Ramachandran, and Charles Jennings. *Development of Smartphone Technology To Monitor Disease Progression in Multiple Sclerosis (P01. 144)*. 2012.

[276] Jeremy Hobart, Donna Lamping, Ray Fitzpatrick, Afsane Riazi, and Alan Thompson. "The multiple sclerosis impact scale (MSIS-29) a new patient-based outcome measure". In: *Brain* 124.5 (2001), pp. 962–973.

[277] C McGuigan and M Hutchinson. "The multiple sclerosis impact scale (MSIS-29) is a reliable and sensitive measure". In: *Journal of Neurology, Neurosurgery & Psychiatry* 75.2 (2004), pp. 266–269.

[278] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge Luis Reyes-Ortiz. "A public domain dataset for human activity recognition using smartphones". In: *Esann*. 2013.

[279] G. M. Weiss, K. Yoneda, and T. Hayajneh. "Smartphone and Smartwatch-Based Biometrics Using Activities of Daily Living". In: *IEEE Access* 7 (2019), pp. 133190–133202.

[280] Gary M Weiss, Jessica L Timko, Catherine M Gallagher, Kenichi Yoneda, and Andrew J Schreiber. "Smartwatch-based activity recognition: A machine learning approach". In: *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2016, pp. 426–429.

[281] I. Lamers and P. Feys. "Assessing upper limb function in multiple sclerosis". In: *Mult Scler* 20.7 (2014), pp. 775–84.

[282] Lutz-Peter Erasmus, Stefania Sarno, Holger Albrecht, Martina Schwecht, Walter Pöllmann, and Nicolaus König. "Measurement of ataxic symptoms with a graphic tablet: standard values in controls and validity in Multiple Sclerosis patients". In: *Journal of Neuroscience Methods* 108.1 (2001), pp. 25–37.

[283] M. Memedi, A. Sadikov, V. Groznik, J. Zabkar, M. Mozina, F. Bergquist, A. Johansson, D. Haubenberger, and D. Nyholm. "Automatic Spiral Analysis for Objective Assessment of Motor Symptoms in Parkinson's Disease". In: *Sensors (Basel)* 15.9 (2015), pp. 23727–44.

[284] M-P Dubuisson and Anil K Jain. "A modified Hausdorff distance for object matching". In: *Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on*. Vol. 1. IEEE, 1994, pp. 566–568.

[285] Daniel P. Huttenlocher, Gregory A. Klanderman, and William J Rucklidge. "Comparing images using the Hausdorff distance". In: *IEEE Transactions on pattern analysis and machine intelligence* 15.9 (1993), pp. 850–863.

[286] R. C. Veltkamp. "Shape matching: similarity measures and algorithms". In: *Proceedings International Conference on Shape Modeling and Applications*. 2001, pp. 188–197.

[287] Josien PW Pluim, JB Antoine Maintz, and Max A Viergever. "Mutual-information-based registration of medical images: a survey". In: *IEEE transactions on medical imaging* 22.8 (2003), pp. 986–1004.

[288] Du-Yih Tsai, Yongbum Lee, and Eri Matsuyama. "Information Entropy Measure for Evaluation of Image Quality". In: *Journal of Digital Imaging* 21.3 (2008), pp. 338–347.

[289] Virginia Clark, OJ Dunn, and RM Mickey. *Applied statistics, analysis of variance and regression*. Wiley, 1974.

[290] Alexandre Belloni and Victor Chernozhukov. "Least squares after model selection in high-dimensional sparse models". In: *Bernoulli* 19.2 (2013), pp. 521–547.

[291] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

[292] N. Yozbatiran, F. Baskurt, Z. Baskurt, S. Ozakbas, and E. Idiman. "Motor assessment of upper extremity function and its relation with fatigue, cognitive function and quality of life in multiple sclerosis patients". In: *J Neurol Sci* 246.1-2 (2006), pp. 117–22.

[293] Poonam Zham, Dinesh K. Kumar, Peter Dabnichki, Sridhar Poosapadi Arjunan, and Sanjay Raghav. "Distinguishing Different Stages of Parkinson's Disease Using Composite Index of Speed and Pen-Pressure of Sketching a Spiral". In: *Frontiers in Neurology* 8.435 (2017).

[294] Hickey Aodhán, Din Silvia Del, Rochester Lynn, and Godfrey Alan. "Detecting free-living steps and walking bouts: validating an algorithm for macro gait analysis". In: *Physiological Measurement* 38.1 (2017), N1.

[295] G. M. Lyons, K. M. Culhane, D. Hilton, P. A. Grace, and D. Lyons. "A description of an accelerometer-based mobility monitoring technique". In: *Medical Engineering and Physics* 27.6 (2005), pp. 497–504.

[296] J. H. Hollman, E. M. McDade, and R. C. Petersen. "Normative spatiotemporal gait parameters in older adults". In: *Gait & Posture* 34.1 (2011), pp. 111–8.

[297] Jr. Bassett D. R., B. E. Ainsworth, S. R. Leggett, C. A. Mathien, J. A. Main, D. C. Hunter, and G. E. Duncan. "Accuracy of five electronic pedometers for measuring distance walked". In: *Med Sci Sports Exerc* 28.8 (1996), pp. 1071–7.

[298] Nicole A. Capela, Edward D. Lemaire, and Natalie Baddour. "Novel algorithm for a smartphone-based 6-minute walk test application: algorithm, application development, and evaluation". In: *Journal of NeuroEngineering and Rehabilitation* 12.1 (2015), p. 19.

[299] Hwan-hee Lee, Suji Choi, and Myeong-jin Lee. "Step Detection Robust against the Dynamics of Smartphones". In: *Sensors* 15.10 (2015), p. 27230.

[300] A. Godfrey, S. Del Din, G. Barry, J. C. Mathers, and L. Rochester. "Instrumenting gait with an accelerometer: A system and algorithm examination". In: *Medical Engineering and Physics* 37.4 (2015), pp. 400–407.

[301] P. S. Addison, J. Walker, and R. C. Guido. "Time–frequency analysis of biosignals". In: *IEEE Engineering in Medicine and Biology Magazine* 28.5 (2009), pp. 14–29.

[302] Christopher Torrence and Gilbert P Compo. "A practical guide to wavelet analysis". In: *Bulletin of the American Meteorological society* 79.1 (1998), pp. 61–78.

[303] Siddhartha Khandelwal and Nicholas Wickström. "Novel methodology for estimating Initial Contact events from accelerometers positioned at different body locations". In: *Gait & posture* 59 (2018), pp. 278–285.

[304] Patrice Abry. *Ondelettes et turbulences: multirésolutions, algorithmes de décomposition, invariance d'échelle et signaux de pression*. Diderot éd., 1997.

[305] Stephane Mallat. *A wavelet tour of signal processing: the sparse way*. Academic press, 2008.

[306] P. Ren, S. Tang, F. Fang, L. Luo, L. Xu, M. L. Bringas-Vega, D. Yao, K. M. Kendrick, and P. A. Valdes-Sosa. "Gait Rhythm Fluctuation Analysis for Neurodegenerative Diseases by Empirical Mode Decomposition". In: *IEEE Transactions on Biomedical Engineering* 64.1 (2017), pp. 52–60.

[307] Norden E Huang, Zheng Shen, Steven R Long, Manli C Wu, Hsing H Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung, and Henry H Liu. "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis". In: *Proceedings of the Royal Society of London A: mathematical, physical and engineering sciences*. Vol. 454. The Royal Society, 1998, pp. 903–995.

[308] M. Costa, C. K. Peng, Ary L. Goldberger, and Jeffrey M. Hausdorff. "Multiscale entropy analysis of human gait dynamics". In: *Physica A: Statistical Mechanics and its Applications* 330.1 (2003), pp. 53–60.

[309] Max A Little, Patrick E McSharry, Stephen J Roberts, Declan AE Costello, and Irene M Moroz. "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection". In: *Biomedical engineering online* 6.1 (2007), p. 23.

[310] John Platt. "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods". In: *Advances in large margin classifiers* 10.3 (1999), pp. 61–74.

[311] Catrine Tudor-Locke and David A. Rowe. "Using Cadence to Study Free-Living Ambulatory Behaviour". In: *Sports Medicine* 42.5 (2012), pp. 381–398.

[312] R. Senden, B. Grimm, I. C. Heyligers, H. H. C. M. Savelberg, and K. Meijer. "Acceleration-based gait test for healthy subjects: reliability and reference data". In: *Gait Posture* 30 (2009).

[313] Silvia Del Din, Alan Godfrey, and Lynn Rochester. "Validation of an accelerometer to quantify a comprehensive battery of gait characteristics in healthy older adults and Parkinson's disease: toward clinical and at home use". In: *IEEE journal of biomedical and health informatics* 20.3 (2016), pp. 838–847.

[314] Alper Köse, Andrea Cereatti, and Ugo Della Croce. "Bilateral step length estimation using a single inertial measurement unit attached to the pelvis". In: *Journal of neuroengineering and rehabilitation* 9.1 (2012), p. 9.

[315] S. Del Din, A. Godfrey, C. Mazza, S. Lord, and L. Rochester. "Free-living monitoring of Parkinson's disease: Lessons from the field". In: *Mov Disord* 31.9 (2016), pp. 1293–313.

[316] Minh H. Pham, Morad Elshehabi, Linda Haertner, Silvia Del Din, Karin Srulijes, Tanja Heger, Matthis Synofzik, et al. "Validation of a Step Detection Algorithm during Straight Walking and Turning in Patients with Parkinson's Disease and Older Adults Using an Inertial Measurement Unit at the Lower Back". In: *Frontiers in Neurology* 8.457 (2017).

[317] R. C. González, D. Alvarez, A. M. López, and J. C. Alvarez. "Ambulatory estimation of mean step length during unconstrained walking by means of COG accelerometry". In: *Computer Methods in Biomechanics and Biomedical Engineering* 12.6 (2009), pp. 721–726.

[318] Qu Wang, Langlang Ye, Haiyong Luo, Aidong Men, Fang Zhao, and Changhai Ou. "Pedestrian Walking Distance Estimation Based on Smartphone Mode Recognition". In: *Remote Sensing* 11.9 (2019), p. 1140.

[319] Ahlem Hajjem, François Bellavance, and Denis Larocque. "Mixed effects regression trees for clustered data". In: *Statistics & probability letters* 81.4 (2011), pp. 451–459.

[320] Max A. Little, Gael Varoquaux, Sohrab Saeb, Luca Lonini, Arun Jayaraman, David C. Mohr, and Konrad P. Kording. "Using and understanding cross-validation strategies. Perspectives on Saeb et al". In: *GigaScience* 6.5 (2017), pp. 1–6.

[321] Francisco Javier Ordóñez and Daniel Roggen. "Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition". In: *Sensors (Basel, Switzerland)* 16.1 (2016), p. 115.

[322] Daniele Ravì, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, and Guang-Zhong Yang. "Deep learning for health informatics". In: *IEEE journal of biomedical and health informatics* 21.1 (2016), pp. 4–21.

[323] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. "A survey on deep learning in medical image analysis". In: *Medical image analysis* 42 (2017), pp. 60–88.

[324] Mohammad Abu Alsheikh, Ahmed Selim, Dusit Niyato, Linda Doyle, Shaowei Lin, and Hwee-Pink Tan. "Deep activity recognition models with triaxial accelerometers". In: *arXiv preprint arXiv:1511.04664* (2015).

[325] Jianbo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiaoli Li, and Shonali Krishnaswamy. "Deep convolutional neural networks on multichannel time series for human activity recognition". In: *Ijcai*. Vol. 15. Citeseer, 2015, pp. 3995–4001.

[326] Kaixuan Chen, Dalin Zhang, Lina Yao, Bin Guo, Zhiwen Yu, and Yunhao Liu. "Deep learning for sensor-based human activity recognition: overview, challenges and opportunities". In: *arXiv preprint arXiv:2001.07416* (2020).

[327] Francisco Javier Ordóñez Morales and Daniel Roggen. *Deep convolutional feature transfer across mobile activity recognition domains, sensor modalities and locations*. Proceedings of the 2016 ACM International Symposium on Wearable Computers. Heidelberg, Germany: Association for Computing Machinery, 2016, pp. 92–99.

[328] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. "Deep learning for sensor-based activity recognition: A survey". In: *Pattern Recognition Letters* 119 (2019), pp. 3–11.

[329] Terry Taewoong Um, Franz Michael Josef Pfister, Daniel Pichler, Satoshi Endo, Muriel Lang, Sandra Hirche, Urban Fietzek, and Dana Kulić. "Data augmentation of wearable sensor data for Parkinson's disease monitoring using convolutional neural networks". In: *arXiv preprint arXiv:1706.00527* (2017).

[330] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. "How transferable are features in deep neural networks?" In: *Advances in neural information processing systems*. 2014, pp. 3320–3328.

[331] Sinno Jialin Pan and Qiang Yang. "A survey on transfer learning". In: *IEEE Transactions on knowledge and data engineering* 22.10 (2009), pp. 1345–1359.

[332] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. "Learning and transferring mid-level image representations using convolutional neural networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 1717–1724.

[333] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. "Decaf: A deep convolutional activation feature for generic visual recognition". In: *International conference on machine learning*. 2014, pp. 647–655.

[334] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. "Transfer learning for time series classification". In: *IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 1367–1376.

[335] M. T. Martinez and P. De Leon. "Falls Risk Classification of Older Adults Using Deep Neural Networks and Transfer Learning". In: *IEEE Journal of Biomedical and Health Informatics* (2019), pp. 1–1.

[336] Xin Qin, Yiqiang Chen, Jindong Wang, and Chaohui Yu. "Cross-dataset activity recognition via adaptive spatial-temporal transfer learning". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3.4 (2019), pp. 1–25.

[337] Fernando Andreotti, Huy Phan, Navin Cooray, Christine Lo, Michele TM Hu, and Maarten De Vos. "Multichannel sleep stage classification and transfer learning using convolutional neural networks". In: *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 171–174.

[338] Huy Phan, Oliver Y Chén, Philipp Koch, Zongqing Lu, Ian McLoughlin, Alfred Mertins, and Maarten De Vos. "Towards More Accurate Automatic Sleep Staging via Deep Transfer Learning". In: *arXiv preprint arXiv:1907.13177* (2019).

[339] Matthew D Zeiler and Rob Fergus. "Visualizing and understanding convolutional networks". In: *European conference on computer vision*. Springer, 2014, pp. 818–833.

[340] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1026–1034.

[341] Francois Chollet. *Deep Learning mit Python und Keras: Das Praxis-Handbuch vom Entwickler der Keras-Bibliothek*. MITP-Verlags GmbH & Co. KG, 2018.

[342] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, and Matthieu Devin. "Tensorflow: Large-scale machine learning on heterogeneous distributed systems". In: *arXiv preprint arXiv:1603.04467* (2016).

[343] Fazle Karim, Somshubra Majumdar, Houshang Darabi, and Shun Chen. "LSTM fully convolutional networks for time series classification". In: *IEEE Access* 6 (2018), pp. 1662–1669.

[344] Amina Adadi and Mohammed Berrada. "Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI)". In: *IEEE Access* 6 (2018), pp. 52138–52160.

[345] David S Watson, Jenny Krutzinna, Ian N Bruce, Christopher EM Griffiths, Iain B McInnes, Michael R Barnes, and Luciano Floridi. "Clinical applications of machine learning algorithms: beyond the black box". In: *Bmj* 364 (2019).

[346] Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. "Interpretable machine learning in healthcare". In: *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. 2018, pp. 559–560.

[347] Davide Castelvecchi. "Can we open the black box of AI?" In: *Nature News* 538.7623 (2016), p. 20.

[348] Moritz Böhle, Fabian Eitel, Martin Weygandt, and Kerstin Ritter. "Layer-Wise Relevance Propagation for Explaining Deep Neural Network Decisions in MRI-Based Alzheimer's Disease Classification". In: *Frontiers in Aging Neuroscience* 11.194 (2019).

[349] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. "A survey of methods for explaining black box models". In: *ACM computing surveys (CSUR)* 51.5 (2018), pp. 1–42.

[350] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. "Feature visualization". In: *Distill* 2.11 (2017), e7.

[351] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. "Visualizing higher-layer features of a deep network". In: *University of Montreal* 1341.3 (2009), p. 1.

[352] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473* (2014).

[353] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. "Effective approaches to attention-based neural machine translation". In: *arXiv preprint arXiv:1508.04025* (2015).

[354] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need". In: *Advances in neural information processing systems*. 2017, pp. 5998–6008.

[355] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. "Show, attend and tell: Neural image caption generation with visual attention". In: *International conference on machine learning*. 2015, pp. 2048–2057.

[356] Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y Chén, and Maarten De Vos. "SeqSleepNet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 27.3 (2019), pp. 400–410.

[357] Farah E Shamout, Tingting Zhu, Pulkit Sharma, Peter J Watkinson, and David A Clifton. "Deep interpretable early warning system for the detection of clinical deterioration". In: *IEEE journal of biomedical and health informatics* (2019).

[358] Patrick Schwab, Gaetano C Scebba, Jia Zhang, Marco Delai, and Walter Karlen. "Beat by beat: Classifying cardiac arrhythmias with recurrent neural networks". In: *Computing in Cardiology (CinC)*. IEEE, 2017, pp. 1–4.

[359] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps". In: *arXiv preprint arXiv:1312.6034* (2013).

[360] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. "Striving for simplicity: The all convolutional net". In: *arXiv preprint arXiv:1412.6806* (2014).

[361] Ruth C Fong and Andrea Vedaldi. "Interpretable explanations of black boxes by meaningful perturbation". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 3429–3437.

[362] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. "Learning deep features for discriminative localization". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2921–2929.

[363] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.

[364] Fernando Andreotti, Huy Phan, and Maarten De Vos. "Visualising convolutional neural network decisions in automatic sleep scoring". In: *CEUR Workshop Proceedings*. 2018, pp. 70–81.

[365] Sebastian Lapuschkin, Alexander Binder, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. "The LRP toolbox for artificial neural networks". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 3938–3942.

[366] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. "Evaluating the visualization of what a deep neural network has learned". In: *IEEE transactions on neural networks and learning systems* 28.11 (2016), pp. 2660–2673.

[367] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models". In: *arXiv preprint arXiv:1708.08296* (2017).

[368] Irene Sturm, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. "Interpretable deep neural networks for single-trial EEG classification". In: *Journal of Neuroscience Methods* 274 (2016), pp. 141–145.

[369] Maximilian Alber, Sebastian Lapuschkin, Philipp Seegerer, Miriam Hägele, Kristof T Schütt, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Sven Dähne, and Pieter-Jan Kindermans. "iNNvestigate neural networks". In: *Journal of Machine Learning Research* 20.93 (2019), pp. 1–8.

[370] François Petitjean, Alain Ketterlin, and Pierre Gançarski. "A global averaging method for dynamic time warping, with applications to clustering". In: *Pattern Recognition* 44.3 (2011), pp. 678–693.

[371] Donald J Berndt and James Clifford. "Using dynamic time warping to find patterns in time series." In: *KDD workshop*. Vol. 10. 16. Seattle, WA, USA: 1994, pp. 359–370.

[372] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. Vol. 793. John Wiley & Sons, 2019.

[373] A Rogier T Donders, Geert JMG Van Der Heijden, Theo Stijnen, and Karel GM Moons. "A gentle introduction to imputation of missing values". In: *Journal of clinical epidemiology* 59.10 (2006), pp. 1087–1091.

[374] Lei Clifton, David A Clifton, Marco AF Pimentel, Peter J Watkinson, and Lionel Tarassenko. "Gaussian processes for personalized e-health monitoring with wearable sensors". In: *IEEE Transactions on Biomedical Engineering* 60.1 (2012), pp. 193–197.

[375] Wiebren Zijlstra and At L. Hof. "Assessment of spatio-temporal gait parameters from trunk accelerations during human walking". In: *Gait & Posture* 18.2 (2003), pp. 1–10.

[376] Massimiliano Pau, Silvia Caggiari, Alessandro Mura, Federica Corona, Bruno Leban, Giancarlo Coghe, Lorena Lorefice, Maria Giovanna Marrosu, and Eleonora Cocco. "Clinical assessment of gait in individuals with multiple sclerosis using wearable inertial sensors: Comparison with patient-based measure". In: *Multiple Sclerosis and Related Disorders* 10 (2016), pp. 187–191.

[377] Johan van Beek, Rita Freitas, Corrado Bernasconi, Xavier Montalban, Butzkueven, et al. "FLOODLIGHT Open–A Global, Prospective, Open-Access Study to Better Understand Multiple Sclerosis Using Smartphone Technology". In: *Annual Meeting of the Consortium of Multiple Sclerosis Centers*. CMSC. 2019.

[378] Mikaela L Frechette, Brett M Meyer, Lindsey J Tulipani, Reed D Gurchiek, Ryan S McGinnis, and Jacob J Sosnoff. "Next steps in wearable technology and community ambulation in multiple sclerosis". In: *Current neurology and neuroscience reports* 19.10 (2019), p. 80.

[379] Silvia Del Din, Alan Godfrey, Brook Galna, Sue Lord, and Lynn Rochester. "Free-living gait characteristics in ageing and Parkinson's disease: impact of environment and ambulatory bout length". In: *Journal of NeuroEngineering and Rehabilitation* 13.1 (2016), p. 46.

[380] Michelle Norris, Ross Anderson, Robert W Motl, Sara Hayes, and Susan Coote. "Minimum number of days required for a reliable estimate of daily step count and energy expenditure, in people with MS who walk unaided". In: *Gait & Posture* 53 (2017), pp. 201–206.

[381] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[382] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. "Speech recognition with deep recurrent neural networks". In: *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 6645–6649.

[383] Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzel. "Learning to diagnose with LSTM recurrent neural networks". In: *arXiv preprint arXiv:1511.03677* (2015).

[384] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling". In: *arXiv preprint arXiv:1803.01271* (2018).

[385] Li Deng and Dong Yu. "Deep learning: methods and applications". In: *Foundations and trends in signal processing* 7.3–4 (2014), pp. 197–387.

[386] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, and Jianfei Cai. "Recent advances in convolutional neural networks". In: *Pattern Recognition* 77 (2018), pp. 354–377.

[387] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[388] Clayton R Pereira, Danilo R Pereira, Gustavo H Rosa, Victor HC Albuquerque, Silke AT Weber, Christian Hook, and João P Papa. "Handwritten dynamics assessment through convolutional neural networks: An application to Parkinson's disease identification". In: *Artificial intelligence in medicine* 87 (2018), pp. 67–77.

[389] Pedram Khatamino, İsmail Cantürk, and Lale Özyılmaz. "A Deep Learning-CNN Based System for Medical Diagnosis: An Application on Parkinson's Disease Handwriting Drawings". In: *6th International Conference on Control Engineering & Information Technology (CEIT)*. IEEE, 2018, pp. 1–6.

[390] Manuel Gil-Martín, Juan Manuel Montero, and Rubén San-Segundo. "Parkinson's disease detection from drawing movements using convolutional neural networks". In: *Electronics* 8.8 (2019), p. 907.

[391] Amina Naseer, Monail Rani, Saeeda Naz, Muhammad Imran Razzak, Muhammad Imran, and Guandong Xu. "Refining Parkinson's neurological disorder identification through deep transfer learning". In: *Neural Computing and Applications* 32.3 (2020), pp. 839–854.

[392] Junyuan Xie, Ross Girshick, and Ali Farhadi. "Unsupervised deep embedding for clustering analysis". In: *International conference on machine learning*. 2016, pp. 478–487.

[393] Xifeng Guo, Xinwang Liu, En Zhu, and Jianping Yin. "Deep clustering with convolutional autoencoders". In: *International conference on neural information processing*. Springer, 2017, pp. 373–382.

[394] Trevor J Hastie and Robert J Tibshirani. *Generalized additive models*. Vol. 43. CRC press, 1990.

[395] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

[396] Frederick N Fritsch and Ralph E Carlson. "Monotone piecewise cubic interpolation". In: *SIAM Journal on Numerical Analysis* 17.2 (1980), pp. 238–246.

[397] George EP Box and David R Cox. "An analysis of transformations". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 26.2 (1964), pp. 211–243.

[398] RM Sakia. "The Box-Cox transformation technique: a review". In: *The statistician* (1992), pp. 169–178.

[399] Kai Kunze, Paul Lukowicz, Kurt Partridge, and Bo Begole. "Which way am I facing: Inferring horizontal device orientation from an accelerometer signal". In: *2009 International Symposium on Wearable Computers*. IEEE, pp. 149–150.

[400] Zhi-An Deng, Guofeng Wang, Ying Hu, and Di Wu. "Heading estimation for indoor pedestrian navigation using a smartphone in the pocket". In: *Sensors* 15.9 (2015), pp. 21518–21536.

[401] Ulf Blanke and Bernt Schiele. "Sensing location in the pocket". In: *Ubicomp Poster Session* (2008), p. 2.

[402] Saba Rezvanian and Thurmon E Lockhart. "Towards real-time detection of freezing of gait using wavelet transform on wireless accelerometer data". In: *Sensors* 16.4 (2016), p. 475.

[403] Gabriel Rilling, Patrick Flandrin, and Paulo Goncalves. "On empirical mode decomposition and its algorithms". In: *IEEE-EURASIP workshop on nonlinear signal and image processing*. Vol. 3. NSIP-03, Grado (I), 2003, pp. 8–11.

[404] Gang Wang, Xian-Yao Chen, Fang-Li Qiao, Zhaohua Wu, and Norden E Huang. "On intrinsic mode function". In: *Advances in Adaptive Data Analysis* 2.03 (2010), pp. 277–293.

[405] DN Joanes and CA Gill. "Comparing measures of sample skewness and kurtosis". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 47.1 (1998), pp. 183–189.

[406] Robert M Gray. *Entropy and information theory*. Springer Science & Business Media, 2011.

[407] Steven M Pincus. "Approximate entropy as a measure of system complexity". In: *Proceedings of the National Academy of Sciences* 88.6 (1991), pp. 2297–2301.

[408] Norbert Marwan, M Carmen Romano, Marco Thiel, and Jürgen Kurths. "Recurrence plots for the analysis of complex systems". In: *Physics reports* 438.5-6 (2007), pp. 237–329.

[409] Lawrence R Rabiner. "Digital processing of speech signal". In: *Digital Processing of Speech Signal* (1978).

[410] Jan W Kantelhardt, Stephan A Zschiegner, Eva Koscielny-Bunde, Shlomo Havlin, Armin Bunde, and H Eugene Stanley. "Multifractal detrended fluctuation analysis of nonstationary time series". In: *Physica A: Statistical Mechanics and its Applications* 316.1-4 (2002), pp. 87–114.

[411] J. M. Hausdorff, Y. Ashkenazy, C. K. Peng, P. C. Ivanov, H. E. Stanley, and A. L. Goldberger. "When human walking becomes random walking: fractal analysis and modeling of gait rhythm fluctuations". In: *Physica A* 302.1-4 (2001), pp. 138–47.

[412] Sotirios Damouras, Matthew D Chang, Ervin Sejdić, and Tom Chau. "An empirical examination of detrended fluctuation analysis for gait data". In: *Gait & posture* 31.3 (2010), pp. 336–340.

[413] Srimonti Dutta, Dipak Ghosh, and Sucharita Chatterjee. "Multifractal detrended fluctuation analysis of human gait diseases". In: *Frontiers in physiology* 4 (2013), p. 274.

[414] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[415] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[416] James F Kaiser. "On a simple algorithm to calculate the'energy'of a signal". In: *International conference on acoustics, speech, and signal processing*. IEEE, 1990, pp. 381–384.

[417] Madalena Costa, Ary L Goldberger, and C-K Peng. "Multiscale entropy analysis of complex physiologic time series". In: *Physical review letters* 89.6 (2002), p. 068102.

[418] Jian Yu, Junyi Cao, Wei-Hsin Liao, Yangquan Chen, Jing Lin, and Rong Liu. "Multivariate Multiscale Symbolic Entropy Analysis of Human Gait Signals". In: *Entropy* 19.10 (2017), p. 557.

[419] Joshua S Richman and J Randall Moorman. "Physiological time-series analysis using approximate entropy and sample entropy". In: *American Journal of Physiology-Heart and Circulatory Physiology* 278.6 (2000), H2039–H2049.

[420] S. Wang, P. G. Bain, T. Z. Aziz, and X. Liu. "The direction of oscillation in spiral drawings can be used to differentiate distal and proximal arm tremor". In: *Neurosci Lett* 384.1-2 (2005), pp. 188–92.

[421] Frederik Maes, Andre Collignon, Dirk Vandermeulen, Guy Marchal, and Paul Suetens. "Multimodality image registration by maximization of mutual information". In: *IEEE transactions on Medical Imaging* 16.2 (1997), pp. 187–198.

[422] J-C Yoo and CW Ahn. "Image matching using peak signal-to-noise ratio-based occlusion detection". In: *IET image processing* 6.5 (2012), pp. 483–495.

[423] Manesh Kokare, BN Chatterji, and PK Biswas. "Comparison of similarity metrics for texture image retrieval". In: *TENCON 2003. Conference on convergent technologies for Asia-Pacific region.* Vol. 2. IEEE, pp. 571–575.

[424] Wei Yang, Luhui Xu, Xiaopan Chen, Fengbin Zheng, and Yang Liu. "Chi-squared distance metric learning for histogram data". In: *Mathematical Problems in Engineering* 2015 (2015).

[425] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. "Image quality assessment: from error visibility to structural similarity". In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.