Abstract

This thesis presents our research on using machine learning for digital phenotyping. We attempt to address some of the associated data science challenges in using digital phenotypes for building a vision for Predictive, Personalised, Participatory and Preventative (P4) Medicine. Some of the most significant data science challenges associated with P4 medicine lie in identifying and engineering feature representations for raw digital data and then identify and learning predictive models for disease-related outcomes. This is a significant challenge due to the variety and volume of Big Health Data generated by digital technologies but also because of the diversity in disease-related phenotypes that can be associated.

Our research has used accelerometry data generated from wearable digital activity trackers applied to Type-2 Diabetes as a case study. Type-2 Diabetes is a chronic disease that is becoming more prevalent in the UK and other countries known to be closely associated with low levels of physical activity, poor sleeping patterns and a sedentary lifestyle. Accelerometry data are an objective and reliable measure of physical activity.

The key hypothesis in our research is that physical activity traces collected by digital accelerometers can be used to build machine learning models that can predict Type-2 Diabetes related outcomes and used to characterise a person’s digital phenotype. The goal of this case study is to then demonstrate and address some of the data science challenges of predictive modelling using ubiquitous digital devices. Two accelerometer datasets were used for our experimentation test beds: The UK Biobank and IMI DIRECT.

We first begin by using features extracted with a state-of-the-art tool for accelerometer data analysis, GGIR, to learn predictive models for Type-2 diabetes related target clinical outcomes. This gave us a baseline on which to improve upon in our research. The first stage of our research involved using methods from neuroscience and gait analysis are adopted in another approach to build a new representation using the features generated from Human Activity Recognition (HAR) to learn predictive models for Type-2 Diabetes outcomes. Human Activity Recognition is one of the main areas where machine learning models are applied to accelerometry data. Furthermore, we devise a strategy for refining the training data which helped to enhance predictive performance and address some of the associated data science challenges.

In the next stage of our research, advanced deep learning techniques were used to construct a latent representation using an unsupervised autoencoder approach thereby removing the need for manual feature engineering. This latent representation was then also used to learn predictive models for Type-2 diabetes related clinical target outcomes. This, however, produced a poor level of performance. We then sought to validate the representations for digital accelerometry data we developed in our research.
We used the DIRECT dataset’s longitudinal studies over time to then evaluate whether physical activity changes are demonstrated in clinical disease progression over time. The goal of this set of analyses was to determine whether these two phenotypes are associated with one another, since we know type-2 diabetes is closely associated with physical activity levels.

The final set of experiments in our research focused on unsupervised machine learning approaches to explore and validate our representations for digital accelerometry data. This is to demonstrate and validate how the representations for digital accelerometer data we developed in our research can be used to characterise physical activity patterns by clustering them into groups that exhibit similar behaviours. The goal of this work was to demonstrate how our representations produced meaningful characterisations of physical activity patterns. This would demonstrate a challenge in the data-driven P4 vision for medicine by illustrating how important it is to choose the optimal set of clinical target outcomes for learning predictive models for disease.

Overall, our research developed and evaluated various representations of digital accelerometer data. Although the prediction results have been weak for our case study of Type-2 Diabetes, we were able to validate our representations of digital accelerometer data through unsupervised machine learning approaches. We have been successful in illustrating the significance of the data science challenges that still need to be addressed before a truly data-driven, predictive and personalised vision for the future of healthcare can be fully realised.
Acknowledgements

The first person I must thank is my PhD supervisor, Professor Paolo Missier, for giving me the opportunity to pursue a PhD under his guidance and for always being on my side. Paolo has been a source of inspiration from whom I have learned so much. His counsel, extensive knowledge, patience and advice have been invaluable to my experience and development. Words cannot express how grateful I am.

I have also been fortunate enough to work closely with Ossama Alshabrawy and Jaume Bacardit during my time in the School of Computing. Their deep knowledge has been critical in support of my research. Outside the School of Computing, I also received the valuable support and help from the clinical experts including Professor Michael Catt and Professor Michael Trenell in the Faculty of Medical Sciences at Newcastle University, and Sophie Cassidy in the Faculty of Medicine at the University of Sydney. I also worked extensively with partners outside of Newcastle University, including Robert Koivula at the University of Oxford, and other researchers in the IMI DIRECT consortium. I want to thank them too for all their help and expertise, and for giving me the access to critical resources that enabled my research.

I would also like to thank Professor Paul Watson, Jennifer Wood and Andrew Turnbull for giving me the opportunity to be part of the wonderful community that is the Centre for Doctoral Training in Cloud Computing for Big Data, and all for all their support, both financially and morally, in my time with the CDT throughout the years. I also want to thank all my colleagues and fellow PhD students in the CDT for their friendship through the last four years. I also thank the Engineering and Physical Sciences Research Council (UK) for funding my PhD. I also want to thank Professor John Fitzgerald who is one of my earliest mentors since my undergraduate studies and who inspired me to go into further postgraduate research.

Most importantly, I want to express my biggest thanks to my family for all their unconditional love and support: my father, Dr Kam-Kei Lam; my mother, Ai-Ping Lee; my brother, Jonathan; and my grandmother, Yau-Hing Tsang. And, finally, although sadly no longer with us, I want to extend a special thanks to my grandfather Chan-Wing Lam and my grandfather Shu-Cheng Lee for their unwavering love and devotion to our family. We love you and miss you everyday. May you both rest in peace.
I dedicate this thesis to my family.
## Contents

### List of Figures

ix

### List of Tables

xv

### 1 Introduction

1. Introduction ................................................................. 1
2. P4 Medicine ................................................................. 2
   2.1 Challenges of an Ageing Population ................................ 3
   2.2 Big Data and P4 Medicine ........................................... 3
3. Thesis Key Contributions ............................................... 4
4. Research Aims and Outline ............................................ 5
5. Publication ................................................................. 7

### 2 Background Literature and Related Works

1. Introduction ............................................................. 9
2. P4 Medicine ............................................................. 10
3. Digital Wearable Technology in Clinical Science .................. 11
   3.1 Potential for Digital Biomarkers .................................. 11
   3.2 Digital Wearable Technology in Healthcare Applications .... 12
   3.3 Digital Wearables Technology: Challenges ..................... 13
4. Foundations of Machine Learning ................................... 15
   4.1 Machine Learning Basics ......................................... 15
   4.2 Supervised vs Unsupervised Machine Learning ................ 16
   4.3 Supervised Machine Learning: Classification .................. 16
   4.4 Supervised Machine Learning: Regression ...................... 19
   4.5 Unsupervised Machine Learning: Clustering .................... 20
   4.6 Performance Evaluation in Machine Learning ................. 23
5. Foundations of Deep Learning ....................................... 27
   5.1 The Multilayer Perceptron ........................................ 27
   5.2 Training and Optimising Neural Networks ...................... 28
   5.3 Recurrent Neural Networks and Long Short Term Memory .... 33
   5.4 Autoencoders ....................................................... 36
6. Human Activity Recognition .......................................... 38
   6.1 Overview of Human Activity Recognition ....................... 38
   6.2 Preprocessing Accelerometry Data for HAR tasks ............ 39
   6.3 Feature Engineering for HAR tasks ............................. 40
   6.4 Classification Methods for HAR tasks .......................... 40
   6.5 Oxford Accelerometer Analysis Tool ............................ 42
2.2 Signal-level epoch features ................................. 98
2.3 Long Short-Term Memory Autoencoder Architecture ... 100
2.4 Supervised learning for regression on clinical target outcomes from latent representation ............... 103
2.5 Comparing with available Physical Activity features from DIRECT dataset .......................... 107

3 Results and Discussions ........................................ 107
3.1 Results from predictive modelling ....................... 107
3.2 Principal Findings ......................................... 111
3.3 Implications of Study ...................................... 111
3.4 Limitations ................................................... 112
3.5 Conclusions ................................................. 113

6 Analysis of longitudinal physical activity and clinical progression over time ............................. 115
1 Introduction ...................................................... 116
2 Motivation and Background ................................. 116
3 Materials and Methods ....................................... 117
3.1 DIRECT dataset follow-ups ............................... 117
3.2 Calculating the similarity of physical activity over different time points .................................. 117
3.3 Phenotypes for clinical disease progression .............. 119
3.4 Determining the relationship between physical activity change and change in clinical variables .... 120
3.5 Analysis of distribution of physical activity features over time ............................................. 121
3.6 Analysis from most extreme individual clinical and physical activity changes ...................... 122

4 Results and Discussion .......................................... 125
4.1 Correlation analysis results ............................... 125
4.2 Analysing the distribution of physical activities over time ................................................. 126
4.3 Analysing the distributions for the individuals with the greatest disease progression ............... 130
4.4 Principal Findings ......................................... 134
4.5 Implications of Study ...................................... 134
4.6 Limitations ................................................... 135
4.7 Conclusions ................................................. 136

7 Unsupervised clustering from representations of accelerometer data ..................................... 137
1 Introduction ...................................................... 138
2 Methods and Materials ....................................... 138
2.1 Datasets ...................................................... 138
2.2 k-means clustering ......................................... 139
2.3 Hierarchical agglomerative clustering ................. 141
2.4 DBSCAN .................................................... 142
2.5 Affinity Propagation ....................................... 142
2.6 Spectral clustering ......................................... 143
2.7 Clustering validity indices: Calinski-Harabasz, Silhouette, Davies-Bouldin ......................... 143
List of Figures

2.1 Illustration of the random forest algorithm ........................................... 17
2.2 Diagram illustrating the maximum-margin hyperplane for the SVM algorithm. The support vectors are also labelled. This SVM has two classes discriminated by the hyperplane ........................................... 18
2.3 Structure for the neuron unit of a neural network ................................... 27
2.4 Multilayer perceptron with two-dimensional input, two hidden layers (with 3 units, 4 units respectively) and output layer with one unit ................................................................. 28
2.5 A recurrent neuron (LHS) unrolled through time (RHS) ......................... 33
2.6 Illustration of backpropagation through time. Dashed arrow represents forward pass. Black full arrow represents the propagation backwards of cost function gradients. $C$ is the cost function being applied to the last three outputs ......................................................... 34
2.7 LSTM cell [76, 81] ................................................................................ 35
2.8 Autoencoder structure ............................................................................ 37
2.9 Example of Oxford human activity classification output plotted along daily timeline ................................................................. 43
2.10 Summary of major components of the data collected in the UK Biobank [25] ................................................................. 45
3.1 Histogram for GGIR feature $acc_n_noc_AD_T5A5_sd$ ............................. 59
3.2 Histogram for GGIR feature $acc_eff_AD_T5A5_sd$ ................................. 59
3.3 Histogram for GGIR feature $acc_n_noc_AD_T5A5_mn$ ......................... 60
3.4 Histogram for GGIR feature $acc_n_sibd_AD_T5A5_sd$ ......................... 60
3.5 Histogram for GGIR feature $acc_n_sibd_WD_T5A5_sd$ ....................... 61
3.6 Histogram for GGIR feature $acc_timeinbed_AD_T5A5_mn$ ................. 61
3.7 Histogram for GGIR feature $acc_wake_AD_T5A5_mn$ ........................... 62
4.1 Training set selection criteria for Type-2 Diabetes negative and Type-2 Diabetes positive individuals. (EHR = Electronic Health Records), (QC = Quality Control) ......................................................... 71
4.2 Construction of feature spaces for representing UK Biobank accelerometer data. Each level builds on the levels below ............................ 73
4.3 Feature matrix for the new high-level representation of the UK Biobank’s accelerometer cross-section in this study ............................... 74
4.4 Elbow curve for $k$ ............................................................................... 78
4.5 Dendrogram visualisation for hierarchical clustering with selected UK Biobank accelerometer cohort ......................................................... 78
4.6 Histogram for daily average percentage times spent asleep ................... 79
4.7 Histogram for daily average length of sleep bouts in 30-second epochs .................. 80
4.8 Histogram for daily average number of sleep bouts ........................................... 80
4.9 ROC and AUROC score for type 2 diabetes vs Norm-0: High-level activity bout features, sociodemographic and lifestyle features combined ................................................. 82
4.10 ROC and AUROC score for Type-2 Diabetes positive vs Norm-0: High-level activity bout features only ............................................................... 83
4.11 ROC and AUROC score for Type-2 Diabetes positive vs Norm-0: sociodemographic and lifestyle features only .............................................................. 83
4.12 ROC and AUROC score for Type-2 Diabetes positive vs Norm-2: High-level activity bout features, sociodemographic and lifestyle features combined ................................................. 83
4.13 ROC and AUROC score for Type-2 Diabetes positive vs Norm-2: High-level activity bout features only ............................................................... 84
4.14 ROC and AUROC score for Type-2 Diabetes positive vs Norm-2: sociodemographic and lifestyle features only .............................................................. 84
4.15 Visualisation with T-SNE for k-means clustering. 4 overlapping clusters are seen ......................... 86
4.16 Visualisation with T-SNE for hierarchical agglomerative clustering. 3 clear clusters are seen ......................... 86
4.17 Average daily percentage time spent in each activity type for Cluster 1 .................... 90
4.18 Average daily percentage time spent in each activity type, for Cluster 2 .................... 90
4.19 Average daily percentage time spent in each activity type, for Cluster 3 .................... 91
4.20 Daily average number of sleep bouts for all clusters ........................................... 91
4.21 Daily average length of sleep bouts for all clusters ........................................... 91
4.22 Daily average percentage time spent in sleep for all clusters .............................. 92

5.1 Pie chart showing the mean average percentage distribution in physical activities for the pre-diabetes Cohort 2.1 .................................................. 97
5.2 Pie chart showing the mean average percentage distribution in physical activities for the Type-2 Diabetes positive Cohort 2.2 .................................................. 98
5.3 LSTM Autoencoder architecture diagram .................................................. 101

6.1 Histogram showing distribution of moderate-to-vigorous activity levels for subjects with confirmed Type-2 Diabetes at baseline assessment ................................... 126
6.2 Histogram showing distribution of moderate-to-vigorous activity levels for subjects with confirmed Type-2 Diabetes 36 months after baseline assessment ................................... 126
6.3 Histogram showing distribution of percentage of time spent performing light tasks for subjects with confirmed Type-2 Diabetes at baseline assessment ................................... 127
6.4 Histogram showing distribution of percentage of time spent performing light tasks for subjects with confirmed Type-2 Diabetes 36 months after baseline assessment ................................... 127
6.5 Histogram showing distribution of percentage of time spent performing sedentary activity for subjects with confirmed Type-2 Diabetes at baseline assessment. ........................................ 127
6.6 Histogram showing distribution of percentage of time spent performing sedentary activity for subjects with confirmed Type-2 Diabetes 36 months after baseline assessment. .................. 127
6.7 Histogram showing distribution of number of hours asleep for subjects with confirmed Type-2 Diabetes at baseline assessment. .............................. 127
6.8 Histogram showing distribution of number of hours asleep for subjects with confirmed Type-2 Diabetes 36 months after baseline assessment. .................. 127
6.9 Histogram showing distribution of MET levels for subjects with confirmed Type-2 Diabetes at baseline assessment. .............................. 128
6.10 Histogram showing distribution of MET levels for subjects with confirmed Type-2 Diabetes 36 months after baseline assessment. .................. 128
6.11 Histogram showing distribution of percentage of time spent performing light tasks for subjects in pre-diabetes stage at baseline assessment. .................. 128
6.12 Histogram showing distribution of percentage of time spent performing light tasks for subjects in a pre-diabetes stage 48 months after baseline assessment. .................. 128
6.13 Histogram showing distribution of percentage of time spent performing moderate-to-vigorous activity for subjects in pre-diabetes stage at baseline assessment. .................. 128
6.14 Histogram showing distribution of percentage of time spent performing moderate-to-vigorous activity for subjects in a pre-diabetes stage 48 months after baseline assessment. .................. 128
6.15 Histogram showing distribution of number of hours asleep for subjects in pre-diabetes stage at baseline assessment. .............................. 129
6.16 Histogram showing distribution of number of hours asleep for subjects in a pre-diabetes stage 48 months after baseline assessment. .................. 129
6.17 Histogram showing distribution of percentage of time spent performing sedentary activity for subjects in pre-diabetes stage at baseline assessment. .................. 129
6.18 Histogram showing distribution of percentage of time spent performing sedentary activity for subjects in a pre-diabetes stage 48 months after baseline assessment. .................. 129
6.19 Histogram showing distribution of MET levels for subjects in pre-diabetes stage at baseline assessment. .............................. 129
6.20 Histogram showing distribution of MET levels for subjects in a pre-diabetes stage 48 months after baseline assessment. .................. 129
6.21 Spider plot for daily behaviour patterns for the individual in Cohort 2.2 (Type-2 Diabetes positive diagnosis cohort) with the highest change in HbA1c clinical progression at baseline assessment. .................. 131
6.22 Spider plot for daily behaviour patterns for the individual in Cohort 2.2 with the highest change in HbA1c clinical progression at 36 months after baseline. .................. 131

xi
6.23 Spider plot for summarised daily behaviour patterns for the individual in Cohort 2.2 close to the average progression in HbA1c clinical progression at baseline ........................................... 132
6.24 Spider plot for summarised daily behaviour patterns for the individual in Cohort 2.2 close to the average progression in HbA1c clinical progression at 36 months after baseline .......................... 133

7.1 Experimental design plan with the DIRECT accelerometer dataset. This chapter focuses on the unsupervised clustering of the latent representation and the high-level activity bout features. ......................... 139
7.2 Elbow curve for the autoencoder latent representation. It can be seen that when $k = 2$, there is point of inflection in the curve. This would suggest that this is the optimal value for $k$ when using the silhouette score as its metric .......................... 140
7.3 Elbow curve for the high-level granular bout-based representation of the DIRECT accelerometer data. It can be seen that when $k = 4$, there is point of inflection in the curve. This would suggest that this is the optimal value for $k$ when using the silhouette score as its metric. However the value for the silhouette score is very low, this would suggest that the clusters that form, even with the optimised value set for $k$, would result in poorly defined clusters .......................... 140
7.4 Dendrogram plot of the hierarchical structure for the high-level granular bout-based representation. It is clear that four clusters are identified at the cutoff distance threshold of 3. The Ward linkage function was used ........................................... 141
7.5 Dendrogram plot of the hierarchical structure for the high-level granular bout-based representation. It is clear that four clusters are identified at the cutoff distance threshold of 110. The Ward linkage function was used ........................................... 141
7.6 t-SNE visualisation for $k$-means clustering from autoencoder latent representation ........................................... 145
7.7 t-SNE visualisation for $k$-means clustering from the high-level granular bout-based representation ........................................... 145
7.8 t-SNE visualisation for hierarchical clustering from the autoencoder latent representation ........................................... 146
7.9 t-SNE visualisation for hierarchical clustering from high-level granular bout-based representation ........................................... 146
7.10 t-SNE visualisation for DBSCAN clustering from the autoencoder latent representation ........................................... 147
7.11 t-SNE visualisation for DBSCAN clustering from the high-level granular bout-based representation ........................................... 147
7.12 t-SNE visualisation for Affinity Propagation clustering from latent representation ........................................... 147
7.13 t-SNE visualisation for Affinity Propagation clustering from the high-level granular bout-based representation ........................................... 147
7.14 t-SNE visualisation for spectral clustering from autoencoder latent representation ........................................... 148
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>7.15</strong></td>
<td><em>t</em>-SNE visualisation for spectral clustering from the high-level granular bout-based representation</td>
</tr>
<tr>
<td><strong>7.16</strong></td>
<td>Clinical feature ranking for random forest classification algorithm. Sex, defined as <em>NewGender</em>, seem to be the most powerful predictor by far. BMI is the second most important. Other remaining clinical features gradually decrease in importance for the classification results</td>
</tr>
<tr>
<td><strong>7.17</strong></td>
<td>Clinical feature ranking for XGBoost classification algorithm. Sex, again, seems to be the most powerful predictors and overpower the other features significantly. The remaining clinical features present similar importance in the feature ranking</td>
</tr>
<tr>
<td><strong>7.18</strong></td>
<td>Clinical feature ranking for AdaBoost classification algorithm. BMI and Sex seem to be the most powerful predictors significantly and overpower the other features significantly</td>
</tr>
<tr>
<td><strong>7.19</strong></td>
<td>Histogram showing the distribution of values for OGIS insulin sensitivity across Cluster 0 (Blue cluster)</td>
</tr>
<tr>
<td><strong>7.20</strong></td>
<td>Histogram showing the distribution of values for OGIS insulin sensitivity across Cluster 1 (Red cluster)</td>
</tr>
<tr>
<td><strong>7.21</strong></td>
<td>Histogram showing the distribution of values for Matsuda insulin sensitivity across Cluster 0 (Blue cluster)</td>
</tr>
<tr>
<td><strong>7.22</strong></td>
<td>Histogram showing the distribution of values for Matsuda insulin sensitivity across Cluster 1 (Red cluster)</td>
</tr>
<tr>
<td><strong>7.23</strong></td>
<td>Histogram showing the distribution of values for Stumvoll insulin sensitivity across Cluster 0 (Blue cluster)</td>
</tr>
<tr>
<td><strong>7.24</strong></td>
<td>Histogram showing the distribution of values for Stumvoll insulin sensitivity across Cluster 1 (Red cluster)</td>
</tr>
<tr>
<td><strong>7.25</strong></td>
<td>Histogram showing the distribution of values for body mass index across Cluster 0 (Blue cluster)</td>
</tr>
<tr>
<td><strong>7.26</strong></td>
<td>Histogram showing the distribution of values for body mass index across Cluster 1 (Red cluster)</td>
</tr>
<tr>
<td><strong>7.27</strong></td>
<td>Histogram showing the distribution of values for fasting glucose across Cluster 0 (Blue cluster)</td>
</tr>
<tr>
<td><strong>7.28</strong></td>
<td>Histogram showing the distribution of values for fasting glucose across Cluster 1 (Red cluster)</td>
</tr>
<tr>
<td><strong>7.29</strong></td>
<td>Histogram showing the distribution of values for age across Cluster 0 (Blue cluster)</td>
</tr>
<tr>
<td><strong>7.30</strong></td>
<td>Histogram showing the distribution of values for age across Cluster 1 (Red cluster)</td>
</tr>
<tr>
<td><strong>7.31</strong></td>
<td>Histogram showing the distribution of values for OGTT 120 minutes across Cluster 0 (Blue cluster)</td>
</tr>
<tr>
<td><strong>7.32</strong></td>
<td>Histogram showing the distribution of values for OGTT 120 minutes across Cluster 1 (Red cluster)</td>
</tr>
<tr>
<td><strong>7.33</strong></td>
<td>Histogram showing the distribution of values for sex across Cluster 0 (Blue cluster)</td>
</tr>
<tr>
<td><strong>7.34</strong></td>
<td>Histogram showing the distribution of values for sex across Cluster 1 (Red cluster)</td>
</tr>
<tr>
<td><strong>7.35</strong></td>
<td>Histogram showing the distribution of values for waist circumference across Cluster 0 (Blue cluster)</td>
</tr>
</tbody>
</table>
## List of Tables

2.1 Table showing the number of subjects for each centre at each collection time point in the DIRECT physical activity accelerometer dataset. Study 2.1 are pre-diabetes subjects, Study 2.2 are subjects positive for Type-2 Diabetes ........................................... 48

3.1 Table displaying the number of individuals in each cohort of the GGIR physical activity dataset. .......................................................... 55

3.2 Table displaying the final set of features extracted by GGIR used in our baseline experiment after preprocessing and quality control process. Gives detail on how much data was missing from each variable. ............................................................ 57

3.3 Table showing demographics of the selected final training dataset after preprocessing effort. .......................................................... 58

3.4 Table displaying the values for the results of standard performance metrics for the random forest’s prediction of disease outcome using selected GGIR features. ........................................... 63

3.5 Table displaying the values for the results of standard performance metrics in predicting disease outcome using selected GGIR features with the SVM classifier. ........................................... 64

4.1 Number of participants in each sample according to physical activity impairment severity scores. .................................................. 70

4.2 Sociodemographic, lifestyle, and anthropometric characteristics selected from the UK Biobank baseline assessment for comparison with high-level activity bout features space. ................................ 75

4.3 Classification results measured using area under the receiver operating characteristic curve scores, showing the effect of choice of type 2 diabetes–negatives, Norm-0 (no physical activity impairment) versus Norm-2 (severe physical activity impairment). The values in the cells represent area under the receiver operating characteristic curve scores. .................................................. 81

4.4 Classification results measured using F1, showing the effect of choice of type 2 diabetes-negatives, Norm-0 (no physical activity impairment) versus Norm-2 (severe physical activity impairment). The values in the cells represent F1 scores. .................................................. 82

4.5 Calinski-Harabasz and Silhouette scores for unsupervised, resulting clusters by high-level activity bout features using k-means and hierarchical clustering algorithms. .................................................. 87
4.6 Distribution of discretised sociodemographic, anthropometric and lifestyle features in all clusters. .............................. 90

5.1 Table displaying the clinical target outcomes selected for supervised learning experiments based on clinical knowledge and literature review. There are 21 potentially useful clinical target outcomes. . . . 104

5.2 Table displaying proportion of the clinical target outcome missing values for data. .................................................. 105

5.3 Supervised regression results using the LSTM autoencoder latent representation as training data, with R2 score and MAE reported. .............................................................................. 108

5.4 Supervised regression results using the granular bout-based high-level activity features as training data, with R2 score and MAE reported. .............................................................................. 109

5.5 The results for supervised regression models are displayed for all selected clinical target outcomes. R2 scores and MAE are reported. These are the results of the predictive models that used the DIRECT hand-crafted vm.hpf.mean and enmo.mean variables as the feature set. 110

6.1 Table demonstrating the results for the tests for normality in each of the high-level physical activity summary features at baseline for Cohort 2.1 .............................. 122

6.2 Table demonstrating the results for the tests for normality in each of the high-level physical activity summary features at baseline for Cohort 2.2 .............................. 123

6.3 Table demonstrating the results for the tests for normality in each of the high-level physical activity summary features 36 months after baseline for Cohort 2.2 .............................. 123

6.4 Table demonstrating the results for the tests for normality in each of the high-level physical activity summary features 48 months after baseline for Cohort 2.1 .............................. 124

6.5 Table displaying the results for correlation analysis, using Spearman and Pearson correlation coefficients, between the Euclidean distance changes in the autoencoder latent representation of the physical activity from baseline to 36 months for those individuals belong to Cohort 2.2 (Those with confirmed diabetes) .............................. 125

6.6 Table displaying the results for correlation analysis, using Spearman and Pearson correlation coefficients, between the Euclidean distance changes in the autoencoder latent representation of the physical activity from baseline to 48 months for those individuals belong to Cohort 2.1 (Pre-diabetes) ............................................. 125

7.1 Calculated scores for clustering validity indices for the latent representation of DIRECT physical activity dataset. .............................. 148

7.2 Table demonstrating the number of individuals within each of the two clusters generated by spectral clustering algorithm. .................. 149
7.3 Table displaying the values for scores for precision, recall and f-1 performance evaluation metrics for the classification algorithms used to predict which cluster an individual should belong to based on their clinical feature profiles 152

7.4 Table demonstrating the results for the tests for normality in each of the high-level physical activity summary features 160

7.5 Table displaying the results for $F$-test results in each of the high-level physical activity summary features. All features show approximately equal variance 161

7.6 Table displaying the results for the two-tailed t-test results in each of the high-level physical activity summary features. p-values and t-statistics are reported 162

7.7 Table displaying the results for the t-test results in each of the high-level physical activity summary features. p-values and t-statistics are reported 163
Chapter 1

Introduction
1 Introduction

The concept of an “extended phenotype” [104] was first known to be described by the evolutionary biologist Richard Dawkins in 1982 [51]. It is the idea that phenotypes should not be restricted to only biological processes but also encompass all effects that a gene has on its environment within and outside the body of an organism. Human beings interact with their environments all the time. Examples of this extended phenotype are found in biology, such as manipulation of host behaviour by parasitic organisms [101], or even the human perception of using cosmetic products [66]. These interactions and behaviours are an expression of their genetic make-up and so can be described as part of their phenotype.

Digital and personal technologies, such as wearables, are now increasingly ubiquitous and more and more embedded into our lives. The data generated by these technologies from various different types of sensors could be used to enlarge the definition of the “extended phenotype” even further into the notion of a digital phenotype [104]. That is, the interactions between technology and humans can be considered as part of a person’s extended phenotype.

The use of this digital phenotype has already been documented in many examples, such as disease surveillance for identifying outbreaks of infectious diseases [21] through social media, and identifying suicidal ideations with Google search queries [124]. These examples are an insight into how data generated from digital technologies are a way of contributing to a person’s overall phenotype as a result of their interactions. This paradigm could be exploited further to prevent, treat or diagnose many diseases and enable a vision of Preventative, Participatory, Personalised and Predictive (P4) medicine [68, 100].

2 P4 Medicine

Medicine and healthcare currently exists as a reactive profession. This means that individuals, or groups of individuals, undergo clinical interventions at the point of when disease manifests or has progressed to a more severe state. This is problematic as it is inefficient and expensive for national health services and also leads to more deaths [133].

Chronic diseases such as cancer, diabetes, cardiovascular disease and dementia are increasingly the leading cause of death, disability and a diminished quality of life. More than two thirds of all deaths are caused by one or more of these five chronic diseases: heart disease, cancer, stroke, chronic obstructive pulmonary disease, and diabetes [164] According to the World Health Organization, diabetes (Type-1, Type-2 and gestational) alone caused the deaths of 1.5 million people in the world directly in 2019, nearly half of which occurred in those under the age of 70. Diabetes also caused nearly half a million cases of kidney disease deaths and one in five cardiovascular disease deaths. Many individuals also suffer from multiple chronic diseases at the same time, known as multimorbidity, which further

https://www.who.int/news-room/fact-sheets/detail/diabetes
presents challenges to health services. For example, in the UK, it is estimated by the year 2035, that nearly 17% of the population will suffer from at least four chronic diseases, up from 8-9% in 2019[153] This rise in chronic disease and multimorbidity is partially driven by an ageing population.

Medicine and healthcare therefore needs to shift from a reactive approach to a preventative and personalised paradigm [197, 54] as the challenges of an ageing population are only exacerbated with increasing life expectancy. But this rise in life expectancy does not necessarily correspond to a rise in the quality of life.

2.1 Challenges of an Ageing Population

Life expectancy has increased enormously in recent decades and continues to rise [37]. Global life expectancy increased from 66.8 years in 2000 to 73.9 in 2019. and between 2015 and 2050, the proportion of the world population aged 60 or above is projected to nearly double from 12% to 22%. This means that many countries are now facing an ageing population, meaning that the younger, working-age population represents a diminishing cohort in the overall population. This shift in demographics presents many implications for society, including the future of family care, living arrangements, intergenerational relationships between older and younger members of society, increased unemployment and social pressure for the younger members of the family unit to provide care and support for an increasing number of elderly family members [198].

On a national scale, an ageing population will potentially present greater pressure on health services in countries experiencing this shift and also greater economic burden [102, 200] A preventative, personalised and predictive vision for healthcare practice can alleviate some of this pressure by preventing or alleviating individuals from transitioning from a state of wellness to illness.

2.2 Big Data and P4 Medicine

Responding to the need for such new predictive capabilities, the next generation of healthcare will be increasingly driven by the continuous collection, integration and analysis of a diversity of individuals and population data. Features of these complex datasets are highly diverse, ranging from periodic multi-omics data (genomics, proteomics, metabolomics, and more), clinical measures such as blood, urine and medical imaging to continuous and high rate self-monitoring data from wearable sensors as well as measurable social and physical environmental factors that could affect one’s mental, social and physical well-being.

The increasing availability of data generated from sources such as the UK Biobank physical activity dataset [56] are helping to enable a new future for medicine that is data-driven, predictive, preventive, personalised and participatory. This vision is called P4 medicine [100] Although this goal has existed for a long time [220], the recent advances in data science, technology and relative low cost of personal wearable devices and genetic sequencing have made this a very realistic vision.

https://www.who.int/news-room/fact-sheets/detail/ageing-and-health
the future [100]. If P4 medicine is to be truly realised, it will have a profound impact on society by transforming healthcare systems, reducing expenses, digitising medical practice and creating enormous economic opportunities. However, several data science challenges exist and must be addressed before this vision can be fully realised.

These challenges include building and identifying the best set of feature spaces that can be used to learn predictive models of disease outcomes. Furthermore, challenges exist in evaluating and identifying the best representation spaces to learn these predictive models of disease-related outcomes. This is due to the immense diversity of different types of digital biomarkers (such as physical activity trackers, remote blood pressure monitors) which can then further be complicated by using them in combination with more traditional clinically derived biomarkers such as biochemistry and medical imaging. Our research demonstrates the importance of addressing these challenges to enable a vision of data-driven P4 medicine.

3 Thesis Key Contributions

The hypothesis that underpins this research is that digital biomarkers [208] can be extracted from sensor data generated by self-monitoring digital devices when applied to populations at-scale to support the clinical study of chronic diseases and characterise individual digital phenotypes. Although the number of diseases that could be studied is vast and diverse, in this thesis, we have chosen to focus on Type-2 Diabetes as a case study. There is a good body of work that establishes the importance of dietary factors in the development and progression of Type-2 Diabetes [71], even that low calorie diets can cause Type-2 Diabetes remission [134, 190]. However, Type-2 Diabetes is also known to be associated with low levels of physical activity, poor sleep, and increasing age [194]. This relationship has been established for quite some time already [121, 41].

This disease has become much more prevalent and is rapidly rising globally, especially in parts of the developing world [33]. In the United Kingdom alone, Type-2 Diabetes is estimated to cost the National Health Service around £12 billion a year, both in terms of indirect and direct care [3]. More than 2.7 million people in the UK have been diagnosed with Type-2 Diabetes, but many more have yet to be diagnosed, and even more people will go on to develop this disease due to a combination of an increasingly sedentary lifestyle and ageing [112]. All of this makes Type-2 Diabetes an ideal case study to address both the clinical questions and the data science and engineering questions. Our research demonstrates the value of signals extracted from raw time series traces for studying Type-2 Diabetes disease outcomes and further exploit the digital phenotype by using techniques in data science to characterise individual digital phenotypes and learn predictive models for clinical outcomes.

Self-monitoring digital devices which take track the signals generated by individuals are a powerful tool in understanding a person’s behaviour and phenotype.
These devices can measure many aspects of a person such as glucose monitoring and pulse oximetry. We, however, have chosen physical activity trackers in the form of a wrist-worn accelerometer as the case study for its investigations. Physical activity traces possess many of the limitations that arise when exploiting sensor data generated by many different kinds of self-monitoring, personal devices. Another reason why this specific case study in physical activity tracker data with application to Type-2 Diabetes is that there is a good level of the data availability where both the physical activity traces and the clinical phenotype cohorts appropriate to our study can be accessed.

The challenges presented by both the Type-2 Diabetes clinical focus, and the raw physical activity traces [88, 154], are a powerful and informed case study to address some of the data science challenges in enabling a vision of preventive, personalised and predictive healthcare practice. Our research also addresses some of the biggest data science challenges in enabling a data-driven, ubiquitous and personalised approach to the collection and analytics of Big Health Data. A major challenge lies in finding the best method to represent the signals generated from these sensing devices which are then valid and viable for learning predictive models of targeted clinical disease outcomes or to further the ability to interpret the data generated by these self-monitoring, ubiquitous sensing devices.

This thesis also makes a comprehensive and critical comparison of various methods for representing the raw time series traces for accelerometer data between established and widely used methods for extracting value from raw accelerometer traces, and also the application of methods to this context. Our research work lies at the interface of data science and healthcare research.

4 Research Aims and Outline

Two core aims motivated this research. First is to investigate how methods from data science and in healthcare science could be implemented, adapted and applied for extracting features from signals generated by self-monitoring, personal, ubiquitous devices such as wearables. Second is to deploy empirical methods to ascertain the validity and viability of these methods and the power of these features to either learn predictive models for targeted disease outcomes or to segregate and group individuals, by these features, according to their similarities and to characterise these segregated groups. The implications of this research could support clinical and data science research in this domain, and furthermore, could offer new insight into the way digital, self-monitoring and ubiquitous personal devices could be used to form part of a person’s digital phenotype and included in their wider whole phenotypes.

To achieve these aims, a case study has been chosen to validate and implement these methods. Data from wrist-worn accelerometer devices measuring physical activity and sleep are chosen due to their wide availability, and a comprehensive body of research already existing in using physical activity for disease research for a long period of time [56,53,207]. Although physical activity may be an important lifestyle...
indicator for a variety of diseases, the focus is placed on Type-2 Diabetes. Type-2 Diabetes is a challenging disease as it often develops slowly and progressively.

Chapter 2 explores the methods and study in the current state-of-the-art in exploiting digital biomarker signals and wearables in clinical science and healthcare applications. It outlines some of the challenges in this field of study. Furthermore, this chapter also explores the use of machine learning techniques that were adopted in the research work we undertook. This includes both classical machine learning methods, including both supervised and unsupervised machine learning techniques, as well as more advanced deep learning architectures that were also used in our research. It provides a background into these methods and how some of these methods have been used to address the research questions in this thesis. Chapter 2 also explores one of the key applications of these methods in the field of Human Activity Recognition which was widely used and built upon in our research. The two main datasets used in our study are also described in this chapter.

Chapter 3 uses one of the most widely used and known tools to extract features from raw accelerometer traces called GGIR \cite{205, 206}. The features from this package are widely used in clinical research into physical activity and contains a suite of features such as the Euclidean Norm Minus One (ENMO), periods of estimated sleep, estimated periods of wake and sleeping times, and metadata description (such as number of days, name of weekday etc)\footnote{https://cran.r-project.org/web/packages/GGIR/vignettes/GGIR.html}. This package was used to establish a baseline for which the remaining research works in this thesis sought to improve upon. These features generated from GGIR are then deployed on classical machine learning models to predict whether an individual has cardiovascular disease and/or Type-2 Diabetes, or neither (classed a control subject) to study the predictive power of a subset of GGIR features after preprocessing steps were undertaken. It is the initial exercise undertaken with the UK Biobank, a comprehensive resource containing genotype and phenotypic data from approximately 500,000 individuals, aged between 40 and 69. This includes clinical measurements, lifestyle indicators (including physical activity traces for a large proportion of participants), biochemical and medical imaging data that characterise each individual’s phenotype, as well as their data representing their genotype.

Chapter 4 seeks to improve the baseline achieved with GGIR. A different tool called the Biobank Accelerometer Analysis Toolkit is used to generate labelled and segmented human activity recognition sequences of around 20,000 out of approximately 100,000 accelerometer wearers in the UK Biobank through a rigorous process of quality control and defined selection criteria. This includes discarding individual data from this study where an individual’s physical activity may be impaired due to a disease, condition or a clinical event that is not related to Type-2 Diabetes, such as for example if they have recently had a major surgery or a cancer diagnosis, but would otherwise severely impair their ability to perform physical activity. We then apply methods from neuroscience and gait analysis to study the physical activity patterns based on the labelled activity sequences to extract features at the granular time-of-day, bout-based level. These methods are adapted
from studies for diseases such as Parkinson’s disease and dementia, and also for fall detection in the elderly. These features were then deployed to learn predictive models for Type-2 Diabetes. The results from this series of investigations offered significant improvements compared to GGIR and also highlights the importance of data and information quality and provides useful insights into clinical data collection protocols with free-living cohorts.

Chapter 5 begins a new stage of research where deep learning methods are implemented and applied. After segmenting the raw physical activity traces and feature extraction, an autoencoder architecture is used to learn a latent representation of the low-level epoch signal features of the selected accelerometer dataset. Furthermore, a second test bed accelerometer dataset is used to ensure that methods can be applied generally. The DIRECT diabetes dataset is specific to diabetes\cite{116}: All participants are either in the pre-diabetes stage or already diagnosed with Type-2 Diabetes. DIRECT has also collected data for a cohort of accelerometer wearers, albeit a smaller dataset. However, more rigorous protocols were used in the collection of this accelerometer dataset. The new latent representation for the low-level signal features are then also deployed onto predictive models for targeted disease outcomes, and also compared with the granular, bout-based features from chapter 4.

Chapter 6 investigates the change in physical activity over time and the association of this change with the changes in progression for clinical phenotypes of interest. This exercise demonstrated the impact and predictive power of physical activity on disease progression.

Chapter 7 aims to address the second aim of using features to segregate individuals into distinct groups of similar patterns of physical activity. Unsupervised clustering methods are deployed to achieve this and we measure these with standard clustering validity indices before interpreting these clusters for clinical and statistical significance in terms of physical activity.

Chapter 8 is the final discussions chapter which highlights the key insights and contributions that this thesis has made, as well as the applications in the real world. It also discusses the potential directions for future work.

5 Publication

The work contributing to this thesis has resulted in the following publication:

Chapter 2

Background Literature and Related Works
1 Introduction

In this chapter, we conduct a systematic and critical review of works in the related latest literature. We also provide some background and context to the key ideas in this thesis, including an introduction to machine learning, and deep learning, the role of self-monitored, ubiquitous, free-living digital wearable technologies in the clinical context of healthcare, human activity recognition. The chapter also explores the two datasets used in this thesis as the test beds for experimental analyses. It describes some of the selected works in the related literature, critically discusses its strengths, challenges and limitations, and describes how this thesis addresses these identified limitations.

2 P4 Medicine

In Chapter 1 the idea of P4 medicine was briefly introduced. This is a paradigm for the future of medicine and healthcare which envisions a transformation from a reactive ‘evidence-based’ diagnose-and-treat approach to a predictive, preventative, personalised and participatory vision. Several trends demonstrate that the P4 paradigm of medicine is growing and will grow to replace the current reactive approach [68]. Clinical cohorts previously limited in size are now being replaced by massive lakes of data generated at individual and population scales, and these data have the potential to be processed with advancing Big Data technologies [144]. Furthermore, there are now more cost-effective methods to diagnose and treat diseases at the biochemical and multi-omics level- for example more affordable genome sequencing compared to the past [1]. Healthcare is extending itself beyond just the disease diagnosis and treatment stages to now incorporating the maintenance and enhancement of a state of wellness in free-living settings such as at home and in workplaces. Many private companies and public institutions are beginning to emerge that aim to contribute to this paradigm and they will provide a major driver of economic growth [68]. P4 Medicine will also help to alleviate the challenges that will manifest with the growth of an ageing demographic [58].

P4 Medicine is a very different approach to the current ‘evidence-based’ practice, but this approach still has two main challenges before this paradigm becomes the normal practice. Firstly, a limitation in the current technology means that further advances are also limited until technology catches up to the demands needed for P4 medicine to be realised. The second major challenge is that society at-large may not embrace this approach, which would be problematic in enabling the participatory element of this paradigm as enough individuals need to be engaged in this shift to realise its potential. This involves a change in social attitude towards a data-driven vision of healthcare. Most importantly, trust and consent must be obtained and preserved. Other considerations of societal issues that need to be addressed to increase participation include ethics, legal issues, privacy and confidentiality, patient data accessibility, and data ownership [99].

Societal changes must be implemented to facilitate a paradigm shift from the conventional evidence-based medicinal approach to personalized medicine’s predictive and preventive approach. These societal challenges include the following

1https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data
considerations: ethics, legal, privacy, patient data accessibility, who owns the data, etc.

Technologies that enable a digital phenotype, even including mobile phone applications, provide a means to stratify and visualise chronic diseases which can help to reveal disease heterogeneity and pathology [103]. Systematic reviews carried out in the general field of P4 medicine and its challenges have highlighted notable issues before this paradigm of medicine and healthcare can be fully realised. There is a growing need to consider the socioeconomic and ethnic inequalities, and also ethical considerations, that exist in many societies which could hinder the progress of this paradigm [187]. There is also the need to preserve privacy, anonymity and confidentiality of all participants to maintain trust in the system. Emotional and psychological effects of the growing role digital technology in society is also a critical issue that presents a challenge to be considered [138].

The issues of incorporating and truly exploiting the power of Big Data in Health is well published in the literature for some time now [4] but work still remains to be done to fully address these issues. Major bottlenecks and challenges include the lack of sufficient performance by predictive models, challenges in model stability and interpretability, as noted by Frölich et al [75].

Lastly, one of the difficulties exists in the challenge of insufficient validation mechanisms before statistical and machine learning methods can be used in clinical practice. Machine learning models are often optimistic and fit tightly to training data, but may perform poorly on unseen validation data in the future due to the lack of generalisation. The thesis aims to demonstrate the exigency of some of these issues and will attempt to address them where possible to do so. Issues our research attempts to address are the complex challenges in feature engineering and finding the best representation of raw data that can be used to learn predictive models for clinical outcomes, and also in identifying and optimising these predictive models for clinical outcomes.

3 Digital Wearable Technology in Clinical Science

This section discusses the use of digital wearable technologies in the context of clinical science and healthcare applications. This gives a brief overview of the state-of-the-art and related works. This overview also hopefully helps to delineate where the unique contributions this thesis makes lies and the limitations of the state-of-the-art and literature.

3.1 Potential for Digital Biomarkers

Digital health technologies are one of the key developments that will help to enable a vision of P4 medicine [208]. These technologies generate digital signals which could become digital biomarkers for clinical outcomes. The Biomarkers, EndpointS and other Tools (BEST) glossary developed by the US Food and Drug Administration (FDA) defines a biomarker as “a defined characteristic that is measured as an indicator of normal biological processes, pathogenic processes, or biological responses to an exposure or intervention, including therapeutic
interventions." The US FDA then defines *digital biomarkers* as one of these characteristics, or set of characteristics, that are collected by digital technologies.

In order to extract digital outcomes that could act as digital biomarkers, these digital technologies must satisfy some properties. Firstly, they need a mechanism for input such as a camera, pressure sensor, microphone etc capture digital signals. For example, accelerometer devices are used to measure physical activity levels by generating digital signals for proper acceleration and these could then be later evaluated and proven to be a digital biomarker when they are objective indications of clinical state and can be measured accurately and reproducible. Digital technologies also need to be verifiable, so that the raw digital signals they capture can be translated into a form that is an accurate representation of what the technology is trying to measure (for example acceleration, pulse rate). These technologies need to be valid, *fit-for-purpose* in being operational in capturing signals and translating these signals for target users and context[42].

As digital health technologies become more integrated in the practice of medicine and healthcare, including long-term care, it becomes more necessary to combine digital signals from wearables with other characteristics that include multi-sourced physiological biomarkers (such as blood glucose, blood pressure, heart rate) and self-reported details given by patients (questionnaires) [42] to improve the diagnosis and outcomes in healthcare. The increased integration of digital technologies in healthcare applications also means that it is important to ensure that these technologies must also respect issues such as privacy, anonymity and confidentiality for the patients. This includes cybersecurity risks in the use of technologies, in particular with the increase in Cloud and Internet of Things systems [14], informed consent and the need for transparency to improve trust in technology. Participation is one of the greatest challenges in the increased integration of digital technologies that will be revolutionary in enabling a personalised and participatory vision of medicine and healthcare.

Despite these challenges, digital signals which can be proven as digital biomarkers are still a powerful tool if patients trust in safe and secure participation and provide the means for clinical healthcare providers to screen whole populations, estimate statistical prevalence and enable a scalable, non-invasive and ubiquitous examination of health outcomes in target populations[129].

### 3.2 Digital Wearable Technology in Healthcare Applications

Wearable technologies are smart electronic devices that are worn close to and/or on the surface of the skin, where they can detect, analyse and transmit the digital signals, and potentially allow feedback to the person wearing the device immediately[61]. These technologies are an example of the Internet of Things. Wearables are a powerful tool in generating digital signals which could become digital biomarkers for disease and health assessment, for example in the case of depression screening from digital signals generated using activity tracking devices [174]. They are becoming increasingly an important component in the broader context of digital health [3].

An enormous quantity of applications of these wearable sensing systems exist that include the analysis of healthcare and behavioural patterns. Wearables can be...
used to collect and analyse data on the wearer’s health including their heart rate, calorie expenditure, step count, blood pressure, biochemical release, remote glucose monitoring [170][98][199].

Devices now often have the capacity to collect this data on a single device unit, rather than having several devices for each measure being pursued. Advances in cloud storage technology and increasing affordability of these technologies have enabled a combination of non-invasive wearable technologies and deep learning to achieve a continuous assessment of health risks in the ageing populations [161][151].

In the future, home robotics and wearable technology combined may have the potential to perform a key step in building towards the goal of effectively monitoring of patients in their free-living home settings [18].

combining home robots and wearable technology is likely to be a key step toward achieving the goal of effectively monitoring patients in the home.

3.3 Digital Wearables Technology: Challenges

Many of the challenges in digital health [45] are applicable to the challenges in digital wearables.

According to the literature, feature extraction remains one of the biggest challenges in the context of digital physical activity monitoring from wearable devices. Many efforts have gone into automating this effort. Li et al [127] automated this feature extraction process from the UK Biobank’s accelerometer data with unsupervised machine learning. This method was able to identify genetic loci associated with certain characteristics of sleep patterns. Feature extraction was achieved to identify sleep-wake states with Hidden Markov Models.

Wang et al [214] use a Kalman filter for data fusion from different wearables (this related work uses accelerometers and magnetometer data) and then deploys convolutional neural networks to extract features from wearables for the task of human activity recognition. This approach achieved results with varying degrees of success on five different datasets, with F-1 scores ranging from 44.7% to 93% for recognising activities in supervised learning experiments.

The need for data fusion was also demonstrated in [128] where data from accelerometry, gyroscopes and magnetometers were fused together to learn a hand movement classifier for the detection of resting tremors and bradykinesia. This paper recognises the advantages in minimising the number of devices used for extracting digital signals for disease, but some conditions may entail many symptoms and exhibitions of physical activity and behaviour, so there one should consider deploying multiple devices rather than limit to only those for gait and postural impairments.

These works demonstrate one of the other challenges of digital wearables for phenotyping for disease and physiological parameters, namely the need to incorporate multiple sources of digital parameters. In the case of movement-related tasks, the data from accelerometry, gyroscopes and magnetometers are considered. However, even though these methods can extract useful features for human activity recognition, it does not address the power of these features in predicting disease outcomes, especially for those diseases where habitual behaviour and physical activity plays a subtle role such as in the case of Type-2 Diabetes. This thesis aims to address and investigate this problem.
3.3.1 Validation and Verification

It is vital that the tools and outcomes used are validated and verified. Although it is not certain that this would improve results when processing the tools’ outputs to build our machine learning pipelines, the results would certainly be more robust and reliable given that the tools used are determined to be a reliable and valid representation of reality.

Another challenge of digital medicine is the confusion caused by the diversity of terminology and best practices among different groups. It is common for a single term to be understood to have different meanings among different groups. A framework is therefore fundamental so that digital health technologies are approved by regulatory authorities such as the FDA and EMA. Goldsack et al. [80] propose a three-component framework, the V3 framework, to provide a clarification of terminology and best practices for Biometric Monitoring Technologies (BioMeTs). BioMeTs are digital medicine tools that process data generated from digital sensors to measure behavioural and/or physiological function. The components of this framework are as follows:

- Verification: assesses the performance of a sensor, and the sample-level data it generates, against a set of specified criteria.
- Analytical validation: assesses the performance of the algorithm and its ability to measure, detect, and/or predict behavioural or physiological measures of interests.
- Clinical validation: assesses whether a BioMeT satisfactorily identifies, measures or predicts a clinical, physical, functional or biological state in the stated context of use in a specified population. These are typically performed in clinical trials with cohorts that include participants with and without the phenotype of interest.

Another challenge with wearable technology in clinical applications is the heterogeneity in protocols, sensor technical specifications, data formats, and gold standards. They present challenges in data sharing, validation and reproducibility. Efforts to address these challenges are constantly ongoing, for example in the Mobilise-D initiative [110, 150], where guidelines are being developing in collaboration with regulatory bodies to achieve data standardisation and ensure that solutions are acceptable. Another example is the IDEA-FAST project which is aiming to shift away from more traditional questionnaire-based approaches to assessing symptoms of fatigue and sleep disturbance by identifying digital endpoints which are reliable, sensitive and objective. Useful digital endpoints could then be used to make clinical trials more effective and efficient.

3.3.2 Patient and Public Involvement and Engagement (PPIE)

Patient and Public Involvement and Engagement (PPIE) is also an important factor that must be considered when conducting research in digital health technology and targeted digital outcomes for phenotyping. Public involvement and engagement

\[\text{3https://www.imi.europa.eu/projects-results/project-factsheets/idea-fast}\]
regards the importance of actually enabling members of the public to become part of the research effort, rather than the research being done to, about or for them.

The term *public* can refer to patients, carers, or organisations. Public involvement means that patients can help to shape how research is carried, what research is carried and how it is shared. Public engagement, on the other hand, refers to how research can be shared with the public in a two-way process, which can often give researchers in an opportunity to discuss research in a general sense and consider other important issues such as those in research ethics, confidentiality, and consent.

Furthermore, it also reinforces the need and importance for practitioners of data science and engineering to have an understanding of their relevant clinical application area and what users or patients may ultimately find important in results. Evidence shows that studies that run PPIE activities are often better run [160, 202].

## 4 Foundations of Machine Learning

This section introduces the key concepts and methods from classical machine learning that were used in the study design of this thesis. It includes both supervised and unsupervised machine learning models, regression and classification tasks, how the model performances were evaluated and the mechanisms deployed to optimise model performance.

### 4.1 Machine Learning Basics

Machine learning is a form of artificial intelligence which is concerned with building algorithms that rely on a collection of samples. These samples can be generated by non-human causes, crafted by domain experts in a given field, or constructed from the output of a different algorithm. Each of the samples is organised into a *feature vector* which is a vector where each dimension contains a value that holds a descriptive characteristic, called a *feature*, of the sample [7]. For example, a feature vector for a *Person* sample may contain dimensions for *Height*, *Age*, and *Name*. In a dataset of the given samples, each sample should have a value for each feature.

The learning algorithm is then applied to the dataset in order to derive a model. The goal of the learning algorithm is then to use the dataset to find the optimal values for model parameters, called hyperparameters. The hyperparameters depends on the chosen learning algorithm. Once these optimal values for the hyperparameters are found, the model is derived. This process of optimising the model to attain the optimal values for the hyperparameters is called *training*.

We can divide the dataset into three sets: Test set, validation set, and a training set. The training dataset is used to fit the machine learning model. The test dataset a subset of the training data that is split from the total dataset and is used to evaluate the final model that was fit on the training data. The performance this model is then evaluated by how well the model predicts the labels for these unseen test samples with various metrics. A validation dataset, a subset of the total dataset, is used to provide an evaluation of the model fitted on the training data while tuning hyperparameters.

4.2 Supervised vs Unsupervised Machine Learning

Classical machine learning can be divided into two main categories for the type of learning algorithm: Supervised learning and unsupervised learning. Supervised learning builds a model which then forms its outputs based on training the models with data that has been fully labelled with the ground truth. The learning algorithm then uses these ground truth labels to train the model and performance is measured by how well the learning algorithm is able to correctly label new samples. The most common popular supervised learning algorithms include random forests, decision trees, support vector machines, linear regression, logistic regression and elastic net regression. These supervised learning algorithms have been used in this thesis.

Unsupervised learning deploys a different approach whereby the learning algorithm possesses the ability to learn from the training data without any ground truth labels. There are no error signals to evaluate the output of the learning algorithm. This lack of ground truth means that the learning algorithm can be very good at identifying new patterns and segregate data points into groups, such as in clustering, which had not been considered by a supervised algorithm which tries to fit the output to the training data[177]. Common models for unsupervised learning include $k$-means clustering, hierarchical agglomerative clustering, DBSCAN, and techniques for dimensionality reduction such as principal component analysis and $t$-distributed Stochastic Neighbourhood Embedding (t-SNE). These algorithms have been used in this thesis.

Both methods from supervised and unsupervised machine learning models have been used in this thesis to investigate different representations of raw time series of physical activity traces. Supervised learning models are used to predict clinical Type-2 Diabetes target variables, whereas unsupervised clustering methods were used to assess and segregate individuals that exhibit similar habitual behaviour patterns for deeper analysis, and also for the representation learning task with the deep neural networks, in the form of autoencoders.

4.3 Supervised Machine Learning: Classification

Classification is one of the main tasks in supervised learning. Classification learning algorithms take a collection of labelled examples as inputs, and then produce a predictive model that is able to receive an unlabelled example as the input and then puts out an interpretable label [117]. A predictive model that performs classification is a classifier, where models that can produce two possible labels are called binary classifiers - for example if a label could be Healthy or Not healthy. If there is more than one class, it is called a multiclass classifier but there is only one label for each given example. Classifications works on a one-to-one or many-to-one mapping.

4.3.1 Decision trees

Decision trees are an acyclic graph, a graph which doesn’t form a cycle, and forms a flowchart-like structure. Each internal node of the graph represents one of the predictive features, and then each branch of the node represents the outcome of each feature. Each leaf node, at the end of the tree, forms the classification label when all the decisions have been made going along the tree [175]. Random forests form an
integral part of one of the classifiers that was most used in this thesis: The random forest.

4.3.2 Random forest

The random forest classifier is a type of learning algorithm that uses an *ensemble* approach. This is where multiple classifiers, often weak performing ones, are grouped together to attain a stronger performing classifier. In the case of the random forest, the weak performing classifiers are made up of multiple decision trees [16]. Each decision tree in the forest is set up with a random number of samples, with replacement. Then, a randomly selected subset of features, but without replacement, is then used to split each node of the decision tree. The algorithm used to train each decision tree in the random forest is the Classification And Regression Tree (CART) algorithm [6]. This algorithm is applied to each decision tree which then produces an class label based on the ground truth. Each decision produces one prediction label. The random forest then takes a majority vote (or an average in regression). The final prediction is the one with the most votes (or the average in regression problems).

![Random Forest Diagram](image)

Figure 2.1: Illustration of the random forest algorithm.

Random forests are a powerful supervised machine learning method as it works well on high-dimensional and low-dimensional datasets, as it only works with a subset of the total number of features. As it deploys a method called *bagging*, it means that there is less chance of overfitting. Bagging is a technique where many new copies of the training dataset are created (each copy is slightly different to the other). Also, the algorithm performs automatic feature selection, so it gives an indication to the ranking of feature importance if the model performs well [180]. This is useful for this thesis in identifying which set of features have the greatest effect on the predictive modelling results for a selected target outcome.

The main disadvantages of this algorithm is that if the number of trees in the forest is increased, the computation time for training increases, but this would be only be an issue in extremely large datasets. Furthermore, this algorithm is purely a tool for predictive modelling and not for descriptive analysis.
4.3.3 XGBoost

The eXtreme Gradient Boosting (XGBoost) algorithm \[34\] is another classification algorithm that was extensively used in our research. \[5\] It is an implementation of the gradient boosted decision trees. XGBoost is widely used in machine learning tasks because it is optimised for speed and generally has good levels of performance. The algorithm works by implementing the gradient boosting decision tree algorithm, which is described in section 4.4.

XGBoost has several characteristics that make it a powerful machine learning algorithm that often outperforms other supervised learning models. It was built for efficiency in terms of speed and performance. It is sparse-aware, so can automatically handle missing data and eliminates the need for imputation.

4.3.4 Support Vector Machine

Support vector machines (SVM) are a supervised learning model used widely in classification tasks. This is a totally different supervised learning approach. This learning algorithm operates by discriminating between examples in the dataset with a line (if the data is two-dimensional), or a hyperplane. A hyperplane is a higher-dimensional generalization of lines and planes. The maximum margin is the maximum distance between points of the classes. The decision boundary is where the two classes are separated, this normally corresponds to the location of the hyperplane.

![Diagram illustrating the maximum-margin hyperplane for the SVM algorithm. The support vectors are also labelled. This SVM has two classes discriminated by the hyperplane.](image)

Figure 2.2: Diagram illustrating the maximum-margin hyperplane for the SVM algorithm. The support vectors are also labelled. This SVM has two classes discriminated by the hyperplane.

The aim of the SVM algorithm is to maximise the margin by finding the best hyperplane that achieves this. The examples in the dataset with feature vectors closest to the hyperplane are the support vectors, and these support vectors are key

\[5\] More can be read about XGBoost here: https://github.com/dmlc/xgboost
to maximising the margin

The equation for this hyperplane is given by two parameters, a real-valued vector \( \mathbf{w} \) normal to, and with the same dimensionality as, the input feature vector \( \mathbf{x} \), and finally a real number \( b \) called the offset is given as where \( y_i \) is the ground truth label for \( y \):

\[
\mathbf{w}\mathbf{x} - b = 0 \quad (2.1)
\]

The SVM learning algorithm then attempts to optimise the values for \( \mathbf{w} \) and \( b \) in the equation above. The model should satisfy the following two constraints:

\[
\begin{align*}
\mathbf{w}\mathbf{x} - b &\geq +1 \quad if \quad y_i = +1 \\
\mathbf{w}\mathbf{x} - b &\leq +1 \quad if \quad y_i = -1
\end{align*}
\]

However, it may not be possible to separate the data with a simple line. The data is then non-linearly separable. This could be due to noise in the data, outliers etc. SVM handles this problem by implicitly transforming the original space into a higher-dimensional space with a kernel trick. The resulting algorithm replaced every dot product \( \mathbf{w} \cdot \mathbf{v} \) with a non-linear kernel function. This allows the algorithm to fit the maximum-margin hyperplane in a transformed higher-dimensional space. The kernel transforms the original input space, which could be non-linear, into a higher-dimensional space where a hyperplane can then separate the two classes. The kernels used in classification experiments are the radial basis function, the polynomial and linear kernels.

### 4.4 Supervised Machine Learning: Regression

Regression is the other possible main problem to solve in machine learning. This is where a model attempts to predict real-valued label, called the target, from an unlabelled example given a values for the set of predictor features. Examples of tasks of this kind could be to calculate an individual’s age or their blood glucose level.

Three linear models were implemented, trained and then compared: Ridge regression, simple linear regression and elastic net. Linear models make a prediction by calculating a weighted sum of the input predictors, plus a constant value called the bias as shown in the equation 2.2

\[
y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_i x_i \quad (2.2)
\]

\( y \) is the value to predict, \( i \) is number of input features, \( x_i \) is the \( i \)th feature’s value, \( \theta_0 \) is the bias term. \( \theta_1, \theta_2, ..., \theta_i \) are the feature weights updated during training. Linear regression is trained by setting its bias and weights so that the model fits best to the training dataset and minimises the cost function, the performance evaluation metric for regression models.

Ridge regression, also known as Tikhonov regularisation, is a version of the linear regression model that is regularised. A regularisation term is added to the cost function, which causes the learning algorithm to both fit to the data and ensure weights are minimised. This is achieved by introducing a penalty on the size of the weights. The ridge coefficients minimise a penalised residual sum of square, following Equation 2.3 below:
\[
\min_w ||Xw - y||_2^2 + \alpha ||w||_2^2
\] (2.3)

The key hyperparameter in this model is the regularisation strength, \(\alpha\), which controls how much regularisation is added. If \(\alpha\) is very large, then weights will tend towards zero. The bias term is not regularised.

The final linear model for regression that was selected was the elastic net. This is a regression model that, like ridge, is also regularised but performs some feature selection by reducing the less useful features’ weights down to zero.

The gradient boosted regression model is another ensemble approach that is popular. Gradient boosting operates by adding weak learners to an ensemble, like the random forest, but it does so sequentially so that it corrects its preceding learner. The new learner attempts to train to fit to the residual errors of the preceding learner model.

A gradient boosting ensemble learning algorithm for regression is constructed with first a constant model \(f = f_0\):

\[
f = f_0(x) \overset{\text{def}}{=} \frac{1}{N} \sum_{i=1}^{N} y_i
\] (2.4)

The model then changes the labels of each observation \(i = 1, \ldots, N\) in the training dataset as follows:

\[
\hat{y} \leftarrow y_i - f(x_i)
\] (2.5)

The training set has been modified, where the original ground truth labels have been replaced with the residuals. A decision tree model, \(f_1\) is defined as the following:

\[
f \overset{\text{def}}{=} f_0 + \alpha f_1 + \alpha f_2 + \ldots + \alpha f_M
\] (2.6)

In Equation 2.6, \(\alpha\) is the learning rate hyperparameter. The process of redefining the boosting model is repeated until the predefined maximum number of trees, \(M\), are combined. The current model \(f(x_i)\)’s performance can be evaluated by calculating the residuals as in Equation 2.6. At each iteration, a new tree is trained to rectify the errors of the current model, since the trees are now predicting the residual rather than the original labels. Three hyperparameters need to be tuned in this model: The learning rate, \(\alpha\); the depth of trees; and the number of trees. Overfitting is avoided by tuning these hyperparameters. Gradient boosting is a powerful regression algorithm that usually outperforms random forests in terms of accuracy, but takes much longer to train due to the sequential process of iterating over new boosted models.

4.5 Unsupervised Machine Learning: Clustering

Another set of machine learning algorithms we used in our research are unsupervised clustering methods. Clustering is a task in unsupervised machine learning which involves grouping a set of data points together into groups (called clusters) which are similar in their properties compared to objects in other clusters. Clustering is a useful method for identify well-segregated groups within the data and does not necessarily rely on ground truth and labelled data as it seeks to find the intrinsic natural similarities within the data that may not have been considered by domain experts and analysts.
There are many different approaches to clustering [178], such as graph-based, density-based and centroid-based algorithms. Our research selects a subset of the algorithms in each type of clustering approach.

We chose to adopt a diverse set of clustering approaches in order to build up a comprehensive set of clustering results on which to compare. Some of the methods (\(k\)-means and hierarchical clustering) are classic methods that are often the initial algorithms chosen for clustering exercises. Other density-based approaches such as spectral and DBSCAN are more complex but use a different approach to segregating into clusters. This section presents the methods chosen in our research.

### 4.5.1 K-means clustering

The \(K\)-means clustering algorithm is one of the most widely used clustering models. The algorithm makes an attempt to segregate sample into \(k\) groups.

\(K\)-means has three main steps basically. Intuitively, the first step assigns the number of initial centroids, \(k\), (i.e. the number of clusters to be formed). Centroids are feature vectors placed randomly into the feature space at first. After setting a value for \(k\), the algorithm then iterates through the next two steps. The second step is to assign each sample, \(x\) of the dataset to its nearest centroid, \(c\). The third step is then to re-assign the centroids by taking the mean of all the samples to each previous centroid. The difference between this new centroid and the one from the previous iteration is then calculated. These two steps are then repeated until the difference is less than a threshold (i.e. the centroids have remained largely the same).

The key disadvantage to \(k\)-means is that the \(k\) hyperparameter must be specified before the algorithm’s iterative steps are executed. This can be achieved by the Elbow method which is how the algorithm was optimised in our work. Furthermore, the algorithm tries to minimise the inertia when reassigning centroids but this metric is not normalised and so could be affected by high dimensionality in the data. It also performs best when the clusters are convex (the clusters are circular or spherically shaped) and isotropic (not varying in magnitude regardless of which direction it is measured in) which must be determined beforehand.

### 4.5.2 Hierarchical Agglomerative clustering

Agglomerative clustering builds a hierarchy of clusters from the bottom-up\[143, 105\]. Initially, each sample is treated as an individual cluster. Then, at each iteration of the algorithm, the nearest pairs of clusters are joined together and so forth. A measure of distance between a pair of samples is needed to determine how similar one sample is to another, and then to decide whether to agglomerate or not. The distance metric chosen in our work is the Euclidean distance as it is commonly used.

Furthermore, a linkage function is needed that determines the distance between sets of samples (i.e. distance between clusters). The Ward linkage function \[106\] was used in our work. Other linkage functions may measure the distance directly, but the Ward linkage function instead analyses the variance of clusters. The Ward function is the most suitable for quantitative data, such as in the datasets in this thesis. This linkage function defines the distance between two clusters as the increase in the error sum of squares after two clusters are agglomerated into one cluster.

The main disadvantage of this algorithm is that number of clusters needs to be specified before the clustering takes place. This algorithm however doesn’t assume
that clusters must be spherical-shaped so removes this limitation that \( k \)-means possesses.

### 4.5.3 DBSCAN clustering

Density-based spatial clustering of applications with noise (DBSCAN) \([65]\) is a density-based clustering algorithm. This algorithm is widely used for density clustering \([181]\).

DBSCAN defines a cluster as a set of \textit{core points} close to each other in areas of high density, as well as non-core samples close to a core point. A core point is defined as being a sample where there are at least a specified minimum number of samples within a distance of \( \varepsilon \). Thus, the main hyperparameters to manipulate in this algorithm were \( \varepsilon \), the maximum radius between two data points \( A \) and \( B \) for \( A \) to be considered as being in the neighbourhood of \( B \), and the minimum number of samples in a neighbourhood for a data point to be considered a core point. If a sample meets these conditions, it is said to be a core point. These parameters define the density necessary to form a cluster: A higher minimum number of samples to consider, and a lower \( \varepsilon \) would imply greater density is needed to form a cluster.

A cluster is built by recursively taking a core point, finding all its neighbouring core points, and repeating this process while the \textit{min_samples} and \( \varepsilon \) conditions are met. Non-core samples that are neighbours of a core point (but are not core points themselves) are included in a cluster. These samples are usually on the edges of a cluster. Nearest neighbouring points are found using the \textit{K}-nearest neighbour algorithm, so this is also a distance-based method for unsupervised clustering.

This algorithm does not assume that clusters must be spherical-shaped, therefore it removes this limitation that may hinder the performance of other approaches. However, there is complexity in optimising the values for the \textit{min_samples} and \( \varepsilon \) hyperparameters as the DBSCAN algorithm is extremely sensitive to these.

### 4.5.4 Affinity Propagation

Affinity propagation is a graph-distance based form of clustering. It uses a voting system, where each data point votes for similar data points by sending “messages” to one another to represent them until the algorithm converges \([74]\). Similarity between points is most commonly calculated by the negative Euclidean distance. This iterative algorithm converges when the output gets approaches closer and closer to a specific value with each iteration.

Each “representative”, known as the \textit{exemplar}, and its samples that vote for it then become a clustered group of data points. The advantage with affinity propagation is that there is no need to specify a number of clusters before running the algorithm, unlike the case with some other clustering approaches such as \( k \)-means and makes no assumptions about the internal structure of the data.

The messages sent between two points belong to one of two categories: The \textit{responsibility} and the \textit{availability}. The \textit{responsibility} is the message sent by potential cluster members to candidate exemplars that communicates how suitable the cluster member would be as a member of the candidate exemplar’s cluster.

The \textit{availability} message is sent from candidate exemplars to potential cluster members that communicates how appropriate this candidate would be to its exemplary representative.
This means that a sample votes for another sample to be its exemplar if the candidate is similar enough many other samples and also chosen by many samples to be their exemplars too. The damping factor hyperparameter is introduced to prevent numerical oscillations for the responsibility and availability messages. The other hyperparameter tuned in this algorithm is the preference. This hyperparameter controls how many exemplars are used the algorithm’s iterations.

Affinity propagation has a key advantage in that the user does not need to initially specify the number of clusters to segregate the dataset into. This means that we make fewer assumptions about the inherent nature of the dataset. As it is a graph-based approach to clustering, it could sometimes consume more memory and take longer to run than approaches such as $k$-means clustering.

4.5.5 Spectral Clustering

Spectral clustering [211] is another graph-based method for clustering and works by using the set of eigenvalues, the spectrum, of the similarity matrix, $A$, of the original data to construct a low-dimensional embedding. The algorithm then deploys a standard clustering algorithm, such as $K$-means as in our work, on the components of the eigenvectors in this low-dimensional embedding. This is achieved through a critical step of calculating the Laplacian matrix representation of the graph.

The algorithm first requires the similarity matrix to be derived. Next, the algorithm calculates the graph Laplacian, $L$ on this similarity matrix. Laplacian is calculated by subtracting the diagonal matrix of degrees, $D$, from the similarity matrix, $A$. It then has to compute the first $k$ eigenvectors, which corresponds to the specified number of clusters, on this Laplacian matrix. The matrix formed by the first $k$ eigenvectors are then considered where the $l$-th row defines the low-dimensional embedding’s features for the graph node $l$. The standard clustering algorithm, such as $K$-means, is then applied on this set of embedded features.

This algorithm is good at dealing with high-dimensional data because it performs a step of dimensionality reduction. They are also a good approach when no assumptions are made about the dataset to cluster. However one of the biggest disadvantages is that spectral clustering is that it tends to be slower than more traditional approaches such as $k$-means and it is not very intuitive to explain due to the complex Laplacian calculation. However, this algorithm does often outperform more traditional approaches.

4.6 Performance Evaluation in Machine Learning

4.6.1 Performance Metrics for Classification

Accuracy is the simplest metric. It is calculated by dividing the number of correct predictions by the total number of predictions. It is then usually represented as a percentage.

Precision is calculated by dividing the number of true positives over the number of total positives (both true positives and false positives) predicted, given in the equation below:

\[
Precision = \frac{TruePositives}{TruePositives + FalsePositives}
\]  

23
A precision score of 1 for a class X implies that every instance classified into class X is correctly labelled as a member of class X, but it does not say anything about the number of instances in class X that were incorrectly labelled.

Recall is calculated by dividing the total number of true positives over the number of true positives and false negatives. It gives an indication of the number of correct positive predictions made out of all positive predictions that could have been made:

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{2.8}$$

A Recall score of 1 means that every instance from class X was labelled as belonging to class X. However, it gives no indication as to how many instances from other classes were not correctly labelled also being a member of class X.

F1 score incorporates both recall and precision. It is calculated as the harmonic mean of recall and precision. A high F-1 score indicates high precision and high recall, meaning there is a good balance and gives good performance, even on an imbalanced training dataset. The equation for F-1 scores is given below:

$$F - 1 = \frac{(2 \times Precision \times Recall)}{(Precision + Recall)} \tag{2.9}$$

The last metric combines the True Positive Rate (TPR) and the False Positive Rate (FPR). The TPR corresponds to the number of positive data points correctly predicted as positive by the model. The FPR corresponds to the number of positive predictions incorrectly predicted as positive by the model. TPR and FPR can then be plotted on a curve against each other. This curve is known as the Receiver Operating Characteristic curve, ROC curve, and the metric calculated is the area under this curve, the AUC score. An AUC of 0.5 suggests no discriminating capability. A score close to 1 implies perfect discriminating capability. A score of 0 implies the model is reciprocating the results (i.e. all negatives are being predicted as positives and vice versa).

Specificity, also known as the True Negative Rate (TNR), measures the proportion of negative negatives against the sum of false negative predictions and true negative predictions. The specificity of a model is a useful metric when false positives will incur a high cost. For example, a test with 100% specificity will correctly identify all healthy individuals in a test for a disease as indeed healthy i.e. there are no false positives.

$$Specificity = \frac{TrueNegatives}{TrueNegatives + FalseNegatives} \tag{2.10}$$

### 4.6.2 Performance Metrics for Regression

The mean-squared error (MSE) corresponds to the expected value of the squared error. It is calculated as the following: If \( \hat{y}_n \) is the predicted value of the \( n \)-th sample, then the MSE over \( x \) samples is calculated as:

$$MSE(\hat{y}_n, y) = \frac{1}{x} \sum_{n=1}^{x-1} (y_n - \hat{y}_n)^2 \tag{2.11}$$
The $R^2$ score is the coefficient of determination. It indicates the proportion of variance predictable based on the input variables. This metric gives an estimation of the quality of fit and how likely the model is to predict on unseen test samples. The drawback of this metric is that it cannot be compared across datasets since variance is dataset-specific.

It represents the proportion of variance (of $y$) that has been explained by the independent variables in the model. It provides an indication of goodness of fit and therefore a measure of how well unseen samples are likely to be predicted by the model, through the proportion of explained variance. The best score is 1.0 which would imply the predicted values perfect fit to the ground truth data, whereas a score of 0.0 would imply that the model predicts the same value regardless of the independent variables. If $\hat{y}_n$ is the predicted value of the $n$-th sample, then the $R^2$ over $x$ samples is calculated as:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{n=1}^{x} (y_n - \hat{y}_n)^2}{\sum_{n=1}^{x} (y_n - \bar{y}_n)^2}$$

(2.12)

### 4.6.3 k-fold Cross-Validation

$k$-fold cross-validation was to reduce the probability of model overfitting to the training dataset. This cross-validation splits the training set into ten subsamples and uses nine of these samples as training and one for testing and validation. This cross-validation is repeated ten times until all of the subsamples have been used for validation exactly once. The results from all 10 training exercises are then averaged to produce the final estimations. It is useful because all observations are used for validation and also training.

### 4.6.4 Clustering Validity Indices

Evaluating the performance of clustering tasks is different from supervised classification and regression models. An evaluation metric for clustering algorithms needs to be able to either determine whether clusters are defined similarly to some ground truth set of classes or whether cluster members are similar enough to each other than to members of other clusters according to some similarity metric. In our research work, the clusters formed are totally unsupervised and there is no ground truth labels available for comparing against. So, the cluster evaluation metrics, also called cluster validity indices, [5] must be able to evaluate clustering results without ground truth labels. Three metrics have the capacity to perform this function: Silhouette, Davies-Bouldin and Calinski-Harabrasz.

The Silhouette index [171] is defined for each sample and is composed of two scores:

- **a**: Mean distance between a sample and all other points in the same cluster class.
- **b**: Mean distance between a sample and all other points in the next nearest cluster.

The Silhouette score $s$ is then calculated as:
\[ s = \frac{b - a}{\max(a, b)} \] (2.13)

This score is bound between +1 and -1, where a Silhouette score of -1 implies incorrect clustering, 0 for overlapping clusters and +1 would imply highly dense clusters. This index is higher when the clusters are dense and well-separated [171]. However, it would likely produce lower scores for density-based clustering methods such as DBSCAN as the clusters formed by these algorithms are not necessarily convex shaped.

The Davies-Bouldin index [50] is the second clustering validity index used in our work where there is no ground truth. It is defined as the average similarity between each cluster \( C_i \) for \( i = 1, \ldots, k \) and its most similar one \( C_j \). Similarity, \( R_{ij} \), is defined as a metric that trades off the following:

- \( s_i \), the average distance each sample in cluster \( i \) and the centroid of that cluster.
- \( d_{ij} \), the distance between cluster centroids \( i \) and \( j \) for Cluster \( C_i \) and \( C_j \).

To construct the similarity \( R_{ij} \),

\[ R_{ij} = \frac{s_i - s_j}{d_{ij}} \] (2.14)

The Davies-Bouldin index calculates the average similarity for all clusters and is therefore defined as for \( k \) clusters:

\[ DB = \frac{1}{k} \sum_{i=1}^{k} \max(R_{ij}) \] (2.15)

A Davies-Bouldin score closer to zero implies better clustering performance as it means smaller inter-cluster similarity, but larger intra-cluster similarity. Zero is the lowest possible score. This index, again, however would perform better for convex-shaped clusters than clusters that are not, such as those generated by density-based clustering algorithms as in the case with Silhouette.

The final clustering validity index used in this work is the Calinski-Harabasz score. This is an unbounded index and can be used to compare clustering results relatively, where a higher score implies better performance. It measures the ratio of the sum of inter-cluster dispersion and intra-cluster dispersion for all clusters. Dispersion is the sum of distances squared.

This index however also tends to, again, perform better for convex-shaped clusters than clusters generated by other clustering approaches.

In our work, we used all of three of these clustering validity indices so that we can evaluate clustering methods and different concepts of clustering (graph-based, density-based, centroid-based). Each index uses a different concept for measuring validity, so using all three gives a more comprehensive understanding of the clustering performance. If the results for these indices differ radically for the non-convex shaped clusters and those that are convex-shaped clusters, then other indices will need to be explored, since the chosen clusters tend to perform better for convex shaped cluster than non-convex shaped clusters.
5 Foundations of Deep Learning

This section details the deep learning material necessary for the research. Deep learning is a subfield of Machine Learning which broadly focuses on training and building artificial neural networks (ANNs). Deep learning has been applied to many different tasks across many fields from drug design, medical imaging analysis, automated board game playing, speech recognition and computer vision. This section will give an overview of the basics of deep learning, the foundations of how neural networks are trained and then optimised, and then goes on to detail the more complex neural network architectures (convolutional and recurrent neural networks) that were built and then trained in our research work. We primarily used deep neural network architectures, namely autoencoder architectures, to learn new latent embedded spaces of low-level signal traces extracted from raw accelerometer time series. Our research was about the viability and power of these latent spaces to learn models for predicting and studying Type-2 Diabetes disease states.

5.1 The Multilayer Perceptron

Artificial neural networks, ANNs, are composed of units, or neurons, that are linked to each other through connections with values called weights. The structure of one of these neurons is given in Figure 2.3

![Structure for the neuron unit of a neural network](image)

The input are the set of features fed into the model. The weights are the values for the connections that assign the importance to each feature’s contribution to the learning. The activation function introduces non-linearity in the neuron to enable the neural network to approximate nonlinear functions. There are many activation functions to choose from depending on the task at hand, for example the hyperbolic tangent, rectified linear unit, sigmoid and leaky rectified linear unit. The bias shifts the output of the activation function. When more than one neuron is stacked together into a row, it is called a layer and when there are multiple layers stacked next to each other, we have a multi-layer neural network. This process in illustrated in the vanilla neural network, known as a multilayer perceptron (MLP), in Figure 2.4 or a feedforward neural network.
5.2 Training and Optimising Neural Networks

5.2.1 Backpropagation

The goal of training an feedforward neural network (FNN) is to attain the optimal values for the set of weights in the neural network[173]. The problem with forward propagation is that it there may be errors between the predicted output and the expected value. This error is calculated by the cost function. 

FNNs are trained through the backpropagation algorithm. The goal of backpropagation therefore is to minimise this cost function. Backpropagation is used to efficiently calculate the gradients using the chain rule in calculus and then updating the weights with each iteration.

To summarise how the algorithm works: for each training example, the backpropagation algorithm first makes a prediction (the forward pass step) and measures the error between this prediction and the expected values (the cost function), it then goes through each layer in reverse order to measure how much of the error is contributed by each connection (this is the reverse pass) through the chain rule in calculus, and then finally tunes the connection weights to reduce the error (the gradient descent).

The weights in the neural network are randomly initialised to the break the symmetry of neurons so that backpropagation trains a diverse set of neuron units. It is the job of the optimisation algorithm, such as stochastic gradient descent, RMSProp, AdaGrad and Adam, to use these gradients to update the weights of the neural network.

5.2.2 Stochastic gradient descent (SGD)

Gradient descent is the way to optimise neural networks, making use of the gradients calculated by backpropagation. Stochastic gradient descent (SGD) is an iterative algorithm for the optimisation of a cost function. It takes a random subset of
instances in the dataset at every step and calculates the gradients for this random subset of instances only. SGD shortens the training time as it converges much quicker than gradient descent since only a subset of instances are being used, whereas gradient descent has to run through the whole dataset to do a single update for the weights. However, SGD can also lead to erratic gradients for the cost function due to this approach.

In gradient descent, the learning rate hyperparameter is the most important to tune. The learning rate represents the step size at each iteration as the algorithm tends towards a minimum for the loss function. Intuitively, it controls how often new information overrides old information. The learning rate should start out large (for quicker progress and avoiding a local minima) and then decrease gradually so that the algorithm can settle at a global minimum. To choose the appropriate value for this, standard practice is to conduct training with values for the learning rate at different scales (e.g. $10^{-1}$, $10^{-3}$, $10^{-5}$, ...). This value is set at the initialisation stage. Each full iteration of this process over the dataset (after all the batches) is called an epoch.

The parameter updating rule is given in Equation 2.16. Stochastic gradient descent (SGD) for example $x^{(i)}$ and its corresponding true label $y^{(i)}$. $J$ is the objective cost function to minimise, $\theta$ is the set of parameters at this step, and $\eta$ represents the learning rate.

$$\theta = \theta - \eta \cdot \nabla \theta J(\theta; x^{(i)}; y^{(i)})$$

(2.16)

However, SGD can also leads to erratic gradients for the cost function due to this approach. The algorithm may never converge on a global optimum, but will oscillate once it reaches a close approximation [76].

5.2.3 Optimising the Stochastic Gradient Descent Algorithm

SGD requires the setting of the learning rate at initiation and maintains this learning rate for all parameter updates during training, but SGD forms the foundations of many algorithms for optimising neural networks. Another class of stochastic optimisation algorithms have an adaptive learning rate. An adaptive learning rate is when the learning rate decays faster for steep dimensions than for dimensions with gentler slopes. An advantage of this approach is that it also requires less tuning of the learning rate hyperparameter. Other popular optimisation algorithms include: momentum optimisation, Nesterov accelerated gradient, AdaGrad, RMSProp, and Adam. Some of these algorithms will be presented in this section. Adam was the optimiser used in our research work since it is one of the latest developments in the deep learning state-of-the-art and has since become standard practice to use this optimiser by default. However, an understanding of the concepts from AdaGrad and RMSProp are needed to understand how Adam works.

Momentum optimisation, proposed by Polyak in 1964 [158], takes into the account the values for past gradients and accumulates an exponentially decaying moving average of these gradients, as if it was picking up momentum until it eventually reaches the global optimum. Regular gradient descent takes small, fixed steps down the slopes, so more time is needed for convergence on the global optimum. The momentum algorithm, unlike SGD, considers the previous gradients. It subtracts the local gradient away from the momentum vector $m$. Parameters, or
weights, are updated by adding them to the momentum vector. The algorithm is shown in Equations 2.17 and 2.18 below.

\[
m \leftarrow \beta m - \eta \nabla_\theta J(\theta) \tag{2.17}
\]

\[
\theta \leftarrow \theta + m \tag{2.18}
\]

\(\beta\) is another hyperparameter called the momentum. This is set between 0 and 1, where 1 implies no friction. “Friction” prevents the momentum from growing too large. As in SGD, \(\eta\) is the learning rate. Friction also helps to prevent oscillation where the optimiser overshoots the global optimum and speeds up the rate of convergence. Usually, momentum is set to 0.9 by default and is now standard practice[76]. Nesterov introduced a variant of momentum in 1983 called the Nesterov accelerated gradient (NAG) [81] which measures the gradient of \(J\) slightly ahead of \(\theta\) at position \(\theta + \beta m\). The momentum vector usually points in the right direction, so it will be more accurate to use the gradient slightly ahead than the original \(\theta\). NAG is faster than vanilla momentum optimisation.

AdaGrad, Adaptive Gradient, [60], adapts the learning rate by rescaling them inversely proportional to the square root of the sum of all history squared values of the gradient. The aim of the algorithm is to ensure that it can correct its direction earlier to point towards the global optimum[76]. The first step is to accumulate the square of the gradients into the vector \(s\). The second step is then to scale down the gradient by a factor of \(\sqrt{s + \epsilon}\) where \(\epsilon\) is the smoothing term to avoid divisions by zero (usually a small number). This is all shown in Equation 2.19 and 2.20 below. In this equation, \(\odot\) is element-wise division, \(J\) is the cost function, \(\theta\) is the set of parameters and \(\oplus\) is the element-wise addition.

\[
s \leftarrow s + \nabla_\theta \odot \nabla_\theta J(\theta) \tag{2.19}
\]

\[
\theta \leftarrow \theta - \eta \nabla_\theta J(\theta) \odot \sqrt{s + \epsilon} \tag{2.20}
\]

AdaGrad decays the learning the rate, but it does so faster where dimensions are steeper than for the dimensions which flatter. However, the issue with this algorithm is that it may stop too early as the learning rate gets scaled down so much that the optimisation stops entirely before the global optimum is reached.

RMSProp fixes the problem of halting training too early by accumulating only gradients from the most recent iterations, whereas AdaGrad accumulated all gradients from the beginning of training) [49]. It uses the exponentially decaying average to discard history [81]. This is achieved by exponentially decaying the learning rate in the first step, as seen in Equation 2.21 and 2.22 below:

\[
s \leftarrow \beta s + (1 - \beta) \nabla_\theta \odot \nabla_\theta J(\theta) \tag{2.21}
\]

\[
\theta \leftarrow \theta - \eta \nabla_\theta J(\theta) \odot \sqrt{s + \epsilon} \tag{2.22}
\]

The decay rate, \(\beta\) is usually set to 0.9 and works quite well. RMSProp usually outperforms AdaGrad and so is preferred[76] However, the Adam optimisation algorithm has since become the new preferred optimisation algorithm.
The adaptive moment estimation, Adam, \cite{kingma2014adam} optimisation algorithm for updating the network parameters was used in our research. This algorithm is one of the latest developments in stochastic optimisation and it combines the best ideas from momentum SGD and RMSProp (which further builds on AdaGrad), hence why it is important to understand these algorithms first\cite{dean2012large}.

The Adam algorithm monitors the exponential decay of both the average of past gradients and past squared gradients. The optimiser algorithm steps are shown below:

1. \( \mathbf{m} \leftarrow \beta_1 \mathbf{m} - (1 - \beta_1) \nabla \theta J(\theta) \)
2. \( \mathbf{s} \leftarrow \beta_2 \mathbf{s} + (1 - \beta_2) \nabla \theta J(\theta) \otimes \nabla \theta J(\theta) \)
3. \( \hat{\mathbf{m}} \leftarrow \frac{\mathbf{m}}{1 - \beta_1^t} \)
4. \( \hat{s} \leftarrow \frac{s}{1 - \beta_2^t} \)
5. \( \theta \leftarrow \theta + \eta \hat{\mathbf{m}} \otimes \sqrt{\hat{s} + \epsilon} \)

In the algorithm above, the \( t \) represents the iteration number (which starts at 1). \( \beta_1 \) and \( \beta_2 \) are the exponential decay rates for the momentum and the scaling respectively. These values are usually set as \( \beta_1 = 0.9 \) and \( \beta_2 = 0.999 \). The smoothing term \( \epsilon \) is usually set to a very small number.

### 5.2.4 Vanishing and Exploding Gradients

In backpropagation and gradient descent techniques for optimisation, each of the neural network’s weights receives an update proportional to the partial derivative of the error function with respect to the current weight. It is possible that the gradient becomes smaller and smaller as the training process goes on. This could eventually cause the gradient to become so small that it basically vanishes and stops the weight from changing its value \cite{glorot2010understanding}. Sometimes, the activation function being used could actually cause the gradient to become larger and larger which can result in the opposite problem: An exploding gradient.

Some light was shed on what could cause these unstable gradients in Glorot and Bengio’s 2010 paper \cite{glorot2010understanding}. Unstable gradients can be affected by choice of activation functions and weight initialisation. They determined that the choice of this can lead to the variance of the outputs of each layer being larger than the variance of its inputs which further increases after each layer until the activation function is saturated \cite{glorot2010understanding}. For example, the sigmoid activation function saturates at 0 or 1 when the inputs are large, and so would give no gradient to backpropagate for the lower layers as the gradients only become more and more diluted.

There are techniques that are used to reduce the possibility of these problems manifesting. In \cite{glorot2010understanding}, the authors propose methods to alleviate these unstable gradients. The notion behind these strategies is that the flow of signal should not die out or explode and saturate the activation function when gradients are being backpropagated. To variance of the outputs of each layer should be equal to the variance of its inputs to ensure that the signal flows properly. It is now standard
practice to use the Glorot initialisation strategy when using certain activation functions, including the hyperbolic tangent and sigmoid. This strategy was deployed in our research work.

To handle the exploding gradient problem, where gradients become exponentially larger with each step, we can use a technique called gradient clipping\textsuperscript{152}. This is where gradients are forbidden from exceeding a given threshold (i.e. gradients are clipped). This threshold is a hyperparameter that can be used in the optimiser. The optimiser clips all the components of the gradient vector to a value between the values set so that all partial derivatives of the loss function (with regard to each and every parameter) \textsuperscript{76} is clipped between these values. This approach can however change the direction of the gradient vector, for example if the original gradient vector was $[0.9, 100.0]$, the gradient clearly points mostly in the second component axis, but if the gradient was clipped by threshold 1.0, then you get $[0.9, 1.0]$. This can be corrected by clipping the gradients if the L2 norm exceeds the set threshold for clipping\textsuperscript{76} which would clip the gradients to $[0.0090,0.99]$, preserving its orientation but eliminating the first component and keeping the directions the same as the original gradient vector.

5.2.5 Combatting Overfitting in Neural Networks

Neural networks often have a huge number of parameters which means they have the capacity to fit a huge variety of datasets with differing complexity. But this also means that neural networks can overfit onto a dataset. So, regularisation is needed to reduce the likelihood of this happening. Regularisation includes methods that force the learning algorithm to build a less complex model that is less likely to overfit on a dataset.

Two of the most widely types of regularisation are L1 and L2 regularisation. L1 regularization is used for sparsity (feature vectors consisting mostly of zeros). This is useful for high-dimensional feature spaces. Neural networks can also apply both L1 and L2 regularisation at each during training \textsuperscript{76}. L1 regularization introduces a penalty to the sum of absolute values of the weights, whereas L2 regularization introduces penalty to the sum of squares of the weights.

Another popular technique to regularise neural networks is to use dropouts\textsuperscript{94, 188}. This technique has proven to be highly successful in improving neural network accuracy. The dropout approach is simple: At every training step, every neuron (including the input but excluding the output neurons) is assigned a probability $p$ of being dropped out, meaning the neuron will be totally ignored during this training step, but the same neuron could be reactivated in the next step. $p$ is called the dropout rate and is usually set between 30% to 40% depending on the neural network being trained.

Dropouts reduce regularisation because it forces neurons to become independent and useful on their own since neurons now cannot become reliant on each other and co-adapt with neighbouring neurons\textsuperscript{76}. This results in a neural network becoming less sensitive to changes in the inputs and produces a neural network that is less prone to overfitting and better generalisation.

Another way in which the neural network was regularised was through the use of early stopping. This technique halts the training process at the epoch where the model stops improving on a hold out validation dataset. If the neural network is
trained on too many epochs, not only can it be time-consuming on large datasets but also the neural network may overfit.

5.3 Recurrent Neural Networks and Long Short Term Memory

Recurrent neural networks (RNNs) are a type of neural network for processing sequential data such as time series. A sequence is a matrix, where each row is a feature vector and the order of rows matters. RNNs are different from CNNs and feed-forward NNs in one fundamental respect: They are not feed-forward. They have connections that point backward. In the most basic RNN, shown in Figure 2.5, a single recurrent neuron unit receives inputs, produces an output and sends that output back to itself. At each time step \( t \) (a frame), this recurrent neuron receives the inputs \( x_t \), as well as the output from the previous time step, \( x_{(t-1)} \). The information is the hidden state, which represents the previous inputs. These recurrent neurons can be used to form the layers in a RNN.

![Figure 2.5: A recurrent neuron (LHS) unrolled through time (RHS)](image)

The outputs of a layer of recurrent neurons for all instances in a mini-batch is given in

\[
Y(t) = \phi(X(t) W_x + Y_{(t-1)} W_y + b) \tag{2.23}
\]

In equation 2.23, all the inputs at time step \( t \) are placed in matrix \( X_{(t)} \). \( \phi \) is the activation function. The hyperbolic tangent activation function (tanh) tends to be the preferred choice of activation for RNNs. In a RNN, the weights are unchanged and stay the same at each time step while the outputs of gradient descent continue to increase slightly at each time step. As a result, the gradients may explode. A non-saturating activation function like ReLU would not prevent this, so a saturating activation function like tanh is preferred.

\( b \) is the bias vector of size equal to the number of neurons. \( Y_{(t)} \) in a \( m \times n_{\text{neurons}} \) matrix representing the layer’s outputs for each example in the mini-batch (of size \( m \)) at time step \( t \). \( X_{(t)} \) is a matrix of \( m \times n_{\text{inputs}} \) that represents the layer’s inputs for all examples (with \( n_{\text{inputs}} \) number of input features). \( W_x \) (size \( n_{\text{inputs}} \times n_{\text{neurons}} \)) and \( W_y \) (size \( n_{\text{neurons}} \times n_{\text{neurons}} \)) contain the connection weights for inputs at time step \( t \) and the connection weights for outputs at time step \( t - 1 \), respectively. The equation shows that \( Y_{(t)} \) is a function of \( X_{(t)} \) and \( Y_{(t-1)} \), which itself is a function of \( X_{(t-1)} \) and...
\(Y_{t-2}\). This indicates that \(X_t\) is a function of all the inputs since \(t=0\). When \(t=0\), there are no previous inputs so they are assumed to be all zeros.

### 5.3.1 Backpropagation through time (BPTT)

RNNs are capable of capturing the patterns of a time series. But RNNs have a short-term memory and can only learn the hidden states of a time series for a few time steps due to the instability in gradients caused by the vanishing gradient problem or the exploding gradient problem, both described in 5.2.4. The vanishing gradient problem in RNNs is due to the nature of backpropagation through time, the version of backpropagation used to train RNNs. The gradient values exponentially shrink with every time step and so the RNN may not learn the long-range dependence and dynamics across time steps, especially the earliest time steps. Not being able to learn from earlier time steps causes the network to have a short-term memory. This is the main drawback of a recurrent neural network.

![Figure 2.6: Illustration of backpropagation through time. Dashed arrow represents forward pass. Black full arrow represents the propagation backwards of cost function gradients.](image)

As mentioned, a RNN is mostly trained through backpropagation through time (BPTT) as it faster than other techniques. The idea is to unroll the RNN through time and then use backpropagation as you would for FFNNs. BPTT begins making a first forward pass through the unrolled RNN. Then the output sequence is evaluated using a cost function. The gradients of the cost function are then propagated backwards through the unrolled RNN. The model parameters are updates using the calculated gradients.

### 5.3.2 Long Short-Term Memory (LSTM)

The instability in gradients for RNNs can be reduced through choosing different optimisers, weight initialisation, dropouts and using different activation functions. Some information is lost at each time step when a RNN is being trained with
BPTT. As mentioned, this causes the short-term memory problem in RNNs. To solve this, various cells have been proposed with long-term memory capacities. The most popular of these is the Long Short-Term Memory cell (LSTM), proposed by Sepp Hochreiter and Jurgen Schmidhuber in 1997 [97]. LSTM recurrent neural networks were used in our research because of their ability to handle sequential data and use in accelerometer data analysis in human activity recognition tasks [87, 83]. These neural networks form part of the autoencoder architecture in Chapter 5 to learn an embedded latent representation of the accelerometer traces that were used in predictive modelling and clustering analyses.

LSTM cells can help with converging faster during training and detecting the long-term dependencies in sequential data. If viewed as a black box, then there are two states in a LSTM cell: \( h(t) \) and \( c(t) \) where \( h(t) \) can be thought of as the short-term state, like in the recurrent neuron in 2.5, and \( c(t) \) as the long-term state.

The structure of a LSTM cell is shown in Figure 2.7. As \( c_{t-1} \) passes forward in the network from left to right, it goes through the forget gate, where some of the memories from previous time steps are “forgotten”. It then adds some new memories through an addition operation, represented by \( A \) in Figure 2.7. \( A \) adds new memories selected by the input gate in the cell. The output of this cell at time step \( t \), \( C(t) \), is then transmitted out of the cell. In this process, at each time step, some memories are forgotten and some new ones are added. The long-term state is copied after the addition operation at \( A \) and is then fed into the tanh activation function where the result is then passed to the output gate. The output gate produces the short-term state \( h_t \), equal to the cell’s output for this time step (\( y(t) \)).

The input vector \( x(t) \) and the short-term state from \( t-1 \) (\( h(t-1) \)) are fed into the four fully connected layers.

Layer \( g \) is the main layer. It analyses the current input \( x(t) \) and the short-term state \( h_{t-1} \). The LSTM cell selects the most important temporal dependencies and dynamics in the long-term state and throws away the rest from this layer’s output.
Layers $f$, $i$ and $o$ are the gate controllers that use the sigmoid activation function, denoted as $\sigma$, with outputs in the range 0 to 1, and then their outputs are passed to the element-wise multiplication operations, as seen in Figure [2.7]. If the outputs are 0s, then the gate is closed and if the outputs are 1, then the gate opens.

The operation of the three gates are as follows:

- The forget gate, controlled by layer $f_t$, determines which parts of the long-term state is forgotten (fed in at $C_{(t-1)}$).
- The input gate, controlled by layer $i_t$, determines which parts of $g_t$ should be added to the long-term state after the $\tanh$ activation function has been applied to the long-term state $c(t)$.
- The output gate, controlled by layer $o(t)$, determines which of the dependencies in the long-term state should be read and output to both $h_t$ and $y_t$.

The LSTM cell makes six computations to attain the long-term state ($c_t$, its short-term state $h(t)$, and the outputs for each of the hidden layers at each time step $t$ as well as its final output $y(t)$ for a single instance. Each of the computations are shown in the Equations 2.24 to 2.29 below:

\[
i(t) = \sigma(W_{xi}x(t) + W_{hi}h_{(t-1)} + b_i) \quad (2.24) \\
f(t) = \sigma(W_{xf}x(t) + W_{hf}h_{(t-1)} + b_f) \quad (2.25) \\
o(t) = \sigma(W_{xo}x(t) + W_{ho}h_{(t-1)} + b_o) \quad (2.26) \\
g(t) = \tanh(W_{xg}x(t) + W_{hg}h_{(t-1)} + b_g) \quad (2.27) \\
c(t) = f(t) \odot c(t-1) + i(t) \odot g(t) \quad (2.28) \\
y(t) = h(t) = o(t) \odot \tanh(c(t)) \quad (2.29)
\]

In the equations above, $W_{xi}$, $W_{xf}$, $W_{xo}$ and $W_{xg}$ are the weight matrices for the connections between the four hidden layers of the LSTM cell and the input $x(t)$. $W_{hi}$, $W_{hf}$, $W_{ho}$ and $W_{hg}$ are the weight matrices for the connections between the four hidden layers of the LSTM cell and the previous short-term state $h_{(t-1)}$. $b_i$, $b_f$, $b_o$, and $b_g$ are the bias terms for each of the four hidden layers.

We can see from all this that the LSTM cell is a powerful tool that has the ability to learn important temporal dependencies and dynamics over the long-term, preserve these dependencies and then select the appropriate dependencies and add to the developing long-term state on an ad hoc basis. LSTM cells are one of the primary constituent parts in constructing the recurrent autoencoder used in our research. We built the autoencoder using LSTM cells due to its ability to store long-term dependencies in a long sequential time series.

### 5.4 Autoencoders

The autoencoder architecture [95] was the main deep neural network architecture used in our research. These architectures make use of convolutional and recurrent neural network architectures. They are a form of FFNN designed to reconstruct the data passed as the input- so they ‘learn the data’. Autoencoders have three
The encoder function, $h = f(x)$, codes the input data and compresses this into a latent representation after the input data has passed through all the hidden layers and is fed to the bottleneck layer. The decoder function, $g$, attempts to make a reconstruction, $r = g(h)$ of the input from the latent representation. The bottleneck layer lies between the encoder and decoder functions and it contains the embedding of the $D$-dimensional original input data. Usually, this embedded representation (or latent representation) has less units than $D$. However, an autoencoder is not useful if it simply acts as an identity function for the training data, so if $g(f(x)) = x$. Ideally, the autoencoder should be able to reconstruct the original input only approximately as it should force the model to prioritise inputs that resemble the original data\[81\]. Unlike most other machine learning applications, autoencoders are trained with recirculation \[93\].

There are a variety of autoencoder architectures: Sparse, denoising, variational, recurrent and convolutional. These architectures were chosen due to their wide use in learning representations of time series data and in Human Activity Recognition tasks\[82, 146, 216\].

Denoising autoencoders \[210\] receive a corrupted input with perturbation added. The encoder attempts to address risks of the autoencoder becoming an identity-function by denoising the corrupted input data. It does this by attempting to reconstruct the original input from the corrupted input. Perturbation, or noise, is added and expressed as a percentage of the input. Usually, Gaussian noise is added to each feature. For each feature $j$ of the input feature vector $x$ the noise value $n^{(j)}$ is sample from the Gaussian distribution:

$$n^{(j)} = \mathcal{N}(\mu, \sigma^2)$$  \hspace{1cm} (2.30)
Autoencoders minimise a loss function $L$ which penalises the reconstruction process $g(f(x))$ for being dissimilar from the original input $x$. Therefore, autoencoders minimise $L(x, g(f(x)))$. However, denoising autoencoders, instead minimise $L(x, g(f(\bar{x})))$ where $\bar{x}$ is the corrupted input of $x$.

6 Human Activity Recognition

Human Activity Recognition, HAR, is a broad field of study and a challenging task of classifying time series into interpretable and meaningful classes that correspond to a form of human movement, or activity- for example walking or sleeping. The aim is to use predictive models to label an individual based on sensor data, usually using domain expertise and methods from signal processing. Both convolutional neural network and recurrent neural network architectures have demonstrated the ability to learn representations from raw accelerometry data to achieve highly accurate results.

HAR techniques were used extensively in our research. New feature spaces were built to represent accelerometer data which were then used to learn predictive models for Type-2 Diabetes related outcomes and for characterising individual’s digital phenotypes. This section therefore provides some background and explores the state-of-the-art in the applications of machine learning for HAR tasks.

6.1 Overview of Human Activity Recognition

Sensor-based data, such as the low-level readings from a person’s wear-time accelerometer trace, can be classified into high-level labelled movements representing standard activities such as walking, moderate-to-vigorous activity such as running, sedentary behaviour such as sitting and sleep\cite{215}. The data generated by digital sensors has grown massively due to the advent of smartphones and other personal self-monitoring devices. These are often relatively affordable and have grown in popularity. Enabling a P4 vision for the future of medicine must clearly exploit this technology. Several challenges in this field of study exist\cite{48,151}. This section discusses this field of applied machine learning in healthcare and how it ties into the research work in this thesis.

There are three main phases in the human activity recognition workflow: Training data collection, feature extraction, and learning/inference\cite{123}. The training phase requires a time series to be derived from a raw data collected by the sensor device. The time series must contain the measured features that represent the performance of different activities. The time series is split into time windows and then the feature extraction steps are applied to each time window. This involves filtering the relevant information from the raw signal (for example removing noise) and normalising this filtered signal before the time series is broken up (segmentation) into fixed-length epochs. Then the classifier models are learned to generate a model that can classify using the features extracted. The feature extraction process can include basic statistics about the filtered signal, time domain and frequency domain features such as the mean, standard deviation, signal vector magnitude, pitch, roll, raw, and fast Fourier transform coefficients\cite{213,229,64}. This would be an exercise in applying classical machine learning methods in HAR models and is broadly followed by most machine learning based methods for HAR.
tasks.

The process followed to build HAR systems is a relevant component in our research work because we use this system to derive labelled time series sequences for feature extraction in our own analyses and we also use the low-level signal features for representation learning with autoencoder architectures. HAR systems incorporate many of steps necessary in our analyses and our research builds on the works in HAR applied in the clinical context.

6.2 Preprocessing Accelerometry Data for HAR tasks

As described in the 6.1 above, the raw sensor data needs to be preprocessed before it is passed to the classifier model. This includes, in this order, the filtering of the signal, segmentation of the time series into fixed-length frames, and then normalising this signal. Filtering is the first preprocessing step and seeks to remove noise and unwanted features of the signal. This process usually works in the frequency domain and removes certain frequencies of data.

Calibration is often the first preprocessing steps undertaken, such as in [56, 205]. The signals are usually calibrated to local gravity. This is an important step to ensure that devices produce outputs similar to other device types given similar conditions in the local environment [15, 130].

One of the most important HAR steps is filtering the raw signal to remove noise and unwanted components [48]. There are four main types of filters for signal processing: low-pass, high-pass, notch and band-pass filters. Although there are different one of the most popular filtering techniques is the Butterworth filter [24]. The range of frequency values passed by the filter is the passband and the band it rejects is the stopband. In a low-pass filter, the low frequencies are set in the passband and thus passed, and higher frequencies are set in the stopband and thus attenuated. A Butterworth filter mathematically attempts to ensure that the frequency response is as flat as possible in the passband. The Butterworth filter was also the filter used in the works in the literature that our research work builds upon [56], so is the relevant filter in our research. Other filters also exist including median filter, discrete wavelet package shrinkage, and Kalman filter are used to filter out noise in accelerometer signals [48, 217, 163].

The next component of the preprocessing steps is the segmentation of the filtered time series into fixed-length windows, depending on the sampling rate of the device. This enables the classification model to assign each window, or frame, with a label for its activity class (for example walking or moderate-to-vigorous activity). Window size may have an impact on HAR segmentation, but there is no clear consensus on what is a preferable window size [9]. Smaller window sizes could enable faster activity detection, but larger window sizes may be needed for more complex activities to be recognised. Ideally, each frame should have only one activity class label for that frame of activity in the time series. Some HAR systems do use smaller frame sizes, such as 3 seconds in [230] and 1 second in [22]. The choice depends on a variety of factors, such as the sampling rate, and could be optimised through a grid search.
6.3 Feature Engineering for HAR tasks

As with most classical machine learning tasks, feature engineering is one of the main challenges. Features that are crafted by analysts and scientists require, possibly extensively, the knowledge of the domain concerned.

For HAR tasks, feature extraction can be based on the time-domain or the frequency-domain [230], terms introduced in the 1950s [126]. The time domain features measure the properties that can be calculated from time windows directly. Traditional time-domain features for HAR tasks include the Signal Vector Magnitude, mean, correlation, standard deviation, covariance, roll, pitch, yaw, the Signal Magnitude Area (SMA) and auto regression coefficients [201, 63].

Frequency domain features are concerned with the frequency of the signals within a range of frequencies. A transformation is necessary in the frequency to determine derive this set of features, such as Fast Fourier Transforms, Discrete Wavelet Transforms and Discrete Cosine Transforms [91]. These kinds of transforms have been used widely in HAR systems [230, 209].

Time-domain and frequency-domain features include the Signal Vector Magnitude [90], its mean, standard deviation, coefficient of variation, median, minimum, maximum, 25th & 75th percentiles, mean amplitude deviation, mean power deviation, kurtosis and skewness, and Fast Fourier Transform (FFT) 1–15Hz [222].

These sets of features have been extracted by the tools that our research work has used, based on the literature, are similar to the features extracted by most HAR systems. Extracting features in a domain-driven way may fail to capture unforeseen properties of the data that could improve performance but was not considered by the researcher. Deep learning based methods for HAR can eliminate the need for feature extraction altogether and only requires the normalised data [35] and is another paradigm for HAR tasks.

HAR systems may then apply dimensionality reduction feature selection techniques to decrease the number of redundant features to prevent the problem of overfitting and the curse of dimensionality issue frequently encountered in feature engineering. Techniques for this deployed in HAR tasks include principal component analysis [223] such as in Balli et al [8] and in Hassan et al’s works [90] where Kernel PCA was deployed. Autoencoders are also good at learning features for HAR tasks in a compressed way and can be used for dimensionality reduction [137, 31].

Some other HAR systems incorporate features extracted from the signals of multiple sensors, for example on different locations, and fused these together [209]. These systems perform very well (over 90% accuracy, recall, precision and F-1) but the UK Biobank and DIRECT datasets, described in Section [8] did not collect the multi-sensor data that would enable this.

6.4 Classification Methods for HAR tasks

6.4.1 HAR Classification with Traditional Supervised Machine Learning

The final step of the training process is the classification part of the model. Both deep learning and classical machine learning approaches have been used to achieve this. Standard classical machine learning approaches use the hand-crafted features
extracted from the raw signal for classification of windows into labelled activity classes. Common supervised classifiers are usually the first port-of-call, including the SVM, random forests, k-nearest neighbours. The ground truth labels are usually either self-reported by the subject, or labelled from camera recordings, as in [222].

The work by Willets et al [222] and in [57] use an approach with balanced random forests and Hidden Markov Models (HMMs) [62]. This is a popular methodology also used in other HAR systems for example by Kabir et al [107].

Although balanced random forests are good at handling non-temporal data points, they are not able to handle sequential time series data. A HMM encodes the temporal structure of the sequence of classes so that it can be used for classification with the random forest model.

A HMM performs time smoothing. It is a state space model consisting of a sequence of hidden discrete states. A stochastic sequence of states \( z = \{z_1, z_2, ..., z_t-1, z_t, z_{t+1}, ... \} \) has the Markov property if, for any given time, the conditional distribution of future states of the stochastic process given present and past states depends only on the present state and not at all on the past states. This is the memoryless property:

\[
p(z_t | z_1, z_2, ..., z_{t-1}) = p(z_t | z_{t-1})
\]  

(2.31)

For each time-step \( t \) in Equation 2.31, the hidden states \( z_t \) is assigned to one of \( n \) classes \( \{c_1, c_2, ..., c_n\} \).

The dynamics of transitions between class \( i \) and class \( j \) can be described by the transition matrix in Equation 2.32 below:

\[
T_{ij} = p(z_t = c_j | z_{t-1} = c_i)
\]  

(2.32)

We do not observe the hidden states at each time step. Instead, there is an observed stochastic emission \( y_t \) at each time step that is observed. These are drawn from a probability distribution \( p(y_t | z_t, \phi) \), where \( \phi \) represents the parameters for the probability distribution [222]. Emission probabilities are calculated from the votes of the trees in the random forest model. In random forests, each tree is trained on a subset of the training data and so by iterating through each tree, the training data subset that was not used for training will give an estimate of the error of the forest, thus directly giving the probability of correctly predicting the class given the ground truth label. In Willets et al’s work on human activity recognition from accelerometer data [222], the Viterbi algorithm [70] is then used to find the most likely sequence of states given the sequence of observed emissions from the random forest. This is a dynamic programming algorithm for obtaining the maximum a posteriori probability estimate of the most likely sequence of hidden states.

6.4.2 HAR Classification with Deep Learning

Traditional machine learning methods such as random forests and SVMs are powerful tools with high accuracy, but are dependent on the selection and extraction of features. Features are hand-crafted from domain knowledge, such as in Oxford accelerometer analysis tool and GGIR in Chapter 3. Deep learning techniques can eliminate this dependence by automatically learning and extracting features.
Deep learning has been used extensively in HAR tasks with a variety of DNN architectures [166, 157].

RNNs have been used to carry out HAR tasks in examples such as [87] where RNNs with LSTM units have been used, and also in [83] where strong LSTM learners are ensembled together to improve performance. CNN architectures have also been used in HAR tasks, such as [147] which uses a denoising convolutional autoencoder for feature learning and also in [224] that used a LSTM-CNN combination where the raw signals are fed into two-layer LSTM followed by convolutional layers. The authors report extremely high performance for recall, outperforming other approaches including CNNs and DeepConvLSTM architectures. They tested these recognition models on several benchmark datasets that are normally used in HAR tasks for training and testing.

There is a clear foundation in adopting deep learning for HAR tasks from across many works in the literature, primarily due to its advantages over traditional machine learning. Our research work incorporates both classical machine learning and deep learning techniques from the outputs of a state-of-the-art HAR system and then uses that to build new representations of physical activity for digital phenotyping individuals for Type-2 Diabetes.

6.5 Oxford Accelerometer Analysis Tool

Our research made significant use of a toolkit for analysing large sets of accelerometer data: the Biobank Accelerometer Analysis tool developed at Oxford University [56]. This software tool can be used to construct representations of accelerometer data that are useful where activity classification is needed to answer clinical questions. It also uses some of the procedures from GGIR, as described in [3] to calibrate the signals to local gravity [205].

The raw accelerometer traces undergo a standard process to minimise error and bias to calibrate, resample, and summarise the accelerometer traces [56, 212, 57, 222]. It then produces a series of outputs, including a signal-level feature set at the epoch level which is then used to undergo a activity classification process that outputs a one-hot encoded time series where each 30-second epoch is labelled as one of several activity types, namely Moderate-to-Vigorous Physical Activity (MVPA), Sedentary, Sleep and Light Tasks. These activities are indicated by different values for Metabolic Equivalent Task, a measurement for energy expenditure. METs are a ratio of a person’s working metabolic rate relative to that person’s resting metabolic rate. A person’s metabolic rate is the rate of energy that person expends per unit of time [159]. To put it in perspective, a brisk walk at 3 or 4 miles per hour has a value of around 5 METs. Jumping rope, which is a more vigorous activity, has a MET value of around 12 [135].

An example of the activity classification output is provided in the Figure 2.9 below, taken from the Oxford accelerometer analysis toolkit manual [6].

Figure 2.9: Example of Oxford human activity classification output plotted along daily timeline.

This tool uses some of the most standard and state-of-the-art techniques in the literature for the analysis of accelerometer data and human activity recognition tasks and was used extensively in our research.

7 Methods from Neuroscience and Gait Analysis

Gait analysis is the study of human locomotion by analysing the patterns of limb movements during locomotion, usually walking. A reduction and deterioration in the quality of movement, while walking, is closely associated with a multitude of diseases and health events related to ageing, such as Parkinson’s disease and the risk of falls[20,168]. They have been widely used as a biomarker of movement-related diseases.

Existing works have studied physical activity patterns exhibited by individuals for these diseases in a more granular way. Works that demonstrates this include other studies of Parkinson’s disease and also studies for dementia [55, 89]. These studies make an observation on or represent physical activity on a bout-scale, meaning that they study the distribution, variability, duration and intensities of bouts of activity, such as in [55,168].

Although techniques of analysing accelerometer data, including gait analysis, are useful in these diseases, they are limited in that they have only been applied to diseases known to cause abnormal exhibitions of physical activity. In a subtle and multi-faceted chronic such as Type-2 Diabetes, there are challenges that remain. There is a known association between these movement-related chronic diseases such as Parkinson’s disease and dementia [36,191] and the development or risk of Type-2 Diabetes. Given this underlying relationship, it justifies the notion of adopting some
of these bout-based analyses of physical activity in the context of Type-2 Diabetes. Furthermore, this would also tie up with existing works in HAR classification tasks, discussed in section [6] as it enables us to analyse the distribution, frequency and variability of different activity types performed based on the labelled time series sequences. It also develops on GGIR which was much more domain-driven and built based on bespoke requirements.

The works proposed in the literature search helped to guide the formation of the first of our own contributions. We adopted some of the techniques discussed in this section in the context of Type-2 Diabetes due to the established link between physical activity and Type-2 Diabetes. Emerging scientific and clinical knowledge is beginning to appreciate the value of this level of bout-based analysis for Type-2 Diabetes and further justifies this approach for our first set of experimental analyses[29, 162]. Our research inspired by these methods are discussed in Chapter 4.

8 Datasets

We worked mainly with two sources of accelerometer data and Type-2 Diabetes cohorts in our research: The UK Biobank and the IMI DIRECT datasets. This section provides an overview of these two datasets and which resources were used in our research.

8.1 The UK Biobank

The UK Biobank is a large cohort biobank study based in the UK. A biobank is a repository of biological samples for use in research. Over 500,000 individuals aged between 40 and 69 at the time of baseline assessment were selected[192]. It is unique in its size and scope and is one of the most well-studied resources of this kind. For each participant, there is a large variety of measurements taken to characterise a person’s phenotype, their genotype [25], and other health-related information, such as physical measurements including BMI and age, indicators for lifestyle and behaviour such as smoking habit and physical activity levels, biochemistry and medical imaging results. This resources available in the UK Biobank is summarised in the illustration shown in Figure 2.10. Further to this, there is a follow-up through linkage with Electronic Health Records linked to hospital records. The UK Biobank is also linked with GP primary care records[47]. This is a useful tool in longitudinal studies [227].
There is a wide body of diverse research works using the resources of the UK Biobank. While conducting this literature search, it was discovered that a lot of examples of these research works only went as far as prospective and cross-sectional exploratory analyses that described the association between different variables of the UK Biobank, in particular genomics.

This was especially the case in the study of the UK Biobank’s physical activity data analyses. The other main body of work was in the classification task of human activity recognition, already described in Section [6].

We used the UK Biobank’s physical activity dataset, which consists of over 100,000 volunteer subjects but this was significantly reduced in our study due
to quality controls and selection criteria adopted. The UK Biobank’s massive repository of other data, including cardiometabolic variables, sociodemographic and anthropometric features were also used in our analyses. The UK Biobank was used mostly in Chapters 3 and 4. The remaining chapters primarily used the DIRECT dataset, partially due to the limitations we discovered with the UK Biobank.

### 8.1.1 Physical Activity Monitoring in the UK Biobank

Physical activity was measured at 100Hz with a median wear-time of 6.9 days with a wrist-worn accelerometer device. Physical activity is one of the main lifestyle behaviours that are known to contribute the improvement of quality of life and increasing longevity. Studies with the UK Biobank have shown the association between physical activity and multimorbidity, a state where an individual suffers from two or more long-term chronic diseases. However, exploration of this subtle relationship has been limited to cross-sectional stratification exercises or risk analyses for different diseases. Techniques from machine learning have been deployed before with the UK Biobank data. In the last few years, there has been a high volume of research and analyses conducted for COVID-19 given the urgency of public health needs.

Both objective and subjective measurements were used to collect data to characterise individuals. This includes the use of questionnaires for measuring individual physical activity patterns, similar to other examples outside the UK Biobank. This is a subjective measure which is prone to human error or confirmation biases, for example, so objective measures such as through the accelerometer device tracings would be more useful in assessing actual patterns of behaviour. This was also included in our UK Biobank work since it is a useful complementary set of features and but can also serve as a validation on both fronts: Whether the objective accelerometer measurements correspond to subjective reporting and vice versa.

Physical activity is monitored in the UK Biobank cohorts with the Axivity AX3 accelerometer device. This device samples 25 to 400Hz, although it is designed to sample for at least 2 weeks on a full battery charge at 100 Hz in the full power mode. A trade-off exists in balancing sampling rate and battery life. Efforts have been undertaken to validate the reliability of this device. In Godinho et al’s work, compared the AX3 against other state-of-the-art devices and came to the conclusion that this device be deemed as “recommended”. Clarke et al attempted to validate the use of the AX3 on older functionally adults by validating it with the RT3 device and also reached the conclusion that this device is acceptably valid.

### 8.2 IMI DIRECT

The second dataset used in our research work was from DIRECT. Type-2 Diabetes was identified by the World Health Organization as a priority disease and was then selected as one of the diseases to be researched under the Innovative Medicines Initiative (IMI). Similar to the UK Biobank, it is also a repository containing a large variety of diverse data from individual volunteers. They collected and conducted analyses on clinical, molecular, biochemical and MRI data as well as

7https://axivity.com/userguides/ax3/settings/
lifestyle indicators including physical activity, diet across cohorts over Europe, rather than a UK focus as in the UK Biobank.

There were three main differences between this dataset and the UK Biobank. The first is the size of the cohorts collected: The DIRECT population size is much smaller than the UK Biobank as can be seen when comparing the number subjects displayed in Table 2.1 and Figure 4.1 in Chapter 4. The second is the nature of the cohorts: DIRECT exclusively focused on collecting and analysing data from individuals confirmed to have a diagnosis of diabetes and individuals who are in a pre-diabetes stage of disease. This meant that the protocols adopted in selecting individuals to partake in the data collection were rigorous and more robust to noisy controls and the impact of non-diabetes related conditions. Finally, the third was the rigorous protocols in following up subjects over time with comprehensive collection, including accelerometer data collection efforts.

The aim of IMI DIRECT was to explore the heterogeneity of Type-2 Diabetes and the impact of this on diabetes-related outcomes such as progression and responses to different drugs. The hope is to discover biomarkers for the deterioration of glycaemic control before and after the onset of Type-2 Diabetes. This would help to further enable the vision of personalised and predictive medicine for people suffering from Type-2 Diabetes. The pre-diabetes and post-diabetes cohorts of volunteer subjects are labelled as 2.1 and 2.2 respectively. The dataset is open to access to researchers providing approval from the regulators is given.

Table 2.1 below shows the number of individuals at each data collection time point (i.e when a subject wore the fixed wrist-worn accelerometer device for 10 days) from baseline to 18 months after for both Cohort 2.1 and 2.2, and 36 months after baseline for Cohort 2.2 and 48 months after for Cohort 2.1.

It is clear from Table 2.1 that the baseline had the highest number of participants, with the number gradually decreasing over time. Cohort 2.1, the cohort of subjects in a pre-diabetes stage of disease, has a much higher number of participants than Cohort 2.2, the cohort of subjects with a confirmed diagnosis of Type-2 Diabetes.

It can be seen that the cohorts gradually decrease in size as the length of time elapsed from the baseline assessment increases. This trend is demonstrated in both Cohort 2.1 (Pre-diabetes) and Cohort 2.2 (Diabetes confirmed). The difference in cohort sizes over time presents a problem whereby it limits the dataset sizes that can be used for training machine learning models to an increasingly smaller size as the longitudinal study extends. This was an issue in Chapter 6.
<table>
<thead>
<tr>
<th>Study</th>
<th>Centre</th>
<th>Country</th>
<th>Baseline</th>
<th>18 months after baseline</th>
<th>36 months after baseline</th>
<th>48 months after baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Vrije Universiteit Medical Center</td>
<td>The Netherlands</td>
<td>439</td>
<td>345</td>
<td>N/A</td>
<td>285</td>
</tr>
<tr>
<td>2.1</td>
<td>Lund University</td>
<td>Sweden</td>
<td>139</td>
<td>127</td>
<td>N/A</td>
<td>90</td>
</tr>
<tr>
<td>2.1</td>
<td>University of Copenhagen</td>
<td>Denmark</td>
<td>283</td>
<td>194</td>
<td>N/A</td>
<td>104</td>
</tr>
<tr>
<td>2.1</td>
<td>University of Eastern Finland</td>
<td>Finland</td>
<td>1068</td>
<td>759</td>
<td>N/A</td>
<td>682</td>
</tr>
<tr>
<td>2.2</td>
<td>University of Exeter</td>
<td>UK</td>
<td>165</td>
<td>145</td>
<td>127</td>
<td>N/A</td>
</tr>
<tr>
<td>2.2</td>
<td>University of Dundee</td>
<td>UK</td>
<td>168</td>
<td>136</td>
<td>138</td>
<td>N/A</td>
</tr>
<tr>
<td>2.2</td>
<td>Newcastle University</td>
<td>UK</td>
<td>131</td>
<td>61</td>
<td>57</td>
<td>N/A</td>
</tr>
<tr>
<td>2.2</td>
<td>Vrije Universiteit Medical Center</td>
<td>The Netherlands</td>
<td>150</td>
<td>126</td>
<td>99</td>
<td>N/A</td>
</tr>
<tr>
<td>2.2</td>
<td>Lund University</td>
<td>Sweden</td>
<td>101</td>
<td>88</td>
<td>75</td>
<td>N/A</td>
</tr>
<tr>
<td>2.2</td>
<td>University of Copenhagen</td>
<td>Denmark</td>
<td>46</td>
<td>26</td>
<td>7</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Table 2.1: Table showing the number of subjects for each centre at each collection time point in the DIRECT physical activity accelerometer dataset. Study 2.1 are pre-diabetes subjects, Study 2.2 are subjects positive for Type-2 Diabetes.

The wrist-worn tri-axial accelerometer device used in DIRECT is the ActiGraph GT3X+ created by Actigraph LLC, Pensacola, FL, USA. It was fit to a participant’s non-dominant wrist and was set to record at 30Hz with the manufacturer’s sleep mode disabled. The ActiGraph GT3X+ device itself is sampled by a 12 bit analog to digital converter at rates ranging from 30 Hz to 100 Hz and stored in a raw, non-filtered/accumulated format in the units of gravity (G’s). The device includes both a micro-electro-mechanical system (MEMS) based accelerometer and an ambient light sensor. The ambient light data is sampled and stored to memory at a 1Hz rate.

Efforts to validate this device have been undertaken in accordance with the V3 Framework by Goldsack et al [80]. The ActiGraph GT3X+ device was used as part of a study by Smith et al [186] in validating their performance for sleep and wake indices and comparing with polysomnography. They reported high sensitivity, low specificity and also recognised importance of the body location where the device is worn, such as the wrist and the hip in measuring sleep timing and sleep quality. In Aadland et al [1], the reliability of this device was established but they also recognised the importance of adequate wear time length which is another challenge that has to be addressed by data collectors. Their study stretched out over several days.

To compare this device with the Axivity AX3, the validity of ActiGraph GT3X+ and Axivity AX3 accelerometer devices was evaluated by Hedayatrad et al [92] with
around 74-96% accuracy, with the Axivity AX3 performing slightly better (4% higher, p <0.01). However, it was recognised that both devices had their respective greatest errors when contrasting sitting/standing, sedentary/light intensity activity, and moderate/light intensity.

DIRECT has carried out some analyses with this accelerometer dataset that was found in the literature. The work found in the literature was the work completed by Koivula et al [115] which noted that the extent to which physical activity has an effect on the metabolic homeostasis under the twin-cycle hypothesis is unclear. The twin-cycle hypothesis, proposed in [195], basically postulates that positive calorie balance from dietary intake over the long-term would cause fat to accumulate in the liver and this would then spill over into the pancreas, eventually causing the inhibition of insulin secretion. Thus this would lead to the onset of hyperglycaemia (high blood glucose). Koivula et al’s work hand-crafted features that characterised the physical activity of individuals into a single variable: mean of the high-pass filtered vector magnitude \((hpfVM)\). This variable does not take into account the nuances, the granular or the different types of physical activities that can be used to characterise individual behaviour. The work in this thesis builds on this investigation by exploring other methods of representing physical activity that may offer better insights into characterising individual behaviour and thus enabling the digital phenotyping from accelerometry data.

9 Conclusions

This chapter has provided an overview of the state-of-the-art by reviewing some of the works in the literature. It provides a foundation into the key theories and methods drawn from data science and machine learning that were used in our research in an attempt to explore the validity of our hypotheses. It reviews the current state-of-the-art in recent uses of digital wearables in the clinical and healthcare sphere of applications. The unique contributions of our research build upon the material in this chapter.

We have established that there is a breadth of knowledge and research in applying machine learning methods for tasks related to recognising human activities, including some of the classical machine learning and then deep learning approaches adopted. We also explored the two datasets that were used as our test beds for experimenting with a comprehensive suite of predictive modelling and analytics tasks. There is also a review of some of the research in neuroscience and gait analysis as applied to disease research which have also been adopted into our research.
Chapter 3

Establishing a baseline with GGIR: An accelerometry analysis framework
1 Introduction

This is a relatively short chapter that presents the work conducted to establish a baseline for the remaining work that contributed to our research at its beginning. Although it was short, it was a vital component to gain a comprehensive understanding and critically appraise the work that is at the forefront in the body of knowledge of accelerometry data analysis for disease. The resulting baseline is needed because it produces a standard for comparison of methods developed later as part of our research. The technical goal is to derive a baseline on which we can then compare our work against. The other methods and techniques which contribute in our research aims to improve on this baseline or to complement it to further research in this space.

GGIR, which stands for GENEActiv and GENEa data In R, is an open-source software package written in the R programming language and has been widely used in other related research works in physical activity and sleep science to analyse multi-day raw accelerometer traces.

Because this package is at the forefront of analysing sleep and physical activity data, it has been used to establish a baseline. This chapter describes the process of how GGIR was used and then critically analyses its performance.

Using GGIR as an initial research starting point was an important task in establishing what the current state-of-the-art methods for extracting features to represent accelerometry data can be used for predictive modelling experiments for Type-2 Diabetes related outcomes. It also provided an initial exercise in experimenting with our test bed dataset: The UK Biobank.

The hypothesis that this chapter is seeking to investigate is that features extracted using the GGIR package can be used to learn predictive models for cardiometabolic disease outcomes.

2 An Overview of GGIR

Raw accelerometer data implies that the data is expressed in terms of gravitational acceleration as opposed to brand specific units. Raw accelerometer data are stored in different binary formats. The data used in our research have been stored in .cwa and .gt3x formats.

The primary aim of our research was on extracting signals from self-monitoring, personal digital wearable devices that could be used to support the study of Type-2 Diabetes. One of the most well-known and widely used tools in the study of extracting meaningful and interpretable features from raw accelerometry data is GGIR, which stands for GENEActiv and GENEa data In R. GENEActiv and GENEa are two standard and widely used wrist-worn devices measuring acceleration for the purposes of studying physical activity and sleep. Raw accelerometry data refers to the data being presented in m/s², or gravitational
acceleration. GGIR processes multi-day raw accelerometry data \[137\]. This tool is used as a method to build a baseline, since it is well-studied and widely used in this area of research.

The GGIR packages automates many of the standard preprocessing steps needed before data analysis can take place, this includes automatic calibration, detecting abnormally high values, non-wear periods and calculating the magnitudes of acceleration.

The functionality of GGIR is divided into five separate "parts". Each "part" builds upon the output of the previous "part". A brief description of each "part" is provided here:

Part 1: This part auto-calibrates the signal based on local gravitational acceleration and detects non-wear periods. Although this is the most time-consuming part as it performs many of the preprocessing steps that are used in the subsequent parts. Part 2 to 5 consume much less time and computational resources.

Part 2: Analyses the quality of data and provides a low-level, basic descriptive set of interpretable features which summarises the activity signal on a per day and per data file basis.

Part 3: Detects sustained, prolonged periods of inactivity which would be indicative of sleep.

Part 4: Sleep detection. Labels the prolonged periods of inactivity as sleep, during night time, or inactivity during day time. This is also completed per day and per data file.

Part 5: Uses the output from parts 2, 3 and 4 to derive sleep and physical activity characteristics to merge them together. Part 5 also calculates the total time in different intensity categories (light, moderate and vigorous intensities), number of bouts, time spent in bouts and average acceleration.

GGIR offers a broad and comprehensive set of functionalities and operations which range from quality control to per day and per data file characterisation of physical activity and sleep. It is also relatively easy to use without requiring much programming expertise as the functionality can be called from a shell function. The outputs of parts 2, 4 and 5 could be saved as a separate CSV file.

We use GGIR for our initial experiments by processing the selected subset of raw individual physical activity traces from the UK Biobank. This performs the GGIR feature extraction process which are then used to learn predictive models for appropriate disease-related outcomes (i.e whether an individual has neither Type-2 Diabetes or cardiovascular, is suffering from only one of these two diseases, or is suffering from both diseases.)

There are works in the literature that have been undertaken to validate the GGIR package and its related outcomes. This is an important task given how

\[1\] https://cran.r-project.org/web/packages/GGIR/vignettes/GGIR.html
widely GGIR is used in the physical activity analysis space. In Rowlands et al’s work [172], ActiGraph GT3X+ and Axivity AX3 accelerometer devices were specifically compared against each other with GGIR, making this a very relevant study for our research. Both devices demonstrated a 95% accuracy with GGIR’s sleep classification algorithm. Validation of GGIR has not been well established in with either the UK Biobank or IMI Direct accelerometry datasets which means that there is no direct validation of the package using the cohorts relevant to the study in this chapter. Kerr et al [111] compared GGIR’s raw vector magnitude algorithm against machine learning algorithms and single cut axis points. Their results showed variance by as much as 52% across techniques. This shows that validation and verification of accelerometer analysis tools, including GGIR, remains a very significant challenge given the lack of standardisation. Femiano et al’s work [67] tried to verify the accuracy of step-counting algorithms in accelerometer analysis tools, of which GGIR was one they included. They found that different types of walking and running activities resulted in varying levels of accuracy, thus further highlighting the difficulty in validating tools. They also evaluated the Oxford accelerometry analysis toolkit that was described in Chapter 2’s subsection 6.5.

3 UK Biobank physical activity dataset with GGIR

3.1 Training dataset selection

All roughly 94,000 individuals for whom the UK Biobank collected the one-week period of recorded physical activity underwent processing with GGIR, which will henceforth be called the UK Biobank GGIR dataset, and the outputs for Parts 2, 4 and 5 were merged together to produce the most complete range of features available. From this population of the UK Biobank GGIR dataset, only those possessing an inferred status of being diagnosed with either Type-2 Diabetes or Cardiovascular, or both diseases, or confirmed as being negative, were selected for our study in this initial set of analyses. This reduced the roughly 94,000 individuals down to 46,711 individuals. Those with a null inferred status for Type-2 Diabetes and cardiovascular disease were dropped and discarded from this initial analysis. Individuals identified as being positive cases for Type-2 Diabetes, or cardiovascular disease, or suffering from both diseases, and negative control individuals without either disease were selected for the experiments in establishing a baseline. Inferred disease status was derived from Cassidy et al’s work [27]. Cardiovascular disease was included at this stage due to the existing works in the research literature with GGIR that did not study Type-2 Diabetes exclusively at this stage.

There are four disease cohorts for the baseline experiments as a result from Cassidy et al’s work [27] and are described in Table 3.1 giving the size of the cohort and the nature of their health status.
Table 3.1: Table displaying the number of individuals in each cohort of the GGIR physical activity dataset.

This subset of the entire UK Biobank GGIR dataset, with the complete available information from the UK Biobank at baseline gives around 50,000 individuals for the initial baseline experiments.

In order to establish the baseline on which to build the rest of this thesis upon, it was important to learn predictive models for these disease outcomes using the GGIR features derived for the selected UK Biobank participants as the training and test data. However, the four cohorts were heavily imbalanced with the negative controls heavily outnumbering the Type-2 Diabetes and/or cardiovascular disease positive case cohorts. This would create a situation where a predictive model could achieve reasonable accuracy by simply predicting the negative control class [85].

To rectify and alleviate this issue, class balancing must be adopted. This can be achieved by having a more simplified definition of the classes to be predicted by the model. We can define two classes: Positive and Negative. The positive class includes all of the selected individuals in the training dataset that were positive for Type-2 Diabetes, positive for cardiovascular disease, or positive for both diseases. The negative class would include only the negative controls identified by the UK Biobank. This balanced the classes into a roughly 50:50 ratio of positive and negative, forming a more balanced binary classification problem.

### 3.2 Preprocessing the UK Biobank GGIR physical activity training dataset

Some features were removed that did not serve any purpose to this exercise, mostly metadata and timestamp variables. These were relevant in this study which looks exclusively at GGIR’s physical activity and sleep features.

Another issue was that combining all the features from GGIR’s parts 2, 4 and 5 outputs resulted in 438 features. However, since each part of GGIR builds on the previous part, it is assumed that many of these 438 features would likely be highly correlating with each other, thereby introducing redundant features which could only cause greater complexity to models and increase possibility of overfitting [228]. A model that overfits tends to take all of the features into consideration when

<table>
<thead>
<tr>
<th>Disease cohort</th>
<th>Number of individuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neither positive for Type-2 Diabetes nor cardiovascular disease (Negative controls)</td>
<td>23,283</td>
</tr>
<tr>
<td>Positive for cardiovascular disease</td>
<td>21,186</td>
</tr>
<tr>
<td>Positive for Type-2 Diabetes</td>
<td>1,605</td>
</tr>
<tr>
<td>Positive for both Type-2 Diabetes and cardiovascular disease</td>
<td>637</td>
</tr>
</tbody>
</table>
predicting, even the features which have little effect on the output, or are actually meaningless noise.

This process of feature elimination reduced the set of features from 438 total features from GGIR parts 1, 2, 3, 4 and 5 combined down to a subset of 83 features, a huge reduction in redundancy and thereby reducing the risk of overfitting. Pearson calculation was used to calculate which features were highly correlated to one another, with a threshold of 0.8 [119].

The next step was to detect outliers in the dataset and to devise a strategy of eliminating these. This was done to reduce the skewness and kurtosis of features where possible. Skewness in the data could result in the tail regions acting as outliers and would then affect any predictive model’s performance, in particular for any regression-based models.

The GGIR documentation was used to guide the manual filtering out of non-valid records. The selection criteria for inclusion into the training dataset are as follows:

- The clipping score for any row is 0
- There is between 6.9 to 7.3 days of measurements
- Less than one hour of non-wear per day of the accelerometer device

These criteria ensured that all records were consistent and the data values they stored would be satisfactory to the movement scientist. The number of rows removed was quite large, reducing the dataframe further from 46,711 records to 32,440 records.

The selected UK Biobank GGIR training dataset had significant portions of missing data after the raw UK Biobank accelerometer data files underwent processing with GGIR. Data imputation was a significant step in cleaning the dataset. Features where more than 20% of the proportion of data values were missing were removed. This then reduced the number of GGIR features from 83 to 33. All remaining missing data were then imputed using the sci-kit learn library’s IterativeImputer class [155]. The IterativeImputer class in scikit-learn is an implementation of the Multivariate Imputation by Chained Equations (MICE) algorithm [203].

Finally, feature scaling was the final preprocessing step where feature values are normalised. This was achieved by using the StandardScaler in scikit-learn. This final dataframe with 33 features and 32,440 records was used for the baseline experiments in this chapter.

The final list of 33 GGIR physical activity features are displayed in Table 3.2 and shows how much of each feature was missing which were then imputed with the IterativeImputer in Scikit-learn, as the proportion of data missing was relatively small. Full details of the individual features are available in the GGIR manual.²

²https://cran.r-project.org/web/packages/GGIR/vignettes/GGIR.html
<table>
<thead>
<tr>
<th>GGIR feature name</th>
<th>Percentage of data missing (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>n_nights_acc</td>
<td>0.018</td>
</tr>
<tr>
<td>acc_timeinbed_AD_T5A5_mn</td>
<td>0.0075</td>
</tr>
<tr>
<td>acc_timeinbed_AD_T5A5_sd</td>
<td>2.17</td>
</tr>
<tr>
<td>acc_dur_noc_AD_T5A5_mn</td>
<td>0.015</td>
</tr>
<tr>
<td>acc_dur_noc_AD_T5A5_sd</td>
<td>2.07</td>
</tr>
<tr>
<td>acc_eff_AD_T5A5_mn</td>
<td>0.0038</td>
</tr>
<tr>
<td>acc_eff_AD_T5A5_sd</td>
<td>1.69</td>
</tr>
<tr>
<td>acc_dur_sibd_AD_T5A5_mn</td>
<td>0.72</td>
</tr>
<tr>
<td>acc_dur_sibd_AD_T5A5_sd</td>
<td>1.43</td>
</tr>
<tr>
<td>acc_n_noc_AD_T5A5_mn</td>
<td>0.0076</td>
</tr>
<tr>
<td>acc_n_noc_AD_T5A5_sd</td>
<td>1.4</td>
</tr>
<tr>
<td>acc_n_sibd_AD_T5A5_sd</td>
<td>0.94</td>
</tr>
<tr>
<td>acc_dur_msibd_AD_T5A5_mn</td>
<td>0.66</td>
</tr>
<tr>
<td>acc_dur_msibd_AD_T5A5_sd</td>
<td>1.44</td>
</tr>
<tr>
<td>acc_n_days_w_sibds_AD_T5A5</td>
<td>0</td>
</tr>
<tr>
<td>acc_onset_AD_T5A5_mn</td>
<td>0.19</td>
</tr>
<tr>
<td>acc_onset_AD_T5A5_sd</td>
<td>1.74</td>
</tr>
<tr>
<td>acc_wake_AD_T5A5_mn</td>
<td>0</td>
</tr>
<tr>
<td>acc_wake_AD_T5A5_sd</td>
<td>1.62</td>
</tr>
<tr>
<td>acc_timeinbed_WD_T5A5_sd</td>
<td>2.59</td>
</tr>
<tr>
<td>acc_dur_noc_WD_T5A5_mn</td>
<td>2.92</td>
</tr>
<tr>
<td>acc_eff_WD_T5A5_sd</td>
<td>2.57</td>
</tr>
<tr>
<td>acc_dur_sibd_WD_T5A5_sd</td>
<td>2.27</td>
</tr>
<tr>
<td>acc_n_noc_WD_T5A5_sd</td>
<td>2.28</td>
</tr>
<tr>
<td>acc_n_sibd_WD_T5A5_sd</td>
<td>1.94</td>
</tr>
<tr>
<td>acc_dur_msibd_WD_T5A5_mn</td>
<td>1.22</td>
</tr>
<tr>
<td>acc_dur_msibd_WD_T5A5_sd</td>
<td>2.39</td>
</tr>
<tr>
<td>acc_n_days_w_sibds_WD_T5A5</td>
<td>0</td>
</tr>
<tr>
<td>acc_onset_WD_T5A5_mn</td>
<td>0.81</td>
</tr>
<tr>
<td>acc_onset_WD_T5A5_sd</td>
<td>2.37</td>
</tr>
<tr>
<td>acc_wake_WD_T5A5_mn</td>
<td>0.53</td>
</tr>
<tr>
<td>acc_wake_WD_T5A5_sd</td>
<td>2.03</td>
</tr>
<tr>
<td>acc_n_days_w_sibds_WE_T5A5</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3.2: Table displaying the final set of features extracted by GGIR used in our baseline experiment after preprocessing and quality control process. Gives detail on how much data was missing from each variable.

Table 3.3 below gives a description of the demographics for the population of the final training dataset used in our baseline analysis.
<table>
<thead>
<tr>
<th></th>
<th>Category</th>
<th>Negative control</th>
<th>Positive T2D/CVD case</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Population size</strong></td>
<td></td>
<td>19532</td>
<td>12903</td>
</tr>
<tr>
<td><strong>Age (years)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40 to 49</td>
<td></td>
<td>4588</td>
<td>2932</td>
</tr>
<tr>
<td>50 to 59</td>
<td></td>
<td>7921</td>
<td>5204</td>
</tr>
<tr>
<td>60 to 70</td>
<td></td>
<td>7023</td>
<td>4767</td>
</tr>
<tr>
<td><strong>Number of males</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Body fat percentage (Male)</strong></td>
<td>9468</td>
<td>6967</td>
<td></td>
</tr>
<tr>
<td>6 to 13</td>
<td></td>
<td>273</td>
<td>105</td>
</tr>
<tr>
<td>14 to 20</td>
<td></td>
<td>582</td>
<td>952</td>
</tr>
<tr>
<td>21 to 24</td>
<td></td>
<td>12467</td>
<td>4214</td>
</tr>
<tr>
<td>Over 25</td>
<td></td>
<td>6210</td>
<td>7632</td>
</tr>
<tr>
<td><strong>Body fat percentage (Female)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 to 13</td>
<td></td>
<td>253</td>
<td>175</td>
</tr>
<tr>
<td>14 to 20</td>
<td></td>
<td>578</td>
<td>386</td>
</tr>
<tr>
<td>21 to 24</td>
<td></td>
<td>2757</td>
<td>1552</td>
</tr>
<tr>
<td>25 to 31</td>
<td></td>
<td>6593</td>
<td>2038</td>
</tr>
<tr>
<td>Over 31</td>
<td></td>
<td>9351</td>
<td>8752</td>
</tr>
<tr>
<td><strong>Ethnicity</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td></td>
<td>18993</td>
<td>12528</td>
</tr>
<tr>
<td>Mixed</td>
<td></td>
<td>54</td>
<td>69</td>
</tr>
<tr>
<td>Asian</td>
<td></td>
<td>167</td>
<td>147</td>
</tr>
<tr>
<td>Black</td>
<td></td>
<td>154</td>
<td>113</td>
</tr>
<tr>
<td>Chinese</td>
<td></td>
<td>28</td>
<td>14</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td>136</td>
<td>32</td>
</tr>
</tbody>
</table>

Table 3.3: Table showing demographics of the selected final training dataset after preprocessing effort.

The histograms in the Figures 3.2, 3.3, 3.4, 3.7, 3.1, 3.5, 3.6 illustrate the distributions of a subset of 7 of these 33 GGIR physical activity features after feature scaling. These distributions are presented because they demonstrate the typical distribution for all the GGIR extracted physical activity features after all the preprocessing steps.
Figure 3.1: Histogram for GGIR feature \textit{acc\_n\_noc\_AD\_T5A5\_sd}

Figure 3.2: Histogram for GGIR feature \textit{acc\_eff\_AD\_T5A5\_sd}
Figure 3.3: Histogram for GGIR feature $acc_n_{noc\_AD\_T5A5\_mn}$

Figure 3.4: Histogram for GGIR feature $acc_n_{sibd\_AD\_T5A5\_sd}$
Figure 3.5: Histogram for GGIR feature $acc_{\_n\_sibd\_WD\_T5A5\_sd}$

Figure 3.6: Histogram for GGIR feature $acc_{\_timeinbed\_AD\_T5A5\_mn}$
Figure 3.7: Histogram for GGIR feature acc_wake_AD_T5A5_mn

The histograms above give an indication into some of the distributions for each of the final list of GGIR features, most of the other features follow a similar bell-shaped distribution as the one for Figure 3.6 and Figure 3.2 with notable exceptions being in Figures 3.5 and 3.4. This set of features are identified as being not demonstrating high levels of missing data, correlation with other features and having a high number of outlier values.

4 Supervised learning for predicting disease outcomes

After exploratory data analysis and preprocessing the selected UK Biobank GGIR dataset, predictive models were learned to establish the baseline. The baseline we establish is a performance level with the GGIR feature space for the selected subset of the UK Biobank’s accelerometer cohort for binary classification of whether an individual belong to the Positive cases or Negative controls classes.

The performance from these models guides the next chapter of work in this thesis. This task is a binary classification task where the refined set of GGIR features are used to train and then validate the predictive models for disease outcome. The target outcomes are the POSITIVE and NEGATIVE class label defined in 3.1 after class balancing. The scikit-learn and pandas libraries in Python were used extensively in this set of baseline experiments.
4.1 Binary classification model selection and training process

Traditional classification models are first used due to the high levels of interpretability and performance. These two classifier models were selected so that this exercise can be applied to two very different approaches of classification which are appropriate to this dataset. We use the random forest and support vector machine, as described in Chapter 2, to predict the \textit{POSITIVE/NEGATIVE} disease class labels defined in \ref{sec:problem-definition}.

There are several options for kernel functions to deploy, including the radial basis function (RBF), polynomial and linear kernels. The choice of kernel is one of the main hyperparameters to fine tune in the training process for the SVM classifier. The radial basis function is the default kernel in \textit{scikit-learn}'s implementation of the support vector machine. This was kept as one of the kernel choices in this exercise. The other choice was the linear kernel due to the relative ease and speed of training. The SVM algorithm takes the same input as the random forest model.

The UK Biobank GGIR dataset was then trained with 10-fold cross-validation.

4.2 Training process and performance evaluation

When the models were trained and then used for prediction of their disease class label, several different standard machine learning metrics were calculated to evaluate the performance of the models. The main metrics which were evaluated were accuracy, precision, recall as well as Receiver Operating Characteristic (ROC) curves and their respective Area-Under-the-Curve (AUROC) metric.

5 Results

The metrics for evaluating the performance of binary classification models as described in section 4.2 were used to evaluate our predictive models’ performance. These metrics are standard and used widely in machine learning exercises.

<table>
<thead>
<tr>
<th>Number of trees in random forest model</th>
<th>Accuracy</th>
<th>Recall</th>
<th>Precision</th>
<th>Area under ROC curve</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.59</td>
<td>0.59</td>
<td>0.59</td>
<td>0.63</td>
</tr>
<tr>
<td>100</td>
<td>0.63</td>
<td>0.63</td>
<td>0.63</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Table 3.4: Table displaying the values for the results of standard performance metrics for the random forest’s prediction of disease outcome using selected GGIR features.
6 Discussion

Table 3.4 demonstrates that the features from GGIR have a very modest level of predictive power, with around 60% for AUC, precision, recall and accuracy. The predictive power of a model refers to the strength of the model’s ability to generate testable predictions. The AUC score would imply that the model performs slightly better than a model that totally predicted its results randomly. For recall and precision metrics, the results are also very modest. It can be seen however that an increase in the number of trees for the random forest increases performance on all metrics. But the difference between using 10 trees and 100 trees offered very little improvement in the performance metrics, so it implies that the random forest’s performance has most likely been exhausted even if the number of trees were significantly increased.

The SVM model performs even worse than the random forest model with around 40% accuracy, precision, recall and AUC scores. This is seen in Table 3.5. This is regardless of choice of kernel. This comprehensive predictive modelling exercises allow us to critically analyse the quality of GGIR features being used to train predictive models for disease.

It is clear from section 5 that the features derived from the GGIR framework are not powerful predictors of cardiometabolic disease when analysing the GGIR outputs from a large epidemiological dataset like the UK Biobank. This is true regardless of model choice. It is unclear, however, whether this is a result of the nature of the data itself or whether there are issues with the predictive models themselves. Furthermore, the output of GGIR requires significant efforts in preprocessing, training set refinement efforts and domain expertise and knowledge. The features from GGIR, if without documentation, are not easy to interpret and require a deep understanding of physical activity research and knowledge before it can be easily understood. A limitation of our baseline analysis is that we only used one accelerometer dataset, the UK Biobank, to establish the baseline for comparison in later experiments of our research.

Since the GGIR features are built upon clinical and biomedical domain knowledge, the features are very specific and crafted based on what researchers may request. This means that the scope is limited and may not capture other ways of describing physical activity and sleep patterns and other behaviour phenotypes. So,

<table>
<thead>
<tr>
<th>Choice of kernel for support vector machine classifier</th>
<th>Accuracy</th>
<th>Recall</th>
<th>Precision</th>
<th>Area under ROC curve</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radial basis function</td>
<td>0.49</td>
<td>0.49</td>
<td>0.46</td>
<td>0.44</td>
</tr>
<tr>
<td>Linear</td>
<td>0.45</td>
<td>0.44</td>
<td>0.44</td>
<td>0.42</td>
</tr>
<tr>
<td>Polynomial</td>
<td>0.39</td>
<td>0.42</td>
<td>0.42</td>
<td>0.40</td>
</tr>
</tbody>
</table>

Table 3.5: Table displaying the values for the results of standard performance metrics in predicting disease outcome using selected GGIR features with the SVM classifier
a new approach may be more useful and more powerful in allowing us to predict
disease states based on physical activity patterns and sleep. Additionally, there was
not a range of feature selection techniques deployed in this set of baseline analyses
which could potentially have enhanced the baseline results, such as Variable
Importance in Projection (VIP). This is a potential limitation of our results.

It is also important to state that this exercise was very rudimentary. There
were no quality controls in the selection of negative control subjects, nor was there
an assessment of whether the positive cases may be suffering from confounding
clinical events that may affect their physical activity levels. For example, someone
may be negative for cardiovascular disease and Type-2 Diabetes at the baseline
assessment when they participated in the UK Biobank, but they may have some
neurological disorder which significantly impairs their physical activity or they may
suffer severe insomnia which affects their sleep patterns. These confounding factors
could significantly impact the performance of any predictive models that aim to
predict disease outcomes and then to assess the power of digital phenotypes for
disease. This highlights the importance of robust and rigorous quality controls in
the selection of the training set.

Furthermore, GGIR is also limited in that its variables are engineered by hand
and does is limited in capturing properties of the data that represent those specific
variables. A more unsupervised approach can capture hidden representations in the
data that may not have been considered before.

On the other hand, in terms of data collection and quality controls, one of the
limitations could be that were not controlled in our selection of the training dataset.
For example, we did not account for age or working status (retired or not) which
may have confounded our results.

Our approach to balancing the classes in this rudimentary baseline analysis may
also have been too strict in its definition and could have impacted the robustness
of our predictive modelling performance by manipulating the results in terms of
precision and recall in favour of improving performance accuracy.

In view of these results, we formulated an enhanced experimental plan that uses
methods from data science and clinical applications to improve on this baseline.
We look at some of the features that GGIR extracts from raw accelerometer data
and adapt some of these into a new feature space to represent physical activity
data and then used to learn predictive models for disease outcomes. Furthermore,
a more advanced process in selecting our training dataset was developed. This is
described in the next chapter. In later parts of our research, we address the limited
nature of GGIR by studying new representations that uses advanced deep learning
techniques to build a latent representation of accelerometry data to learn predictive
models similar to the ones described in this chapter.
Chapter 4

High-level activity bout features for predicting Type-2 Diabetes
1 Introduction

This chapter of the thesis introduces a new representation space we developed in our research for the raw accelerometer data. As with the goals from our predictive modelling exercise with the features derived from the GGIR package, the aim of our research is to use this new representation space for the UK Biobank’s accelerometer dataset to discriminate between individuals with Type-2 Diabetes and without. However, we also use a more selective criteria to refine our training dataset, in particular our selection of the negative control class by considering other health conditions and clinical events that may affect an individual’s physical activity and excluding those with highly impaired physical activity due to noise.

In our research, we used an established tool to apply and build upon the work done in neuroscience research for other chronic diseases such as dementia, and Parkinson’s, as well as events that affect health such as fall detection. This methodology is applied in the context of Type-2 Diabetes. The tool produces a labelled activity sequence from raw accelerometry traces. The objective of this work is to extract signals generated by methods adapted from neuroscience and gait analysis to construct digital biomarkers that can be used to group individuals who are suffering from Type-2 Diabetes and individuals who are normoglycaemic, individuals whose fasting blood glucose levels are in the normal range.

After we extracted our new representation of the physical activity traces, we carried two independent efforts in our analyses of the data. The first set of experimental analyses was to use the new high-level feature space to learn predictive models for binary classification into either Type-2 Diabetes positive or Type-2 Diabetes negative control. Successful results would illustrate key differences in exhibited behaviour between individuals who have Type-2 Diabetes and those who do not.

The second part of our analyses took an unsupervised route where we deployed various clustering methods on the high-level physical activity feature space. This work was to reinforce the conclusions we reached in the first part of our study but also to demonstrate that Type-2 Diabetes positive individuals exhibit behaviours which are similar and would show that digital biomarkers derived from accelerometry may be useful in identifying populations at risk of this disease.

2 Materials and Methods

2.1 Hypotheses

One of the hypothesies made in this study is that physical activity profiles, when presented in sufficient detail, differ significantly between individuals who have Type-2 Diabetes and those of the general population. An assumption is made here that those in the general population are not affected by other conditions or health events that may severely impair their ability and capacity to perform physical activity. Examples of this would be individuals who are recovering from major surgery or individuals who, while they do not Type-2 Diabetes, but might be suffering from Parkinson’s disease at the time of when they wore their accelerometer device.

The second part of our study deployed clustering methods on our representation for the physical activity data. The hypothesis in this research effort was that individuals in our dataset could be clustered into well-separated groups that exhibit behaviours which are then associated with phenotypes related to Type-2 Diabetes individuals. To prove this hypothesis, we would need demonstrated by projecting their Type-2 Diabetes status and looking at whether Type-2 Diabetes positive individuals are represented in one particular cluster or not.

2.2 Refining the training dataset with health records

There were 500,000 participants in the UK Biobank. Between 2013 and 2015, approximately 100,000 UK Biobank participants were sent wrist-worn accelerometers to record their activity over a seven-day period. At the time of baseline assessment, 2755 individuals were identified to have Type-2 Diabetes through self-reported data or from the participant's primary care Electronic Health Record. The same criteria as described in the study by Schüssler-Fiorenza Rose et al [182] was used to determine inclusion based on self-reported data. This involved selecting those with an explicit Type-2 Diabetes diagnoses at baseline assessment. The individuals who started taking insulin within the first year of their diagnosis, and those who were less than 35 years of age were excluded to reduce the chances that individuals who had type-1 diabetes were included. Type-1 diabetes is not related to obesity, increasing age or low levels of physical activity as is the case with Type-2 Diabetes.

When this study was conducted, approximately 245,000 (45%) of the 502,644 UK Biobank participants had available Electronic Health Record data for primary care. These were used to identify individuals who did not develop Type-2 Diabetes at the time of baseline assessment, but did develop the disease before their accelerometer data collection time. EHR analysis identified a further group of individuals who should be selected for inclusion by using the work by Darke et
A total of 4076 individuals in the UK Biobank accelerometer cohort were found to have developed Type-2 Diabetes before wearing the accelerometer device.

The prevalence of Type-2 Diabetes in the UK Biobank was very low compared to the much larger group of individuals who did not have Type-2 Diabetes (4076 individuals were identified as having Type-2 Diabetes, whereas 99,636 did not). This would mean that a classifier for the Type-2 Diabetes status of an individual could have reasonable accuracy by just predicting the majority negative control class. Therefore, to alleviate this possibility, we rebalance the control class (those who did not develop Type-2 Diabetes before wearing the accelerometer) with the case class (those who developed Type-2 Diabetes before wearing the accelerometer device). We also use F1 scores as a performance metric as this metric is better at providing robust results for imbalanced datasets.

Random sampling for the negative control class was not the selection technique chosen for this task, as it was observed that the normoglycaemic control class included individuals who did not have Type-2 Diabetes but whose ability to perform physical activity might be severely impaired. Excluding these individuals from the study was desirable to remove noise in the negative control class.

An algorithm to judge and score the severity of an individual’s physical activity impairment was used to make an effort into selecting control groups based on Electronic Health Records. It is assumed that the training dataset that is produced from this effort is no worse than using random selection of controls. The selection process then involved a further analysis of Electronic Health Records for a period of time before the accelerometers were worn to identify non-diabetes conditions or other clinical events that may have caused severe physical activity impairment. As only 20,000 individuals in the accelerometer cohort had available Electronic Health Records, this part of the training dataset selection process was limited to this subset.

This analysis of selecting the appropriate controls resulted two control samples: Norm-0 are the negative controls without Type-2 Diabetes and without expected severe physical activity impairment, Norm-2 are the negative controls without Type-2 Diabetes and with expected severe physical activity impairment (1666 individuals). The remaining 9,723 Type-2 Diabetes negative control individuals, for whom there were available Electronic Health Records and were also not classed into Norm-0 or Norm-2 subpopulations are not considered in the supervised machine learning exercises. The number of participants in the Norm-0 and Norm-2 subpopulations are shown in Table 4.1 below.

<table>
<thead>
<tr>
<th>Impairment score</th>
<th>Total number of participants, N</th>
<th>With sufficient accelerometer wear time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Norm-0</td>
<td>11,019</td>
<td>8,463</td>
</tr>
<tr>
<td>Norm-2</td>
<td>3,355</td>
<td>1,666</td>
</tr>
</tbody>
</table>

Table 4.1: Number of participants in each sample according to physical activity impairment severity scores.

Having these two control groups allows us to build two training separate datasets: Training Set 1: Type-2 Diabetes positive cases vs Type-2 Diabetes negative controls
with no expected physical activity impairments, and Training Set 2: Type-2 Diabetes positive cases vs Type-2 Diabetes negative controls with expected severe physical activity impairment. Training Set 1 was used to investigate the hypotheses made for this study. Training Set 2 was used to quantify the effect of possible non-diabetes related health conditions that may impair a negative control subject’s physical activity performance abilities and as a potential source of noise.

Furthermore, a small number (151 out of 3101) individuals with Type-2 Diabetes are also observed to have high physical activity impairment scores. But this small subset was decided to be included in the training data for the positive case class. The reason for this is that Type-2 Diabetes is complex and multifaceted disease which can cause many complications or other co-morbidities with other conditions, for example cardiovascular diseases or cancer. In order to capture all behaviours and activity that may be caused or linked to Type-2 Diabetes, it is therefore necessary to include this small subset of individuals who have Type-2 Diabetes and are severely impaired in terms of their physical activity performance capabilities.

The process for selecting the individuals to be included, or excluded, from Training Set 1 and Training Set 2 is outlined and summarised in [4.1]

![Diagram](image-url)
2.3 Feature extraction: Representing high-level activity bout-based activity patterns

Physical activity is often completed in distinct continuous periods of sustained activity, known as a bout. Bouts of activity are of interest to the study of some chronic age-related diseases such as Parkinson’s disease and dementia. GGIR also extracts some features of its own representing physical activity at the bout-level. The labelled activity classification sequences can allow us to extract another space of representation for physical activity. Bouts of activity give an indication into the regularity, intensity and length of physical activity. We built a new representation for the physical activity traces by using tools from Human Activity Recognition.

We generate a new high-level granular representation based on physical activity bouts. A physical activity bout is defined as a single epoch, or an uninterrupted consecutive series of epochs in which a single activity type is performed. The features extracted for this part of the study are at the level of physical activity bouts of each activity type: their frequency of bouts for each activity type, the average length of bouts for each activity type, and the percentage of time spent in each activity type. This method is adopted from neuroscience and gait analysis where these techniques are applied to movement-related diseases such as Parkinson’s disease and forms of dementia. Notable case studies that demonstrate this include Del Din et al’s work [53, 55] and McArdle et al [131]. In Del Din et al’s work [53], the impact of ambulatory bout (bouts of walking) lengths are studied as a marker for ageing and Parkinson’s disease and identified significant differences in gait characteristics and ambulatory bout lengths between negative control subjects and subjects who were diagnosed with Parkinson’s disease. Similar results were found when analysing the gait of individuals who frequently fall [55] (perhaps due to Parkinson’s disease or being older adults) and those diagnosed with dementia [131].

The Biobank accelerometer analysis toolkit [222], developed at Oxford University, was used to extract representations of the raw accelerometer data in the selected UK Biobank accelerometer wearers. The tool annotated and labelled all the raw activity traces. Each 30-second epoch of the time series was labelled as one of five activity classes: Sedentary, moderate, sleep, walking and light tasks. These activity classes correspond to different levels of Metabolic Equivalent Task: Sedentary (1.5); moderate (4.9); walking (3.2); sleep (1.0); and light tasks (2.2). This was done for the individuals that were selected for the training set, as described in subsection 2.2. We used this labelled sequence of activity classification time series to construct the new high-level activity bout representation.

The sequence of processes undertaken, using the Oxford accelerometer analysis tool, to extract the high-level activity bout-based features from the raw accelerometer data in the UK Biobank is shown in Figure 4.2. Each level builds on the level below.

---

The top of this hierarchy, the activity bout-level feature extraction, is the new contribution from our research and was used for the studies in this chapter.

A personalised analysis of daily activities is performed to extract these features for the selected individuals. This involves personalising and accommodating for different individual sleeping habits by identifying the average night-sleep time boundaries for each individual across the total wear time. This was achieved by identifying the longest continuous bout of sleep activity over a 24-hour period. The remaining time of the 24-hour day is divided into 3 phases to represent morning, afternoon and evening times. This set of features shall be referred to as the high-level activity bout features in this thesis.

This extraction process gives a total of 60 activity-bout level features for each individual: 3 features extracted for each of the 5 types of activity for 4 periods of the day. This process is extended across the 7-day week of accelerometer wear time and aggregated by taking the average for each feature. The feature matrix extracted from the Oxford human activity recognition tool is displayed in Figure 4.3 below.
### 2.4 Selection of socio-demographic, anthropometric, and lifestyle variables from the UK Biobank

In order to quantitatively evaluate the relative importance of the new high-level activity bout features, traditional socio-demographic, anthropometric variables and lifestyle indicators which are commonly used in the study of Type-2 Diabetes are selected. The choice of these are inspired by the work of Cassidy et al. [28].

The table below shows the selected features for socio-demographic, anthropometric, and lifestyle characteristics.

![Feature matrix for the new high-level representation of the UK Biobank’s accelerometer cross-section in this study.](image)

**Table 4.3: Selected features for socio-demographic, anthropometric, and lifestyle characteristics.**

<table>
<thead>
<tr>
<th>Activity type</th>
<th>Time of day</th>
<th>Morning</th>
<th>Afternoon</th>
<th>Evening</th>
<th>Sleep time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walking</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Light tasks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moderate activity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sedentary activity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Personalized to each individual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of time spent in activity type during time period</td>
</tr>
<tr>
<td>Average length of bout for activity type during time of day</td>
</tr>
<tr>
<td>Number of bouts of activity type during time of day</td>
</tr>
<tr>
<td>Sociodemographic, lifestyle and anthropometric characteristic</td>
</tr>
<tr>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td>Sex</td>
</tr>
<tr>
<td>Age at assessment centre</td>
</tr>
<tr>
<td>Current alcohol drinking status</td>
</tr>
<tr>
<td>Current smoking status</td>
</tr>
<tr>
<td>Body fat percentage</td>
</tr>
<tr>
<td>Waist circumference</td>
</tr>
<tr>
<td>Sleep duration</td>
</tr>
<tr>
<td>Time spent watching television</td>
</tr>
<tr>
<td>Townsend deprivation index</td>
</tr>
<tr>
<td>Duration of walking activity</td>
</tr>
<tr>
<td>Duration of vigorous activity</td>
</tr>
<tr>
<td>Duration of moderate activity</td>
</tr>
</tbody>
</table>

Table 4.2: Sociodemographic, lifestyle, and anthropometric characteristics selected from the UK Biobank baseline assessment for comparison with high-level activity bout features space.

Furthermore, self-reported physical activity indicators in the UK Biobank are also selected. These are subjective measurements based on the accelerometer wearer’s self-reporting. In contrast, the physical activity features extracted for this study are not subjective and are derived from objective accelerometer traces. The objective physical activity features could also help to validate these subjective measurements.

The International Physical Activity Questionnaire-Short Form [125, 43] was used for these self-reported features and contains the participant’s self-assessed input for sleep duration, television viewing times, durations of walking, vigorous and moderate activities. Unfortunately one of the issues with these sociodemographic, anthropometric and lifestyle variables was that they were not fully complete and contained missing data. These were imputed using the $k$-nearest neighbour imputing algorithm from scikit-learn in Python. This calculates the missing value using the mean of $k$ nearest neighbours using the Euclidean distances, thereby preserving the distribution of the original data.

### 2.5 Supervised learning from high-level activity bout-based physical activity features

Once the training dataset was formed, a set of supervised machine learning experiments are conducted to investigate whether there is evidence that validates the hypotheses made for this study. This is achieved through a series of binary
classification models, where the Type-2 Diabetes status acts as the target binary outcome: Either the individual has Type-2 Diabetes, or does not have Type-2 Diabetes. Both training sets 1 and training sets 2 were used in this part of the study, as introduced in 2.2 in two separate sets of predictive modelling experiments. Additionally, different combinations of predictive features were used for each of the two training sets:

1. Using only high-level activity bout features extracted, as described in section 2.3
2. Using only socio-demographic, anthropometric and lifestyle features, as described in section 2.4
3. Using a combination of both the high-level activity bout features and the sociodemographic, anthropometric and lifestyle features

A combination of different feature sets were used to learn our predictive models in order to have a comprehensive suite of results on which to compare the performance of the high-level granular activity bout features on their own and in combination with standard sociodemographic, anthropometric and lifestyle features. This resulted in a total of six training datasets to train the binary classification models. Three binary classifier models were tested on these datasets due to their wide use high levels of performance: random forests, logistic regression and Extreme Gradient Boosting (XGBoost). XGBoost [34] is perhaps less well known and is relatively new. It uses gradient boosting, an ensemble method that builds stronger classifiers by adding models with weaker performance on of each other, iteratively, until a good level of predictive performance is achieved.

18 classification exercises were conducted using these combinations of 6 training datasets and 3 classification algorithms. Standard 10-fold cross-validation was used to avoid overfitting to the training dataset. A grid search was also conducted to tune the hyperparameters for each classification model. Furthermore, since the number of Norm-0 negative controls without Type-2 Diabetes (8,463 individuals) still vastly outnumbered the number of Type-2 Diabetes positive cases (3,103 individuals), a random selection of half of the Training Set 1 negative control class was undertaken to balance the negative control and positive case classes.

Standard machine learning practice reports on several performance evaluation metrics and these are replicated in this study. The metrics reported on are F1-scores, precision, recall, and the area under the receiver operating characteristic curve (AUROC) scores. F1 is a metric for the balance between precision and recall, a value between 0 and 1, where a score of 1 would imply perfect precision and recall. It is calculated using the harmonic mean of precision and recall. The AUROC score is a metric between the values 0 and 1 for how well a binary classifier is capable of discriminating between two classes. A value of 1 would imply perfect positive measure of discrimination, where was a value of 0.5 implies no discrimination capability and a value less than 0.5 would imply negative measure of discrimination.

In order to prove the hypotheses made for this study, model performances were evaluated to assess the following properties: The differences in predictive power
between Training Set 1 and Training Set 2; the effect of negative controls with high physical activity impairments and the potential for noise; and the best choice of binary classification model.

2.6 Unsupervised clustering from high-level activity bout-based physical activity features

The second part of this study involved conducting a clustering analysis. This is essentially an exercise to discover the high-level activity bout features naturally separate into different classes/clusters. These clusters should identify groups of individuals that exhibit similar behaviours, regardless of Type-2 Diabetes or severity of physical activity impairment. The high-level activity bout features were used to cluster individuals. Clustering is a set of techniques in unsupervised learning, where labels are not used to indicate if individuals belong to one group or not. The clustering study in this chapter is conducted on all the accelerometer participants identified who have Electronic Health Records, including those with Type-2 Diabetes positive case individuals and Type-2 Diabetes negative control individuals regardless of level of physical activity impairment or none.

2.6.1 Selecting and optimising clustering algorithms

$k$-means clustering and agglomerative hierarchical clustering were implemented in this part of this chapter’s study.

$K$-means often produces stronger and cohesive clusters by minimising the intra-cluster variation, but it requires the number of clusters to be optimised to ensure the best clustering result. This hyperparameter value, $k$, can be optimised and pre-defined by using the elbow method. This is hyperparameter grid search method where different values for $k$ are tried out and then this is plotted against a chosen clustering performance evaluative metric. In this case, the silhouette score was chosen to evaluate the quality of clustering, although others, such as distortion and the Calinski-Harabasz score, could also be used. When this is plotted in the graph, as shown in 4.4, the point of inflection can be found when $k$ is equal to 4. The point of inflection is also very clearly shown, suggesting that the data is highly clustered.

The agglomerative hierarchical clustering algorithm, on the other hands, uses a bottom-up approach where each data point begins in its own cluster, and pairs of clusters are merged as one moves up the hierarchical tree. The Ward linkage function was used. This linkage function minimises the sum of squared differences within all the clusters. The dendrogram is shown in the diagram in 4.5. It can be seen in the dendrogram that three distinct clusters are formed at the cutoff distance of 300.
2.6.2 Cluster interpretation

This study interprets these clusters by projecting their Type-2 Diabetes status and looking at whether Type-2 Diabetes positive individuals are represented in one particular cluster or not. This would then give an indication as to how individuals with Type-2 Diabetes exhibit similar behaviour in terms of the high-level activity bout features. Additionally, this study also projects the sociodemographic, lifestyle and anthropometric features described in subsection 2.4 onto these clusters by categorising and binarising these features where appropriate. Tests for statistical significance are then conducted to determine if there is enough confidence to
determine whether clusters separate participants into statistically significant groups.

3 Results

3.1 Distribution of High-Level Activity Bout Features

To summarize the distribution of the Type-2 Diabetes-positive and Norm-0 Type-2 Diabetes-negative populations, the high-level activity bout features were aggregated for a 24-hour period and averaged across both populations.

On average, both the Type-2 Diabetes-positive and Type-2 Diabetes-negative populations do not undertake significantly different quantities of each activity type aggregated to the level of the 24-hour day with approximately 5% moderate activity, 42% sedentary activity, 38% asleep, 5% light tasks, and 10% walking. However, the high-level activity bout features also offer an insight into the regularity and length of activity bouts. The values for these features do offer some discrimination between the Type-2 Diabetes-positive and Norm-0 Type-2 Diabetes-negative populations. The histograms below demonstrate an example of this by showing the distribution of daily averages for bout length, the number of bouts, and the percentage of times spent on sleep activity.

The histograms in Figure 4.6, Figure 4.7, and Figure 4.8 show noticeable differences between the 2 populations in the features that we have developed, when aggregated out to a day. Breaking the daily patterns into 4 distinct times of day (morning, afternoon, evening, and during sleep) would further demonstrate the differences in activity bout patterns for the 2 populations by virtue of the granularity. The combined effect of all these granular-level activity bout features produces high model accuracy, as reported below.

![Figure 4.6: Histogram for daily average percentage times spent asleep.](image-url)
These distributions are presented here to illustrate how activity bout length, frequency and percentage of daily time spent in each activity type are different between the Type-2 Diabetes positive cohort and the Type-2 Diabetes negative control cohort selected in this chapter’s study.
3.2 Results from supervised learning for predicting diabetes status

A summary and performance comparison across the 18 models built for this study is presented in Table 4.3 for AUC scores and Table 4.4 for the F1 scores, where AUC measures are obtained by averaging over 10 models using cross-validation for robustness. The receiver operating characteristic (ROC) curves and AUC scores are shown in Figure 4.9, Figure 4.11, Figure 4.10, Figure 4.12, Figure 4.13, and Figure 4.14.

AUC and F1 scores measure the performance of the binary classifiers differently and so are shown in separate tables. AUC is the area under the ROC curve (a plot of True Positive Rate against False Negative Rate). This is calculated at thresholds between the True Positive Rate and the False Positive Rate. This metric does well on more balanced datasets such as our training dataset. Imbalanced datasets may cause false positive predictions. A high AUC score would show us that the model is good at discriminating between the positive and negative classes.

On the other hand, F1 is a calculation that involves the recall and precision of the model’s predictions. F1 maintains a balance between these two metrics and so a high score would a high performance when balanced between these two metrics.

We report on both AUC and F1 scores of our predictive modelling experiments to give a more comprehensive set of results to compare from as they describe different aspects of a model’s predictive power.

<table>
<thead>
<tr>
<th>Predictive model</th>
<th>High-level activity bout features</th>
<th>Sociodemographic, anthropometric and lifestyle features</th>
<th>High-level activity bout features + sociodemographic, anthropometric and lifestyle features combined</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Norm-0</td>
<td>Norm-2</td>
<td>Norm-0</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.80</td>
<td>0.68</td>
<td>0.83</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>0.79</td>
<td>0.70</td>
<td>0.83</td>
</tr>
<tr>
<td>Extreme Gradient Boost (XGBoost)</td>
<td>0.78</td>
<td>0.66</td>
<td>0.80</td>
</tr>
</tbody>
</table>

Table 4.3: Classification results measured using area under the receiver operating characteristic curve scores, showing the effect of choice of type 2 diabetes–negatives, Norm-0 (no physical activity impairment) versus Norm-2 (severe physical activity impairment). The values in the cells represent area under the receiver operating characteristic curve scores.
<table>
<thead>
<tr>
<th>Predictive model</th>
<th>Type-2 Diabetes Status</th>
<th>High-level activity bout features only</th>
<th>Sociodemographic and lifestyle features only</th>
<th>High-level activity bout features + sociodemographic and lifestyle features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Norm-0</td>
<td>Norm-2</td>
<td>Norm-0</td>
</tr>
<tr>
<td>Random forest</td>
<td>Positive</td>
<td>0.65</td>
<td>0.70</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>0.78</td>
<td>0.54</td>
<td>0.78</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>Positive</td>
<td>0.66</td>
<td>0.72</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>0.77</td>
<td>0.54</td>
<td>0.79</td>
</tr>
<tr>
<td>Extreme Gradient Boosting (XGBoost)</td>
<td>Positive</td>
<td>0.66</td>
<td>0.68</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>0.77</td>
<td>0.52</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Table 4.4: Classification results measured using F1, showing the effect of choice of type 2 diabetes-negatives, Norm-0 (no physical activity impairment) versus Norm-2 (severe physical activity impairment). The values in the cells represent F1 scores.

Figure 4.9: ROC and AUROC score for type 2 diabetes vs Norm-0: High-level activity bout features, sociodemographic and lifestyle features combined.
Figure 4.10: ROC and AUROC score for Type-2 Diabetes positive vs Norm-0: High-level activity bout features only.

Figure 4.11: ROC and AUROC score for Type-2 Diabetes positive vs Norm-0: Sociodemographic and lifestyle features only.

Figure 4.12: ROC and AUROC score for Type-2 Diabetes positive vs Norm-2: High-level activity bout features, sociodemographic and lifestyle features combined.
When performance is measured using AUC, stronger results are achieved when using high-level activity bout features and sociodemographic and lifestyle in combination, as expected. Using high-level activity bout features on their own reduces performance (approximately 7%-8%). However, high-level activity bout features provide almost the same performance as traditional sociodemographic and lifestyle features on their own.

Models were also generated using alternate training data sets, where 151 Type-2 Diabetes-positive individuals with high physical activity impairment severity scores were excluded. These models exhibit very similar performance to those presented above, suggesting that physically impaired (Norm-2) Type-2 Diabetes-positive individuals can be used as part of the Type-2 Diabetes positives in the training set.

F1 measures in Table 4.4 reveal differences in classification accuracy between
Type-2 Diabetes against Norm-0 controls, and Type-2 Diabetes against Norm-2 controls. When using Norm-0 controls, negatives are more accurately predicted than Type-2 Diabetes, presumably because of class imbalance (4178 vs 3103). It is also clear that excluding physically impaired negatives improves the results.

When Norm-2 is used, however, Type-2 Diabetes is more accurately predicted than negatives, perhaps because in this case, Norm-2 is the minority class (1666 vs 3103) and because of potential diversity within the highly impaired control population. This will be investigated in a future study.

In all cases, the combination of high-level activity bout features and sociodemographic and lifestyle variables gives better results than using either set of features on their own, as expected. The performances of both feature sets are largely independent of the choice of the learning algorithm, as seen by the overlapping ROC curves.

3.3 Results from unsupervised clustering of individuals by high-level activity bout-based physical activity features

The key hypothesis of our clustering work was that individuals who are positive for Type-2 Diabetes would exhibit similar physical activity patterns with the new high-level activity bout features that are then shown by being clustered into a distinct cluster of their own. Our clustering work was successful in demonstrating this.

The most significant and interesting discovery from this unsupervised clustering exercise was that the fraction of each cluster that is Type-2 Diabetes positive was found to be 2,930 out of 2,936 (99.8%) in Cluster 2, with much smaller distributions in Clusters 1 and 3. This means that all the individuals identified to have Type-2 Diabetes at baseline or before the accelerometer device was worn exhibit similar behaviours that naturally distinguish them from other cohorts. This is also reflected in the broader context of the other socio-demographic and lifestyle features, where phenotypes associated with increased risk of Type-2 Diabetes (increasing age, high body fat percentage and a sedentary lifestyle) are also highly expressed in Cluster 2. This cluster was formed purely using the high-level physical activity feature space developed in this chapter’s study.

We used t-SNE as the technique used for dimensionality reduction to produce a 2D rendering of both k-means and hierarchical clusters, shown in Figure 4.15 and Figure 4.16 respectively. Different colours indicate different clusters. The clusters are much better defined using hierarchical clustering than for k-means, where there is significant qualitative overlap, with poor separation for dimensionality reduction to produce a 2D rendering of both k-means and hierarchical clusters, as seen in Figure 4.15 and Figure 4.16. A description of what these clusters represent are presented in Table 4.6.
Figure 4.15: Visualisation with T-SNE for k-means clustering. 4 overlapping clusters are seen

Figure 4.16: Visualisation with T-SNE for hierarchical agglomerative clustering. 3 clear clusters are seen.

This observation has been quantified using Silhouette coefficients and the Calinski-Harabrasz scores. Silhouette measures similarity of observations within each cluster (cohesion) compared to those in other clusters (separation). Values range from -1 to +1, where a higher value implies the observation is more matched to its own cluster and less so to neighbouring clusters. The Calinski-Harabrasz metric is well-suited where the ground truth is not known. It measures the ratio of the sum of intra-cluster dispersion and of inter-cluster dispersion for all clusters. Dispersion is defined as the squared sum of distances between any two points in a cluster.
There is no cut-off value for this performance metric as it is used for comparison between different clustering methods. A higher Calinski-Harabasz score implies better defined clusters without needing to compare with a set of ground truth labels (which we do not have given the unsupervised nature of this part of the study). Values for these the clustering validity indices are reported in Table 4.5 showing that hierarchical clustering produced better defined and cohesive clusters than k-means. So, only the results of this algorithm were further analysed.

Table 4.5: Calinski-Harabasz and Silhouette scores for unsupervised, resulting clusters by high-level activity bout features using k-means and hierarchical clustering algorithms

<table>
<thead>
<tr>
<th></th>
<th>Calinski-Harabasz</th>
<th>Silhouette</th>
</tr>
</thead>
<tbody>
<tr>
<td>k-means</td>
<td>2390.53</td>
<td>0.105</td>
</tr>
<tr>
<td>Hierarchical</td>
<td>2720.14</td>
<td>0.207</td>
</tr>
</tbody>
</table>

The fraction of males, 1,884 out of 2,936 (64.2%), is higher in Cluster 2 than in the other two clusters, as are the number of individuals in the highest age bracket (60 to 70). This trend is seen in most of the other socio-demographic and lifestyle features which are associated with exhibiting poorer health.

The chi-square tests in $p$-value columns also show a significant correlation between most of the features and Cluster membership, in particular those relating to anthropometric measures such as age, sex, body fat percentage and waist circumference. Table 4.6 provides an interpretation of clusters in terms of socio-demographic and lifestyle features, showing a non-random distribution of the variables within each cluster in all cases. Again Cluster 2 stands out. Note that numerical variables such as waist circumference were discretized into categorical features.

Table 4.6 also gives the $p$-values for each of the discretised sociodemographic, anthropometric and lifestyle variables are also given. The $p$-values for all of the sociodemographic, lifestyle and anthropometric features, with the notable exception of “Duration of walking activity” where the $p$-value is 0.37, allow us to reject the null hypothesis which, in this case, is that these clusters are randomly allocated and have no statistical significant difference in their means. This is because the $p$-values for all the features, with the exception of “Duration of walking activity”, are less than our critical value of $p < 0.05$. We accept the alternative hypothesis which, in this case, is that the clusters are not randomly allocated for all sociodemographic, anthropometric and lifestyle variables other than “duration of walking activity” where no significant effect is observed.

<table>
<thead>
<tr>
<th>Category</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster size</td>
<td>1119</td>
<td>2936</td>
<td>13306</td>
<td>$&lt;0.001$</td>
</tr>
<tr>
<td>Number of Type-2 Diabetes positives</td>
<td>162</td>
<td>2,930</td>
<td>339</td>
<td>$&lt;0.001$</td>
</tr>
</tbody>
</table>

Continued on next page
Table 4.6 – Continued from previous page

<table>
<thead>
<tr>
<th>Category</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of males</td>
<td>564</td>
<td>1884</td>
<td>6107</td>
<td>P &lt;0.001</td>
</tr>
<tr>
<td>Age (n)</td>
<td></td>
<td></td>
<td></td>
<td>P &lt;0.001</td>
</tr>
<tr>
<td>40 to 49</td>
<td>249</td>
<td>255</td>
<td>2697</td>
<td></td>
</tr>
<tr>
<td>50 to 59</td>
<td>411</td>
<td>854</td>
<td>4583</td>
<td></td>
</tr>
<tr>
<td>60 to 70</td>
<td>459</td>
<td>1826</td>
<td>6029</td>
<td></td>
</tr>
<tr>
<td>Body fat percentage - Male - (%)</td>
<td></td>
<td></td>
<td></td>
<td>P &lt;0.001</td>
</tr>
<tr>
<td>6 to 13</td>
<td>23</td>
<td>13</td>
<td>347</td>
<td></td>
</tr>
<tr>
<td>14 to 17</td>
<td>56</td>
<td>54</td>
<td>817</td>
<td></td>
</tr>
<tr>
<td>18 to 25</td>
<td>434</td>
<td>609</td>
<td>5487</td>
<td></td>
</tr>
<tr>
<td>Over 25</td>
<td>607</td>
<td>2260</td>
<td>6654</td>
<td></td>
</tr>
<tr>
<td>Body fat percentage - Female - (%)</td>
<td></td>
<td></td>
<td></td>
<td>P &lt;0.001</td>
</tr>
<tr>
<td>10 to 13</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>14 to 20</td>
<td>13</td>
<td>6</td>
<td>169</td>
<td></td>
</tr>
<tr>
<td>21 to 24</td>
<td>36</td>
<td>23</td>
<td>467</td>
<td></td>
</tr>
<tr>
<td>25 to 31</td>
<td>210</td>
<td>116</td>
<td>2692</td>
<td></td>
</tr>
<tr>
<td>Over 31</td>
<td>857</td>
<td>2791</td>
<td>9973</td>
<td></td>
</tr>
<tr>
<td>Current smoking status</td>
<td></td>
<td></td>
<td></td>
<td>P &lt;0.001</td>
</tr>
<tr>
<td>Never</td>
<td>607</td>
<td>1332</td>
<td>7706</td>
<td></td>
</tr>
<tr>
<td>Past</td>
<td>416</td>
<td>1377</td>
<td>4749</td>
<td></td>
</tr>
<tr>
<td>Current</td>
<td>95</td>
<td>227</td>
<td>852</td>
<td></td>
</tr>
<tr>
<td>Current alcohol drinking status</td>
<td></td>
<td></td>
<td></td>
<td>P &lt;0.001</td>
</tr>
<tr>
<td>Never</td>
<td>36</td>
<td>132</td>
<td>357</td>
<td></td>
</tr>
<tr>
<td>Past</td>
<td>29</td>
<td>132</td>
<td>446</td>
<td></td>
</tr>
<tr>
<td>Current</td>
<td>1054</td>
<td>2672</td>
<td>12637</td>
<td></td>
</tr>
<tr>
<td>Ethnic group</td>
<td></td>
<td></td>
<td></td>
<td>P &lt;0.001</td>
</tr>
<tr>
<td>White</td>
<td>1101</td>
<td>2866</td>
<td>13214</td>
<td></td>
</tr>
<tr>
<td>Mixed</td>
<td>4</td>
<td>5</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>0</td>
<td>15</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>Black African</td>
<td>4</td>
<td>18</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>Chinese</td>
<td>1</td>
<td>10</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>8</td>
<td>22</td>
<td>64</td>
<td></td>
</tr>
</tbody>
</table>

Continued on next page
Table 4.6 – Continued from previous page

<table>
<thead>
<tr>
<th>Category</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time spent watching television</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less than 1 hour</td>
<td>275</td>
<td>390</td>
<td>3158</td>
<td></td>
</tr>
<tr>
<td>1 to 2 hours</td>
<td>330</td>
<td>709</td>
<td>3860</td>
<td></td>
</tr>
<tr>
<td>2 to 3 hours</td>
<td>234</td>
<td>1065</td>
<td>3147</td>
<td></td>
</tr>
<tr>
<td>Over 3 hours</td>
<td>279</td>
<td>772</td>
<td>3142</td>
<td></td>
</tr>
<tr>
<td>Sleep duration</td>
<td></td>
<td></td>
<td></td>
<td>P &lt; 0.001</td>
</tr>
<tr>
<td>Less than 7 hours</td>
<td>238</td>
<td>775</td>
<td>2883</td>
<td></td>
</tr>
<tr>
<td>7 to 8 hours</td>
<td>810</td>
<td>1873</td>
<td>9632</td>
<td></td>
</tr>
<tr>
<td>Over 8 hours</td>
<td>71</td>
<td>287</td>
<td>790</td>
<td></td>
</tr>
<tr>
<td>Waist circumference (Male)</td>
<td></td>
<td></td>
<td></td>
<td>P &lt; 0.001</td>
</tr>
<tr>
<td>Less than 94cm</td>
<td>504</td>
<td>617</td>
<td>6612</td>
<td></td>
</tr>
<tr>
<td>94cm to 102cm</td>
<td>265</td>
<td>667</td>
<td>3307</td>
<td></td>
</tr>
<tr>
<td>Over 102cm</td>
<td>350</td>
<td>1652</td>
<td>3388</td>
<td></td>
</tr>
<tr>
<td>Waist circumference (Female)</td>
<td></td>
<td></td>
<td></td>
<td>P &lt; 0.001</td>
</tr>
<tr>
<td>Less than 80cm</td>
<td>507</td>
<td>289</td>
<td>6390</td>
<td></td>
</tr>
<tr>
<td>80cm to 88cm</td>
<td>271</td>
<td>415</td>
<td>3136</td>
<td></td>
</tr>
<tr>
<td>Over 88cm</td>
<td>341</td>
<td>2232</td>
<td>3780</td>
<td></td>
</tr>
<tr>
<td>Duration of walking activity</td>
<td></td>
<td></td>
<td></td>
<td>P = 0.37</td>
</tr>
<tr>
<td>0 to 20 mins</td>
<td>361</td>
<td>1035</td>
<td>4533</td>
<td></td>
</tr>
<tr>
<td>21 to 30 mins</td>
<td>272</td>
<td>694</td>
<td>3168</td>
<td></td>
</tr>
<tr>
<td>31 to 60 mins</td>
<td>329</td>
<td>864</td>
<td>3884</td>
<td></td>
</tr>
<tr>
<td>Over 60 mins</td>
<td>157</td>
<td>343</td>
<td>1720</td>
<td></td>
</tr>
<tr>
<td>Duration of moderate activity</td>
<td></td>
<td></td>
<td></td>
<td>P &lt; 0.001</td>
</tr>
<tr>
<td>0 to 15 mins</td>
<td>211</td>
<td>688</td>
<td>2484</td>
<td></td>
</tr>
<tr>
<td>16 to 30 mins</td>
<td>403</td>
<td>997</td>
<td>4765</td>
<td></td>
</tr>
<tr>
<td>31 to 60 mins</td>
<td>337</td>
<td>822</td>
<td>4139</td>
<td></td>
</tr>
<tr>
<td>Over 60 mins</td>
<td>168</td>
<td>428</td>
<td>1917</td>
<td></td>
</tr>
<tr>
<td>Duration of vigorous activity</td>
<td></td>
<td></td>
<td></td>
<td>P &lt; 0.001</td>
</tr>
<tr>
<td>0 mins</td>
<td>9</td>
<td>17</td>
<td>94</td>
<td></td>
</tr>
<tr>
<td>1 to 20 mins</td>
<td>360</td>
<td>1267</td>
<td>4511</td>
<td></td>
</tr>
<tr>
<td>21 to 45 mins</td>
<td>410</td>
<td>896</td>
<td>4430</td>
<td></td>
</tr>
</tbody>
</table>

Continued on next page
Table 4.6: Distribution of discretised sociodemographic, anthropometric and lifestyle features in all clusters.

<table>
<thead>
<tr>
<th>Category</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>46 to 180 mins</td>
<td>341</td>
<td>757</td>
<td>4279</td>
<td></td>
</tr>
</tbody>
</table>

In Figure 4.17, Figure 4.18, Figure 4.19 we can see that the daily relative proportions of time spent in each activity type for each cluster does not vary greatly between different clusters. But if these physical activity features include the average number of bouts and length of bouts for activity, then it can be seen there are differences in each cluster. The example for sleep activity shown in Figure 4.21, Figure 4.20 and Figure 4.22. Figure 4.22 demonstrates different aggregated distributions for the average daily number of sleep bouts, daily average length of sleep bouts, and average daily percentage time spent sleeping. This trend is seen in the other activity types.

Figure 4.17: Average daily percentage time spent in each activity type for Cluster 1

Figure 4.18: Average daily percentage time spent for each activity type, for Cluster 2
Figure 4.19: Average daily percentage time spent for each activity type, for Cluster 3

Figure 4.20: Daily average number of sleep bouts for all clusters

Figure 4.21: Daily average length of sleep bouts for all clusters
4 Discussion

4.1 Principal Findings

Using data from the UK Biobank, this study supports the hypothesis that individuals with diagnosed Type-2 Diabetes exhibit physical activity patterns that are significantly different from those of normoglycemic controls, thus providing novel ways to detect Type-2 Diabetes, that is, through appropriate analysis of physical activity patterns. Although most previous studies, particularly using UK Biobank, are limited to self-reported physical activity levels [27, 182, 115], here we have demonstrated the benefits of extracting a more objective and granular representation of physical activity from raw accelerometry traces data, namely, by activity type and time of day or sleep time. Using these features, either on their own or in combination with a selected set of sociodemographic, anthropometric, and lifestyle variables, we have shown that appropriately trained machine learning models were able to discriminate between the 2 cohorts with good predictive accuracy.

4.2 Implications of Study

These findings suggest that it may be possible to use continuous or periodic self-monitoring of individuals at risk of Type-2 Diabetes, specifically those in a prediabetes state, for screening and early detection of disease progression. This is particularly important as evidence shows that reversal of Type-2 Diabetes is possible, with a higher success rate when interventions are undertaken within the first 5 years of the disease [196, 132, 225, 183].

However, early detection is still an unsolved problem, with recent figures
reporting that over 190 million people worldwide live with undiagnosed diabetes [30]. Risk scores that are routinely used for screening, such as the Leicester score, are easy to obtain but not very accurate [11].

This suggests that self-monitoring of physical activity patterns, such as those presented in this study, may complement risk scores to help with the early detection of Type-2 Diabetes, especially in high-risk individuals. Today, this can be achieved at a low cost using readily available technology [189], including internet-enabled data loggers that do not require participants to return devices, such as smartphones, periodically. However, further research is required to establish the quality and significance of physical activity data for this specific purpose.

4.3 Limitations

In principle, it may be possible to try and detect early signs of Type-2 Diabetes using specific fingerprint patterns found in physical activity traces, where an example of a pattern may be a person who takes short bouts of low or moderate activities with frequent sedentary breaks in between. However, in practice, we found no evidence in the UK Biobank data set that strong correlations exist between specific physical activity patterns and Type-2 Diabetes. Thus, what the machine learning approach has to offer may be limited to the strong indication demonstrated in this work, namely, that granular features extracted from the raw traces, taken together, are indeed good predictors and usefully augment the more traditional sociodemographic set of variables.

Although the UK Biobank is the largest known public accelerometry data set where a Type-2 Diabetes cohort can be identified, detecting differences between Type-2 Diabetes and controls remains challenging because of their low prevalence in the population, which is reflected in this study with the relatively small data set available for training when using supervised machine learning. Simultaneously, this data set was subject to noise for two reasons. First, because no formal quality assurance protocol was enforced during data collection, and second, because of the limited knowledge about other non-Type-2 Diabetes–related conditions among the controls, which may contribute to reduced physical mobility or a more sedentary routine. We have shown how EHRs can be used to overcome this limitation.

Another potential limitation in the results of the study of this chapter is that we did not make an attempt to extract non-linear, pattern-based features to measure physical activity, such as the Alpha-Gini index which is more widely used in economics [185]. This type of feature would study percentage of long bouts of activity. Although it cannot be certain that this would improve the results of the experiments in our study, it would be a more comprehensive appraisal of our analyses since we would be study a broader range of differentiating physical activity metrics. An example of this would be that two subjets with the same volume/amount of physical activity, for example 60 minutes of walking, may exhibit different patterns— for example 2 bouts of 30 minutes of walking or six 10-minute bouts. This is definitely a potential route for future work.
4.4 Conclusions

This study motivates further research into the use of granular physical activity measures as a form of digital phenotype for Type-2 Diabetes. It also suggests that more rigorous protocols on wearing physical activity loggers are required to improve the quality of the data and the signal-to-noise ratio, along with stringent inclusion and exclusion criteria or at least comprehensive knowledge of clinical conditions that may affect the signal in the traces. This is also reflected in other studies [221, 167]. When such quality criteria are met, it should be possible to repeat the analysis presented here using data sets from large-scale deployment of physical activity loggers to validate the hypothesis that early detection of Type-2 Diabetes is scientifically and technically feasible.
Chapter 5

Latent representation learning of signal-level accelerometer data with autoencoders
1 Introduction

Extracting features from raw data such as raw accelerometry traces is a substantial and complex data science challenge that may require deep levels of domain expertise and knowledge. Tools such as GGIR are a good option that make use of domain expertise to extract features. These extracted variables are often of interest to scientists studying physical activity traces from accelerometers, however, it is albeit a limited set of features. Another approach is to use representation learning, also called feature learning, which are a set of techniques that enable the automatic discovery of feature spaces from raw accelerometry traces, bypassing the need for domain knowledge-based feature engineering. Representation learning is also useful for dimensionality reduction. Deep learning, particularly autoencoder architectures, are well suited to this task. Other related works have made use of this technology to extract features for disease outcome prediction [226, 108].

This chapter explores the use of autoencoders, a type of artificial neural network, to learn efficient encodings of data in an unsupervised and unlabelled way [118]. The aim is to extract a new representation of the accelerometer traces as a means of digital phenotyping from accelerometry for Type-2 Diabetes and to evaluate this representation in learning predictive models for disease outcomes.

The goal of this chapter is to learn a latent representation of second accelerometry dataset, the DIRECT diabetes dataset, which will then be used to learn predictive models for Type-2 Diabetes related target outcomes, and then compare these to the existing physical activity representations in DIRECT. The hypothesis that this chapter investigates is that the latent representation learned by an autoencoder architecture are a viable feature set for predictive models for Type-2 Diabetes related target outcomes.

2 Methods and Materials

2.1 DIRECT Diabetes Dataset

The UK Biobank was used in the previous chapters of this thesis as the initial test bed dataset for experimentation. In Chapter 4, it was established that the careful and calculated selection of the training set is an important task before predictive model learning. However the features extracted for the tasks with the UK Biobank in Chapter 4 may not capture hidden variables and representations of the physical activity data as all the features extracted so far have been engineered manually and with some domain guidance and knowledge based on existing works in the literature.

The DIRECT diabetes consortium [1] collected the accelerometer traces from

1https://directdiabetes.org/overview-of-the-project/
volunteer subjects. This dataset formed the second experimentation test bed was used for the work in this chapter. We decided to use this dataset due to the rigorous protocols that were used to assess and collect data from volunteer subjects. Furthermore, since the consortium is specifically specialised to studying Type-2 Diabetes, it could be assumed that the clinical knowledge would have been accounted for correctly in the collection and analysis of this dataset and that appropriate quality control protocols were used.

Although similar to the UK Biobank, there are several significant differences with the DIRECT dataset. Firstly, the volunteer subjects resided in different geographic locations across Europe, not just the UK alone. Each subject wore the accelerometer devices for period of 10 24-hour days with strict protocols on quality control and assessment, rather than 7 days as in the UK Biobank. The problem of non-wear, as was the case with significant numbers of participants in the UK Biobank’s accelerometer dataset, should be much more alleviated in DIRECT because the devices were fixed more tightly so that removing them would be harder than the UK Biobank. Furthermore, all subjects were either known to already have Type-2 Diabetes or in a state of pre-diabetes. Pre-diabetes is where their blood glucose levels would indicate a potential progression into the disease state at the time of when their data was collected.

The pie charts in Figures 5.1 and 5.2 illustrate the physical activity distribution between the pre-diabetes (Cohort 2.1) and Type-2 Diabetes confirmed (Cohort 2.2) cohorts respectively. These pie charts are plotted using the mean average of the distribution of physical activities for each of the two cohort. It is observed that habitual behavioural and sleep patterns do not vary greatly between the two cohorts on a high-level scale such as those seen in the pie charts. However, our research work aims to identify representations of this physical activity that offer an insight into the heterogeneity of physical activities and sleep patterns exhibited by the individuals, using the DIRECT accelerometer dataset as a test bed.

Figure 5.1: Pie chart showing the mean average percentage distribution in physical activities for the pre-diabetes Cohort 2.1

97
Another difference between the UK Biobank and DIRECT is the number of participants and scale of the collected data in terms of population size. The training dataset available for DIRECT is more robust and strictly controlled, but the population size is much smaller (2,168 at baseline) compared to the UK Biobank’s selected subpopulation (11,566 as per Table 4.1 and Figure 4.1).

The Oxford accelerometer analysis toolkit, which was used extensively for the work in Chapter 4, is capable of processing both .cwa and .gt3x binary file formats. Thus, we can use the Oxford accelerometer analysis toolkit to process the raw accelerometer traces data files from the DIRECT dataset.

The raw accelerometry .gt3x data files were processed with the Oxford accelerometer analysis toolkit in the same High-Performance Computing environment as in Chapter 4’s processing of the .cwa files with the UK Biobank. However, in this chapter, we used the signal-level features for each of the 30-second epoch, as seen in the feature space hierarchy illustrated in Figure 4.2 in Chapter 4. We used these signal-level epoch features to build a new representation in this chapter using a deep neural network.

The set of features that represent the accelerometer traces at this level of the feature pyramid in Chapter 4’s Figure 4.2 are more low-level than the activity labelled sequences in that same pyramid. The hypothesis is that this level of representation can be used to learn new features which may capture hidden states of behaviour that are not captured at the high-level representations. This involves an experiments and implementation of deep neural networks to learn new representations of the raw accelerometer traces at this level.

2.2 Signal-level epoch features

In Figure 4.2 of Chapter 4 we can see that, before the Oxford accelerometer analysis tool labels the activity time series with activity classes, the tool extracts a set of features above the raw triaxial time series called the *epoch-level signal extraction*. 

![Pie chart showing the mean average percentage distribution in physical activities for the Type-2 Diabetes positive Cohort 2.2](image)
This is a feature vector with 120 dimensions for every non-overlapping 30-second epoch of the raw accelerometer time series trace. 30 seconds is the default epoch period on which the models in this tool were trained. These signal-level features are in both the time domain and the frequency domain. Examples of features which are extracted at the signal-level are the vector magnitude (the Euclidean Norm Minus One), its mean, standard deviation, kurtosis, skewness, Fast Fourier transforms at 1 Hz to 15Hz.

Additionally, for each of the x, y and z axes, the tool also extracts the following features: Mean, range, standard deviation, covariance, and Fast Fourier transforms at 1Hz to 15Hz. Other features which are also extracted are roll, pitch, yaw, x/y/z correlations, frequency, average arm angles, and power bands. The precise process for extracting the signal-level features and data processing that the Oxford accelerometer analysis tool uses are described in Doherty et al [56] and Willetts et al’s work [222]. Some other features included metadata descriptors such as temperature and clipping scores which are not objective features for the accelerometer traces themselves so were discarded in our study.

2.2.1 Preprocessing the signal-level epoch features

All the .gt3x raw accelerometer traces data files of the DIRECT dataset underwent processing with the Oxford analysis tool.

Another task which needed to be completed was to remove those signal-level epoch outputs that did not meet the required length of wear time (10-days). This is to ensure that only those individual epoch outputs with the correct full 10-day wear time underwent the next part of the unsupervised feature learning process. We undertook this step to ensure that we have consistent lengths of accelerometer wear times in our research. To achieve this, all those signal-level epoch output files without 28,880 time steps were discarded because there are 28,880 30-second epochs in 10 full 24 hour days.

Some of the features in the signal-level epoch outputs contained constant values. These were discarded to avoid introducing complexity to the autoencoder. This step reduced the number of features from 120 to 113.

The last preprocessing step was to normalise all the signal level epoch feature outputs. This is an important step in transformation the training data as deep neural networks do not perform well when input values have different scales due to increasing complexity in training which may cause a model to be unstable. The feature values of the signal-level epoch data are all normalised to a range between 0 and 1. The data is now ready to be used for representation learning with deep neural networks.

The input training data now consists of 2608 samples, each with 28,799 time steps and 113 features.

99
2.3 Long Short-Term Memory Autoencoder Architecture

The neural network architecture that was implemented for learning a latent representation of the signal level epoch representations of the accelerometer traces was a recurrent LSTM autoencoder, an architecture already described in Chapter 2. This is a type of autoencoder that is widely used in dimensionality reduction and unsupervised learning tasks for sequential data, such as in this case. The model takes an input sequence, encodes it into a hidden (latent) representation which can be extracted at the bottleneck layer. It then decodes this latent representation and then reconstructs it. A reconstruction loss function is then used to measure the loss between the reconstructed output at the decoder and the original input passed into the encoder unit of the architecture.

The bottleneck layer is the hidden layer that often has a lower number of dimensions. This is where the encoding is produced. The bottleneck layer has a lower number of nodes and the number of nodes in the bottleneck layer also gives the dimension of the encoded output.

2.3.1 Autoencoder Architecture

The recurrent autoencoder was implemented in the Keras framework. Figure 5.3 is a diagram displaying the structure of the recurrent LSTM autoencoder.
Figure 5.3: LSTM Autoencoder architecture diagram

The architecture in Figure 5.3 is a standard architecture for recurrent autoencoders that learn from multidimensional data.

This recurrent autoencoder is capable of processing sequences of length 28,800 time steps (Totalling the equivalent of 10 full 24-hour days of 30-second epochs) with 113 dimensions (number of signal-level epoch features) per time step. It has three LSTM layers in the encoder part of the autoencoder, and three LSTM layers in the decoder part of the layer for the reconstruction.

This architecture reduces the dimensionality from 113 dimensions down to 25 dimensions for the latent representation. We added three hidden layers in this architecture due to the high dimensionality of the input signal-level representations. This architecture gradually reduces the dimensionality by 50% as we pass through the three hidden layers (from the original 113 dimensional signal feature vector input to 100, then 50 and then 25). The number of units in each layer correspond to the dimensionality of the layer’s outputs.

Optimising these hyperparameters is often an unsupervised exercise that depends on a hyperparameter grid search. This was a major challenge for this chapter of our research given the length of time it took to train this neural network for a large dataset and the computational resources available.
We extract the latent representation from the bottleneck layer (The \textit{RepeatVector} layer in the architecture above) after training. The \textit{RepeatVector} layer in Keras is the first layer of the decoder. It ensures that the input vector from the antecedent layer is passed into the decoder at each time step.

The \textit{TimeDistributed} layer at the end of this architecture is a wrapper that allows a layer to be applied to every temporal slice of the input. In \textit{5.3} the Dense layer is applied to every slice of shape (28799, 100). The Dense layer is a layer of neurons where each neuron receives input from all the neurons of previous layer.

2.3.2 Training the Autoencoder

The autoencoder is trained on all the preprocessed signal-level epoch outputs for all subjects with the correct length of wear time at baseline. We measure the reconstruction loss between the original input dataset and the reconstructed output from the decoder part of the autoencoder architecture. The binary cross-entropy loss function is used because our input values are scaled between 0 and 1. The autoencoder training phase is performed with backpropagation through time (BPTT), as is the traditional algorithm for training recurrent neural networks for updating weights in the network.

The Adam optimisation algorithm is used for learning in all the layers as it combines some of the best qualities of the other optimisation algorithms. Optimisation aims to minimise the loss function. The learning rate is 0.0001, chosen from a hyperparameter grid search from the range $[0.01, 0.001, 0.0001, 0.00001]$ to identify a good value for this learning rate.

Overfitting is prevented by stopping the training with early stopping algorithm in Keras. This stops the training process when the loss function has been minimised and no longer improves. The maximum number of epochs has been set to 100, which is usual standard practice.

Dropout is also introduced in the hidden LSTM layers of the autoencoder with a hyperparameter grid search at $(0.1, 0.25, 0.5)$ and then finely tuned for optimal performance. This regularisation technique often results in better accuracy [188]. The notion behind this technique is that random units are dropped with a probability $p$. This $p$ is the hyperparameter that is tuned in our neural network architecture. This technique results in many versions of the neural network being trained and the final resulting neural network could be seen as an averaging ensemble of the different combinations of neural networks. We do not apply dropout to the output bottleneck layer in the autoencoder.

$L_2$ and $L_1$ regularisation was also introduced in the hidden layers in different versions of the model when we were optimising. $L1$ regularisation is used to build a more sparse model by setting weights equal to 0. $L2$ regularisation is used to restrict the connection weights. This helps to deal with independent variables being highly correlated. We added these two types of regularisation as a hyperparameter search and exploration exercise. We evaluate the action of adding the regularisation on its
performance on the reconstruction loss function. These regularisation techniques were later removed after this architecture exploration exercise when it was found that it did not cause an improvement to the reconstruction loss function.

The activation function in the hidden layers was the hyperbolic tangent \( \tanh \). This choice was made because \( \tanh \) is zero centered and the gradients are not restricted to move in a certain direction as would be in the sigmoid activation function. Training was also conducted with the Rectified Linear Unit (ReLU) activation function, but the \( \tanh \) activation function was still found to perform better at minimising the reconstruction loss function. As a result, the \( \tanh \) activation function was used in our final architecture.

2.4 Supervised learning for regression on clinical target outcomes from latent representation

DIRECT collected and analysed clinical, molecular, biochemical, diet, exercise and MRI data from diabetic and pre-diabetic volunteer subjects from all over Europe. Vast repositories of data was collected that encode the volunteer subjects’ phenotypes: from multi-omics data (proteomics, genomics), medical imaging, blood tests, HbA1c results etc. This means there is a wide selection of potential clinical target outcomes to use in predictive modelling experiments. The objective of the DIRECT consortium was to use this data to help explore the heterogeneity of type 2 diabetes (T2D), and how this might impact diabetes outcomes like progression and drug response.

To evaluate and assess the usefulness of the latent representation extracted using the LSTM autoencoder, we can learn predictive models for various disease-related outcomes. We can evaluate the viability of the latent representation by assessing their performance in predicting these disease-related outcomes. Identifying the best clinical target outcomes is driven by domain knowledge.

2.4.1 Selection and preprocessing of clinical target outcomes

A subset of potential clinical target outcomes were selected from the wide array of phenotypes available in the IMI DIRECT consortium’s repositories. We chose these clinical target outcome variables based on the previous works in DIRECT [116].
<table>
<thead>
<tr>
<th>Target outcome</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘X2.h.OGIS’,</td>
<td>Oral Glucose Insulin Sensitivity index.</td>
</tr>
<tr>
<td>‘Matsuda’</td>
<td>Matsuda index, another index for insulin sensitivity.</td>
</tr>
<tr>
<td>‘Stumvoll’</td>
<td>Stumvoll index, another index for insulin sensitivity.</td>
</tr>
<tr>
<td>‘glu.sens’</td>
<td>Glucose sensitivity. The mean slope over the observed glucose range.</td>
</tr>
<tr>
<td>‘rate.sens’</td>
<td>Rate sensitivity. Represents the dependence of insulin secretion on the rate of change of glucose concentration</td>
</tr>
<tr>
<td>‘PFR1’</td>
<td>Potentiation factor ratio. A measure of beta cell function. Beta cells are cells in the pancreas responsible for insulin secretion. Individuals with Type-2 Diabetes have an increased resistance to insulin.</td>
</tr>
<tr>
<td>‘bmi’</td>
<td>Body mass index. A measure of a person’s classification of their weight based on their height-weight ratio.</td>
</tr>
<tr>
<td>‘taat’</td>
<td>Total abdominal adipose tissue. Adipose (fatty) tissue within the abdomen in its totality (Both intra-abdominal and abdominal subcutaneous)</td>
</tr>
<tr>
<td>‘iaat’</td>
<td>Intra abdominal adipose tissue. Adipose tissue within the abdomen area.</td>
</tr>
<tr>
<td>‘asat’</td>
<td>Abdominal subcutaneous adipose tissue. The adipose tissue you can feel if you pinch excess skin and tissue around the middle of the abdomen.</td>
</tr>
<tr>
<td>‘liver.fat’</td>
<td>Amount of adipose tissue around the liver. A liver with more fatty adipose tissue around it results in a higher risk of Type-2 Diabetes.</td>
</tr>
<tr>
<td>‘panc.fat’</td>
<td>Amount of adipose tissue around the pancreas. A fattier pancreas often results in higher risk of Type-2 Diabetes, or worse prognosis.</td>
</tr>
<tr>
<td>‘fasting.Glucose’</td>
<td>Fasting glucose. Glucose measured after an overnight fast.</td>
</tr>
<tr>
<td>‘ogtt.120.Glucose’</td>
<td>2-hour oral glucose tolerance test. Glucose level two hours after drinking oral glucose solution as part of test</td>
</tr>
<tr>
<td>‘age’</td>
<td>Age of volunteer subject at time of assessment.</td>
</tr>
<tr>
<td>‘gender’</td>
<td>Gender of volunteer subject.</td>
</tr>
<tr>
<td>‘IGR’</td>
<td>Impaired glucose regulation. Measures how much the blood glucose levels are above the normal range but are not high enough for the diagnosis of type 2 diabetes mellitus</td>
</tr>
</tbody>
</table>

Table 5.1: Table displaying the clinical target outcomes selected for supervised learning experiments based on clinical knowledge and literature review. There are 21 potentially useful clinical target outcomes.

Many of these clinical target outcomes had missing data which required imputation in the preprocessing steps. The proportion of data missing is shown in Table 5.2.
Table 5.2: Table displaying proportion of the clinical target outcome missing values for data

Table 5.2 shows that the target outcomes for fatty adipose tissue deposits in the abdomen, pancreas and liver had significant portions of data missing. This is due to the high levels of effort needed to collect this data using expensive and time-consuming procedures. So these features are discarded from the supervised learning experiments due to the high proportions of data missing. The rest of features are imputed using the $k$-nearest neighbour data imputation algorithm in Python’s scikit-learn as per section 2.4 in Chapter 4.

Although we know that HbA1c is a key indicator for Type-2 Diabetes as it is linked to blood glucose level, there was a significant portion of this data missing because of the difficulty in collecting this data at the time with a dried blood spot test which meant it was not fully complete at the baseline assessment time.

We have three different indices for insulin sensitivity (Stumvoll, Matsuda and OGIS). These three were available in DIRECT and we used these three to have a more diverse comparison of different ways of measuring the same clinical variable (insulin sensitivity). The choice of these three indices are also typical in the literature, such as in [148] that compared these indices to a reference, and also in [139] where the three indices were used as an indicator for vascular damage.
2.4.2 Predictive regression models: Linear model selections

Supervised learning experiments provides the foundation on which the performance of the latent representation as a suitable predictive set of features can be assessed. A suite of supervised learning models for regression are trained and then used as predictive models of clinical target outcomes. Regression models are appropriate since all the clinical target outcomes selected for this task are continuous variables.

Both linear models and ensemble regression models are chosen so that there is a diverse range of models being evaluated and thus model choice can be eliminated as a cause for poor performance, as this exercise aims to isolate the effect of using the latent representation learned from the autoencoders as the input set of features.

The first set of predictive regression models we used were linear models. Elastic net was preferred over the Lasso regression model because it is more robust in general. Elastic net performs variable selection which would automatically select the most useful of the latent representation’s variables to be used for predictions.

2.4.3 Predictive regression models: Ensemble approaches

The second set of regression models were ensemble models. Two of these models were chosen: Random forest regression and the gradient boosted regression.

The random forest regression model works in a similar way as described in Chapter 2. However, the key difference is that all the outputs of the randomised decision trees in the forest are now averaged, instead of majority voted on as is the case with classification tasks. The reasons for choosing the random forest regressor algorithm are similar to those as for classification. Again, the key hyperparameter to tune in this model are the number of trees and maximum number of features that are considered.

2.4.4 Model training and performance evaluation for regression predictive models

The cost function is the metric to evaluate the performance of a regression model. These functions calculate the error between the regression model’s predicted value and the ground truth value.

Two common metrics used for assessing the performance of regressions are the R2 score and the mean absolute error (MAE). These are used in evaluating the predictive power of the regression models in this set of experiments for the thesis. MAE represents the average of the absolute difference between the actual and predicted values. MAE, however, is not respective of scale of the dataset values. It measures the average of the residuals. The R2 score, or the coefficient of determination, represents the proportion of the variance in the dependent variable. It is scaled between 0 and 1.
Models are trained with hyperparameter tuning through a grid search and 10-fold cross-validation to minimise model overfitting. There was no key hyperparameter to tune for linear regression, but for the ridge and elastic net regression models, the key hyperparameter to tune were the regularisation strength $\alpha$. Only the results for the best regression experiments are reported in this thesis.

2.5 Comparing with available Physical Activity features from DIRECT dataset

The DIRECT group have hand-crafted their own set of features to represent physical activity. Their dataset uses the vector magnitude high-pass filter mean ($vm.hpf.mean$) and the Euclidean norm minus one mean ($enmo.mean$), which are widely used in other physical activity data analysis tools, such as GGIR. Thus, the performance of using the autoencoder latent representations can be used to build upon or to compare to these features as a baseline.

A separate suite of supervised learning experiments were undertaken using these two features as the input. This was to compare the handcrafted features in DIRECT to the autoencoder latent representations’ performance. The same models are chosen as in section 2.4.3 and section 2.4.2 in this chapter. The target outcomes are the same as those clinical variables highlighted in Table 5.1, so regression models are suitable for this analysis.

We also compare this autoencoder-generated latent representation to the granular high-level activity bout-level features from Chapter 4. We extract the granular high-level activity bout features for the DIRECT dataset, following the same process described in Chapter 4. The regression experiments are then also conducted using the high-level activity bout feature representation of physical activity to predict the clinical variables in Table 5.1.

3 Results and Discussions

3.1 Results from predictive modelling

The supervised learning experiments were conducted with results attained for the regression models. The metrics to evaluate the performance of these models are discussed in 2.4.4. We trained 180 regression models in total, where 60 regression models were trained for each set of features (the granular high-level activity bout features from Chapter 4, the autoencoder latent representation, and the DIRECT physical activity features as described in section 2.5). These predictive regression models were then evaluated and compared.
<table>
<thead>
<tr>
<th>Clinical target outcome</th>
<th>Linear regression</th>
<th>Ridge regression</th>
<th>Elastic net</th>
<th>Random forest</th>
<th>Gradient boosting</th>
</tr>
</thead>
<tbody>
<tr>
<td>OGIS</td>
<td>57.213</td>
<td>57.901</td>
<td>58.166</td>
<td>57.77</td>
<td>58.10</td>
</tr>
<tr>
<td></td>
<td>0.021</td>
<td>0.0021</td>
<td>-0.000854</td>
<td>0.0078</td>
<td>0.00052</td>
</tr>
<tr>
<td>Matsuda</td>
<td>2.146</td>
<td>2.151</td>
<td>2.160</td>
<td>2.186</td>
<td>2.145</td>
</tr>
<tr>
<td></td>
<td>0.016</td>
<td>0.003</td>
<td>-0.0015</td>
<td>-0.0490</td>
<td>-0.0093</td>
</tr>
<tr>
<td>Stumvoll</td>
<td>2.12</td>
<td>2.14</td>
<td>2.15</td>
<td>2.14</td>
<td>2.14</td>
</tr>
<tr>
<td></td>
<td>0.016</td>
<td>0.0051</td>
<td>-0.00026</td>
<td>0.0058</td>
<td>0.00035</td>
</tr>
<tr>
<td>Glucose sensitivity</td>
<td>41.90</td>
<td>41.93</td>
<td>41.88</td>
<td>42.11</td>
<td>41.57</td>
</tr>
<tr>
<td></td>
<td>-0.0033</td>
<td>0.00051</td>
<td>-0.00018</td>
<td>-0.0112</td>
<td>0.002</td>
</tr>
<tr>
<td>Rate sensitivity</td>
<td>564.54</td>
<td>562.90</td>
<td>565.43</td>
<td>577.46</td>
<td>561.3</td>
</tr>
<tr>
<td></td>
<td>-0.0065</td>
<td>-0.0008</td>
<td>-0.00044</td>
<td>-0.060</td>
<td>-0.0032</td>
</tr>
<tr>
<td>Potentiation factor ratio</td>
<td>0.468</td>
<td>0.465</td>
<td>0.0465</td>
<td>0.464</td>
<td>0.464</td>
</tr>
<tr>
<td></td>
<td>-0.016</td>
<td>-0.0045</td>
<td>-0.0028</td>
<td>-0.0073</td>
<td>-0.0073</td>
</tr>
<tr>
<td>Body mass index</td>
<td>3.20</td>
<td>3.23</td>
<td>3.25</td>
<td>3.21</td>
<td>3.24</td>
</tr>
<tr>
<td></td>
<td>0.015</td>
<td>0.001</td>
<td>-0.0007</td>
<td>0.00046</td>
<td>-0.0011</td>
</tr>
<tr>
<td>Fasting glucose</td>
<td>0.76</td>
<td>0.76</td>
<td>0.77</td>
<td>0.76</td>
<td>1.80</td>
</tr>
<tr>
<td></td>
<td>-0.005</td>
<td>-0.0019</td>
<td>-0.0096</td>
<td>-0.003</td>
<td>0.00082</td>
</tr>
<tr>
<td>2-hour glucose</td>
<td>1.79</td>
<td>1.78</td>
<td>1.79</td>
<td>1.77</td>
<td>1.78</td>
</tr>
<tr>
<td></td>
<td>0.0008</td>
<td>0.0031</td>
<td>-0.0003</td>
<td>0.0018</td>
<td>0.0067</td>
</tr>
<tr>
<td>Age</td>
<td>3.90</td>
<td>3.93</td>
<td>3.96</td>
<td>3.96</td>
<td>3.96</td>
</tr>
<tr>
<td></td>
<td>0.032</td>
<td>0.004</td>
<td>-0.00018</td>
<td>-0.0028</td>
<td>0.002</td>
</tr>
<tr>
<td>Impaired glucose regulation</td>
<td>0.46</td>
<td>0.45</td>
<td>0.45</td>
<td>0.45</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>-0.002</td>
<td>-0.0017</td>
<td>-0.003</td>
<td>-0.012</td>
<td>-0.003</td>
</tr>
<tr>
<td>Waist circumference</td>
<td>8.21</td>
<td>8.25</td>
<td>8.27</td>
<td>8.25</td>
<td>8.25</td>
</tr>
<tr>
<td></td>
<td>0.006</td>
<td>-0.0015</td>
<td>-0.001</td>
<td>-0.0028</td>
<td>-0.0077</td>
</tr>
</tbody>
</table>

Table 5.3: Supervised regression results using the LSTM autoencoder latent representation as training data, with $R^2$ score and MAE reported.

Table 5.3 displays the results for $R^2$ and MAE scores using the latent representation generated by the LSTM autoencoder. The best results for each of the clinical target outcomes for each regression model is highlighted in **bold**. The linear regression model appears to perform the best when compared to the other regression models in this set of experiments. Even the best highlighted results for both MAE and $R^2$, however, are very poor, regardless of model choice when using this set of features to feed train and predict clinical target outcomes. The latent representation doesn’t appear to serve as a useful predictor for any of the listed clinical target variables, despite the models being optimised and all necessary preprocessing steps having been carried out.

It is not clear at this stage whether this poor level of performance is due to the features themselves or the data itself does not support the clinical hypotheses that were made at the start of this thesis. In order to validate this assumption, it reinforces the need to compare this feature with other training data and evaluating the performance of other feature sets. This autoencoder architecture may not capture...
any useful behaviour patterns in its hidden representations.

The high-level granular bout-based representation for physical activity is more interpretable and domain-driven, so could produce a better result for predictive performance.

<table>
<thead>
<tr>
<th>Regression model</th>
<th>Clinical target outcome</th>
<th>Linear regression</th>
<th>Ridge regression</th>
<th>Elastic net</th>
<th>Random forest</th>
<th>Gradient boosting</th>
</tr>
</thead>
<tbody>
<tr>
<td>OGIS</td>
<td>MAE: 36.83</td>
<td>36.83</td>
<td>40.59</td>
<td>34.92</td>
<td>52.24</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R2: 0.54</td>
<td>0.54</td>
<td>0.48</td>
<td>0.58</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td>Matsuda</td>
<td>MAE: 1.87</td>
<td>1.87</td>
<td>1.97</td>
<td>1.8</td>
<td>2.06</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R2: 0.20</td>
<td>0.20</td>
<td>0.13</td>
<td>0.20</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>Stumvoll</td>
<td>MAE: 1.85</td>
<td>1.85</td>
<td>1.99</td>
<td>1.82</td>
<td>2.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R2: 0.19</td>
<td>0.19</td>
<td>0.11</td>
<td>0.20</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>Glucose sensitivity</td>
<td>MAE: 39.79</td>
<td>39.79</td>
<td>40.13</td>
<td>40.04</td>
<td>40.86</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R2: 0.07</td>
<td>0.07</td>
<td>0.06</td>
<td>0.056</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>Rate sensitivity</td>
<td>MAE: 565.55</td>
<td>565.55</td>
<td>565.55</td>
<td>570.55</td>
<td>561.26</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R2: -0.0011</td>
<td>-0.0011</td>
<td>-0.011</td>
<td>-0.019</td>
<td>-0.0007</td>
<td></td>
</tr>
<tr>
<td>Potentiation factor ratio</td>
<td>MAE: 0.457</td>
<td>0.457</td>
<td>0.465</td>
<td>0.448</td>
<td>0.46</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R2: 0.014</td>
<td>0.014</td>
<td>-0.003</td>
<td>0.042</td>
<td>0.0051</td>
<td></td>
</tr>
<tr>
<td>Body mass index</td>
<td>MAE: 3.08</td>
<td>3.08</td>
<td>3.18</td>
<td>3.02</td>
<td>3.18</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R2: 0.104</td>
<td>0.104</td>
<td>0.054</td>
<td>0.12</td>
<td>0.034</td>
<td></td>
</tr>
<tr>
<td>Fasting glucose</td>
<td>MAE: 1.28</td>
<td>1.32</td>
<td>1.51</td>
<td>1.32</td>
<td>1.65</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R2: 1</td>
<td>0.99</td>
<td>0.64</td>
<td>0.99</td>
<td>0.32</td>
<td></td>
</tr>
<tr>
<td>2-hour glucose</td>
<td>MAE: 3.74</td>
<td>3.74</td>
<td>3.86</td>
<td>3.75</td>
<td>3.89</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R2: 0.046</td>
<td>0.046</td>
<td>0.024</td>
<td>0.042</td>
<td>0.019</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>MAE: 0.24</td>
<td>0.24</td>
<td>0.43</td>
<td>0.16</td>
<td>0.38</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R2: 0.76</td>
<td>0.76</td>
<td>0.12</td>
<td>0.83</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>Impaired glucose regulation</td>
<td>MAE: 8.01</td>
<td>8.01</td>
<td>8.09</td>
<td>7.90</td>
<td>8.15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R2: 0.069</td>
<td>0.069</td>
<td>0.048</td>
<td>0.081</td>
<td>0.025</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.4: Supervised regression results using the granular bout-based high-level activity features as training data, with R2 score and MAE reported.

Table 5.4 shows the granular bout-based high-level activity features perform better than using the latent representation learned from the LSTM autoencoder. However, the improvement in performance is modest as the regression models do not have high $R^2$ scores.

The results for the final set of experiments use the $vm.hpf.mean$ and $enmo.mean$ as the training input data are reported in Table 5.5.
Table 5.5: The results for supervised regression models are displayed for all selected clinical target outcomes. R² scores and MAE are reported. These are the results of the predictive models that used the DIRECT hand-crafted `vm.hpf.mean` and `enmo.mean` variables as the feature set.

This set of features also do not perform very well as predictors of the selected clinical target outcomes, in both $R^2$ and MAE scores. The results produced by these supervised regression experiments clearly demonstrate the low level of predictive power across different feature sets that represent physical activity. In fact, the autoencoder latent representation performs the weakest out of the three representations. This allows us to assess that the latent space is not a viable set of features for predicting the selected Type-2 Diabetes related clinical target outcomes on its own. However, none of the available physical activity representations appear to perform strongly in predicting the selected target outcomes.

We judge the predictive performance of the latent space, and the granular bout-based features, as satisfactory, by comparing their R² score performance metric against the R² score performance metric of the physical activity features available from DIRECT (`vm.hpf.mean` and `enmo.mean`). The performance of the autoencoder...
latent space, and the granular bout-based features, should no less than 5% of the DIRECT physical activity features’ performance and ideally well exceed their levels. This condition is not satisfied.

3.2 Principal Findings

The standard steps were taken to ensure that the representation learning process with autoencoder architectures were carried out optimally to derive the best possible latent space, as validated by the minimised loss function. However, the use of this latent representation to learn predictive models for clinical Type-2 Diabetes related outcomes did not bear high levels of performance.

Two other representations of physical activity: the physical activity features that were handcrafted and taken from DIRECT, described in section 2.5 and the high-level activity bout-based features described in Chapter 4 extracted for DIRECT, did not perform substantially better than the latent representation when they were used to learn supervised regression models for predicting the selected clinical disease outcomes.

The DIRECT hand-crafted features and the high-level activity bout-based features are more interpretable and driven by domain knowledge. From this comparative analysis of the three feature sets, we can conclude that none of these physical activity representations appear to be a suitable predictor for clinical target outcomes related to Type-2 Diabetes. This is based on the results of supervised learning experiments and evaluated in terms of $R^2$ and MAE, regardless of regression model choice.

No single representation outperformed another to a significant degree that would justify the judgement that one representation is decisively and conclusively the best.

3.3 Implications of Study

The results from this chapter demonstrate the unsuitability and impracticality of physical activity features by themselves as a useful and powerful predictor of disease outcomes for Type-2 Diabetes. This can be concluded because a comprehensive set of experiments have been conducted in this chapter with different representations of physical activity that all validate this claim. Each of the physical activity feature spaces used in our research measure different aspects of physical activity: from traditional signal-vector magnitude and ENMO, to more high-level activity classification. None of these seemed to demonstrate a clear predictive ability for the selected Type-2 Diabetes related clinical target outcomes. A reason for this could be that poor levels of physical activity and sleep, although closely associated with Type-2 Diabetes disease, may only be one aspect of developing Type-2 Diabetes but other variables, such as diet and genomics, may perform a greater role in developing the disease. Our research focused only on physical activity feature spaces which may be too narrow in scope to build more accurate predictive models for Type-2 Diabetes.
specific outcomes. The relationship between the physical activity feature spaces and the selected Type-2 Diabetes target outcomes may be too weak or non-existent, to independently establish a high level of predictive ability.

However, as the work conducted in this chapter with the deep autoencoder networks did produce representations that minimise the loss function, it provides another avenue for exploration into whether these embedded spaces carry an interpretable and capture a meaningful description of the DIRECT volunteer subjects’ physical activity and sleep patterns which could not be directly deduced from the raw accelerometer traces.

3.4 Limitations

There are some limitations that are recognised in the work conducted in this chapter. Firstly, it makes a heavy reliance on the Oxford accelerometry analysis tool’s outputs, as was the case in Chapter 4. Although this tool is also widely used in other related research works, it is only one tool out of many. Being too dependent on this tool means that the results are vulnerable in case the tool proves to have a fundamental flaw later in the future. However, the steps taken by the Oxford accelerometer analysis tool are standard steps taken by most other Human Activity Recognition frameworks in terms of the choice of signal-level features extracted. So, the Oxford accelerometer analysis tool was chosen given its well-established presence in the literature and provide a reasonable case study for experimentation. For future work, we could overcome this limitation by seeking alternative accelerometer analysis tools or by developing our own technology to achieve this.

The work in this chapter also depends heavily on one dataset for training the models: The physical activity dataset collected by DIRECT. Although this dataset is robust and strictly controlled in terms of protocol, it is only one resource. Given the dataset was solely collected from populations domiciled in mostly Northern and Continental European countries and regions so may be quite limited in capturing physical behaviour, habits and patterns mostly associated with lifestyles in this region of the world. Furthermore, another obvious limitation in this dataset is that the collection wear time only spans 10 consecutive days. This is a very short period of time for conducting a longitudinal characterisation of an individual’s physical activity and sleep patterns, especially when other factors such as life events, or psychological impacts of wearing the device are not considered. After our research work, it could be concluded that the experimental design may have been flawed given the dataset that was available.

Furthermore, another limitation was that only three different physical activity representations were compared which may or may not have been the optimal choice as assumed. A different autoencoder architectures could have produced a latent representation that was more powerful in predicting the target clinical outcomes. However, this would be entirely a data-driven and unsupervised exercise and may not bear results that correspond to the hypotheses made given that none of the available representations for physical activity were good input features to predictive regression
models, and also that none of the representations appeared to outperform the others substantially.

Finally, there was little investigation into the minimum amount of days/time needed for a stable measure of physical activity as the study rather exhaustively only included those with full sequences of the standardised accelerometer wear time length (10 days in IMI DIRECT) which may have cut out potentially viable samples from the feature learning process and made our features more robust. This again stretches the need for good data collection and quality control protocols.

### 3.5 Conclusions

This chapter motivates further work into how physical activity can be used to digitally phenotype a person’s characteristics. It has also demonstrated the limitation of using different representations, both domain knowledge-driven and totally unsupervised learning driven, in predicting clinical disease outcomes on their own. Accelerometry may not be the optimal digital signal of choice for predicting disease outcomes related to chronic, metabolic diseases with slow progression such as Type-2 Diabetes. It may be useful in combination with other examples of clinical phenotypes. This highlights the importance of some of the data science challenges that need to be addressed before digital technologies can be used to realise a predictive and personalised vision of healthcare.
Chapter 6

Analysis of longitudinal physical activity and clinical progression over time
# 1 Introduction

This chapter makes use of the additional accelerometer data that was collected at various time points following the baseline collection period. It is a exploratory, longitudinal study of the volunteer subjects in the DIRECT accelerometer dataset. The subjects were followed up after a period of time elapsed since they wore the accelerometer device at the baseline assessment period. We use the follow-up accelerometer data to investigate if there is a significant change in behavioural patterns between the baseline and the data collection efforts at later points in time, and whether these changes correspond to clinical variables and disease outcomes for Type-2 Diabetes.

The hypothesis that this chapter investigates is that the patterns of progression in clinical phenotypes for Type-2 Diabetes are correlated with the progression in physical activity progression over two different time points.

# 2 Motivation and Background

Type-2 Diabetes is a slow-moving, multi-faceted, chronic and progressive metabolic disease[69]. It is a disease that is known to be associated with increasing the risk of developing or suffering a worse prognosis for other diseases, [84, 149, 141]. It is already known that low levels of physical activity, poor sleep and a sedentary lifestyle are associated with the development of Type-2 Diabetes [122, 109, 204]. Physical activity is often advised to prevent or to manage Type-2 Diabetes after diagnosis.

This chapter explores this relationship by exploiting another source of data of the DIRECT's accelerometer datasets: The follow-up wear time data for the volunteer subjects of the accelerometer cohort. DIRECT has also deposited phenotypes in its data matrices that describe an individual’s progression in their Type-2 Diabetes disease state, in terms of their HbA1c and fasting glucose [17] progression over time. It is also known that changing lifestyle can affect the impact that Type-2 Diabetes has on people already confirmed to have the disease [193], and physical activity is known to help manage the disease. Given this existing work in identifying biochemical biomarkers of progression in Type-2 Diabetes, this can be built upon with our research by using physical activity representations to explore if physical activity also deteriorates at a rate which corresponds to the deterioration in clinical phenotypes. Determining the rate of change in physical activity and sleep levels across a significant period of time can be useful in determining whether accelerometer-derived physical activity data is a useful collection effort. If the change in physical activity levels is large across two time points, and this is reflected in the change in clinical phenotypes, then it directly demonstrates an association between the two variables. If the change between these two variables do not correspond to each other, and the levels of physical activity vary little across different time points, then it suggests that the link between clinical phenotypes for Type-2 Diabetes and physical activity may be very weak and that other biomarkers
need to be considered.

3 Materials and Methods

3.1 DIRECT dataset follow-ups

The work for this chapter makes use of the follow-up accelerometer data. These are the subsequent collection efforts for DIRECT accelerometer volunteer subjects across different periods of time. The number of individuals at each time point, for each cohort at each assessment centre, can be seen in Table 2.1. The individuals identified being in a pre-diabetes state, Cohort 2.1, were followed up at 18 months, and then 48 months after the baseline assessment. Those identified as having been diagnosed with Type-2 Diabetes, Cohort 2.2, were assessed at 18 months, and then 36 months after the baseline assessment. However, the number of volunteer participants decreased steadily over the total study collection period.

These data files undergo the same feature extraction process with the Oxford accelerometer analysis tool, as in Chapter 5. All the samples of follow-up raw accelerometer are processed with the Oxford accelerometer analysis tool described in Chapter 4. The same features at the signal-level epoch level and the labelled activity sequences are extracted.

After the Oxford accelerometer analysis tool was used to extract the signal-level epoch features and labelled activity recognition sequences, the framework developed in Chapter 4 was used to, again, extract the high-level activity bout-based features in a High Performance Computing environment.

3.1.1 Representation learning for follow-up data

The model trained on the first baseline set of accelerometer data is used to learn the latent representation for the later time points in the follow-up accelerometry datasets.

3.2 Calculating the similarity of physical activity over different time points

One of the main challenges in this chapter is to represent the magnitude of change in physical activity of an individual volunteer subject to the same individual’s physical activity at a later point in time when their accelerometer was collected again.

Both the latent representations learned from the autoencoder and the high-level granular bout-based features from Chapter 4 are represented as multidimensional feature vectors in a $n$-dimensional Euclidean space.
Therefore, it is possible to calculate the distance between two data points in this \( n \)-dimensional Euclidean space.

Euclidean distance is used widely in machine learning and data science for calculating the distance between two observations, such as in \( k \)-means clustering and hierarchical clustering for identifying similar points \([145]\), and as a metric for similarity \([184]\). Using the Euclidean can risk the \textit{curse of dimensionality}. However this is reconciled in this study as the number of samples (1,161 for Cohort 2.1 and 503 for Cohort 2.2) far outnumbers the number of latent features generated by the autoencoder (25).

As a result, the Euclidean distance is calculated for individual volunteer subjects between the time of their baseline assessment and the later points in time when more accelerometer data was collected. This calculated distance value gives an indication between the similarity of physical activity at different time points.

However, this metric does not give an indication as to which \textit{direction} the similarity occurs i.e is an individual more active at time point \( t_1 \) than at time point \( t_2 \), or less active. The Euclidean distance only tells you the absolute difference of physical activity levels between two time points i.e a measure of the magnitude of the similarity only. But this is not an issue that required consideration for this chapter’s work because the objective was to identify if there is a direct association between the difference in the clinical phenotypes in \([3,3]\) and the physical activity levels at two different time points. This problem can be approached by identifying and only calculating the similarity of the clinical phenotype and physical activity levels. The task is then to determine whether the magnitude of the similarity in the clinical phenotypes and the physical activity performance similarity, are broadly symmetric in the data itself.

There are other methods to calculate distances in \( n \)-dimensional Euclidean space, such as Minkowski distance, Cosine distance and Manhattan distance. However, Euclidean distance was chosen to measure the similarity at different time points because it is the most standard method for calculating similarity distance. Additionally, the data that we are calculating similarity do not suffer high dimensionality to the extent where it would render calculating this distance meaningless \([218]\).

Dimensionality reduction techniques such as PCA and t-SNE were not applied on the physical activity representations developed in our research so far since the autoencoder already performed dimensionality reduction when learning its features on the signal-level epoch features. In the case of the high-level activity bout-based features, dimensionality reduction was not applied because it was desirable to maintain the ability to interpret the features, which techniques such a PCA and t-SNE would eliminate.

If this result is reflected after the experiments in this chapter, then it would be an approach to demonstrate the value in collecting digital signals from accelerometer devices to identify digital biomarkers of Type-2 Diabetes progression.
3.3 Phenotypes for clinical disease progression

It is important to have a metric to determine the magnitude of similarity in the progression of an individual with Type-2 Diabetes at one time point to the next time point in the accelerometer data collection effort. The same procedure as outlined in section 3.2 was used as similarity metric between time points. The variables which would indicate clinical disease progression for Type-2 Diabetes were selected for this task. These variables are the same as those in Bizzotto et al’s work [17] which are the three indices for insulin sensitivity (OGIS, Matsuda and Stumvoll), body mass index, HbA1c. The fasting plasma glucose (FPG) is also selected, also using work by Bizzotto et al.

For each of the accelerometer cohort subjects, for whom we have the required baseline and follow-up accelerometer data, the absolute difference for each of the clinical variables at baseline and each of the later follow-up time points are calculated. This, as in 3.2 above, is a simple metric for the scale of similarity between two data points at two different time points. The size of the difference would be proportional to the scale of similarity. The absolute difference is calculated, instead of the Euclidean distance, because each clinical variable is one single variable and considered independently of the others.

Rate of glycaemic deterioration can often be determined by looking at the change in HbA1c haemoglobin [59]. This variable was described in Table 5.1 as captured by DIRECT. HbA1c is useful as it gives the three-month average of blood glucose levels. This gives an indication into the glycaemic control of individuals with Type-2 Diabetes, or as a diagnostic tool [1]. This has been widely established in Type-2 Diabetes literature.

DIRECT has focused its study on glycaemic deterioration over time on the HbA1 variable [17, 116, 219]. Their efforts in measuring the deterioration of HbA1c was achieved using a conditional linear mixed effects model [17]. They also considered and adjusted the HbA1c progression phenotype for changes in body mass index and diabetes medications. These two traits are assumed to have the greatest proportionality to the progression in HbA1c. They investigated the role of physical activity through the \( hpf.vm.mean \) variable described in Chapter 5 to determine if physical activity had an independent role. The progression phenotypes for fasting glucose and HbA1c developed by DIRECT are also used in the analysis conducted for the work in this chapter as a baseline on which to compare the results of our own representations of DIRECT’s accelerometer cohorts.

However, this effect was not observed in their limited study of the effect of physical activity since they only used the single variable \( hpf.vm.mean \) as a representation of physical activity. Furthermore, this approach is more complicated and is driven by clinical domain knowledge when developing the linear mixed effects models to calculate progression. The representations developed in Chapter 4 and Chapter 5 are different ways of representing different aspects of physical activity.

\[ \text{Use of Glycated Haemoglobin (HbA1c) in the Diagnosis of Diabetes Mellitus Abbreviated Report of a World Health Organization Consultation: https://www.ncbi.nlm.nih.gov/books/NBK304267/} \]
activity which could demonstrate a proportional relationship between disease progression change and physical activity change. This is explored in this chapter and is compared to method DIRECT used to measure the progression in disease over time

3.4 Determining the relationship between physical activity change and change in clinical variables

3.4.1 Correlation coefficient analysis

There are a variety of ways in which the physical activity change and the change in the selected clinical variables can be demonstrated to correspond and be associated with one another.

This problem can be viewed as determining dependence on bivariate data, thus standard correlation analysis methods can be applied as an initial starting point. The first approach would be to calculate correlation coefficients such as Pearson correlation coefficient, or just Pearson’s $\rho$, and Spearman’s rank correlation coefficient, or just Spearman’s $\rho$.

Pearson’s $\rho$ calculated by the equation in (6.1) for a given pair of random variables $(X,Y)$, where $cov$ is the covariance of $X$ and $Y$. $\sigma_X$ is the standard deviation for $X$ and $\sigma_Y$ is the standard deviation for $Y$.

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \quad (6.1)$$

The results for this coefficient is always bounded between -1 and +1, where -1 implies a perfectly negative correlation and +1 implies a perfectly positive correlation, 0 meaning no correlation. However, this measure only works for linear relationships between $X$ and $Y$, and not other types of relationships.

The other correlation coefficient is the Spearman rank correlation. This is a nonparametric measure of dependence between the rankings of two variables. It evaluates how well two variables can be described using a monotonic function, a function which preserves a given order of rankings. This correlation coefficient is similar to the Pearson’s $\rho$, but instead of taking the random variables $X$ and $Y$ on their own, it measures the correlation between the rank variables of $X$ and $Y$, $R(X)$ and $R(Y)$.

For a sample of size $n$, the $n$ data points $X_i$ and $Y_i$ are converted to rank variables $R(X_i)$ and $R(Y_i)$, and the Spearman correlation is calculated as the following:

$$\rho_s = \rho_{R(X)R(Y)} = \frac{cov(R(X)R(Y))}{\sigma_{R(X)} \sigma_{R(Y)}} \quad (6.2)$$

Spearman’s correlation coefficient where $cov$ is the covariance of $R(X)$ and $R(Y)$. $\sigma_{R(X)}$ is the standard deviation for $R(X)$ and $\sigma_{R(Y)}$ is the standard deviation for $R(Y)$.
The fundamental difference between Pearson’s coefficient and Spearman’s coefficient is that Pearson coefficient works well with linear correlations between two variables, whereas Spearman’s coefficient works well with monotonic relationships \[142\]. The linear relationship between two relationships move in the same direction at a constant rate, whereas monotonic relationships between two variables tend to move in the same relative direction but not necessarily at a constant rate. In our analyses, we used both Spearman and Pearson coefficients to have an insight into the relationship between clinical phenotype progression and the progression in physical activity. This is repeated for high-level activity bout features from Chapter 4 and the latent representation produced by the autoencoder in Chapter 5.

### 3.4.2 Summary of Approach

To summarise, we have developed a metric to calculate the similarity between individual’s physical activity at one time point to their physical activity at another time point. This is achieved by calculating the Euclidean distance between these paired time points for the autoencoder latent representation developed in Chapter 4 and 5.

We then measure the similarity between variables that are known to represent the clinical progression of Type-2 Diabetes, a subset of those from Table 5.1 in Chapter 5 by calculating the absolute difference between two instances of a variable at each time point.

To evaluate whether the clinical progression for Type-2 Diabetes is then reflected in the physical activity over two time points, we then perform a correlation analysis using standard statistical techniques. The hypothesis is that a significant change in the clinical progression of Type-2 Diabetes disease between two time points is then reflected and shown by correlation analysis in the physical activity change between the same two time points.

### 3.5 Analysis of distribution of physical activity features over time

Correlation analysis may demonstrate that physical activity and clinical progression of Type-2 Diabetes are reflected. However, we can also study physical activity changes on their own to evaluate how physical activity patterns change significantly over time. This is implemented by an exploratory analysis by inspecting the distributions of physical activity types in one time point with the distributions of physical activity types in a later time point.

We can use the granular bout-based high level activity features from Chapter 4 to visualise and explore how much physical activity varies between two time points. This involves a distribution analysis of the high-level granular bout-based physical activity features from Chapter 4 and also the latent representation generated by the autoencoder architecture in Chapter 5 over two time points. If there is little difference over two time points, it demonstrates that physical activity does not change much,
especially if this change is negligible relative to the changes in Type-2 Diabetes progression (determined through the clinical progression phenotypes).

The distribution of the granular bout-based physical activity features are represented in histograms summarising these features to the scale of 24-hour days, rather than distinctly defined periods of the day as in Chapter 4. The features were brought up to this scale of time because it attains a more generalised habitual behaviour pattern on a longer time scale. They were used to summarise a person’s physical activity and sleep patterns. This can be demonstrated by observing correlations between the differences in physical activity across two time points and also inspecting the histograms between the physical activity distributions across two time points.

This would build upon and reaffirm the conclusions reached by Bizzotto et al [17] when they used hpf.vm.mean as the proxy feature for physical activity. Bizzotto et al identified that physical activity did not have an independent role. Our results are a more comprehensive confirmation of this observation since more aspects of physical activity (for example activity bouts, labelled activity types, latent features from deep autoencoders) are all considered.

3.6 Analysis from most extreme individual clinical and physical activity changes

The physical activity features are quite normally distributed, which is confirmed by performing statistical normality tests such as D’Agostino, Shapiro and Anderson tests. To perform these normality establishment tests, we project the high-level daily summary physical activity features from Chapter 4 onto the latent representation for the signal-level epoch accelerometer data extracted by the autoencoder in Chapter 5. We then perform the normality tests on a random sample of 40 subjects on four population samples: two at baseline for Cohort 2.1 and Cohort 2.2, and then for 48 months after baseline for Cohort 2.1 in Table 6.4, and for 36 month after baseline for Cohort 2.2 in Table 6.3.

<table>
<thead>
<tr>
<th>Feature</th>
<th>D’Agostino</th>
<th>Shapiro</th>
<th>Pearson</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of time spent performing MVPA daily</td>
<td>Not normal</td>
<td>Not normal</td>
<td>Not normal</td>
</tr>
<tr>
<td>Percentage of time spent performing sedentary activity daily</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
</tr>
<tr>
<td>Percentage of time spent performing light tasks daily</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
</tr>
<tr>
<td>Average number of hours spent sleeping daily</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
</tr>
<tr>
<td>Average daily MET level (Energy expenditure)</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
</tr>
</tbody>
</table>

Table 6.1: Table demonstrating the results for the tests for normality in each of the high-level physical activity summary features at baseline for Cohort 2.1.
<table>
<thead>
<tr>
<th>Feature</th>
<th>D’Agostino</th>
<th>Shapiro</th>
<th>Pearson</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of time spent performing MVPA daily</td>
<td>Not normal</td>
<td>Not normal</td>
<td>Not normal</td>
</tr>
<tr>
<td>Percentage of time spent performing sedentary activity daily</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
</tr>
<tr>
<td>Percentage of time spent performing light tasks daily</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
</tr>
<tr>
<td>Average number of hours spent sleeping daily</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
</tr>
<tr>
<td>Average daily MET level (Energy expenditure)</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
</tr>
</tbody>
</table>

Table 6.2: Table demonstrating the results for the tests for normality in each of the high-level physical activity summary features at baseline for Cohort 2.2

<table>
<thead>
<tr>
<th>Feature</th>
<th>D’Agostino</th>
<th>Shapiro</th>
<th>Pearson</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of time spent performing MVPA daily</td>
<td>Not normal</td>
<td>Not normal</td>
<td>Not normal</td>
</tr>
<tr>
<td>Percentage of time spent performing sedentary activity daily</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
</tr>
<tr>
<td>Percentage of time spent performing light tasks daily</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
</tr>
<tr>
<td>Average number of hours spent sleeping daily</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
</tr>
<tr>
<td>Average daily MET level (Energy expenditure)</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
</tr>
</tbody>
</table>

Table 6.3: Table demonstrating the results for the tests for normality in each of the high-level physical activity summary features 36 months after baseline for Cohort 2.2
Table 6.4: Table demonstrating the results for the tests for normality in each of the high-level physical activity summary features 48 months after baseline for Cohort 2.1

From these tests, it could be the case that a correlation between the progression of physical activity, and also clinical progression phenotypes can only be observed by the individuals who exhibit the greatest change in physical activity or clinical progression. From this assumption, therefore, we can conduct an analysis of distribution of physical activity for the individuals at the furthest edges of the distribution curves that demonstrate the largest extent to which clinical disease progression and physical activity change are associated and then assess and validate if there is any dependence between these two measures.

To identify those with the greatest clinical progression change, the DIRECT clinical variables are used again. The individuals are sorted in ascending order according to the size of their clinical progression change for HbA1c. The 20 individuals at the top of this ordered set of individuals are analysed for their summarised granular high-level activity bout features, extracted as described in 3.5. Only 20 individuals are chosen due to the smaller population size available to conduct this study fully (with the full follow-up data across time points that are required).

The hypothesis is that if there is an associated change between clinical disease progression and physical activity, then the distribution of summarised high-level activities of these selected subsample of individuals would, in theory, be the most pronounced relative to all the other individuals that fit into the normal part of the distribution.
4 Results and Discussion

4.1 Correlation analysis results

The similarity of physical activity at two time points was measured by calculating the Euclidean distance at each time point for the physical activity representations. To define and determine the dependence between physical activity change and the clinical progression change, correlation analysis is performed to determine this.

The results for Pearson and Spearman correlation coefficients are displayed in Tables 6.6 and 6.5. This displays the results for calculating the change in physical activity using only the latent representations of the autoencoder defined in Chapter 5. The column “Number of pairs” in Tables 6.6 and 6.5 corresponds to how many longitudinal difference pairs there were. For example, there were 358 pairs of points between the baseline and 36 months after baseline for the BMI variable in the cohort of individuals with confirmed Type-2 Diabetes (Cohort 2.2).

<table>
<thead>
<tr>
<th>Clinical phenotype</th>
<th>Spearman</th>
<th>Pearson</th>
<th>Number of pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>HbA1c</td>
<td>-0.015</td>
<td>-0.028</td>
<td>358</td>
</tr>
<tr>
<td>Fasting glucose</td>
<td>0.074</td>
<td>0.087</td>
<td>358</td>
</tr>
<tr>
<td>BMI</td>
<td>0.032</td>
<td>0.045</td>
<td>358</td>
</tr>
<tr>
<td>OGIS</td>
<td>0.054</td>
<td>0.062</td>
<td>308</td>
</tr>
<tr>
<td>Matsuda</td>
<td>0.033</td>
<td>0.035</td>
<td>322</td>
</tr>
<tr>
<td>Stumvoll</td>
<td>0.059</td>
<td>0.065</td>
<td>312</td>
</tr>
</tbody>
</table>

Table 6.5: Table displaying the results for correlation analysis, using Spearman and Pearson correlation coefficients, between the Euclidean distance changes in the autoencoder latent representation of the physical activity from baseline to 36 months for those individuals belong to Cohort 2.2 (Those with confirmed diabetes)

<table>
<thead>
<tr>
<th>Clinical phenotype</th>
<th>Spearman</th>
<th>Pearson</th>
<th>Number of pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>HbA1c</td>
<td>-0.079</td>
<td>-0.066</td>
<td>273</td>
</tr>
<tr>
<td>Fasting glucose</td>
<td>-0.057</td>
<td>-0.054</td>
<td>273</td>
</tr>
<tr>
<td>BMI</td>
<td>-0.0089</td>
<td>-0.0077</td>
<td>273</td>
</tr>
<tr>
<td>OGIS</td>
<td>-0.038</td>
<td>-0.031</td>
<td>238</td>
</tr>
<tr>
<td>Matsuda</td>
<td>0.011</td>
<td>0.011</td>
<td>239</td>
</tr>
<tr>
<td>Stumvoll</td>
<td>0.021</td>
<td>0.024</td>
<td>238</td>
</tr>
</tbody>
</table>

Table 6.6: Table displaying the results for correlation analysis, using Spearman and Pearson correlation coefficients, between the Euclidean distance changes in the autoencoder latent representation of the physical activity from baseline to 48 months for those individuals belong to Cohort 2.1 (Pre-diabetes)

The results in Tables 6.5 and 6.6 demonstrate that there are weak correlations, when calculating with both Spearman and Pearson coefficients between the change in the physical activity and the change in the clinical progression for those selected variables. It can thus be concluded that, since both measures of statistical dependence demonstrate a similar result, the relationship between change in physical activity and clinical progression across two time points are very weak. This
implies there is no linear correlation. Although this result is disappointing, it further reaffirms the conclusion reached by the DIRECT study in [17] that physical activity plays a little independent role in progression of clinical outcomes. It reaffirms this conclusion by demonstrating that even when other representations of physical activity are considered, rather than only the \textit{hpf.vm.mean} physical activity variable, the results are similar in that it demonstrates the little effect the change in physical activity has on the clinical progression of Type-2 Diabetes.

4.2 Analysing the distribution of physical activities over time

The previous section shows that we were not able to establish a clear, strong correlation between progression of clinical phenotypes and their corresponding progression of physical activity over two time points. This section illustrates the results for assessing the distribution of different physical activity over two time points for the same set of selected individuals.

We use the high-level activity bout features from Chapter 4 to summarise individual physical activity patterns and interpret their behaviour before visualising the distribution of physical activities across two time points. One histogram can be plotted to represent the frequency density across the different values for each of the high-level granular bout-based activity features at baseline, and then a different histogram plotted for the other later time point. Some of these histograms are shown below.

Figure 6.1: Histogram showing distribution of moderate-to-vigorous activity levels for subjects with confirmed Type-2 Diabetes at baseline assessment.

Figure 6.2: Histogram showing distribution of moderate-to-vigorous activity levels for subjects with confirmed Type-2 Diabetes 36 months after baseline assessment.
Figure 6.3: Histogram showing distribution of percentage of time spent performing light tasks for subjects with confirmed Type-2 Diabetes at baseline assessment.

Figure 6.4: Histogram showing distribution of percentage of time spent performing light tasks for subjects with confirmed Type-2 Diabetes 36 months after baseline assessment.

Figure 6.5: Histogram showing distribution of percentage of time spent performing sedentary activity for subjects with confirmed Type-2 Diabetes at baseline assessment.

Figure 6.6: Histogram showing distribution of percentage of time spent performing sedentary activity for subjects with confirmed Type-2 Diabetes 36 months after baseline assessment.

Figure 6.7: Histogram showing distribution of number of hours asleep for subjects with confirmed Type-2 Diabetes at baseline assessment.

Figure 6.8: Histogram showing distribution of number of hours asleep for subjects with confirmed Type-2 Diabetes 36 months after baseline assessment.
Figure 6.9: Histogram showing distribution of MET levels for subjects with confirmed Type-2 Diabetes at baseline assessment.

Figure 6.10: Histogram showing distribution of MET levels for subjects with confirmed Type-2 Diabetes 36 months after baseline assessment.

Figure 6.11: Histogram showing distribution of percentage of time spent performing light tasks for subjects in pre-diabetes stage at baseline assessment.

Figure 6.12: Histogram showing distribution of percentage of time spent performing light tasks for subjects in a pre-diabetes stage 48 months after baseline assessment.

Figure 6.13: Histogram showing distribution of percentage of time spent performing moderate-to-vigorous activity for subjects in pre-diabetes stage at baseline assessment.

Figure 6.14: Histogram showing distribution of percentage of time spent performing moderate-to-vigorous activity for subjects in a pre-diabetes stage 48 months after baseline assessment.
From visual inspection alone, it is quite clear that most of these distributions, with the only exception being the histogram for moderate-to-vigorous activity in Figures 6.13, 6.14, 6.1 and 6.2, follow normal Gaussian distributions.
People performing moderate-to-vigorous physical activity tend to spend less than 4% of the day doing this intensity of physical activity. However, the distributions also do not change over two time points (baseline to 36 months for individuals in Cohort 2.2, baseline to 48 months for individuals in Cohort 2.1). Furthermore, distributions also do not vary greatly across Cohort 2.1 (pre-diabetes) and Cohort 2.2 (confirmed diabetes) by themselves for all features.

The distribution analysis demonstrates how changes in physical activity does not correspond to changes in the clinical phenotypes for progression. It could be deduced that individuals in both cohorts of the DIRECT dataset have a similar pattern of habitual behaviour with a relatively small number of outliers.

### 4.3 Analysing the distributions for the individuals with the greatest disease progression

It can be seen in the histograms that most of the granular bout-based activity features follow a distribution that is mostly normal or Gaussian, with the exception for the percentage of time spent in moderate-to-vigorous activity. The inference from this distribution is that most of the individuals behave similarly towards a normal. However, some individuals exhibit behaviours that place them at the tails and edges of this distribution. This is seen, for example, at the tails of the histograms in Figures 6.11 and 6.3.

The individuals with the greatest scale of disease progression can be identified and then their physical activity patterns can be explored over the same two time points. We first sort these individuals by the size of their progression for HbA1c between the two time points. Then, we plot the spider diagrams, such as the ones in Figures 6.22 and 6.21, helps to visualise the difference in habitual behaviour patterns over time for the individuals who have progressed the most in their disease state.
Figure 6.21: Spider plot for daily behaviour patterns for the individual in Cohort 2.2 (Type-2 Diabetes positive diagnosis cohort) with the highest change in HbA1c clinical progression at baseline assessment.

Figure 6.22: Spider plot for daily behaviour patterns for the individual in Cohort 2.2 with the highest change in HbA1c clinical progression at 36 months after baseline assessment.

The spider diagram above is the example of the individual calculated to have had the largest change in clinical progression based on DIRECT’s progression phenotype for HbA1c between two time points. The spider diagram in Figure 6.21 displays the daily average habitual behaviour distribution at baseline and Figure 6.22 is 36 months after Figure 6.21.

The distribution in daily habitual behaviour patterns varies only very slightly between the two time points. The most visible, albeit still modest, difference...
is observed for the average number of hours asleep, which has decreased, and percentage of time spent performing light tasks daily, which has increased slightly. However, the proportion of the day occupied by sedentary activity increases to nearly 60% for the assessment 36 months after baseline, whereas this was assessed to be less than 40% at baseline. The average number of hours asleep daily also appears to present some of the greatest change where the average number of hours decreases. This would correspond to the clinical literature regarding the effect of poor levels of sleep increasing the risk of developing Type-2 Diabetes or worsening its effects \[204,10\]. This may contribute to the deterioration in clinical progression.

Figure 6.23: Spider plot for summarised daily behaviour patterns for the individual in Cohort 2.2 close to the average progression in HbA1c clinical progression at baseline
When observing the spider plot for an individual with a value for the HbA1c progression phenotype developed by DIRECT that is close to the mean average of this value we observe a more evenly spread distribution for their daily average for physical activities. An example of such an individual is displayed in Figures 6.23 at baseline assessment and then 6.24 above for 36 months after baseline.

The summarised daily physical activity distributions displayed in the spider diagrams above are displayed for two individuals at different portions of the distribution curve for clinical progression phenotype demonstrate. This analysis could be expanded to the entire cohort.

For individuals with the greatest level of clinical progression, in terms of the DIRECT progression phenotypes, it is observed that individuals tend to spend the majority of their day in sedentary activity (50%-60%) and sleep (35%-45%), with a very small proportion (around 2%-5%) in moderate-to-vigorous activity. The average number of hours asleep daily seem to remain fairly stable between baseline and future assessments.

From this analysis of individuals with the largest progression in terms of clinical phenotypes, it can be observed that there is no substantial difference between habitual activity patterns over time. This analysis further reinforces the conclusion reached by DIRECT with regards to the very limited independent role that physical activity had in the deriving progression over time for Type-2 Diabetes disease.
4.4 Principal Findings

The set of experiments conducted for this chapter have evaluated the viability and usefulness of physical activity, as measured from accelerometer devices, in measuring the progression of Type-2 Diabetes disease. The original hypothesis was that physical activity and clinical disease progression were dependent on each other to an extent.

After conducting a comprehensive suite of different analyses to investigate this hypothesis, there was no strong dependence that was identified even when different representations of physical activity were used. All of the different analyses completed in this chapter further reinforce and demonstrate the conclusion that was reached by DIRECT in [17] where no dependent role was identified between progression in key disease biomarkers for progression such as HbA1c and fasting glucose and physical activity.

However, the experimental analyses completed in this chapter was able to demonstrate this in a different way compared to DIRECT’s work. This chapter showed this lack of dependence not just through the mean signal vector magnitude, which is the usual measure used to represent physical activity, but it was also able to analyse and get to this result with other aspects of physical activity. The bout-based, granular high-level representations were used as a way to summarise daily habitual behaviour patterns and then projected onto both cohorts at baseline and the furthest follow-up time point available. This approach helps to gain a more interpretable understanding of the physical activity patterns.

Also, this chapter measured the change in physical activity through the changes in the latent representations extracted from the autoencoder architectures in Chapter 5 and was able to demonstrate the very weak dependence between the size of change to the change in the progression of clinical phenotypes. This further reinforced this conclusion which refutes the original hypothesis.

Some key differences were identified in the proportions of sedentary activity and average number of hours spent sleeping in the progression over time.

4.5 Implications of Study

Although the results from the analysis conducted for this chapter did not affirm the original hypothesis we made, its implications further reinforce the knowledge in this field thus far by offering new insights through the new representations of physical activity. Work in the clinical literature have clearly established a key role for physical activity in the development of Type-2 Diabetes [40], but this was not established in the work completed in this chapter.

However, in the literature, the work does not make use of digital devices that objectively measure physical activity to explore this association. Longitudinal studies were warranted in [86] to study the long-term effect of physical activity and Type-2 Diabetes. Therefore this chapter’s results highlights the need to objectively
study this association through more work with objective physical activity data, and that this association should be further studied on a long-term basis. It may be that the progression of Type-2 Diabetes requires an even longer period of monitoring than from baseline to a maximum of 48 months after baseline. It may be that the deterioration of clinical progression can be more shown to be associated with physical activity through longer-term studies. It is already established that Type-2 Diabetes takes several years to progress [176]. Therefore, the longitudinal study into physical activity and Type-2 Diabetes should also correspond to this by monitoring over longer periods of time, particularly for cohorts such as those studied for the DIRECT and the UK Biobank who are beyond the age where Type-2 Diabetes tends to progress quicker than younger individuals [12].

### 4.6 Limitations

There may have been some limitations associated with conducting the analyses for this study. This chapter only considered physical activity representations without taking the variations of the time series into its analyses. Our representations eliminate the temporal dimension by summarising physical activities in the frequency domain. It is unknown whether performing an analysis with the time dimension would indeed be more powerful. Additionally, the method used to measure the change in physical activity over two time points may have been unsuitable and a more advanced metric may need to be devised that takes into account the directional change over time.

Furthermore, other limitations are related to the data that was used for this analysis. As in previous chapters, the free-living accelerometry data collected in the large epidemiological cohorts may need to be extended across greater time periods rather than just 10 days in a participant’s life in the case of DIRECT. This is a severe limitation of the dataset as it limited the scope of our study to just those 10 days where the participant wore the accelerometer device. Also, it could be that the individuals in the DIRECT study all exhibit similar behaviours that could not distinguished on a significant scale. This is clearly seen in the histograms in the results section of this chapter whereby most people fit within the bell of the curve.

Other confounding factors were not considered, such as geographic location of assessment centres, the time of the year in which accelerometers were worn by the subjects. It could be that these factors may explain the certain exhibitions of habitual behaviour but this was not established and could outline the future work that could be conducted for a study like this which would aid in establishing better protocols for self-monitoring, free-living data collection efforts.

Aspects of a participant’s health may be included in the decline of physical activity which are not however captured in the clinical variables that were selected as part of this study. This could be investigated further as part of future work.
4.7 Conclusions

In conclusion, this chapter explored the potential association between the longitudinal progression of clinical disease biomarkers over time with the longitudinal progression of performing physical activity. The hypothesis we made was that the two progressions would be correlated with one another. The results do reaffirm the conclusions reached in previous studies but also do so in a comprehensive approach that looks at the different ways to represent physical activity from earlier parts of our research, and progression phenotypes found in the existing literature. This demonstrates the limited predictive power that physical activity has on the case of Type-2 Diabetes, which is a slow-progressing, chronic and multi-faceted disease. Physical activity may be a useful complement to studying other biomarkers of disease that may have a more powerful impact Type-2 Diabetes. Furthermore, the lack of change in physical activity over the time period where this was monitored may suggested that longitudinal studies need to be extended over longer periods.
Chapter 7

Unsupervised clustering from representations of accelerometer data
1 Introduction

The results from the supervised learning experiments for predicting clinical target outcomes, as seen in Chapter 5, did not yield results with very high F1 scores or high AUC scores. This shows that the representations in both the granular bout-based feature set and the embedded space generated by the LSTM autoencoder do not demonstrate a very high performance for predicting outcomes related to Type-2 Diabetes. The results from the longitudinal progression study in Chapter 6 also reaffirmed this weak link between clinical Type-2 Diabetes biomarkers and physical activity.

This chapter will investigate the hypothesis that latent representations learned by the autoencoder can form a description of the physical activity traces not captured by domain knowledge-based feature extraction by being able to be clustered into groups with properties exhibiting differences in behaviour. The raw accelerometer data from the DIRECT study’s collection efforts is used again for the training in this chapter.

This chapter stipulates that the latent representations of the accelerometer data are able to be used as a tool to distinctly segregate individuals into groups and find natural patterns of behaviour within groups using unsupervised machine learning methods, namely methods for clustering individuals based on similar behaviour. It then evaluates these clusters through established clustering validity indices and then seeks to interpret the statistical significance of these clusters. It studies two types of approaches to unsupervised clustering: centroid-based clustering and density-based clustering.

2 Methods and Materials

2.1 Datasets

The dataset in the work for this chapter is the latent representation learned from the autoencoder architectures in Chapter 5. In Chapter 5, the latent representation was used to learn predictive models of clinical target outcomes selected from the clinical literature.

Our research in this Chapter also reproduces the clustering work on the high-level granular bout-based features developed in Chapter 4, as a means for comparison and for studying physical activity segregation when the representation is more high-level and able to be interpreted as that set of features represent specific activities.

Only the accelerometer data collected at baseline was used since this had the largest number of participants. We saw that the number of subjects gradually decreased in the longitudinal follow-ups in Chapter 6. Also, the longitudinal study in Chapter 6 showed that individuals do not vary greatly in activity pattern across the time points considered by DIRECT, so including these follow-ups in the analysis in
this chapter would possibly be redundant.

Figure 7.1: Experimental design plan with the DIRECT accelerometer dataset. This chapter focuses on the unsupervised clustering of the latent representation and the high-level activity bout features.

Figure 7.1 displays the work flow diagram for the work completed with the DIRECT dataset. It makes a clear division between the unsupervised clustering analysis and the supervised predictive modelling experiments. The unsupervised clustering experiments were not dependent on the clinical variables, whereas the supervised regression analyses were.

2.2 \textit{k}-means clustering

\textit{k}-means was one of the algorithms already used for the work conducted in Chapter 4 when studying the clustering of the high-level granular bout-based activity features. This algorithm is centroid-based by determining \textit{k} centroids in the data and then clusters data points by assigning them to the nearest centroid.
Figure 7.2: Elbow curve for the autoencoder latent representation. It can be seen that when $k = 2$, there is point of inflection in the curve. This would suggest that this is the optimal value for $k$ when using the silhouette score as its metric.

Figure 7.3: Elbow curve for the high-level granular bout-based representation of the DIRECT accelerometer data. It can be seen that when $k = 4$, there is point of inflection in the curve. This would suggest that this is the optimal value for $k$ when using the silhouette score as its metric. However the value for the silhouette score is very low, this would suggest that the clusters that form, even with the optimised value set for $k$, would result in poorly defined clusters.

The value for $k$, the number of clusters, is the key hyperparameter that must be optimised with this algorithm. This is set to $k = 4$ for the high-level granular bout-based representation, and $k = 2$ for the autoencoder latent representation after optimisation with the elbow method.
2.3 Hierarchical agglomerative clustering

Hierarchical agglomerative clustering is the other clustering algorithm that was used for the unsupervised learning task in Chapter 4. This is used again for this chapter on both the autoencoder latent representation and the high-level granular bout-based representation for the baseline accelerometer cohort for DIRECT. The same approach as in Chapter 4 was used. The dendrograms were plotted for both the high-level granular bout-based representation and the autoencoder latent representation of the DIRECT baseline accelerometer data. These are shown in Figures 7.4 and 7.5.

Figure 7.4: Dendrogram plot of the hierarchical structure for the high-level granular bout-based representation. It is clear that four clusters are identified at the cutoff distance threshold of 3. The Ward linkage function was used.

Figure 7.5: Dendrogram plot of the hierarchical structure for the high-level granular bout-based representation. It is clear that four clusters are identified at the cutoff distance threshold of 110. The Ward linkage function was used.

These dendrograms show that there are the same optimal number of clusters as were calculated by using the elbow method in the \( k \)-means clustering algorithm. For both approaches, we identify 2 clusters for hierarchical agglomerative clustering.
algorithm and 4 clusters for the $k$-means clustering algorithm.

### 2.4 DBSCAN

Centroid-based clustering algorithms such as $k$-means do not have a concept of outliers in their implementations. All data points are assigned to the nearest centroid. This is problematic when trying to detect data points which are outliers since the outlier would be classed into a cluster containing points that do belong to that cluster.

The next set of unsupervised clustering analyses involve a different type of clustering called density-based clustering. These algorithms identify “dense” groups of points which are similar to one another.

DBSCAN was the first density-based clustering algorithm that was learned for this approach. This algorithm was described in Chapter 2, including its key hyperparameters $\varepsilon$. To optimise the value for $\varepsilon$, the $k$-nearest neighbour distances are used. This is a supervised machine learning clustering algorithm that clusters new data points based on their distance from other labelled data points. The $K$-nearest neighbour is trained on labelled data to determine which cluster each data point belongs to, and then applies this to the test dataset.

The value for the minimum number of samples in a neighbourhood to be considered a core point is usually set to the size of dimensionality of the dataset being learned, which would 25 in the case of the autoencoder latent representation. The visualisation for this is shown in Figure 7.10.

DBSCAN is also implemented for the high-level granular bout-based representation. We followed the same optimisation process for this feature set as we did when we implemented DBSCAN clustering on the autoencoder latent representation. Below is the $k$-nearest neighbour distance curve for the high-level granular bout-based representation.

The optimised number for minimum samples is also set to 60, the size of dimensionality for this representation.

A hyperparameter grid search was implemented to reinforce the decisions made for these tuned values for the hyperparameters. The grid search trained the DBSCAN clustering algorithm with different values for the minimum sample and the $\varepsilon$ hyperparameters two models: one that uses the high-level activity bout feature space and one that uses the autoencoder latent representation. Only the best results are displayed in Figures 7.11 and 7.10 respectively in this chapter’s section 3.1.3.

### 2.5 Affinity Propagation

Affinity propagation is another clustering algorithm used in this chapter’s analyses. The complexity of the algorithm is quadratic $O(x^2)$, so would be impractical for massive datasets. But the training time with the baseline DIRECT latent representation did not take too long to complete, so this algorithm did not present
this disadvantage to the analysis in this chapter.

The two hyperparameters tuned in this algorithm were the preference and damping factor, as described in Chapter 4.5.4. The damping factor is bounded between 0.5 and 1.0 in scikit-learn. This is optimised through a hyperparameter grid search with these initial values [0.5, 0.6, 0.7, 0.8, 0.9, 1.0]. At this point, the damping factor was then further tuned, arriving at 0.9. The maximum number of iterations before convergence was set to 500. The algorithm reached convergence after 369 iterations. The preference was set to -0.2 after performing another hyperparameter grid search with initial values set between [-1.0, -0.5, 0.0, 0.5, 1.0], and then further tuned to reach -0.2 for the preference.

### 2.6 Spectral clustering

The final unsupervised clustering algorithm used was the spectral clustering. This takes a similarity matrix between the samples, which in this case are each instance of the latent representation extracted by the autoencoder from chapter 5 and also the high-level granular bout-based features from chapter 4. The algorithm then builds a low-dimensional embedded space from this similarity matrix. It then uses another clustering algorithm in this low-dimensional embedded similarity matrix. The scikit-learn library uses the K-nearest neighbour algorithm so this was also used in this analysis.

Only one hyperparameter was tuned in this algorithm, the number of clusters $k$. The value for $k$ was set to be 2 because of the elbow method as seen in 7.2.

### 2.7 Clustering validity indices: Calinski-Harabrasz, Silhouette, Davies-Bouldin

Three clustering validity indices were used to evaluate the performance of these analyses with the different approaches to the clustering algorithms: Calinski-Harabrasz, Silhouette and Davies-Bouldin. Although there are many other clustering validity indices, such as the ones described by Arbelaitz et al [5], these three performance metrics were chosen because it did not require the knowledge of ground truth labels before the clustering exercise was performed. The results for these scores are reported in Table 7.1.

However these metrics only give an indication into the quality of the clustering themselves, but not an insight into why the clusters may have formed in the way that they did. But validating the clustering results with these metrics reinforces the implementations in this chapter.
2.8 Cluster interpretation with statistical testing

After the clustering algorithms were implemented and then evaluated with the standard clustering validity indices, it is necessary to explain and interpret these clusters to identify why they might have been formed in the way they were. This is necessary to identify whether the clusters produced by the various clustering algorithms held any meaning in terms of habitual physical activity and sleep patterns. This can be achieved by statistical hypothesis tests, a standard method of statistical inference to determine whether the given data supports a particular hypothesis or not. Statistical tests for normality are first performed to determine that the clusters are indeed normally distributed and therefore these hypothesis tests can be applied to these clusters.

This task would be independent of the clinical disease biomarkers which has so far produced weak results with the latent representation. But a positive result for this would imply that learning embedded representations of physical activity capture hidden qualities and descriptions of accelerometer that cannot be found in higher level and interpretable representations.

2.9 Predicting cluster membership with clinical variables

This set of analyses will revisit the clinical target outcomes for Chapter 5 and makes use of the supervised machine learning algorithms again. We use the same clinical target outcomes from Chapter 5’s supervised learning experiments, as seen in Table 5.1.

The aim of this set of experimental analyses is to evaluate where there is a dependence between the clinical target outcomes and the clustering results. It only focuses on the spectral clustering results as this algorithm performed the best, as seen in the clustering validity indices results in Table 7.1. Classification algorithms are used to predict individuals which cluster an individual belongs to based on their profile for the clinical target outcomes.

Four classification algorithms are trained and evaluated: Logistic regression, random forest classifier, XGBoost classifier and the AdaBoost classifier. Logistic regression, random forest classifier and XGBoost classifier were already used in the classification analyses in Chapter 4 to use the high-level granular bout-based representation to classify individuals as either Type-2 Diabetes positive or Type-2 Diabetes negative.

The AdaBoost classifier [73], Adaptive Boost, was also implemented. This classification algorithm is an ensemble method. This ensemble model works by iteratively fitting a weak classifier on the dataset where the cluster labels assigned by the best performing algorithm, which in our case was the spectral clustering algorithm as seen in 7.1 would be treated as the target outcome.

The clinical variables from Chapter 5 are now treated as the classifier input feature set. The algorithm then fits copies of this classifier on the same dataset but
the weights of the data points that were incorrectly classified are tuned so that the subsequent classifier focuses on those cases. The predictions from all of these weak learners are then combined together and the final prediction label is generated by a weighted majority vote.

All the classification models were trained with 10-fold cross validation. A hyperparameter grid search was also implemented to optimise the model performance.

3 Results

3.1 Visualising the clustering analyses

$t$-distributed stochastic neighbour embedding ($t$-SNE) was used for dimensionality reduction from high dimensional data into two-dimensional space with two components. Each observation in each cluster is given a location in this new two-dimensional space. This allows us to plot these location points onto a map and colour each point a different colour to represent a different cluster. Ideally, points in the same cluster would be well-separated out from other clusters, there would be little if any overlapping points. Intra-cluster cohesion and inter-cluster separation should be maximised where possible. The visualisations below display the results of the different clustering experiments using the embedded latent representation that was extracted from the LSTM autoencoder as described in [5].

3.1.1 $k$-means clustering visualisation

Figure 7.6: $t$-SNE visualisation for $k$-means clustering from autoencoder latent representation

Figure 7.7: $t$-SNE visualisation for $k$-means clustering from the high-level granular bout-based representation

Figure 7.6 shows that two distinct clusters are formed using the $k$-means clustering algorithm—one indicated by the blue points and the other by the red points. There is a modest level of separation of the clusters, but very little overlap between the
points. The visualisations demonstrate a clear difference in the clustering results. The embedded representation produces much more well-defined clusters than the high-level granular bout-based representation in Figure 7.9. The \( k \)-means algorithm clusters which are not well separated and many points overlap with one another. It can be inferred from this result that the embedded representation captures properties of the physical activity traces that are not captured in a higher-level interpretable representation as seen in Figure 7.7.

### 3.1.2 Hierarchical clustering visualisation

![Figure 7.8: t-SNE visualisation for hierarchical clustering from the autoencoder latent representation](image1)

![Figure 7.9: t-SNE visualisation for hierarchical clustering from high-level granular bout-based representation](image2)

Hierarchical agglomerative clustering produces similar results to the \( k \)-means algorithm where the high-level granular bout-based representation performs much more poorly for the clustering compared to the autoencoder latent representation. Figure 7.8 again shows two clear clusters, with modest level of separation but a high degree of non-overlap, whereas 7.9 again shows a very high degree of overlap between the clusters and there is no separation at all. This could further indicate the scale of homogeneity of the population. In this clustering approach, the high-level granular bout-based representation produce clusters which are overlapping, with little cohesion.

With t-SNE visualisation, when we reduce dimensions down to two components for both representations of physical activity, it can be seen that the t-SNE mapping plots distinct separate groupings of individuals for the autoencoder latent representation, whereas the granular bout-based representation produces a single cloud-like structure towards the middle. This may imply that the high-level representations which, although more meaningful and more understandable, do not allow an effective way of segregating individuals by the t-SNE algorithm.
3.1.3 DBSCAN clustering visualisation

Two clusters were formed for the autoencoder latent representation with the DBSCAN algorithm, shown in Figure 7.10. This is similar to the results for the $k$-means clustering and hierarchical clustering results. However, only one cluster was identified in Figure 7.11 when the high-level granular bout-based representation was used. This result is also poor. The performance for $k$-means and hierarchical clustering were also poor. This further suggests that the density of data points for the high-level granular bout-based representation was very high in that most of the individuals fit within this region of density, meaning that no distinguishable patterns of habitual behaviour can be identified in this representation, unlike the autoencoder latent representation.

3.1.4 Affinity Propagation

Two clusters were formed for the autoencoder latent representation with the DBSCAN algorithm, shown in Figure 7.10. This is similar to the results for the $k$-means clustering and hierarchical clustering results. However, only one cluster was identified in Figure 7.11 when the high-level granular bout-based representation was used. This result is also poor. The performance for $k$-means and hierarchical clustering were also poor. This further suggests that the density of data points for the high-level granular bout-based representation was very high in that most of the individuals fit within this region of density, meaning that no distinguishable patterns of habitual behaviour can be identified in this representation, unlike the autoencoder latent representation.
This clustering algorithm produced results that are similar to the other unsupervised clustering algorithms already used in this chapter. Two clusters, as demonstrated by the red and blue coloured points on the $t$-SNE visualisation in Figure 7.12 are formed. However, there is some overlapping as seen by the mixture of red and blue coloured points.

The high-level granular bout-based representation, again, produced a very poorly defined set of clusters with significant overlapping. The blue cluster dominates the dataset, with the green cluster running along the edge.

### 3.1.5 Spectral clustering

![Figure 7.14: $t$-SNE visualisation for spectral clustering from autoencoder latent representation](image)

![Figure 7.15: $t$-SNE visualisation for spectral clustering from the high-level granular bout-based representation](image)

### 3.2 Clustering validity indices

The clustering validity indices used to evaluate the performance of the unsupervised clustering analysis were calculated and are shown below in Table 7.1.

<table>
<thead>
<tr>
<th>Clustering algorithm</th>
<th>Silhouette</th>
<th>Calinski-Harabraz</th>
<th>Davies-Bouldin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affinity Propagation</td>
<td>0.634</td>
<td>2220.021</td>
<td>0.895</td>
</tr>
<tr>
<td>Spectral clustering</td>
<td>0.677</td>
<td>2600.836</td>
<td>0.839</td>
</tr>
<tr>
<td>DBSCAN</td>
<td>0.428</td>
<td>825.376</td>
<td>1.468</td>
</tr>
<tr>
<td>Hierarchical</td>
<td>0.466</td>
<td>2292.27</td>
<td>0.879</td>
</tr>
<tr>
<td>$k$-means</td>
<td>0.482</td>
<td>2617.19</td>
<td>0.839</td>
</tr>
</tbody>
</table>

Table 7.1: Calculated scores for clustering validity indices for the latent representation of DIRECT physical activity dataset.

The clustering algorithm that produced the best scores for the Silhouette, Calinski-Harabraz and Davies-Bouldin cluster validity indices is spectral clustering. The worst performing clustering algorithm was DBSCAN across all three clustering.
validity indices. However, the performance between all of these algorithms did not vary greatly. The results did not demonstrate that one algorithm was significantly outperformed the other algorithms.

The remaining work in this study for this chapter will focus on the results of the spectral clustering since it performed the best despite. First, we carry out hypothesis testing only on the results of spectral clustering. The goal of this exercise is to determine and then demonstrate that the unsupervised clustering of the latent representation generated by the autoencoder can be explained in terms of either the physical activity characteristics and/or clinical variables, and not just random allocation of clustering.

<table>
<thead>
<tr>
<th>Number of samples in each cluster</th>
<th>Cluster 0 (Blue)</th>
<th>Cluster 1 (Red)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>905</td>
<td>1207</td>
</tr>
</tbody>
</table>

Table 7.2: Table demonstrating the number of individuals within each of the two clusters generated by spectral clustering algorithm.

As seen in Table 7.2, the classes had very similar numbers where one class did not significantly outnumber the minority class.

### 3.3 Supervised learning for clusters with clinical variables

The results for the binary classification experiments are displayed in Table 7.3 below.

The bar charts below in Figures 7.16, 7.17 and 7.18 show the relative Gini importance of each clinical feature variable. The importance is shown on the y-axis, and the clinical feature is shown as the x-axis. The Gini importance is defined as the total decrease in the node impurity and this is weighted by the probability of reaching this node (this is estimated by calculating the proportion of samples arriving at this node in the tree model). This probability is averaged for all trees of the ensemble [136].
Figure 7.16: Clinical feature ranking for random forest classification algorithm. Sex, defined as NewGender, seem to be the most powerful predictor by far. BMI is the second most important. Other remaining clinical features gradually decrease in importance for the classification results.

Figure 7.17: Clinical feature ranking for XGBoost classification algorithm. Sex, again, seems to be the most powerful predictors and overpower the other features significantly. The remaining clinical features present similar importance in the feature ranking.
Figure 7.18: Clinical feature ranking for AdaBoost classification algorithm. BMI and Sex seem to be the most powerful predictors significantly and overpower the other features significantly.

From these feature importance rankings, it can be seen that Sex (‘NewGender’) and the body mass index (‘bmi’) appear to have the greatest importance when used in the classification into which cluster an individual belongs to. Although the importance of sex and BMI are relative, we can see that these two clinical variables outweigh the other clinical variables quite significantly, particularly NewGender (Sex). This would imply that the clusters may capture differences in physical activity patterns and habitual behaviours associated between females and males. The implication is that whether an individual is female or male had the greatest relative importance in which cluster they belong to—essentially as a sex classifier when considering relative feature importance.

The effect of body mass index on cluster membership is not a surprising result given how closely body mass index and physical activity levels are linked \[169, 26\].

It is important, however, to state that the predictive power of the selected clinical variables in predicting which cluster an individual belongs to is still weak. This is shown in the scores for the metrics outlined in Table 7.3.

Furthermore, our results also suggest that the specific Type-2 Diabetes related clinical features, such as the indices for insulin sensitivity and fasting glucose, play a smaller independent role than would be assumed.

Table 7.3 displays the scores for the precision, accuracy, recall and f-1 scores of these classification algorithms as a means to measure their performance.
<table>
<thead>
<tr>
<th>Classifier model</th>
<th>Accuracy</th>
<th>Cluster</th>
<th>Precision score</th>
<th>Recall score</th>
<th>F-1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>0.57</td>
<td>0</td>
<td>0.52</td>
<td>0.59</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>0.56</td>
<td>0.49</td>
<td>0.52</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.57</td>
<td>0</td>
<td>0.54</td>
<td>0.57</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>0.59</td>
<td>0.56</td>
<td>0.57</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.50</td>
<td>0</td>
<td>0.58</td>
<td>0.63</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>0.63</td>
<td>0.58</td>
<td>0.60</td>
</tr>
<tr>
<td>AdaBoost</td>
<td>0.58</td>
<td>0</td>
<td>0.58</td>
<td>0.63</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>0.63</td>
<td>0.58</td>
<td>0.60</td>
</tr>
</tbody>
</table>

Table 7.3: Table displaying the values for scores for precision, recall and f-1 performance evaluation metrics for the classification algorithms used to predict which cluster an individual should belong to based on their clinical feature profiles.

From the results above in Table 7.3, we can see that AdaBoost performed slightly better than the other algorithms with approximately 50-60% for all performance metrics (accuracy, precision, recall and F-1). The other classification algorithms however yielded values similar to this. Clearly, a 50-60% accuracy for a binary classification problem is a poor result, despite hyperparameter tuning for optimising model performance. The clusters are roughly of equal size, so there is not an issue of imbalanced classes that would favour the prediction of one cluster over the other. The results actually imply that the classifiers are almost randomly classifying clinical profiles into their cluster membership.

The overall conclusion from this set of classifier experiments is that there is a very weak dependence, if any at all, between a which physical activity cluster an individual belongs to and their Type-2 Diabetes related clinical profiles. It could be concluded that most individuals in the cohorts of DIRECT dataset exhibit similar clinical profiles.

### 3.4 Clustering analysis interpretation

#### 3.4.1 Distribution of clinical variables among physical activity clusters

Two clusters were formed with the spectral clustering, as seen in Figure 7.14 on the autoencoder latent representation. These clusters however did not seem to produce a very strong association with the clinical variables from Chapter 5 as established by the supervised learning analysis seen in Table 7.3. This can be investigated further to establish why there is so little dependence between clinical variables and these physical activity clusters. To achieve this, the distribution of the selected clinical variables were studied. These distributions are shown in the figures below for each of the two clusters generated from spectral algorithm.
Figure 7.19: Histogram showing the distribution of values for OGIS insulin sensitivity across Cluster 0 (Blue cluster)

Figure 7.20: Histogram showing the distribution of values for OGIS insulin sensitivity across Cluster 1 (Red cluster)

Figure 7.21: Histogram showing the distribution of values for Matsuda insulin sensitivity across Cluster 0 (Blue cluster)

Figure 7.22: Histogram showing the distribution of values for Matsuda insulin sensitivity across Cluster 1 (Red cluster)

Figure 7.23: Histogram showing the distribution of values for Stumvoll insulin sensitivity across Cluster 0 (Blue cluster)

Figure 7.24: Histogram showing the distribution of values for Stumvoll insulin sensitivity across Cluster 1 (Red cluster)
Figure 7.25: Histogram showing the distribution of values for body mass index across Cluster 0 (Blue cluster)

Figure 7.26: Histogram showing the distribution of values for body mass index across Cluster 1 (Red cluster)

Figure 7.27: Histogram showing the distribution of values for fasting glucose across Cluster 0 (Blue cluster)

Figure 7.28: Histogram showing the distribution of values for fasting glucose across Cluster 1 (Red cluster)

Figure 7.29: Histogram showing the distribution of values for age across Cluster 0 (Blue cluster)

Figure 7.30: Histogram showing the distribution of values for age across Cluster 1 (Red cluster)
Figure 7.31: Histogram showing the distribution of values for OGTT 120 minutes across Cluster 0 (Blue cluster)

Figure 7.32: Histogram showing the distribution of values for OGTT 120 minutes across Cluster 1 (Red cluster)

Figure 7.33: Histogram showing the distribution of values for sex across Cluster 0 (Blue cluster)

Figure 7.34: Histogram showing the distribution of values for sex across Cluster 1 (Red cluster)

Figure 7.35: Histogram showing the distribution of values for waist circumference across Cluster 0 (Blue cluster)

Figure 7.36: Histogram showing the distribution of values for waist circumference across Cluster 1 (Red cluster)
Figure 7.37: Histogram showing the distribution of values for potentiation factor ratio (PFR) across Cluster 0 (Blue cluster)

Figure 7.38: Histogram showing the distribution of values for PFR across Cluster 1 (Red cluster)

Figure 7.39: Histogram showing the distribution of values for impaired glucose regulation (IGR) across Cluster 0 (Blue cluster)

Figure 7.40: Histogram showing the distribution of values for IGR across Cluster 1 (Red cluster)

Figure 7.41: Histogram showing the distribution of values for glucose sensitivity across Cluster 0 (Blue cluster)

Figure 7.42: Histogram showing the distribution of values for glucose sensitivity across Cluster 1 (Red cluster)
These histograms demonstrate that clinical variables were distributed very similarly across the two clusters, Cluster 0 and Cluster 1. However, a difference can be seen in the distribution for fasting glucose. For Cluster 0, it appears that there is a much narrower distribution in the fasting glucose variable whereas Cluster 1 displays a more evenly spread out normal distribution. The clinical features that were determined to be most important for classification were sex and body mass index for most of the binary cluster classifiers, as seen from Figures 7.18 and 7.17. But when studying the distributions of these two clinical features in histograms 7.33 7.25 7.26 and 7.34 it can be seen that the distribution of sex was virtually identical (where ‘1’ represents ‘male’, ‘0’ was ‘female’). For body mass index, it can be seen that the Cluster 1 (The red cluster) had a greater range of body mass index values compared to Cluster 0 (The blue cluster).

Insulin sensitivity seems to almost have identical distributions across both clusters. This is surprising since insulin sensitivity is closely associated with physical activity, as in the clinical literature. However, this observation was not shown in these distributions for physical activity clusters.

In conclusion, the distributions for the different clinical variables did not exhibit strong differences between the two clusters formed by spectral clustering on the latent representation generated by the autoencoder architecture in Chapter 5. This offers an explanation for the weak performance on the predictive models for classifying cluster membership using the clinical variables. This further reinforces the findings reached in our research that physical activity and clinical Type-2 Diabetes biomarkers have a very weak dependence on one another, despite this being widely reported in the literature.

However, the performance of the clustering is stronger when evaluated with the clustering validity indices. Although this could not be observed with the clinical features, it may be possible to interpret these clusters in terms of physical activity by revisiting the high-level activity bout features from Chapter 4 and identifying whether the clusters can be interpreted and explained with these features that are independent of the clinical variables.
3.4.2 Distribution of high-level summarised physical activity among the clusters

This final set of analyses in our research studies the physical activity clusters in their own right. It looks at the distribution of physical activity types between the two clusters and aims to explain the reason as to why the clusters have formed the way they did. To perform this analysis, the high-level granular bout-based representation, from Chapter 4, for the DIRECT baseline accelerometer dataset were summarised by bringing up to the scale of daily patterns, rather than specific times of day. These were extracted for the work in Chapter 6 when studying the change in physical activity over time and thus used in the following set of analyses.

The histograms for these high-level physical activity daily summaries are plotted below.

Figure 7.45: Histogram showing the distribution of values for percentage of time daily spent performing moderate-to-vigorous activity across Cluster 0 random sample (Blue cluster)

Figure 7.46: Histogram showing the distribution of values for percentage of time daily spent performing moderate-to-vigorous activity across Cluster 1 random sample (Red cluster)

Figure 7.47: Histogram showing the distribution of values for percentage of time daily spent performing sedentary activity across Cluster 0 random sample (Blue cluster)

Figure 7.48: Histogram showing the distribution of values for percentage of time daily spent performing sedentary activity across Cluster 1 random sample (Red cluster)
The histograms above for the summarised high-level physical activity features show that the physical activity among the red and blue clusters appear to be virtually identical. This further suggests that this simple analysis on distribution of physical
activity on these two clusters does not yield results that show a clear distinguishable
behavioural pattern between them. It also suggests that for a cohort of subjects
collected in DIRECT, the variation physical activity among the subjects does not
vary greatly so it is difficult to capture clear distinguishable features of behaviour
that allow characterisation and phenotyping with their accelerometer device-derived
activity pattern measurements. This was already demonstrated in Chapter 6 when
analysed over longer periods of time.

3.4.3 Statistical hypothesis tests for significance of clustering results

The deep autoencoder architectures in Chapter 5 may capture a descriptive
representation of the physical activity that cannot be explained by these high-level
summary features. But this descriptive representation may however be captured
by the clustering algorithms. In order to verify this, statistical hypothesis tests are
conducted to determine whether there is a statistical significance in the difference
between the means of Cluster 1 (Red cluster) and Cluster 0 (Blue cluster). This
analysis continues to make further use of the high-level physical activity daily
summary features.

Before t-tests can be conducted, it is important to ensure that these high-level
physical activity summary features follow a normal distribution on a random sample
of both clusters. Although this can be visually seen in the histograms for most of the
features, the normal distribution can be verified through statistical normality tests.
Three tests are conducted and reported on to confirm this: D’Agostino, Shapiro and
Anderson. Results are given in Table 7.4.

<table>
<thead>
<tr>
<th></th>
<th>D’Agostino</th>
<th>Shapiro</th>
<th>Pearson</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of time spent</td>
<td>Not normal</td>
<td>Not normal</td>
<td>Not normal</td>
</tr>
<tr>
<td>performing MVPA daily</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percentage of time spent</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
</tr>
<tr>
<td>performing sedentary</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>activity daily</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percentage of time spent</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
</tr>
<tr>
<td>performing light tasks daily</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average number of hours</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
</tr>
<tr>
<td>spent sleeping daily</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average daily MET level</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
</tr>
<tr>
<td>(Energy expenditure)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7.4: Table demonstrating the results for the tests for normality in each of the
high-level physical activity summary features.

These tests were performed on a random sample of 40 subjects each from
Cluster 1 and Cluster 0 of the clusters formed by spectral clustering of the latent
representation for the signal-level epoch accelerometer data extracted by the
autoencoder architectures in Chapter 5. According to the results seen in Table
7.4, normal distributions were confirmed for all of the high-level physical activity
summary features with the only exception of percent_time_MVPA_daily, the
percentage of time spent performing moderate-to-vigorous activity daily. This was also clearly observed in the histograms in Figures 7.45 and 7.46.

Homogeneity of variance also needs to be established before it is valid to conduct t-tests. Variances between Cluster 0 and Cluster 1 need to be approximately similar. This was assessed with the F-test for equality of variance, where the null hypothesis $H_0$ is that the variances of the two samples are equal. If the $F$-test value is less than 0.05, then the null hypothesis is rejected, or it is accepted if the $F$-test value is greater than 0.05. The result for this test for each high-level physical activity summary feature is demonstrated in Table 7.5.

<table>
<thead>
<tr>
<th>Percentage of time spent performing MVPA daily</th>
<th>F-test value</th>
<th>Reject/Accept $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of time spent performing sedentary activity daily</td>
<td>0.41</td>
<td>Accept</td>
</tr>
<tr>
<td>Percentage of time spent performing light tasks daily</td>
<td>0.45</td>
<td>Accept</td>
</tr>
<tr>
<td>Average number of hours spent sleeping daily</td>
<td>0.24</td>
<td>Accept</td>
</tr>
<tr>
<td>Average daily MET level (Energy expenditure)</td>
<td>0.32</td>
<td>Accept</td>
</tr>
</tbody>
</table>

Table 7.5: Table displaying the results for $F$-test results in each of the high-level physical activity summary features. All features show approximately equal variance.

It is clear from these results that all features have equal variance and so all the assumptions for T-tests are met. The null hypothesis, $H_0$, is that the true difference between the means of Cluster 0 and Cluster 1 is zero (i.e the means of Cluster 1 and Cluster 0 are equal). The alternative hypothesis, $H_1$, is that the true difference between the means of Cluster 0 and Cluster 1 is not zero (i.e. the means of Cluster 1 and Cluster 0 are not equal). If the null hypothesis can be rejected, then it implies that Cluster 0 and Cluster 1 formed by the spectral clustering algorithm of the embedded latent representation of the signal-level epoch accelerometer data is not due to chance for that specific high-level physical activity summary feature.

Two-tailed t-tests are conducted so that both the positive and negative tails of the distribution are used in the test. This is so that both negative and positive differences are tested for. The result for this two-tailed homoscedastic (equal variance) t-tests are reported in Table 7.6. A critical value of 0.05 is used to reject or accept the null hypothesis.
<table>
<thead>
<tr>
<th></th>
<th>t-statistic</th>
<th>p-value</th>
<th>Reject/Accept H&lt;sub&gt;0&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of time spent performing MVPA daily</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Percentage of time spent performing light tasks daily</td>
<td>2.409</td>
<td>0.018</td>
<td>Reject</td>
</tr>
<tr>
<td>Percentage of time spent performing sedentary daily</td>
<td>0.471</td>
<td>0.639</td>
<td>Accept</td>
</tr>
<tr>
<td>Average number of hours spent sleeping daily</td>
<td>-2.624</td>
<td>0.0104</td>
<td>Reject</td>
</tr>
<tr>
<td>Average daily MET level (Energy expenditure)</td>
<td>2.302</td>
<td>0.024</td>
<td>Reject</td>
</tr>
</tbody>
</table>

Table 7.6: Table displaying the results for the two-tailed t-test results in each of the high-level physical activity summary features. p-values and t-statistics are reported.

The results in Table 7.6 show that the means in three of the four high-level physical activity summary features are statistically different enough to reject the null hypothesis, H<sub>0</sub>. For the percentage of time spent performing light tasks daily, average number of hours spent sleeping daily and the average daily MET level (energy expenditure). Rejecting the null hypothesis implies that the clusters are not formed by random chance and that the latent embedded representation of the signal-level epoch data, as extracted by the autoencoder architectures in Chapter 5, capture a hidden description for the habitual behavioural patterns that are specific to average number of hours spent sleeping daily, daily average energy expenditure (measured through MET levels) and light task activities.

The t-test was repeated on both of the clusters in their entirety, randomly selecting enough individuals from Cluster 1 to match the size of Cluster 0, and the results for this is repeated in Table 7.7.
Table 7.7: Table displaying the results for the t-test results in each of the high-level physical activity summary features. p-values and t-statistics are reported.

<table>
<thead>
<tr>
<th></th>
<th>t-statistic</th>
<th>p-value</th>
<th>Reject/Accept H(0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of time spent performing MVPA daily</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Percentage of time spent performing light tasks daily</td>
<td>3.592</td>
<td>$3.414 \times 10^{-4}$</td>
<td>Reject</td>
</tr>
<tr>
<td>Percentage of time spent performing sedentary daily</td>
<td>1.194</td>
<td>$2.326 \times 10^{-1}$</td>
<td>Reject</td>
</tr>
<tr>
<td>Average number of hours spent sleeping daily</td>
<td>-2.521</td>
<td>$1.182 \times 10^{-2}$</td>
<td>Reject</td>
</tr>
<tr>
<td>Average daily MET level (Energy expenditure)</td>
<td>4.848</td>
<td>$1.402 \times 10^{-6}$</td>
<td>Reject</td>
</tr>
</tbody>
</table>

The results in Table 7.7 reinforce the result with the random sample shown in Table 7.6 to a greater extent. The p-values are even smaller, with 95% confidence interval.

Sleep is already known to be closely related to Type-2 Diabetes outcomes [204], as is energy expenditure. The autoencoders were able to capture a hidden description of sleep and energy expenditure, that could not be explained by higher-level representations, such as DIRECT’s own `hpf.vm.mean` and the high-level granular bout-based representation. This difference was significant enough that it enabled segregation into two groups by unsupervised clustering algorithms. This therefore corresponds to behavioural patterns that are already known to be related to Type-2 Diabetes, such as poor sleep and low energy expenditure.

4 Discussion

4.1 Principal Findings

The aim of this chapter was to shift away from the dependence on clinical Type-2 Diabetes biomarkers and to analyse the latent representations extracted by the autoencoder architectures in Chapter 5 in their own right. This was achieved through unsupervised clustering on the latent representations, and the high-level granular representation from Chapter 4 for comparison.

This analysis implemented and compared different approaches to unsupervised clustering (k-means, hierarchical, DBSCAN, spectral and affinity propagation) and was able to achieve a similar result for all chosen algorithms. Efforts to ensure that the clustering algorithms were optimised by hyperparameter tuning.
was undertaken. Three different clustering validity indices were used to evaluate clustering performance. The spectral clustering algorithm performed the best relative to the other clustering algorithms based on these indices and was thus analysed further to interpret why these clusters formed the way they did.

Supervised classifier models were constructed to predict which cluster an individual belongs using the selected clinical biomarkers as input. This analysis produced quite poor results and demonstrated the weak dependence between the individual physical patterns and habitual behaviour within both DIRECT cohorts and their clinical profiles. This was further reinforced by analysis the distributions of these clinical biomarkers.

The distribution analysis of the selected clinical biomarkers from Chapter 3 and the high-level physical activity summary features was able to demonstrate that these two separate sets of variables were distributed in a normal distribution with little difference between the two clusters despite the strong performance in the reported clustering validity indices.

However the results from the clustering analysis clearly identified hidden description of the physical activity at the epoch signal-level that enabled an unlabelled segregation into distinct groups. To verify this, statistical hypothesis tests were undertaken to determine this for each of the high-level physical activity summary features independently. The null hypothesis in these tests is that the means for Cluster 0 and Cluster 1 were not statistically significantly different and thus could be due to random chance, whereas the alternative hypothesis states that there was a statistically significant difference in Cluster 0 and Cluster 1 means. For these t-tests, the reported $p$-values were less than a critical value of 0.05 for three of the five high-level physical activity summary features: Percentage of time spent carrying out light tasks daily, the average number of hours spent sleeping daily and the average daily MET level (a measure for energy expenditure). This means, with a 95% confidence interval, that there was a discernible difference in the individuals in Cluster 0 and Cluster 1, formed by the spectral clusters, for these three features of summarised high-level daily physical activity.

### 4.2 Implications of study

This chapter’s analyses demonstrate the significant extent of work that still needs to be done to enable the use of collected digital signals from devices such as accelerometers in self-monitored, free-living large-scale cohort studies such as DIRECT and the UK Biobank to study the progression and learning predictive models for slow-moving, complex and multi-faceted chronic diseases such as Type-2 Diabetes. A comprehensive set of different critical analyses were carried out to reinforce this further.

Furthermore this chapter’s results has helped to establish the advantages of adopting techniques from deep learning to learn new representations of digital signals from accelerometry traces. This chapter has been able to show that representation learning can learn new features that represent hidden behavioural
patterns not captured by features engineered manually through domain knowledge, as demonstrated by the results of the unsupervised clustering analysis.

The results from this chapter are very different from the unsupervised clustering results in Chapter 4. In Chapter 4, the clustering analysis clearly found three distinct clusters where one cluster contained exclusively all the individuals positive for Type-2 Diabetes, as seen in the column for Cluster 2 in Table 4.6. Although the clustering analyses with the UK Biobank and the DIRECT accelerometer datasets is completed with two different representations of physical activity (The high-level granular bout-based in the UK Biobank, the autoencoder extracted latent representation in the DIRECT dataset), the clustering analyses with the UK Biobank was able to characterise distinctly different sociodemographic, lifestyle and anthropometric phenotypes whereas the DIRECT dataset was not able to do so.

All the subjects of the DIRECT cohorts were either already positive for Type-2 Diabetes or in a pre-diabetes state. There were no negative controls. Furthermore, most of the selected clinical biomarkers were very normally distributed, as were the higher-level physical activity features, including the ones crafted by the DIRECT consortium. This would imply that most of the individuals in DIRECT exhibit similar patterns of habitual behaviour and in their clinical biomarkers. So, it would be difficult to capture differences in groups of individuals when there is high similarity in terms of physical activity amongst the population. But the clustering analysis in the DIRECT latent representation did capture significant differences in physical activity between two distinct groups of individuals. We were able to identity from hypothesis tests that there were significant statistical differences between the two clusters in terms of performing light tasks, energy expenditure (measured through MET levels), and sleep patterns which caused individuals to be grouped into one cluster over the other.

The implication of this result is that it also highlights the fact that some representations of physical activity do not capture all the elements that could be used to digitally phenotype individuals. In this chapter’s analyses, the latent representation extracted with deep learning techniques from low-level signals provided more detailed insight than the higher-level representations from before.

The methods in this chapter can be extended and generalised to other diabetes-related accelerometer cohorts as long as the raw accelerometer data can be processed by the Oxford accelerometer analysis toolkit and produce the signal-level epoch feature files that are used to learn the latent space, and also the high-level activity classification sequences used to extract the high-level granular bout-based features. This would aid in the analysis of other cohorts by studying the clustering of their latent spaces. The analyst can then interpret these clusters by seeing if each cluster exhibits differences in physical activity. Results would obviously be data-driven.
4.3 Limitations

One of the main limitations in this study was that the population size of the DIRECT cohorts were much smaller compared to the UK Biobank, so the results may not be as robust given the training data size. This meant that the analyses conducted could not be as robust as the clustering analysis conducted in Chapter 4.

Furthermore, another limitation was there were no negative controls for whom accelerometer data was collected as all subjects were either Type-2 Diabetes positive or in a pre-diabetes state. Having negative controls from DIRECT would be a useful comparison and would allow us to reinforce further that the clustering results in this chapter are meaningful. It was considered before that we could use negative controls from the UK Biobank in combination with the pre-diabetic or type-2 diabetes positive populations of the DIRECT dataset. However, given the differences in data collection protocols, such as the difference in wear-time length between the datasets, it would not be as robust since the protocols may introduce unknown factors that could cause negative controls and positive cases to be easily distinguished simply as a nature of the data collection protocols.

Another limitation of this study was the dependence, as was the case in Chapter 6 and Chapter 5, on the Oxford accelerometer analysis tool’s outputs to learn the latent representation used for the clustering analysis. However, as explained in Chapter 5, other tools are available but perform similar functions to this when it comes to extracting signal-level features often used in other tasks such as human activity recognition. This would mean that using the Oxford accelerometer analysis toolkit’s technology would serve as a good case study. But it is recognised that the study could be more comprehensive and robust if other tools were also used as comparison or we developed our own analysis tool.

Another noted limitation was that, as in Chapter 6, confounding factors may have had an impact on the clustering results but were not considered in the interpretation analyses. This could include weather conditions, geographic location or other societal and environmental factors that may have affected an individual’s behavioural patterns that introduced noise or invalid data to the accelerometer traces.

There are other clustering approaches that exist that were not explored in this chapter’s analyses. However, this was judged to be unnecessary and may in fact be redundant since the results for the approaches used to clustering encompassed a range of approaches to clustering (density-based, centroid-based, hierarchical etc) and all these clustering algorithm choices proved to show similar results.

4.4 Conclusions

This chapter conducted analyses that were independent of the selected clinical target outcomes and only used these outcomes as a means of analysing distributions and for interpretation. Instead, it was an exercise in unsupervised learning whereby the latent representations learned from the autoencoder architectures in Chapter 5 were clustered using several approaches to unsupervised learning. The spectral clustering
approach performed the best, relative to the other algorithms, and its results were thus analysed and interpreted further. This algorithm produced two clusters which were visualised with t-SNE.

The results from clustering were analysed by projecting the clinical variables into each of the two clusters formed and assessing their distributions. The distributions of the clinical variables were virtually identical across the two clusters. The analysis also made use of supervised learning by building supervised classifiers for predicting which cluster an individual belongs to based on their clinical variables. This produced poor classification results and further highlighted the weak dependence between clinical Type-2 Diabetes biomarkers and physical activity.

This distribution analysis was then repeated with high-level features that summarised the physical activity on a daily average. Statistical t-tests were then conducted using these high-level physical activity features. The t-tests were able to reject the null hypothesis that these clusters have no statistically significantly different means. This implies that these two clusters, produced by the spectral clustering method, have statistically significant different means for light task activity, average number of hours asleep daily and the average daily energy expenditure measured through MET levels.

This result shows that the latent representations learned from deep autoencoder architectures produce useful representations of the physical activity that capture a hidden description of the physical activity that cannot be captured by higher-level, hand-crafted features normally driven by domain knowledge. The contribution that this chapter makes is that we have implemented a systematic and thorough comparison, and a critical appraisal of a range of methods for extracting physical activity features spaces and evaluated their viability as a means to generate meaningful interpretations of physical activity. We have devised a way to use the latent representation learned by an autoencoder to build digital phenotypes from accelerometry, but we also demonstrated the challenges of associating this with Type-2 Diabetes related clinical target outcomes.
Chapter 8

General Discussion
1 Introduction

This final chapter reflects on the contributions and the original intended focus of our research. It offers a high-level summary of the work completed and then discusses its contributions to the current state of knowledge as guided by the existing literature. This chapter also discusses the limitations of the research completed. Finally, this chapter also elaborates on the potential applications of the research completed and directions for work in the future. Final conclusions are then provided.

2 Summary of Research Work

Wearable digital technologies and machine learning are used widely in healthcare and medicine to extract signals of biomarkers that can contribute towards building an individual’s \textit{extended phenotype} \cite{104} in the form of their digital phenotype. Our research aimed to study the power to extract representations of an individual’s digital phenotype and evaluating the viability and predictive power of these representations for predicting disease-related outcomes.

Our research work is at the intersection of clinical research and data science and engineering. We focused on one case study, on one chronic disease, Type-2 Diabetes, in our work. The reason for this was the nature of this disease itself: it is a multi-faceted, slow-moving complex disease known to be associated with low levels of physical activity, poor sleep patterns and a sedentary lifestyle, as shown in the clinical literature review. Additionally, type-2 diabetes is growing in prevalence in both developed and developing countries. Altogether, these reasons make type-2 diabetes an ideal case study for the research completed and to address some of the data science challenges associated with realising a predictive and participatory vision for the future of healthcare.

The state-of-the-art in this field used machine learning techniques to extract representations of physical activity traces for the purposes of human activity recognition. There is a good foundation that already exists in this area, but mostly focuses on specific movement analyses, such as gait analysis, for diseases directly related to physical activity, such as Parkinson’s disease, and clinical events such as fall detection. There was little work in the literature that sought to apply these techniques to a slow-moving, multi-faceted complex disease such as type-2 diabetes from physical activity data. We use two datasets as test beds for experimentations. They are both large-scale accelerometer epidemiological cohorts: The UK Biobank and DIRECT. The UK Biobank collected a vast array of data for genotyping and phenotyping of an individual such as multi-\textit{omics}, medical imaging, biochemistry, sociodemographics, anthropometry and lifestyle measures including accelerometry. But this accelerometer dataset for the UK Biobank cohort was not rigorously controlled for type-2 diabetes specific study. However, it served as an appropriate option for an experimentation test bed because it was a large resource and provided a deep insight into the data science challenges of free-living digital sensor data-including the need to consider carefully the selection of negative controls and
positive cases in the preprocessing steps before learning predictive models for disease. DIRECT, on the other hand, was better controlled and more related to type-2 diabetes specific research but this dataset was smaller in size and did not have type-2 diabetes negative control subjects (DIRECT has two cohorts: Cohort 2.1 of individuals in pre-diabetes state and Cohort 2.2 of individuals with confirmed type-2 diabetes).

In Chapter 3, we establish a baseline on which to build the research. It uses a widely used and popular package, GGIR, to process raw accelerometer data from the UK Biobank to extract feature spaces that are then used to learn initial predictive models for type-2 diabetes disease status. This produced meagre results, around 60% accuracy, when attempting to address our research questions. This work was presented as a poster at the Massive Analysis and Quality Control (MAQC) conference in Riva del Garda, Italy in April 2019 and the Alan Turing Institute’s Advances in Data Science conference in Manchester in June 2019.

In Chapter 4, our research then sought to build on the established baseline by adopting techniques from gait analysis and studies in neuroscience. A key toolkit in accelerometer analysis developed at Oxford University was used to extract signal-level epoch representations and labelled activity sequences from the UK Biobank accelerometer data through human activity recognition. A high-level representation was extracted from the labelled activity sequences which measured the percentage of time spent, the frequency and duration of bouts for different activity types across time periods of the day. This new set of features was used to learn predictive models for type-2 diabetes negative/positive disease status. The analysis produced better results after carefully selecting negative controls for whom physical activity levels were not impaired by other clinical causes, such as someone not suffering from type-2 diabetes but suffering from Amyotrophic lateral sclerosis which would obviously impair their physical activity performance. Furthermore, an unsupervised clustering analysis was undertaken in this chapter also and this exercise was successful in segregating individuals with type-2 diabetes from negative control subjects purely based on the new high-level granular bout-based representation. The work from this chapter was published in the Journal for Medical Internet Research Diabetes journal in January 2021.

Chapter 5 first uses an autoencoder architecture to learn this latent representation and then repeat the supervised learning experiments by building predictive regression models for selected clinical target outcomes available in DIRECT and used the autoencoder latent representation as the input. This produced very weak results for regression and demonstrated the weak dependence between physical activity and clinical biomarkers of type-2 diabetes, which follows the clinical literature and state-of-the-art. This was further affirmed using the high-level granular bout representation and physical activity features in DIRECT itself in regression experiments.

Chapter 6 made use of another feature of the DIRECT dataset, the follow-up accelerometer traces of individuals over a period of 18 and 36 months after baseline for Cohort 2.2, and 18 and 48 months after baseline for Cohort 2.1. These samples of

---

follow-up accelerometer data enabled an exploration into the progression of physical activity over time compared to the progression of clinical phenotypes. This chapter devised a simple method to measure the similarity of physical activity between two time points by calculating the Euclidean distance between the latent representation of an individual at one time point and the latent representation at another time point later in the future. A correlation analysis was conducted to determine if there is a correlation between the magnitude in change of the clinical phenotype and the change in physical activity. Distribution analyses were also conducted to see if the distribution of activity types change over time a high-level summary physical activity features, and also the physical activity features available in DIRECT for comparison. It did not matter which representation was chosen, but the results were all similar in showing that levels of physical activity of the individuals in both DIRECT cohorts change very little over time, especially relative to the clinical variables.

Chapter 7 removes the dependence on clinical target outcomes and explores the autoencoder latent representations in their own right. This was achieved by using unsupervised clustering on the latent representations and the high-level granular bout-based representation of all DIRECT baseline subjects. A suite of different clustering algorithms were implemented and then evaluated with clustering validity indices. The best performing clusters were formed by the spectral clustering algorithm. This clustering algorithm produced two distinct clusters and these two clusters were analysed and interpreted further. This was first achieved by performing supervised learning experiments to predict which cluster an individual should be assigned to based on their clinical variables in an attempt to explore any dependence between clinical variables and physical activity. However, this also produced poor results. This chapter then removes the dependence on clinical outcomes by analysing the clusters in terms of physical activity. Statistical hypothesis tests were then completed to determine whether there were statistically different means in each cluster. These t-tests showed that the clusters were statistically different in terms of percentage of time spent performing light tasks daily, average number of hours spent sleeping daily and the average daily MET level (energy expenditure). Moderate-to-vigorous activity and sedentary activity were shown to be relative uniform in both clusters and across the dataset. This contributes to research by showing the power of using deep learning to learn new representations over hand-crafted, domain knowledge driven feature engineering of physical activity for digital phenotyping.

3 Key Research Contributions

The contributions of each research chapter were outlined in greater detail in the Discussion sections of each chapter. This section discusses the key overall research contributions and reflects on its success against the original goals. The overall aim of our research was to investigate and evaluate different techniques for extracting representations of physical activity, as a case study, to learn predictive models and thus enable the digital phenotyping for a chronic, slow-moving and multi-faceted complex disease like type-2 diabetes. This was largely achieved and it was an
attempt to investigate our research hypothesis that digital biomarkers can be extracted from ubiquitous digital sensing devices that can be used to form part of a person’s digital phenotype.

Our research began this by conducting a literature search and using some of the most prominent related works and software frameworks in this field of study to conduct analyses into predictive modelling experiments for type-2 diabetes as a means to establish and then extend upon the state-of-the-art. The research then engineers and extracts new representations of physical activity traces through domain-driven clinical knowledge and based on machine learning frameworks in human activity recognition. This exercise built on existing works by deriving a new high-level of representation which proved to be useful in being able to digitally phenotype individuals specifically positive for type-2 diabetes and negative control individuals without type-2 diabetes.

Chapters 5, 6 and 7 of our research then learns new embedded representations of accelerometer data through the power of deep learning techniques and was able to demonstrate the power of these representations for phenotyping individual habitual behaviour patterns that are well defined enough to be able to be clustered and segregated into distinct groups in an unsupervised way without ground truth labels. Clinically, these chapters demonstrated the significant extent of work still needed to be completed to understand the link between physical activity and biomarkers for type-2 diabetes. This was further reinforced by studying physical activity over significant periods of time and whether there is a correlation with progression in clinical outcomes or not.

A comprehensive and comparative analysis was made using different representations of activity traces in an attempt to construct digital biomarkers from accelerometry for type-2 diabetes and all of these came to a similar conclusion. It highlighted some of the data science challenges of conducting predictive data analytics using digital signals collected in self-monitored, free-living conditions. Our research highlights the potential of digital health for enabling a vision of P4 medicine, but also thoroughly and critically points out limitations and future work needed to make this a usable and relevant paradigm in the clinical world and other healthcare applications.

4 Research Limitations

Some of the limitations in the research have already been discussed in the main research body chapters. Two key limitations are the sample size and design of the cross-sectional study. The UK Biobank is one of the largest collections biobank studies in the world, however the number of participants for whom we were able to perform a cross-sectional study from was ultimately a very small sample of the total UK Biobank population. This meant that the greatest strength of the UK Biobank, namely its great size and diversity of data collected, could not be exploited in data analyses to its fullest extent. Only one in five participants of the UK Biobank wore the wrist-worn accelerometer, and only a small sample of this accelerometer
cohort population had the correct length of wear time and then subsequently had a record associated with their type-2 diabetes disease status either at Baseline or in their electronic health records (i.e. were they diagnosed with the disease or not.) Furthermore, the number of individuals with both a confirmed diagnosis of type-2 diabetes and valid accelerometer data records were small. A larger sample size would clearly enable our results to be more scalable and robust. This issue was even more relevant in the DIRECT dataset since the number of accelerometer cohort subjects was less than 4,000 in total (both in the pre-diabetes and diabetes cohorts). Although labels are not necessarily needed for representation learning, they were needed for the machine learning models in predicting Type-2 Diabetes related clinical target outcomes from the physical activity representation spaces. This was one of the hypotheses that underpinned much of our research. The feature extraction processes themselves were independent of clinical target outcomes.

Another key limitation of the research methods is that the feature extraction methods are entirely dependent on the Oxford accelerometry analysis toolkit since only this Human Activity Recognition tool was used in our research. We have used this tool because of its established use in the literature and also provided us with a case study as it follows the pipeline for general HAR applications. But only using one tool is a huge limitation in our results. Chapter 4's high-level granular bout-based representation was extracted from the labelled activity sequences from this tool, and the latent representations learned by the autoencoders used the signal-level epoch outputs for feature learning. Although this is a point of vulnerability since it depends on the outputs of one tool, the literature review conducted an exploration into a wide range of tools available in extracting features from raw accelerometry. It was able to identify that most of these tools and technologies performed similar feature extraction processes and engineered similar features to the Oxford accelerometer analysis toolkit. As a result, it was decided that, due to the time constraints and to reduce the unnecessary complexity of our research, the Oxford accelerometer analysis tool would serve as a useful case study given its similarity to other technologies and also the results that are derived using this tool would be reflective even with other accelerometer analysis tools.

Related to this limitation is the fact that only two datasets have been used in the research for our research. Our research was heavily, almost entirely, data-driven and used only two accelerometer datasets (The UK Biobank and DIRECT). This means that the methods have only been assessed and evaluated on these two resources. It would be more robust to reproduce the analyses and methods in our research on more candidate type-2 diabetes and accelerometer datasets. However, the selected datasets are widely used in the literature and, especially in the case of DIRECT, employed rigorous protocols guided by clinical experts in its data collection process. Therefore, the two datasets used in our research were judged to be good selections. But it is recognised that reproducing these analyses on more similar datasets could reinforce the reproducibility and robustness in our research’ results.

Furthermore, the latent representations were learned only with two autoencoder architectures. However, there are other popularly autoencoder architectures that could have also been implemented and then evaluated in their capability to produce viable and powerful latent representations of the signal-level epoch data. If this had
been completed, then a more comprehensive appraisal of the power of deep learning to construct viable and powerful representations of accelerometer data.

Additionally, we used the test bed datasets as a secondary source for us, therefore we did not have much involvement or engagement with the actual participants of the cohorts in either UK Biobank or IMI DIRECT. This could be a limitation in our research because public engagement and public involvement often helps researchers to arrive at conclusions or viewpoints that may be more general and away from the scientific and clinical viewpoint, but we did not have this benefit. Furthermore, engaging directly with patients and the public participants would also help us to empower patients to involve themselves in research and, maybe more importantly, stay involved in the long-term. This is an important aspect for the future of participatory and personalised healthcare and medicine.

Finally, our research also has a clinical limitation in that only digital signals of accelerometry were considered. Obviously, there are other digital signals that could be collected that may have had a more powerful predictive capacity or played a complementary part in the analysis of digital phenotyping for type-2 diabetes. Examples could be remote glucose monitoring, dietary considerations amongst others. Some of these resources are were available in DIRECT and the UK Biobank’s vast collections of data, but our research did not exploit them in its analyses given the time constraints. our research made a focus on physical activity given its ease, ubiquitous and non-invasive means of collection.

5 Future Work

The next chapter of work to continue our research is to explore more autoencoder architectures to generate different latent representations which may perform better than the one produced by our LSTM autoencoder. Furthermore, the vast array of data that both DIRECT and the UK Biobank have collected should be investigated further and then assessed to see whether some of the phenotypes could be incorporated to improve predictive modelling results or whether the physical activity representations developed in our research can be used to complement and support to study their effects on type-2 diabetes. This could include exploiting the Electronic Health Records [47] and records for Genome-Wide Association Studies in the UK Biobank repositories [114]. The issues of using machine learning technologies to predict outcomes related to type-2 diabetes been realised in recent literature. The two main issues were the considerable heterogeneity of the population and the lack of interpretability in features used to learn predictive models [72].

The high-level granular representation physical activity from Chapter 4 can be extended further to incorporate other activity types, rather than only the ones allowed due to the Oxford accelerometer analysis toolkit (sedentary, sleep, MVPA, light tasks). The most obvious direction for future work is to explore other tools that are perform similar functionality and extract similar spaces to the Oxford accelerometer

---

2https://sites.manchester.ac.uk/reach/patient-and-public-involvement-and-engagement-ppie/
analysis tool and then applying the methods in our research to those tools which would enable. This is desirable because then our results would be more robust as it would provide a more comprehensive comparison. Examples of such tools may be Pampro, another accelerometer analysis software package for large-scale epidemiological cohorts. Furthermore, we can also revisit the GGIR package in Chapter 3 to extract the GGIR physical activity features from that tool and compare them to the features extracted from the methods developed in our research. Revisiting and comparing the performance of GGIR against the methods in our research would reinforce the decisions and validate the design of the study in our research as well make our justification more comprehensive. These are immediate next steps that could be guide the direction of future work but were not completed in the PhD due to time constraints.

Furthermore, other deep learning paradigms can also be explored such as the Generalised Adversarial Network (GAN) architectures. These neural networks have two components: The generative network and the discriminator network. The generative network produces fake candidates that are passed to the discriminator network. The discriminator then identifies this candidate as either a fake or a real, and the goal is such that the generative network will improve and reduce the error rate to an extent where the discriminator network is fooled and cannot tell the difference. This mechanism could be used to generate new synthetic “individuals” who exhibit behaviours of an individual with type-2 diabetes and a negative control individual, which the UK Biobank would provide a useful resource for, and then exploring the commonalities in these individuals or to add to the training data for the predictive modelling analyses already completed in our research.

Our research made a focus on type-2 diabetes, and all the methods developed were applied to this chronic and complex disease. Another direction for future work in a clinical setting would be to apply the methods developed in our research to a different chronic disease and it may be possible to achieve better results to reproduce these analyses on a disease that is more intimately connected to physical activity performance. The contribution this thesis makes is that it has made a comprehensive and critical comparison of different methods for extracting feature spaces from accelerometry data using large-scale epidemiological cohorts, including deep learning methods. By also applying these methods to Type-2 Diabetes, we also demonstrate the challenges in processing and extracting value from free-living and ubiquitous digital wearable technology in clinical applications outside a controlled environment in a clinic or laboratory.

Finally, as mentioned before, our research makes not only a focus on type-2 diabetes but also only on using physical activity traces as the set of objective measure from ubiquitous digital devices. This is a limitation as devices that measure other aspects of a person’s digital phenotype were not considered in the design of this study, for example data collected by the Apple Watch which can remotely measure a person’s heart rate. These are all within the realms of digital phenotyping and the work in our research could be improved by incorporating and extending them into its analyses as a complementary set of features.

\[\text{Pampro accelerometer analysis package: https://github.com/Thomite/pampro}\]
6 Final Conclusions

We have developed methods for extracting representations from raw accelerometer data derived from ubiquitous and self-monitored digital wearables in our research, and then analysed and evaluated the power of these representations against traditional feature extraction techniques. This was achieved using the techniques and technologies used in data science, especially machine learning.

A comprehensive set of data analyses contributed to fulfilling the overall high-level aim set out in the introduction in Chapter 1. The set of analyses in our research produced results with significant implications for the role of digital, ubiquitous technologies such as wearable accelerometer devices in realising a vision for the future of predictive, personalised, preventative and participatory medicine and healthcare. It highlighted the data science challenges and discussed and demonstrated this in a real-life application using two datasets widely used in other studies in the state-of-the-art in identifying different representations of clinical and digital biomarkers that could be integrated and used to learn predictive models for disease outcomes.

Furthermore, our research also provided insight into the issues of scalability in analysing and modelling from Big Health Data and the necessity in ensuring proper and rigorous quality controls on procurement for large-scale epidemiological cohort data. It has reinforced this conclusion by conducting successful predictive modelling analyses after the careful and rigorous selection of candidate subjects. Our research was also successful in showing the power of deep learning over traditional domain-driven and handcrafted feature extraction techniques, as the representations learned may show hidden variables that demonstrate habitual behaviours not captured or considered in traditional feature engineering.

Our research focused on Type-2 Diabetes. However, the actual preprocessing and feature engineering/feature learning components are an independent task from the predictive modelling, so they could be highly adaptable and applied to other epidemiological accelerometer cohort datasets or to study, or complement the study, of other chronic diseases growing in prevalence due to the growing challenge of ageing populations globally. A multi-faceted, slow-moving and complex chronic illness like Type-2 Diabetes may have been an inappropriate candidate disease for this study with the feature spaces we selected or developed. In conclusion, before a P4 vision for the data-driven future of medicine and healthcare can be realised, significant work still remains to be done to fully exploit the value within the digital signals of ubiquitous technologies that fully enable a reliable digital phenotype for chronic disease.
References


J. D. Bell, J. Beulens, S. Brage, S. Brunak, F. De Masi, E. T. Dermitzakis, 
I. M. Forgie, G. Frost, T. Hansen, T. H. Hansen, A. Hattersley, T. Kokkola, 
A. Kurbasic, M. Laakso, A. Mari, T. J. McDonald, O. Pedersen, F. Rutters, 
J. M. Schwenk, H. J. A. Teare, E. L. Thomas, A. Vinuela, A. Mahajan, 
M. I. McCarthy, H. Ruetten, M. Walker, E. Pearson, I. Pavo, P. W. Franks, 
and IMI DIRECT Consortium. The role of physical activity in metabolic 
homeostasis before and after the onset of type 2 diabetes: an IMI DIRECT 

Giordano, T. H. Hansen, M. Hudson, A. D. M. Koopman, F. Rutters, 
M. Siloaho, K. H. Allin, S. Brage, C. A. Brorsson, A. Y. Dawed, 
F. De Masi, C. J. Groves, T. Kokkola, A. Mahajan, M. H. Perry, S. P. Rauh, 
M. Ridderstråle, H. J. A. Teare, E. L. Thomas, A. Tura, H. Vestergaard, 
T. White, J. Adamski, J. D. Bell, J. W. Beulens, S. Brunak, E. T. Dermitzakis, 
P. Froguel, G. Frost, R. Gupta, T. Hansen, A. Hattersley, B. Jablonka, J. Kaye, 
M. Laakso, T. J. McDonald, O. Pedersen, J. M. Schwenk, I. Pavo, A. Mari, 
M. I. McCarthy, H. Ruetten, M. Walker, E. Pearson, P. W. Franks, and 
for the IMI DIRECT Consortium. Discovery of biomarkers for glycaemic 
deterioration before and after the onset of type 2 diabetes: descriptive 
characteristics of the epidemiological studies within the imi direct consortium. 

review of classification and combining techniques. *Artificial Intelligence 

[118] M. A. Kramer. Nonlinear principal component analysis using autoassociative 


[120] B. Lam, M. Catt, S. Cassidy, J. Bacardit, P. Darke, S. Butterfield, 
O. Alshabrawy, M. Trenell, and P. Missier. Using wearable activity trackers to 
predict type 2 diabetes: Machine learning–based cross-sectional study of the 

[121] M. J. LaMonte, S. N. Blair, and T. S. Church. Physical activity and diabetes 
16103523.

Yeoh, T. Tam, M. C. S. Wong, and G. N. Thomas. Increased leisure-time 
physical activity associated with lower onset of diabetes in 44 828 adults with 
impaired fasting glucose: a population-based prospective cohort study. *British 

using wearable sensors. *IEEE Communications Surveys & Tutorials*, 


