Physical Activity Recognition and Identification System

(PARIS)

Submitted by Joshua George Nicholas Twaites to the University of Exeter

as a thesis for the degree of

Doctor of Philosophy in Sports and Health Sciences

In November 2020

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

Signature: 

Abstract

Background:

It is well-established that physical activity is beneficial to health. It is less known how the characteristics of physical activity impact health independently of total amount. This is due to the inability to measure these characteristics in an objective way that can be applied to large population groups. Accelerometry allows for objective monitoring of physical activity but is currently unable to identify type of physical activity accurately.

Methods:

This thesis details the creation of an activity classifier that can identify type from accelerometer data. The current research in activity classification was reviewed and methodological challenges were identified. The main challenge was the inability of classifiers to generalize to unseen data. Creating methods to mitigate this lack of generalisation represents the bulk of this thesis. Using the review, a classification pipeline was synthesised, representing the sequence of steps that all activity classifiers use.

1. Determination of device location and setting (Chapter 4)
2. Pre-processing (Chapter 5)
3. Segmenting into windows (Chapters 6)
4. Extracting features (Chapters 7,8)
5. Creating the classifier (Chapter 9)
6. Post-processing (Chapter 5)

For each of these steps, methods were created and tested that allowed for a high level of generalisability without sacrificing overall performance.

Results:

The work in this thesis results in an activity classifier that had a good ability to generalize to unseen data. The classifier achieved an F1-score of 0.916 and 0.826 on data similar to its training data, which is statistically equivalent to the performance of current state of the art models (0.898, 0.765). On data dissimilar to its training data, the classifier achieved a significantly higher performance than current state of the art methods (0.759, 0.897 versus 0.352, 0.415). This shows that the classifier created in this work has a significantly greater ability to generalise to unseen data than current methods.

Conclusion:

This thesis details the creation of an activity classifier that allows for an improved ability to generalize to unseen data, thus allowing for identification of type from acceleration data. This should allow for more detailed investigation into the specific health effects of type in large population studies utilising accelerometers.

Acknowledgements

I would like to thank my supervisors Associate Professor Melvyn Hillsdon, Joss Langford, and Professor Richard Everson for giving me the opportunity to conduct this PhD research and for the support and encouragement I have received over the last three years. It is through your efforts that I have achieved anything in the last three years.

I would also like to thank Activinsights for funding this research.

Contents

Declaration

Chapters 4: Accelerometer Placement Location and 6: Segmentation of Acceleration into Windows are based on previously published papers:

Twaites, J., Everson, R., Langford, J. and Hillsdon, M., 2019. Achieving Accelerometer Wrist and Orientation Invariance in Physical Activity Classification via Domain Adaption. *Journal for the Measurement of Physical Behaviour*, pp.1-7

Twaites, J., Everson, R., Langford, J. and Hillsdon, M., 2020. Transition Detection for Automatic Segmentation of Wrist-Worn Acceleration Data: A Comparison of New and Existing Methods. *Journal for the Measurement of Physical Behaviour*, pp.19-28.

Glossary

Acceleration trace: a segment of acceleration data over time is referred to as the trace of the acceleration.

Accelerometer: accelerometers are participant-mounted devices that measure acceleration in 1-3 dimensions depending on the number of axes.

Activation function: the function applied to the weighted sum of the node inputs to determine the output of a node in a neural network.

Activity classes: the different groups of activities categorised by the activity classifier, as defined by the labelling schema used.

Activity classification: using pattern matching algorithms to match acceleration data with the corresponding activities.

Activity protocol: how the activities in the study were performed, hence how the acceleration and the corresponding labels were gathered.

Activity transition: transitioning from one activity to another.

ActivPAL: a thigh-mounted accelerometer in this case measuring at 20Hz. This is used to determine the true labels in the free-living data.

Amalgam approach: a method for achieving wrist orientation invariance, where the classifier was trained on data from both wrists.

Automatic derived features: features that are derived from the data itself with respect to either the classification problem or in a more general sense.

Base Classifier: a classification pipeline based on the work of Chowbury et al (55) which is used as the criterion measure.

Basis vectors: vectors that can recreate a data-set through linear combinations

Bias-variance trade-off: this refers to attempting to minimise two sources of error in supervised learning. The bias refers to the ability of the model to capture the relationship between the training data and labels. Whereas the variance refers to the fluctuation in the relationship caused by minor changes in the training data

Bode Plot: a plot showing the impact of a filter on the frequency of a signal

Bonferroni correction: a method of reducing the likelihood to type 1 errors when testing multiple hypothesis. This entails testing each individual hypothesis at a significance level of $\frac{\alpha}{\mathrm{m}}$, where α is the overall hypothesis level (in this case 0.05) and m is the number of hypotheses.

Bouts: a continuous stretch of physical activity.

Butterworth filtering: a filtering method that can be used for the attenuation of high-frequency data from a time series. A butterworth filter is mathematically optimal for removing the higher frequencies without affecting the lower frequencies in the data.

Change Point Detection (CPD): a data driven method for detecting if the underlying process generating time series data changes.

Class balance: this refers to the proportion of the different classes (activity labels) in the data-set.

Classification pipeline: the six-step process used to develop the classifier

1. Determination of device location and setting
2. Pre-processing
3. Segmenting into windows
4. Extracting features

5. Creating the classifier
6. Post-processing

Classifier: a function that maps input data to the desired output, the outputs being discrete classes or labels.

Cross-validation: when the training data is partitioned into a training set and validation set. The classifier is trained to minimise the error on the training set and the performance is estimated on the validation set.

Data aggregation: combining the three acceleration data streams into one aggregate stream.

Data-set shift: a difference in the training and testing data caused by being from different protocols and participants.

Determinism: a measure of the predictability of the dynamical system modelled by the recurrence matrix, used with recurrent quantification analysis.

Discriminative classifier: when given $D, T = \{d_n, t_n\}_{n=1}^{N}$, with d_n and t_n being the n^{th} data point with the corresponding label, a discriminative classifier attempts to model the conditional probability of T given D or $P(D|T)$.

Divergence: a measure of the predictability of the system, used with recurrent quantification analysis.

Domain adaption: a method of adapting data from the target domain to the source domain so that good performance is achieved, thus mitigating the effect data-set shift.

Domain adaption approach: a method for achieving wrist orientation invariance, where the classifier used domain adaption to align the testing data with the training data.

Dynamical system: a system where how much the current point depends on previous points changes with the value of the current point. It is assumed the physical activity can be modelled as such a system.

Ensemble model: a classification model comprised of combinations of multiple models in order to achieve greater classification performance than any of the constituent models. A random forest is such a model.

Entropy: the average rate at which information is produced by a stochastic process. In the case of a signal, this is a measure of the signal's complexity.

Error function: a function indicating how accurately the classifier can identify the activity labels from the acceleration data when the correct labels are known.

Euclidean Norm Minus One (ENMO): this is a data aggregation method, when all three axes are combined into one. This allows for orientation invariance.

F1-score: the evaluation metric used, the harmonic mean of precision and recall. This metric was chosen instead of simpler metrics such as accuracy because the F1-score is typically more robust to class imbalances.

Feature reduction: reducing the number of features used in the classifier, in this work principal component analysis is used as a feature reduction methodology.

Features: a set of attributes (consistent among all windows) that describe the windows that are identified and used to represent the windows, for example; identifying the mean acceleration for each axis and their standard deviations

Filtering: creating an approximation of the time series that can capture important patterns but is less affected by noise. There are three common forms of filtering: low-pass (removing all frequencies higher than a threshold), high-pass (removing all frequencies lower than a threshold) and band-pass (a combination of high and low, keeping only the frequencies between two thresholds).

Fourier transforms: a way to approximate functions/signals by sums of trigonometric functions/signals.

FreeCV performance: the F1-Score of the classification pipeline when trained and tested on the free-living data using leave one subject out cross validation, this is one of the intra-protocol performances.

Free-Lab performance: the F1-Score of the classification pipeline when trained on the free-living data and tested on the lab-based data, this is one of the inter-protocol performances.

Free-Living: this refers to acceleration data not gathered under a specific activity protocol, thus is more representative of realistic activities.

Frequency: one of the characteristics of physical activity; the number of distinct physical activity events over the measurement period, sometimes of a specific type, that occur.

GENEActiv: a wrist-mounted tri-axial accelerometer.

Generalisability: the ability of a classifier to perform on data different to the training data.

Generative adversarial network: this is a classifier that consists of two neural networks, a generator and a discriminator. The generator learns to create synthetic data that is indistinguishable from real data (the training data); the discriminator attempts to identify if data is real or synthetic, this process allows for a high level of classification performance.

Generative classifier: when given observable variables $D, T = \{d_n, t_n\}_{n=1}^{N}$, with d_n and t_n being the n^{th} data point with the corresponding label, a generative classifier attempts to model the joint probability distribution $P(D, T)$.

Generative-Discriminative pair: two classifiers which exist in generative and discriminative ‘forms’ when the underlying model is the same. For example, logistic regression and naive bayes.

Heuristic Orientation Invariant Transformation (HOIT): a method for allowing orientation invariance without this data loss. This entails transforming the 3-D acceleration data into 9-D orientation invariant data.

Hidden Markov Model (HMM): Hidden markov models are statistical models that can be used to describe the creation of an observable time series, making use of internal factors that are not directly observable. These models can then be used for post-processing.

Hyperparameters: these refer to the many modifiable characteristics of the classification pipeline, such as window size and features used. These are akin to parameters in f but over the entire classification procedure.

Imbalanced Classes: having more of one kind of activity than others in the data, this tends to decrease the performance.

Inclination correction: a procedure for altering the acceleration values when the accelerometers is at the correct orientation but may have moved slightly, shifting the values.

Inclusion/Exclusion criteria: this refers to the criteria that the participants must fulfil to be included in this study. This is to ensure that the participant’s physical activity will not be affected by major health issues.

Intensity: this refers to the energy expenditure of physical activity. This is generally measured in Metabolic Equivalents (METs) which identify multiples of the energy expenditure of the physical activity compared to lying supine. MET values can be grouped into four categories:

1. Sedentary (less than 1.5 METs)

2. Light (between 1.5 and 3 METs)
3. Moderate (between 3 and 6 METs)
4. Vigorous (greater than 6 METs)

Inter-protocol-inter-subject: this is where the test data is from different protocols with different participants to the training data. This is also referred to as just the inter-protocol performance.

Inter-protocol-intra-subject: this is where the test data is drawn from the same participants as the training data, but from two different activity protocols.

Intra-protocol-inter-subject: this is where the test data has different participants from the same protocol as the training data.

Intra-protocol-intra-subject: this is where the test data is from the same participants and protocol as the training data, separated only in time.

Kozina's method: this is a data driven approach for transition detection that identifies points where there is a "significant change between consecutive data samples and divides the data into intervals at that point" (112).

Lab-Based: this refers to data collected in a laboratory setting with participants performing an activity protocol reflecting activities of daily living.

LabCV performance: the F1-Score of the classification pipeline when trained and tested on the lab-based data using leave one subject out cross validation, this is one the intra-protocol performances.

Lab-Free performance: the F1-Score of the classification pipeline when trained on the lab-based data and tested on the free-living data, this is one of the inter-protocol performances.

Leave-One-Subject-Out-Cross-Validation (LOSOCV): this is a method for validating a classifiers performance. It works by training the classifier on all but

one participant, and then evaluating the performance on the remaining participant. This procedure is repeated for all participants and the averaged evaluation metric is reported (although the individual performances are retained for statistical testing). This gives an idea of the performance of the classifier over each participant.

Logistic Regression: this is a classification model that attempts to model the probability conditional distribution $p(Y \mid X)$ given observable variables X (input data, acceleration features) and target variable Y (output labels, activity labels). This and naive bayes for a discriminative-generative pair.

Lyden's method: this is a method of transition detection, it is a data driven method that identifies instances of rapid acceleration/deceleration and divides the data at those points.

Machine learning: this is a form of artificial intelligence that builds a 'classifier' based on data in order to make predictions or decisions.

Margin: when performing transition detection, it was noted that the transitions were not instantaneous, therefore a precision at the sampling rate was deemed unfeasible. A transition was considered to have been correctly detected (a true positive) if it was within a specified temporal "margin" of the labelled true transition between activities, this value was set to 3 seconds.

Mathews Correlation Coefficient (MCC): this was a performance metric for the transition detection. This is a correlation coefficient between the observed and predicted binary classification of a transition that takes into account true and false, positives and negatives and is generally regarded as a balanced metric which can be used even if the class sizes are very different.

Mean Minimum Distance: this is a metric used for transition detection, that reports the mean minimum distance between true and detected transitions.

Metabolic Equivalents (METs): these are units of physical activity intensity which identify multiples of the energy expenditure of the physical activity compared to lying supine.

Metabolic Equivalent-hours: these are a unit of physical activity volume, the average MET values accrued per hour (15 minutes of an 8 MET activity is equivalent to 2 MET hours).

Morphology based features: these are features that are based on the shape of the acceleration trace (the morphology), as opposed to statistical features that describe them.

Moving average: this is a ‘dynamic average calculated across successive segments of data’ (typically of constant size and overlapping) of a series of value.

Moderate-Vigorous physical activity: the time spent in moderate to vigorous intensity, this is part of physical activity guidelines and commonly used in health messages.

Naive bayes: this is a classification model that attempts to model the probability conditional distribution $p(Y \mid X)$ given observable variables X (input data, acceleration features) and target variable Y (output labels, activity labels). This and logistic regression for a discriminative-generative pair.

Neural network: this is a discriminative classification model that is based on an abstraction of human cognition. A neural network attempts to directly model the decision boundary between classes.

Noise: noise refers to one of two things: observational noise: random disturbances in the signal caused by the device (typically ‘gaussian noise’), or additional information in the signal that is not useful for the activity classification.

A 45Hz signal in the acceleration is not random but is not useful in activity classification and will disturb the acceleration values, hence is treated as noise.

Non-Domain adaption approach: a method for achieving wrist orientation invariance. In this approach a classifier was trained from the data of one wrist. The resultant classifier was then used to classify the data from the opposite wrist with no modification, this was used as the control for the other wrist orientation invariance methods.

Normalization: this is a procedure used to ensure that all features have similar variance, meaning that all features have an equal weighting in the data-set. This procedure is often used in activity classification work, although is not required in most cases.

Not-applicable approach: a method for achieving wrist orientation invariance. In this method a classifier was trained from the data of one wrist. Domain adaption was then used with the same wrist data serving as the target domain. The resultant domain adapted classifier was then used to classify the data from the same wrist. This method served to investigate the effect of using domain adaption when it is not required, in circumstances where the wrist placement of the accelerometer is unknown.

Nyquist-Shannon theorem: a theorem that states that for a successful reconstruction data needs to be sampled with at least twice its highest frequency, in the case of 100Hz accelerometry data the cut-off frequency should be between 30-40Hz, and this is the minimum required, often 3 or 4 times the highest frequency is preferable.

Online Bayesian Change Point Detection (OBCPD): this is a method for transition detection that works by “estimating the posterior distribution over the current ‘run length’, or time since the last change point, given the data so far observed”. This means that when the change points are computed, both the

probability that each successive point does not belong to the same distribution as previous points and length of runs are estimated.

Orientation invariance: this means that features computed will be identical regardless of the orientation of the sensor. This is required due to inconsistency in the positioning of the accelerometers on participants.

Over-complete dictionary: this refers to using more basis vectors than the minimum required amount of basis vectors when using sparse feature encoding. This allows for more resistance to noise.

Overfitting: this is when the classifier is overspecialised to the training data-set; decreasing the classification error by modelling the noise of the data-set, as well as the mapping function. Modelling the noise allows for a greater ability to classify the training data but reduces a reduced ability to generalise to unseen data.

Overlap: this is a modification to windowing approaches where the sequential segments used to create the windows are not separate but instead share a portion of their data (they overlap). This overlap is most commonly 50% although other proportions are used.

Oversampling: this is a pre-processing technique which refers to generating synthetic data from the under-populated classes in order to make the number of examples from each class equal.

Participant Adaption via Iterative Relearning (PAIR): this is a post-processing method that attempts to use the participant's own data to retrain the classifier and improve the classification.

Participant characteristics: the anthropometric characteristics of the participants in a data set. These are used for evaluating how closely the participants resemble various populations and each other.

Physical activity: any bodily movement produced by skeletal muscles that requires energy expenditure.

Post-processing: after the acceleration data has been classified, the predicted labels may be processed in order to reduce the number of misclassifications. This is referred to as post-processing. Most post-processing approaches use the sequential nature of activity data to improve performance, making use of the fact that adjacent segments are likely to be the same activity.

Pre-Post-Combined: this refers to the final pre-post processing methods used in this work this was a combination of:

- Using structure preserving oversampling to rebalance the classes
- Using ENMO with X, Y, Z acceleration streams
- Using participant adaptation via iterative relearning
- Using a hidden markov model
- Using a smoother with $n = 11$

Pre-processing: this refers to the preliminary processing of acceleration data before any classification steps are carried out and has a range of uses such as: allowing for rotational invariance and the removal of noise.

Principal Component Analysis (PCA): this is a form of feature reduction that works by projecting high dimensional data (in this case 39) into a smaller number of dimensions - a subspace - while preserving as much variance as possible. The resulting low-dimensional features are linear combinations of the original, high-dimensional features.

Quadratic discriminant analysis (QDA): quadratic discriminant analysis is a generative classification method. The model models the class conditional distribution of the data. This method assumes that all features are normally distributed, which greatly decreases the number of parameters required to be learned in the classification training.

Random forest: random forests are an extension to decision trees that can generalise performance to unseen data (prevent overfitting). Random forests are a combination of multiple decision trees (an ensemble) trained on subsets of the training data. The output of a random forest is the majority predicted classification of all trees.

Ratio of Mean Minimum Distance (RMMD): this is the ratio of the mean minimum distance of a transition detection method and a naive transition detection method that detects the same number of transitions. This is used because mean minimum distance does not penalise false positives.

Ratio of Sensitivity (RoS): this is the ratio of the sensitivity of a transition detection method and a naive transition detection method that detects the same number of transitions. This is used because sensitivity does not penalise false positives.

Recall bias: this refers to the systematic error caused by participants incompletely recalling their physical activity.

Recurrence matrix: this is a binary matrix that model how recurrent a signal is. If the distance between points i, j is less than the threshold ε, then point $R(i, j)$ is 1, else it is 0. This is also referred to as a recurrence plot.

Recurrence rate (RR): this is a metric used in recurrent quantification analysis, it is the density of the recurrence points in the recurrence matrix. This corresponds with the probability that any given state will recur in the signal.

Recurrent quantification analysis (RQA): recurrent quantification analysis is a method of statistically analysing data generated by dynamical systems; specifically, it is a way of analysing recurrence plots of a dynamical system (150). Recurrence plots identify the states at which a system approximately repeats a previous state. These recurrence plots characterise the structure of the dynamical system: simple dynamical systems, such as a limit cycle, have

simple recurrence plots with few points of recurrence, while complex dynamical systems will have many points of approximate recurrence. The features extracted through recurrent quantification analysis describe this recurrence plot and hence the structure of the acceleration data.

Refractory period: during the process of investigating the transition detection methods, a limitation of current methodologies (including OBCPD) was identified. It was found that multiple transitions, in close proximity to one another (within 1 second), were predicted when only a single true transition occurred. It appears that these multiple transitions were identified because of the (incorrect) assumption that the transitions are instantaneous (or occur < 1 second). Hence a post-processing method to supress these additional detected transitions was developed, this was the refractory period. Responsiveness validity: the ability to detect behaviour change over time.

Root Mean Squared Error (RMSE): this is a metric used in transition detection, this is computed by calculating square root of the mean squared time difference between each detected transition and the closest true transition. This evaluation informs how close detected transitions are to the true locations.

Rotation invariance: see orientation invariance.

Salcic's method: this is a method of transition detection; it is a classification-based approach that creates a classifier trained on some labelled training data to identify whether a three second moving window contains a transition. It does this by forming a decision tree based on the absolute mean difference of the acceleration in the three second window.

Sedentary-Stand-Active Labelling: this is the labelling scheme used by the activPals therefore it is also used in with the free-living data. It is comprised of three activities: Sedentary, Standing and Active.

Segmentation: this is when the acceleration values are segmented into short duration windows (typically around 10 seconds) during the classification pipeline. This is because performance is improved when classifying a window of acceleration data rather than a single instantaneous value. Observing multiple acceleration values allow an indication of how the values are changing in time, unlike a single instantaneous value, hence classification is improved

Self-report: this is a physical activity monitoring method whereupon the participants are asked to report their performed physical activity. This is a subjective method and therefore prone to many forms of bias.

Self-training: this is a form of semi-supervised learning utilised in participant adaption via iterative re-learning. It refers to re-training a classifier on the most confident of its own predictions in order to better adapt it to the testing data.

Sensitivity: this is a metric used in transition detection, this metric is used to determine how often the transitions are detected correctly.

Smoothing: this is a post-processing method that refers to applying a modal filter to the predicted labels, such that each predicted label is replaced by the most common label of the n closest labels (inclusive of itself).

Social desirability bias: this is a form of bias where participants respond with answers that will be viewed more favourably by others, in this case this refers to under-reporting physical inactivity and over-reporting physical activity.

Source domain: this is the domain that the training data is said to have come from, typically in machine learning it is assumed that the source and target domains are the same. When this assumption is violated performance worsens, domain adaption can be used to mitigate this issue.

Sparse Feature Encoding (SFE): this is a form of creating morphology-based features. Sparse feature encoding creates features by decomposing the

acceleration segments into simple filters, such that the initial data can be recreated via linear combinations of the filtered signals. Sparsity refers to limiting the number of filters that need to be ‘activated’ to recreate any one signal. The sparsity helps to eliminate noise and ensures that the signals do not simply perform a Fourier decomposition. Features are created by identifying the activations of the individual signals required to recreate the data.

Statistical based features: these are features that attempt to represent the data with a single aggregated value, such as the mean or skewness of the acceleration.

Structure Preserving Oversampling (SPO): this is an oversampling technique that is specific for time series data as it generates synthetic samples while preserving the covariance structure of the data (therefore not weakening the correlation structures as is normally an issue in oversampling).

Subspace alignment: this is a domain adaption method. The underlying idea of the subspace alignment algorithm is to rotate the source data so that it best aligns with the target data; a classifier is then trained on the aligned source data in order to be able make accurate predictions on the target/test data. Prior to alignment, the source and target data are each projected into a subspace defined by their principal components. This identifies the principal directions in the data that should be aligned by rotation.

Supervised Learning: this is a form of machine learning. Supervised learning makes use of a data-set containing both inputs (acceleration values) and desired outputs (activity labels) known as the training data. In supervised learning, the classifier is ‘trained’ to identify the outputs from the inputs, in this case creating a classifier that can identify physical activity type from the acceleration data

Takens' theorem: Takens' theorem (154) states that a dynamical system (D) can be reconstructed from a sequence of observations (o) and the state of the system using a time delay τ and an embedding dimension m, such that:

$$D(i) = \big(o(i), o(i+\tau), o(i+2\tau), \dots, o(i+(m-1)\tau)\big)$$

In the context of this work, the participant's physical activity is the dynamical system and the sequence of observations are the acceleration values observed. For different axis combinations these observed values can be 1-3 dimensional (X, Y, Z). According to Takens theorem, it is possible to recreate the dynamical system (the participants physical activity) from these observed values (the accelerations).

Target domain: this is the domain that the testing data is said to have come from, typically in machine learning it is assumed that the source and target domains are the same. When this assumption is violated performance worsens, domain adaption can be used to mitigate this issue.

Test_SFE: using sparse feature encoding on the test data refers to generating the features from the testing data, then extracting these features from the training data and using this to create a classifier. This classifier then attempts to classify the test data. The classifier is trained on the training data first, but the features used are identified from testing data.

Testing data: this data has the same structure as the training data, with acceleration values and corresponding activity labels. However, this data is not used in training the classifier but instead to test the performance of the classification after all optimisations have taken place. The classifier is used to predict the activity labels from the acceleration data, yielding predicted activity labels. These predicted labels can then be compared to the known true labels to evaluate the performance of the classifier.

Train_SFE: using sparse feature encoding on the training data is (train_SFE) refers to generating the features from the training data, then extracting these features from the training data and using this to create a classifier. This classifier then attempts to classify the test data.

Training (a classifier): training a classifier means identifying the parameters that minimise the error of the classification, finding $\theta^* = \underset{\theta}{\operatorname{argmin}} \sum_i E_f\big(t_i, f(d_i;\theta)\big)$. The error function is often chosen to be the misclassification rate. Once the best parameters are known, they may be used to estimate (or predict) the unknown labels t' corresponding to new observations: $t' = f(d', \theta^*)$.

Training data: this is a time series of acceleration values with corresponding activity labels. This data goes through the classification pipeline, as discussed above, training the classifier. This data is also referred to as ‘seen’ data because the classifier has encountered it.

Transition probabilities: these represent the chances of transitioning from one activity to another in the activity protocol. These probabilities are used in post-processing methods to improve performance (68). These probabilities are computed by first segmenting the data into 12.8 seconds blocks and then computing the transition probabilities (the choice of 12.8 seconds is explained in section 3.5).

Type: this refers to the actual activity performed, such as walking or rowing. The type of physical activity has been shown to affect health outcomes

Under-sampling: this refers to removing samples from the well-populated classes until all populations are equal. As under-sampling reduces the amount of data in the training set it is typically not preferred.

Validation data: this data is drawn from the training data and used for optimising the various hyperparameters of the classification process.

Volume: frequency, intensity and duration are regularly used together to estimate the volume of physical activity undertaken. Volume can be expressed as MET-hours, the average MET values accrued per hour.

Wilcoxon signed rank test: this is a statistical test which tests the null hypothesis that two related paired (by participant ID) samples come from the same distribution. A low p-value $(p < 0.05)$ indicates that the results are statistically significantly different from one another with high confidence.

1. Introduction

It is well established that a physically active lifestyle is associated with numerous health benefits and increased longevity when compared to a sedentary lifestyle (1). Physical activity (PA) is inversely associated with many of the common causes of premature death, including: Coronary Heart Disease (CHD), Cancer, Chronic Lower Respiratory Disease, Stroke, Alzheimers Disease, and Diabetes (2). Globally, the World Health Organisation have estimated that of the 57 million deaths per year, 5.3 million can be attributed to physical inactivity (2).

Although there is consensus that PA is beneficial to health, uncertainty still exists about the precise 'dose' of PA required and how the dose should vary for different populations and disease groups. Additionally, the effects of the different characteristics of PA (such as type, frequency, and intensity) are still uncertain. This is in part due to flaws in the measurement process of PA. Methods of identifying PA, both the total volume and the individual characteristics, are imperfect and do not always capture correct information (3). The majority of studies that provide the underpinning evidence of the benefits of PA rely on self-reporting, which is often imprecise and overly simplistic. Self-reports of PA are prone to a number of biases, including recall and social desirability bias (4,5).

1.1 Physical Activity Characteristics

The term PA comprises a wide range of behaviours and is defined as 'any bodily movement produced by skeletal muscles that requires energy expenditure' (24). The amount of energy expended (often expressed in kilocalories) while undertaking PA is determined by the amount of muscle mass involved and the frequency, intensity and duration of muscular contractions. Typically, PA is reported as a single aggregate measure of volume (25).

Aggregating PA into a single metric is useful for studies that rank groups of people according to their level of activity and test associations with health outcomes. However, it may obscure differential associations with health due to the different contributions of the Frequency (F), Intensity (I), Time (T) and Type (T) characteristics of PA (FITT). The evaluation of the FITT characteristics of a given volume of PA allows for investigation into how they are independently associated with health (26,27).

Frequency refers to the number of distinct PA events over the measurement period, sometimes of a specific type, that occur. In most cases the type of PA is not considered relevant, and studies simply attempt to identify the number of the events. In a study of 97230 participants, it was found that the frequency was associated with incident of CHD risk. However, this association disappeared when controlling for total amount of PA performed (26). An additional study in Canadian adults reported similar findings, that the frequency of PA was associated with the incident of Diabetes, however this association also disappeared when controlling for total PA (28). Similar results were found by O'Donovan et al, who found that the associations of PA with mortality remained the same regardless of if the PA was accrued in two days per week or 5 days per week (29). These results show that despite frequency often being associated with health outcomes, this appears to be a reflection of the correlation between frequency and total volume.

The absolute intensity of PA is generally measured in Metabolic Equivalents (METs) which identify multiples of the energy expenditure of the PA compared to lying supine. MET values can be grouped into four categories: Sedentary (less than 1.5 METs), Light (between 1.5 and 3 METs), Moderate (between 3 and 6 METs) and Vigorous (greater than 6 METs) (30). Some common activities and their MET values, along with the intensity classes can be seen in Table 1.

Activity name	Intensity class	MET value
Sleep	Sedentary	0.9
Watching TV	Sedentary	1.0
Slow walking (1.7mph)	Light	2.3
Moderate walking (3.4mph)	Moderate	3.3
Jogging	Vigorous	7.0

Table 1: Common activities with MET values and intensity class (30).

PA guidelines are mostly based on absolute intensity, e.g., ≥150 minutes of at least moderate intensity. Relative intensity can also be used and represents the energy expenditure of an activity relative to an individual's fitness. Although 3 METs is used as the absolute threshold of moderate intensity, 3 METs may be light relative intensity for people with higher fitness. Absolute intensity measures are used in surveillance studies and may lead to an over or underestimate of the true prevalence of PA in a population. For example, the average time spent in moderate to vigorous intensity (MVPA) may be overestimated in younger, fitter populations but underestimated in older less fit populations.

The effect of intensity of PA events is a contentious issue, one study identified that it was only high intensity PA that was positively associated with health outcomes (31). This contrasts with a review by Chastin et al (32) who identified that short but frequent bouts of light-intensity activity throughout the day reduced postprandial glucose by -17.5%. Additionally, 6 out of 8 prospective observational studies reviewed showed time spent in lower intensity PA was associated with lowered mortality (32). This agrees with other work that has identified that light intensity PA appears to be beneficially associated with important health outcomes after adjustment for MVPA in the adult population (33). The lack of agreement may be because of the use of absolute measures of intensity; the activities labelled as light will actually be a relative moderate intensity for some people and activities labelled as moderate will be a relative light intensity for others (34).

Time refers to the duration of each PA event (usually given in minutes) or the total duration of the PA undertaken in the observation period and is mostly correlated with volume. Time alone does not appear to have any association with health outcomes (27). A meta-analysis of intensity and duration on cardiometabolic risk in children and adolescents reported that duration has no significant association with cardiometabolic risk markers (35). The scientific report underpinning the 2018 US PA guidelines stated that sustained periods of MVPA were of no advantage over intermittent MVPA, hence the removal of a 10 minute minimum bout duration (36). However, total volume of PA was not controlled for in the studies cited as references for this decision and therefore, the independent effect of bout duration may not have been properly tested.

Frequency, intensity and duration are regularly used together to estimate the volume of PA undertaken. Volume can be expressed as MET-hours, the average MET values accrued per hour (15 minutes of an 8 MET activity is equivalent to 2 MET hours).

Type refers to the actual activity performed, such as walking or rowing. The type of PA has been shown to affect health outcomes. Lee et al (37) identified that non-runners who met PA guidelines have a higher risk of developing disease than runners who did not meet PA guidelines (18% higher chance of disease), although both had lower risks than non-runners who didn't meet PA guidelines. Furthermore, Chomistek et al (26) identified that separately, running, tennis, and brisk walking were inversely associated with CHD risk, validating the health effects of type on disease.

While there is clear evidence showing the benefits of both PA overall and the individual characteristics of PA, there is uncertainty in the precise relationships between PA (and its characteristics) and health. This in part due to limitations in the measurement of these characterisations. Another impact of these limitations is the decreased ability to determine prevalence of PA.

1.2 Prevalence of Physical Activity

In the UK it is estimated that 41% of adults are insufficiently active (15), not performing 150 minutes of moderate activity per week (10). Agreement about prevalence in other countries varies widely. This is mainly for two reasons: definitions of 'sufficiently active' and differences in measurement instruments within and between countries. Instruments such as the Global Physical Activity Questionnaire (GPPAQ) (17) and the International Physical Activity Questionnaire (18) were designed to try and harmonise measurement of prevalence between countries. However, even these two instruments provide different estimates of prevalence when employed in the same population. The level of detail on frequency, intensity, time and type of PA vary widely between measures used in population surveillance. Most studies focus on frequency, intensity and duration, giving an estimate of volume, but the level of detail on type varies considerably. For example, the GPPAQ (17) investigates just three types of activity – work, travel and recreation whereas the Health Survey for England (19) enquires about an extensive list of recreational activities. Some studies include work but not active travel and some vice versa. In many studies, including the Health Survey for England, information on type is used to help estimate time spent at different intensities. For example, the intensity coding of sports is based on the type of sport reported and the perceived exertion when undertaking the sport.

Whilst there is consensus that too many people are insufficiently active to benefit their health, there is uncertainty about the precise level of prevalence. As mentioned above, the source of this uncertainty lies in the challenge of measuring PA. Most PA surveillance systems rely on self-reports of PA and are subject to the biases described above.

The multi-dimensional nature of PA and its difficulties in measurement are some of the key reasons for uncertainties in the magnitude of the association between PA and health, prevalence estimates, and the effectiveness of interventions.

1.3 Measurement uncertainty of Physical Activity

The multi-dimensional nature of PA outlined above, and the associated measurement challenges explain much of the uncertainty that still remains in our understanding of the relationship between PA and health and the prevalence of PA. As early as the mid 80's concerns were being raised about the multitude of PA measures being employed and to what extent this explained inconsistencies in findings – mainly regarding the importance of PA intensity (42).

The relative importance given to each dimension of PA will depend on the health outcome being targeted and the population being studied. Intensity may be most important for cardiac health (43) whereas type may be most important for bone health (38).

Improving the precision of the measurement of PA overall and its sub-dimensions would lead to a better understanding of the true magnitude of the relationship between PA and health, more tailored PA guidelines, a better estimate of the population prevalence of PA, more accurate screening instruments (ensuring those most in need get interventions), and a better understanding of which interventions are effective in increasing PA. Identification of each of the characteristics of PA will allow for a greater understanding into how they individually impact health outcomes.

1.4 PA measurement methods

Many methods for the observation and recording of PA exist, with the most common method utilized being a Self-Report Questionnaire (13). This entails a retrospective questionnaire focusing on volume and type of activity performed by the participant over a given time period. A similar method is that of an Activity Diary where the participant records their PA as it is performed (13). Questionnaires and diaries are both a subjective form of measurement which

can be affected by social desirability bias and variations in how the questions and activities themselves are understood (14). Self-report questionnaires can investigate all forms of PA characteristics, although they must be specified in the questions. However, both questionnaires and activity diaries are subjective method of measurement and therefore severely impacted by bias. Additionally, both activity diaries and questionnaires have very low resolutions as asking the participant to record events in less than 5-minute intervals in unfeasible.

Another method for the evaluation of PA is direct observation (DO) (44,45). This involves the researcher directly observing the participant. Unlike self-report methods this is an entirely objective form of measurement and therefore not subject to participant bias. DO can identify and record PA as it happens, which allows for a high resolution, additionally, type, intensity, frequency and time spent in PA is easily identified, allowing for full evaluation of PA characteristics. However, DO is highly intrusive and unfeasibly expensive for large populations (45).

A similar method that does not have this high cost, is the use of participant mounted cameras. These allow for the monitoring of a participant PA without requiring a researcher to directly observe them (46, 47). Due to memory constraints the camera typically records an image approximately every 10 seconds. This means that the PA characteristics are easily identified, with high resolution. However, these cameras are intrusive and identifying the PA events from still image sequences is a highly complex and time-consuming task (48).

Accelerometers offer another potential method for the measurement of PA. Accelerometers are participant-mounted devices that measure the acceleration of the body part they are attached to (54-56). This acceleration is then used as a surrogate for the PA undertaken. Accelerometers are lightweight, low cost methods of PA measurement with a low participant burden and high reliability (58). Accelerometers have a resolution of up to 1000Hz, which can easily capture any changes in a participants PA. The major issue with the use of

accelerometers as a method for PA measurement is the fact that the acceleration is merely a surrogate for the PA undertaken. Therefore, it needs a method of translating the acceleration into PA. This translation typically takes the form of using thresholds for the acceleration values to determine the events being performed; for example, acceleration values over 0.981ms-2 (gathered at the wrist) are considered representative of moderate activity (61). This method does not allow for the identification of the type characteristic. Additionally, the optimal values of these thresholds (as identified by metabolic studies) are not consistent among all populations and therefore represent a major source of potential error (62).

Due to the limitations of these methods of PA measurement the type characteristic has been difficult to investigate in the current research. Subjective measures can identify the type characteristic but suffer from various forms of bias and low resolution. Objective methods of PA measurement are either unfeasible to use in large populations or cannot identify the type characteristic.

1.5 Type

As discussed above, the type of PA has been shown to impact health outcomes independently of total volume of PA (28, 37) but limitations of measurement methods with respect to the type characteristic have made it difficult to investigate in the current research. However, there are clear limitations in not focussing on the type characteristic and reasons that the ability to determine type of PA in an objective manner would be advantageous.

- Surveillance: the determination of type would be beneficial for population surveillance for many reasons. Classification of type allows for a natural partitioning of peoples PA in a way that is not affected by age or fitness (such as intensity and duration) (34) by partitioning on the type of PA performed. Partitioning of PA data is currently a topic of some interest (32), albeit one that is limited by the inability to develop a suitable

partitioning method than does not depend on arbitrary thresholds. An additional benefit of the type characteristic when undergoing population surveillance is that it allows for clearer health messages (walking 10000 steps) to be delivered to the public. Another advantage of a type specific focus on an individual level is that people consciously know what type of activity they are doing but do not tend to know the intensity or the duration unless specifically focussing on it.

- Clinical studies: as discussed above, there have been many studies that have shown that type impacts health outcomes independently of total volume of PA (28, 37). However, many of the have been performed with questionnaires and other subjective measures of PA type, therefore any relationships identified may not be valid or may be weaker than reality due to the deficiency of the measurement methods available. It has been established that specific health outcomes are impacted differently by the type of PA, such as bone health where non-weight bearing activities are not associated with bone health, while weight bearing activities are (38). Thus, an identification of type would be beneficial as there clearly exist some health outcomes that respond only to type.
- Evaluation of trials: in many PA interventions, the intervention involves performing a specific type of activity for a set duration (due to the clearer message) (8). When DO of the intervention is not possible, various measures are used to determine the compliance, however as discussed above, these methods are either not objective or cannot determine type, just that some activity has been performed. Therefore, without the ability to determine type it is not possible to tell if such interventions are successful. The current inability to determine type means that when determining the efficacy of PA interventions, the focus is generally on the total PA (due to it being possible to measure objectively) instead on the prevalence of the intervention behaviour. This means that if there is an increased amount of the intervention behaviour but no increased amount

of total PA (due to a compensatory effect), the intervention will falsely be declared ineffective (8).

- There is also a strong commercial case for the development of methods for detecting activity type. The wearables market is currently worth $32.63 billion in 2019 and is projected to expand 15.9% from 2020 to 2027, over 30% of this valuation comes from activity tracking wearables from companies such as Garmin and Fitbit (9).

Clearly methods for the objective identification of activity type have commercial and clinical merit. Such a method would aid in population surveillance, allowing for simpler health messages to be delivered as well as allowing for an intuitive form of partitioning. It would allow for investigation into how type impacts health outcomes and would increase the validity of any associations found, compared to subjective means. Additionally, such a method would allow for determination of the efficacy of interventions without making use of crude measures of total PA volume. As such, it is clear that a method of objective identification of activity type would be beneficial for: population surveillance, clinical work, the evaluation of trials and the commercial sector.

1.6 Activity classification and its relevance.

The rapidly growing field of activity classification may allow for a method of objective identification of activity type. Activity classification attempts to use pattern matching algorithms to match acceleration data (gathered from accelerometers) with the corresponding activities, in essence identifying the PA type from accelerometers (54). Activity classification is not without disadvantages, it has been noted that pattern matching algorithms created by matching Lab-Based accelerations to their corresponding activity fail to correctly identify the activities performed from acceleration gathered in more realistic scenarios (not Lab-Based) (54).

This inability to ‘generalise’ (perform as well on data different to the training data) is the major disadvantage of activity classification but there are other contentious issues, some of which are addressed below.

The creation of a PA pattern recognition algorithm requires the identification of many different parameters: the size of the segments of acceleration that will be pattern matched (63), the information about the segments that will be used to match with (average values, shape of the acceleration, etc.) (64), the specific pattern matching algorithm (65), and any pre or post processing of the acceleration values prior to the pattern matching . Similarly, choices such as the location of the accelerometers affect the accuracy of the type characterisation (66). Ultimately however, all choices of parameters attempt to maximise the classification ability of the algorithm on both the original data used to create the algorithm and other different data. This question of the so-called bias (performance on the training data) variance (performance on the different data) trade-off, is at the root of all choices made when constructing an activity classification system.

The creation of an activity classifier that can characterise type from acceleration data will allow for the accurate characterisation of type in an objective manner. To do this, the optimal parameters for creating an activity classification system that allows for high classification performance and an ability to generalise to unseen data will need to be identified.

Therefore, the main goal of this thesis is to identify these parameters and to create an activity classifier that can accurately characterise type from acceleration data.

This thesis starts by critically reviewing the literature on PA classification using acceleration sensors, identifying the many methodological concerns in PA classification research. These methodological concerns are then addressed

individually, discussing potential solutions. These solutions are then combined to create an activity classifier.

1.7 Chapter Guide

Chapter 2 – Classification of Physical Activity Type From Raw Acceleration Data: this chapter discusses the underpinning concepts of activity classification. The current work in the field is reviewed to identify methodological challenges in creating an activity classifier with high performance on unseen data.

Chapter 3 – Data-sets and Base Classifier: this chapter identifies the data-sets used in this thesis. Additionally, a classifier is created using the current state of the art research in order to create a criterion that can be used as a gold standard when testing the effectiveness of methods developed in the rest of this work.

Chapter 4 - Accelerometer Placement Location: one of the methodological challenges identified in Chapter 2 is that activity classifiers are typically trained on data obtained from sensors at a set orientation. Changes in this orientation (such as being on a different wrist) result in performance degradation. This chapter investigates a method to obtain sensor location and orientation invariance via a technique known as domain adaptation.

Chapter 5 - Pre and post-processing: one of the methodological challenges identified in Chapter 2 is that there are many methods of pre and post-processing, with no consensus of their efficacy. Additionally, the effect of these methods with regards to the bias variance trade-off is unknown. This chapter investigates some pre and post-processing methods, identifying their efficacy with respect to performance and the bias variance trade-off.

Chapter 6 – Segmentation of Acceleration into Windows: one of the methodological challenges identified in Chapter 2 is that there is no consensus of the optimal window size for activity classification, and that different activities

appear to have different optimal window sizes. This chapter develops a method of automatic segmentation of acceleration data using changepoint detection to identify activity transitions. This allows for the creation of variable length windows, removing the need for fixed windows.

Chapter 7: Recurrence Quantification Analysis: one of the methodological challenges identified in Chapter 2 is that features that allow for a high level of performance on the training data, typically do not allow for a high level of performance of unseen data. This chapter identifies features that allow for a high level of performance on both the training data and unseen data. These features are based on Recurrence Quantification Analysis, a methodology for measuring the recurrence of data.

Chapter 8: Sparse Features: this chapter identifies a different solution for the lack of features that perform well on unseen data. By making use of automatic feature learning, it is possible to learn features on the unseen data, possibly mitigating this issue. The automatic feature learning method in this work is Sparse Feature Encoding, a methodology that has previously shown high performance in the activity classification domain.

Chapter 9: Classifiers Used: the final methodological challenge identified in this work is the lack of consensus about which classification algorithm to use for activity classification. Certain classifiers assign different weightings to the bias-variance trade-off. This chapter investigates different classifiers to find the optimal algorithm for this work.

Chapter 10 – The Final Classification: this chapter details the creation of the final classifier used in this work, using the work from the previous chapters to overcome all methodical concerns identified in Chapter 2. Additionally, the classifier is used on a large population data-set in order to investigate its performance.

Chapter 11 – Conclusion: this chapter discusses each chapter, identifying if its goals were achieved and how it impacted the overall thesis. Additionally, the strengths and limitations of the work in this thesis are identified and some potential future work is discussed. The chapter ends with a concluding statement discussing whether the overall aim of the thesis has been achieved.

2. Classification of Physical Activity Type From Raw Acceleration Data

2.1 Introduction

Chapter 1 highlighted the benefits of being able to identify PA type and identified that current methodologies of PA measurement do not allow for the objective identification of type. Activity classification was identified as a methodology for the identification of type in an objective manner. Activity classification (54,67) is the name given to methods for estimating PA type from acceleration data. As noted in Chapter 1, the main limitation with the use of activity classification lies in the pattern matching algorithms used. These pattern matching methods work well in Lab-Based studies but tend to fail in more realistic conditions (68). The rest of this chapter comprises a review of activity classification methods as well as identifying the methodological challenges associated with characterising PA type with activity classification.

2.1.1 Accelerometers

Accelerometers are participant-mounted devices that measure acceleration in 1-3 dimensions depending on the number of axes (see Figure 1).

Accelerometers can be mounted in many different locations on the body (57), as shown in Figure 2. The placement location determines the direction of the axes, as they are relative to the device. The absolute location of the axes may vary during motion as, for example, the wearer swings their arm. Each location has its own advantages and disadvantages; these will be discussed in more detail later in this chapter.

Wrist-mounted GENEActiv

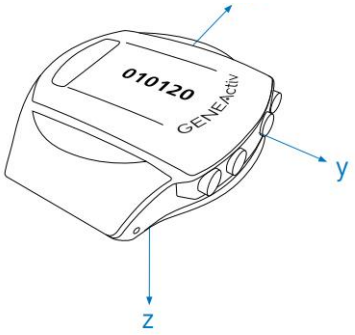

Figure 1: A GENEActiv wrist-mounted triaxial accelerometer illustrating the direction of the X, Y, Z axes.

Accelerometer placement locations

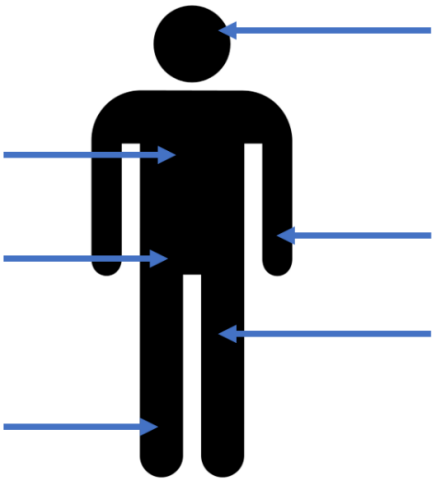

Figure 2: Potential placement locations for accelerometers reported in the literature.

Values of acceleration are given in units of gravitational acceleration (g), 9.81ms^{-2}. A segment of the acceleration data over time is referred to as the trace of the acceleration. The acceleration of each axis is gathered as a time series of acceleration values, resulting in three different time series, each corresponding to an axis when using a triaxial accelerometer. Figure 3 illustrates an acceleration trace over 12 seconds of data gathered at a sampling rate of 100Hz on a triaxial accelerometer. In this work, the acceleration time series are represented by $X = (x_\alpha, x_{\alpha+1}, \dots, x_{\alpha+S}), Y = (y_\alpha, y_{\alpha+1}, \dots, y_{\alpha+S}), Z =$

$(z_\alpha, z_{\alpha+1}, \dots, z_{\alpha+S})$ with S equal to the length of the measurement period multiplied by the sampling rate. All together an acceleration trace is represented by $W = (w_\alpha, w_{\alpha+1}, \dots, w_{\alpha+S})$ where $w_i = (x_i, y_i, z_i)$.

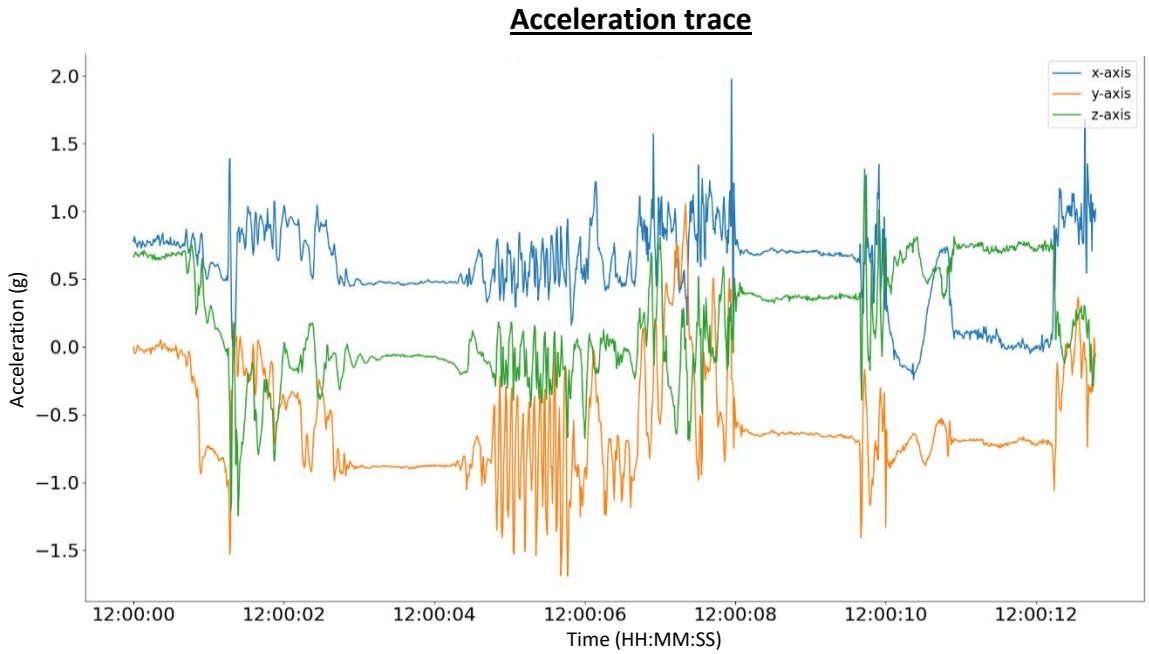

Figure 3: Acceleration trace, showing X, Y and Z accelerations.

2.2 Activity Classification

Activity classification is a method of 'mapping' acceleration values to PA type. This is done by creating a classifier which inputs acceleration values and outputs activity labels, mapping the values to labels. Example acceleration traces from a triaxial wrist-worn accelerometer and their corresponding labels can be seen in Figure 4(a,b,c), the ideal classifier is able to map each trace to its assigned label.

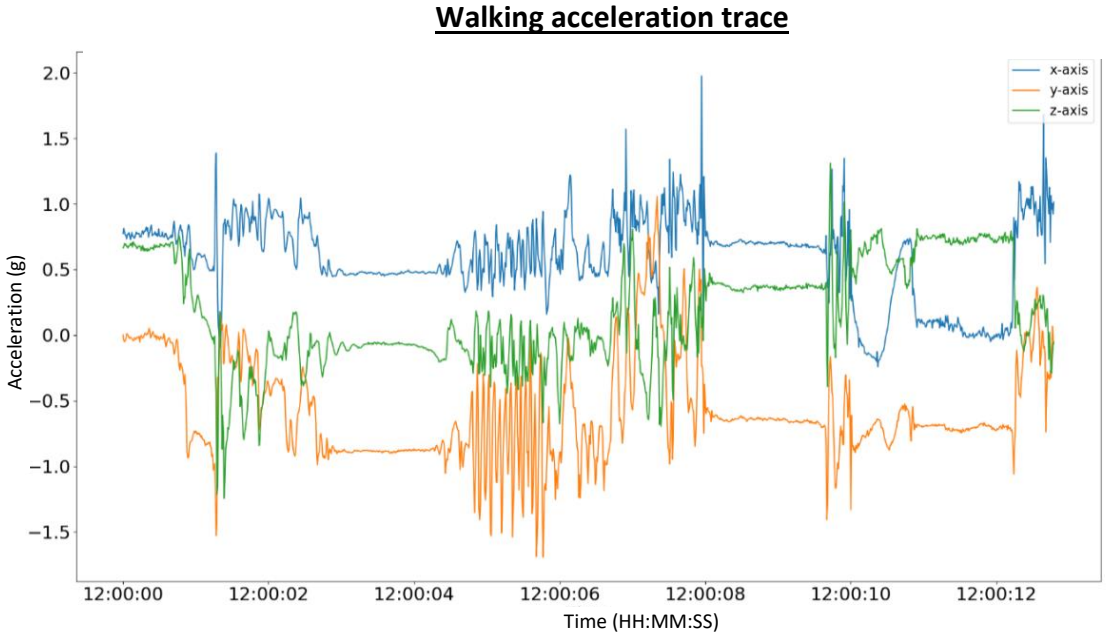

Figure 4a: Acceleration trace of walking.

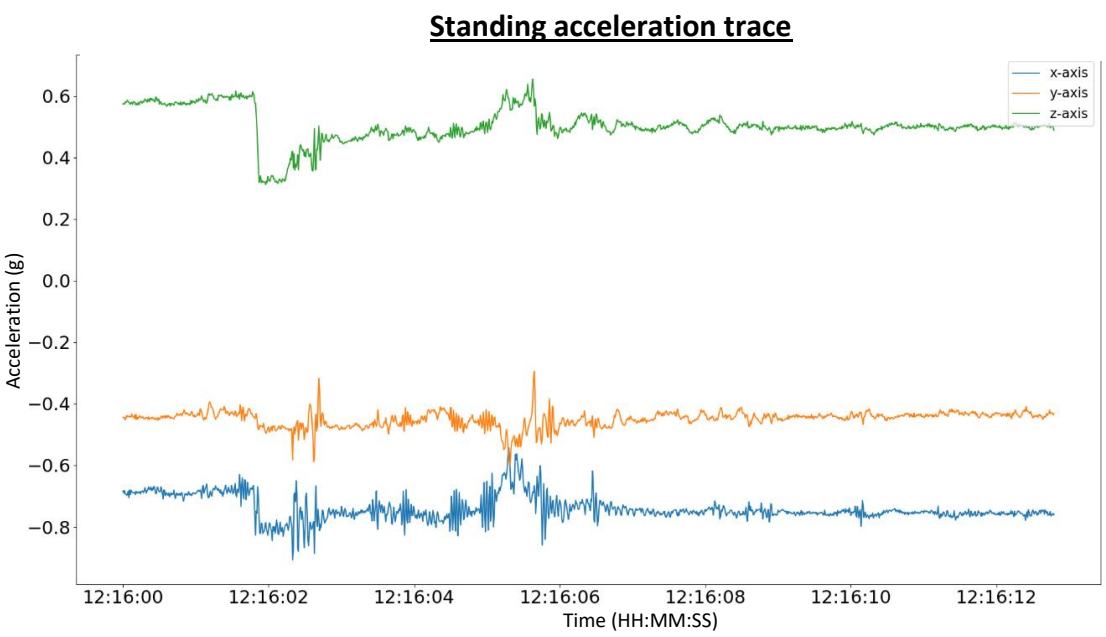

Figure 4b: Acceleration trace of Standing.

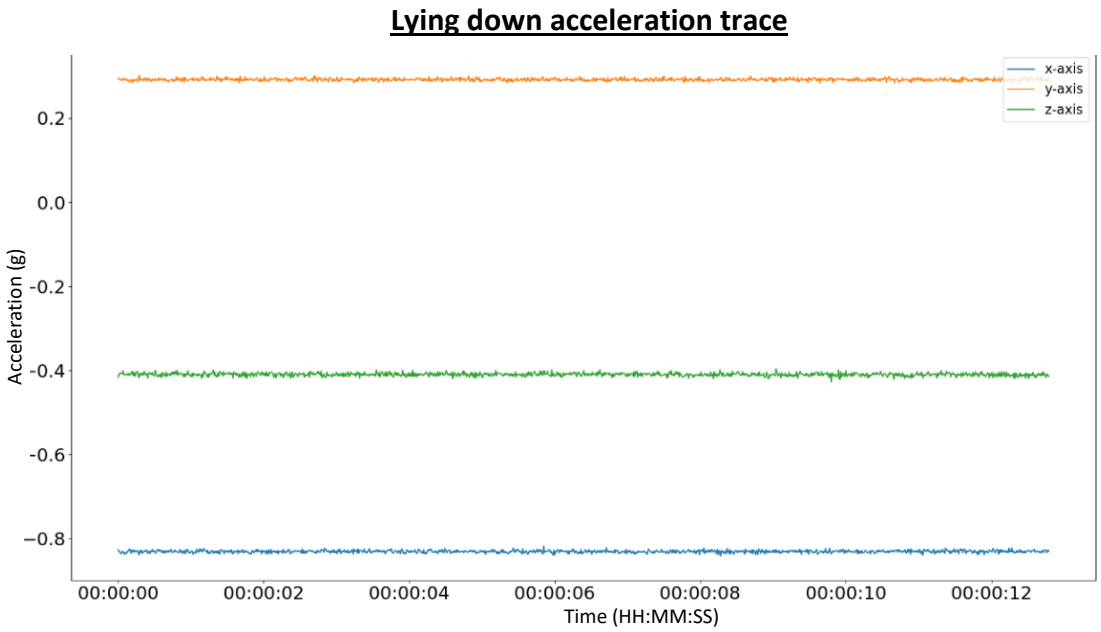

Figure 4c: Acceleration trace of Lying down.

2.3 Machine Learning

The process used to create a classifier in this work is Machine Learning. Machine Learning is a form of artificial intelligence that builds a 'classifier'; a classifier is a function that maps input data to the desired output, the outputs being discrete classes or labels. In the context of this thesis, Machine Learning is used to develop a classifier that can identify PA type from accelerometry data, mapping the acceleration data to the PA types. This classifier can then be used to predict types of PA (the output) from accelerometer data (the input) where these types are not known, allowing for the identification of PA type from accelerometers. More formally a machine is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E. In essence, Machine Learning is building a function that improves performance at a certain task when given more data, as measured by a given performance metric (69).

Activity classification uses a form of Machine Learning called supervised learning. Supervised learning makes use of a data-set containing both inputs (acceleration values) and desired outputs known as the training data. In supervised learning, the classifier is 'trained' to identify the outputs from the inputs, in this case creating a classifier that can identify PA type from the acceleration data. Training a classifier involves adjusting its parameters to minimise an error function, E_f. This is a function indicating how accurately the classifier can identify the activity labels from the acceleration data when the correct labels are known.

Mathematically, supervised learning attempts to identify some parameters θ controlling the behaviour of function f so that the output of the classifier best matches the training label t_i when the corresponding features d_i are input: $t_i = f(d_i;\ \theta)$. $D, T = \{d_n, t_n\}_{n=1}^{N}$, with d_n and t_n being the n^{th} data point with the corresponding label.

Training a classifier means identifying the parameters that minimise the error of the classification, finding $\theta^* = \underset{\theta}{\operatorname{argmin}} \sum_i E_f\big(t_i, f(d_i;\theta)\big)$. The error function is often chosen to be the misclassification rate. Once the best parameters are known, they may be used to estimate (or *predict*) the unknown labels t' corresponding to new observations: $t' = f(d', \theta^*)$.

2.4 Overfitting

An important issue that needs to be addressed in training a classifier is the avoidance of overfitting. This is when the classifier is overspecialised to the training data-set; decreasing the classification error by modelling the noise of the data-set, as well as the classification function. Modelling the noise allows for a greater ability to classify the training data but reduces the ability to classify data with different noise but the same underlying mapping function, which is the purpose of a classifier.

Overfitting can be detected, and hence avoided, via cross-validation. This is when the training data is partitioned into a training set and validation set. The classifier is trained to minimise the error on the training set and the performance is estimated on the validation set. This gives an indication of how well the classifier can generalise to an independent data-set. Typically, multiple partitions are used, and the predicted performances are averaged.

2.5 Classification pipeline

In this thesis, the creation of the classifier was one step of a six-step classification pipeline, as outlined by Bao and Intille (54). Supervised learning typically follows a sequence of steps detailing the creation of a classification pipeline (69). The classification pipeline described here follows this sequence of steps, adapting them specifically for activity classification. The inclusion of segmentation steps is typical in time series data for the reasons described below. Almost all activity classification studies that make use of supervised learning follow this pipeline, although the majority of studies do so implicitly (54,70,71). Additionally, there is lack of consensus about the difference between pre-processing, segmentation and feature extraction. Some studies simply declare the pipeline as pre-processing, classification and post-processing, combining steps 2-4 (72). The six-step pipeline outlined was used as it represents the most granularity.

1. Determination of device location and setting: this refers to the choice of the sampling frequency of the accelerometer (how many measurements it makes in a second) and placement location.
2. Pre-processing: pre-processing refers to the preliminary processing of acceleration data before any classification steps are carried out and has a range of uses such as: allowing for rotational invariance (73) and the removal of noise (74). While this section comes before the choice of window size and features in the ‘pipeline’, it can occur at any point before the classification step.

3. Segmenting into windows: the acceleration values are segmented into short duration windows (typically between 1-60 seconds) (63). This is because performance is improved when classifying a window of acceleration data rather than a single instantaneous value. Observing multiple acceleration values allow an indication of how the values are changing in time, unlike a single instantaneous value, hence classification is improved. This results in a series of acceleration segments with corresponding activity labels, referred to as the windowing stage. Mathematically, with a window size of S, this can be thought of as $W_\alpha = (w_\alpha, w_{\alpha+1}, \dots, w_{\alpha+S})$.
4. Extracting features: once the acceleration data has been segmented into windows, feature extraction is undertaken (54). A set of attributes (consistent among all windows) that describe the windows are identified and used to represent the windows, for example; identifying the mean acceleration for each axis and their standard deviations. This results in feature vectors, one for each window, each with a corresponding activity label.
5. The classifier:
    a. Training the classifier: during the training stage, the classifier is trained with labelled data.
    b. Using the classifier: when the classifier is used for classifying data, the classifier is used to predict labels for unlabelled data.
6. Post-processing: after the acceleration data has been classified, the predicted labels may be processed in order to reduce the number of misclassifications. This is referred to as post-processing. Most post-processing approaches use the sequential nature of activity data to improve performance, making use of the fact that adjacent segments are likely to be the same activity (75).

These six steps result in a classification pipeline that can input acceleration traces and output predicted activity labels, when in the ‘use stage’.

2.6 Activity Data

In this work, the creation and evaluation of the classification pipeline uses three data-sets.

- The training data: this is a time series of acceleration values with corresponding activity labels. See Table 2 for example. This data goes through the classification pipeline, as discussed above, training the classifier. This data is also referred to as 'seen' data because the classifier has encountered it.
- The validation data: this data is drawn from the training data and used for optimising the various hyperparameters of the classification process. Hyperparameters refer to the many modifiable characteristics of the classification pipeline, such as window size and features used. These are akin to parameters in f but over the entire classification procedure.
- The test data: this data has the same structure as the training data, with acceleration values and corresponding activity labels. However, this data is not used in training the classifier but instead to test the performance of the classification after all optimisations have taken place. The classifier is used to predict the activity labels from the acceleration data, yielding predicted activity labels. These predicted labels can then be compared to the known true labels to evaluate the performance of the classifier. Various metrics can be used in this comparison, such as the accuracy or the precision, depending on what aspects are deemed important in the classification. This data is also referred to as unseen, as it has not been encountered by the classifier in the training.

Often training data and test data are collected in a laboratory setting with participants performing an activity protocol reflecting activities of daily living (71). The researcher(s) observe the participant and create a label for each activity with a timestamp. This is generally referred to as Lab-Based data. An alternative approach is to give no participant instructions and allow the

participants to perform activities, in their normal environment; data collected in this manner is referred to as Free-Living data. Free-Living data better reflects the range of activities encountered in daily living but is more costly to gather and accurately labelling the data is a particular challenge, due to the lack of a gold standard measure (54).

X acceleration	*Y acceleration*	*Z acceleration*	*Time*	*Label*
0.132	0.241	-0.562	09:00:00.00	Walking
0.154	0.359	-0.684	09:00:00.01	Walking
0.325	0.259	-0.931	09:00:00.02	Walking
0.156	0.268	-0.354	09:00:00.03	Walking
0.236	-0.254	0.236	09:00:00.04	Standing
0.236	-0.352	0.126	09:00:00.05	Standing

Table 2: Training data, with X, Y, Z acceleration and activity labels, measuring at 100Hz. Test data is identical in format, while unlabelled data, does not have labels.

Typically, classifiers trained on one activity protocol poorly classify data from a different protocol. This generally takes the form of Lab-Based data poorly classifying Free-Living data (68). This lack of an ability to translate the classification from one protocol to another is perhaps the largest roadblock to widespread adoption of accelerometry-based activity type classification as a method of population surveillance.

2.7 Classification performance

When classifying data, it is generally assumed that the relationship between features and labels is consistent among data-sets, so that a mapping function (classifier) trained on one data-set may be used to classify another data-set. In activity classification, this assumption is frequently not true. Different data-sets (from different participants or activity protocols) have very similar relationships between features and labels but they are usually not identical. This can be identified from the four potential ways of categorising test data (in order of performance):

- Intra-subject-intra-protocol: this is where the test data is from the same participant and protocol, separated only in time. This performance is almost always the highest of the four methods. This is because the mapping function is the same (65).
- Inter-subject-intra-protocol: this is where the test data has different participants from the same protocol as the training data. This performance is often high, although typically lower than intra-subject (65,71).
- Intra-subject-inter-protocol: this is where the test data is drawn from the same participants, but from two different activity protocols. This data is typically gathered in the form of two lab visits; therefore, the variation in the protocols is slight. This method generally has a lower performance than inter-subject-intra-protocol, depending on how varied the protocols are (65).
- Inter-protocol-inter-subject: this is where the test data is from different protocols with different participants. This method always reports the lowest performance. This method of identifying the test data gives the best idea of how well the classifier will perform on a real-world population study (where there is no labelled data from either the participants or the protocol) (76,77).

The performance differences between the methods of identifying the test data-set indicate that the underlying mapping function must be changing. It may simply be the case that different protocols have different activities, therefore decreasing the performance. However, even if the training data is from Free-Living and the test data is from a constrained Lab-Based study, performance still drops considerably, as is shown in this work. An additional argument may be that this performance drop is simply a case of not having enough participants to characterise a full range of potential feature-label mappings. This may be the case; however, since gathering labelled acceleration data is expensive it is not possible to simply increase the number of participants in the training data.

Additionally, increasing the number of participants in the training data does not greatly impact the classification when using inter-protocol-inter-subject testing. So, it can be inferred that the lower performance is a result of the mapping function being different for different protocols/participants. This difference in the training and testing data caused by being from different protocols and participants is referred to as data-set shift (78).

Typically studies in activity classification attempt to improve performance of the classification of the test data by optimising the hyperparameters of the classification pipeline. The different ways of identifying the test data affect this optimisation. As the most common form of identifying the test data is inter-subject-intra-protocol, classification pipelines are typically optimised to improve performance when assuming the activity protocol remains consistent (referred to as the intra-protocol performance). This may explain the typical lack of an ability to perform on different protocols (76,77). However, by using inter-protocol-inter-subject testing, hyperparameters may be identified that allow for a high level of performance on different protocols (referred to as inter-protocol performance). This means that any classifier created should have a high performance on all unlabelled data, not just the test data. It is possible to over-optimise inter-protocol performance at the cost of the intra-protocol performance. This occurs when a classification pipeline suffers little performance drop from one protocol to another but the intra-protocol performance is low to begin with (76). As such, it is important to identify hyperparameters that optimise both inter-protocol and intra-protocol performance. The hyperparameters that must be identified are linked to each of the steps of the classification pipeline.

2.8 Comparison of classifiers

The differences in the choice of test-data make comparisons between activity classification models difficult. Additionally, major differences in activity protocols may render comparisons invalid. For instance, identifying between activity

labels such as walking or sleep is much easier than walking or running (54), hence the classification performance will be higher. Because of these reasons, it is not possible to directly compare PA classification research via methods such as meta-analysis. However, it is possible to identify changes caused by a single hyperparameter variation if such results are reported. This is challenging because the large number of hyperparameters in a single classification pipeline makes identifying if a single hyperparameter is optimal in all circumstances difficult due to interactions between the hyperparameters. Therefore, this review will focus on the effects on the performance when modifying a single hyperparameter as opposed to the complete classification pipeline that may have many different hyperparameters.

The evaluation of single hyperparameters will be used to inform the creation of the classification pipeline, as well as identify areas that need further research.

2.9 Determination Of Device Location

The location of the accelerometer placement is a key issue. It is a question of cost/participant burden and protocol adherence versus potential performance. Using multiple sensors means more information is available and therefore a higher performance is likely (54), but multiple sensors increase participant burden and decrease the protocol adherence (79).

2.9.1 Thigh-Mounted/Leg-Mounted

Placing a single accelerometer on the thigh allows for a high level of recognition for activities such as walking, sitting/lying, and cycling (54). Bao et al (54) detect cycling with 96% accuracy (intra-protocol performance). By comparison, a wrist located accelerometer experiences a performance decrease of 31% when compared to a thigh-mounted approach (80). Additionally, a thigh location can detect heel strikes when walking/running and other gait characteristics. A disadvantage of this approach is the increased participant burden compared to

wrist-mounted accelerometers, as the device is typically taped to the leg, leading to possible discomfort, especially when removed (81). A further limitation of thigh located accelerometers is their inability to discriminate between activities that differ only by upper body movements such as standing still and washing windows.

2.9.2 Waist-Mounted

Much research has been carried out on activity classification via waist-mounted accelerometers (68,82–84). This research is added to by activity classification approaches using smartphones, which are usually assumed to be carried in pockets (85). Waist-mounted approaches can obtain very high performance for classification of certain activities (walking, sitting), typically outperforming other potential placements (thigh, wrist) (82). However, waist-mounted approaches struggle to identify activities that involve a large proportion of upper body acceleration, such as basketball/dance and most household chores.

2.9.3 Wrist

Wrist-mounted accelerometers impose the lowest participant burden and have the highest compliance with wear time criteria (86). This reduces the amount of non-wear data points that either must be treated as missing or imputed. Additionally, wrist-mounted approaches are capable of high levels of performance, especially in Free-Living data (77). As a consequence of these factors, wrist-mounted accelerometry has become more popular than waist-mounted in both activity classification work (74,77,87,88) and in broader population studies (58). Despite this widespread use, wrist acceleration imperfectly captures activities that primarily use lower body acceleration such as cycling (80). Additionally, the non-activity related movements from hands (e.g. gesturing) add substantial noise into the acceleration data gathered.

2.9.4 Network-Based Approaches

Most of the early work in activity classification made use of a network of participant-mounted acceleration sensors. The exact placement location varied among studies, but wrists, calf/ankles, waists, and thighs were consistently represented (54). Making use of multiple sensors increases the information available and allows for an increase in performance when compared to single accelerometer approaches (54). Additionally, gathering acceleration from multiple body locations gives a more comprehensive view of total body movement. However, using multiple sensors increases the participant burden and researcher cost in comparison to a single accelerometer approach. While using accelerometer networks does allow for high performance, placing accelerometers on too many positions can be cumbersome, prone to errors, and impractical for participant deployment in Free-Living settings over extended observation periods (79).

2.9.5 Limitations

An issue that has not been fully addressed in the literature is which wrist the sensor should be mounted on when choosing a wrist-worn device. There are conflicting opinions on which wrist should be used in activity classification, some studies championing using the left or right (89), while others use dominant or non-dominant wrists (55). This issue is further exacerbated by poor participant adherence to device wear and orientation guidelines. Dominant and non-dominant wrists obtain different acceleration values when investigating the intensity of the same activity on the same participant (61,90), hence this is an important issue. Furthermore, using activity classifiers where the wearer's wrist is incorrectly specified has been shown to reduce performance by up to 12%, compared to using the correctly specified wrist (89). A related issue occurs when the device is placed upside down, which reverses some of X, Y, Z axes values gathered; this also decreases the performance. Typically, the reduction

in performance when applying a classifier to a different wrist is not remarked upon, and it is assumed that all participants will wear the device on the same wrist or location as advised (58). This assumption is not guaranteed, especially when the participant places the sensor themselves, and the violation of this assumption may partly explain the poor inter-protocol performance of activity classifiers (77).

A possible solution in activity classification is to make use of features derived from the acceleration that are orientation invariant, meaning that the features will be identical regardless of the orientation of the sensor (73). This prevents the performance reduction but limits the available features that can be used in the classification which may compromise performance. Another more effective approach, described by Gjoreski et al (89), involves training the classifier with data from both wrists, resulting in higher performance than using a single wrist. However, data from the second wrist may confuse classification, reducing the performance. An additional limitation of this method is that data from both wrists must be collected in the training stage, increasing the cost of the data gathering process and participant burden.

2.9.6 Conclusion

In conclusion, different body locations that an accelerometer can be mounted on have associated advantages and disadvantages. However, wrist-mounted accelerometers allow for the lowest participant burden while also allowing for a high classification performance, thus will be used in this thesis.

A classifier that allows for wrist invariance but does not depend on training-test data from both wrists remains a gap in the current research and will be addressed in Chapter 4.

2.10 Pre-Processing

In this section, various methods of pre-processing are identified, their role in the literature is briefly discussed and the problems they intend to fix are identified. The precise methodologies behind these methods will be discussed in Chapter 5. It is worth noting that the distinctions between pre-processing methods, segmenting data and feature extraction are somewhat artificial, with different work using different assignments (72).

2.10.1 Data Aggregation

One of the most common forms of pre-processing is to combine the three acceleration data streams (X, Y, Z from a triaxial accelerometer) into a single aggregate data stream (91). This single data stream is used in the classification process as opposed to the X, Y, Z streams. Transformation of the acceleration data, such as computing a single aggregated data stream can be thought of as either a pre-processing technique or a form of feature extraction. In this work, transformations are identified as a form of pre-processing; this is because any form of feature extraction can be carried out on the resulting aggregate data stream. However, other authors may disagree with this definition (72).

Some methods of combining the three axes together allow for rotation invariance, meaning that if the device is rotated, the acceleration values obtained are unchanged. This rotational invariance is advantageous when the participants place the device themselves because incorrectly oriented accelerometers can result in a drop in performance (73). However, this invariance to rotation does remove any information about the rotation or direction of the accelerometer that itself may assist performance. An additional benefit of combining the three accelerations into a single data stream is viewing and understanding the data becomes much simpler.

The most common aggregate measure used in acceleration research is the Euclidean Norm Minus One (ENMO) (90). ENMO is computed by:

$$ENMO_i = max\left(\sqrt{x_i^2 + y_i^2 + z_i^2} - 1{,}0\right)$$

Equation 1: Euclidean Norm Minus One.

Computing the Euclidean norm of the 3 acceleration values $\sqrt{x_i^2 + y_i^2 + z_i^2}$ gives the magnitude of the three accelerations combined into a single value (thereby allowing for rotational invariance). Subtracting one from this magnitude discounts the effects of gravity's downward acceleration and taking the maximum assures that the value has a minimum value of 0. Use of ENMO (or vector magnitude) has reported high levels of performance in many domains (68,84,91–95). However, ENMO is vulnerable to calibration errors and Van Hees recommends an additional device-specific calibration protocol (96) that increases complexity. An alternative method that is seeing increasing usage is to include the aggregate [ENMO] data stream while also using the separate (X, Y, Z) accelerations. This gives the advantages of the aggregate data [ENMO] (orientation/rotational invariance) without complete removal of directional information (97,98).

Figure 5 shows the ENMO data extracted from the X, Y, Z accelerations. The ENMO values are positive for all directions of the acceleration, this is because it is orientation invariant. It can also be seen that periods of inactivity in the acceleration also correspond to periods of inactivity in the ENMO. It is also worth noting that changes of sign of the accelerations (at 12:00:10) do not affect the ENMO value, this is also due to the rotational invariance.

Alternative metrics have also been suggested; Activity Index (99) and Mean Amplitude Deviation (90), however no activity classification research currently makes use of these metrics.

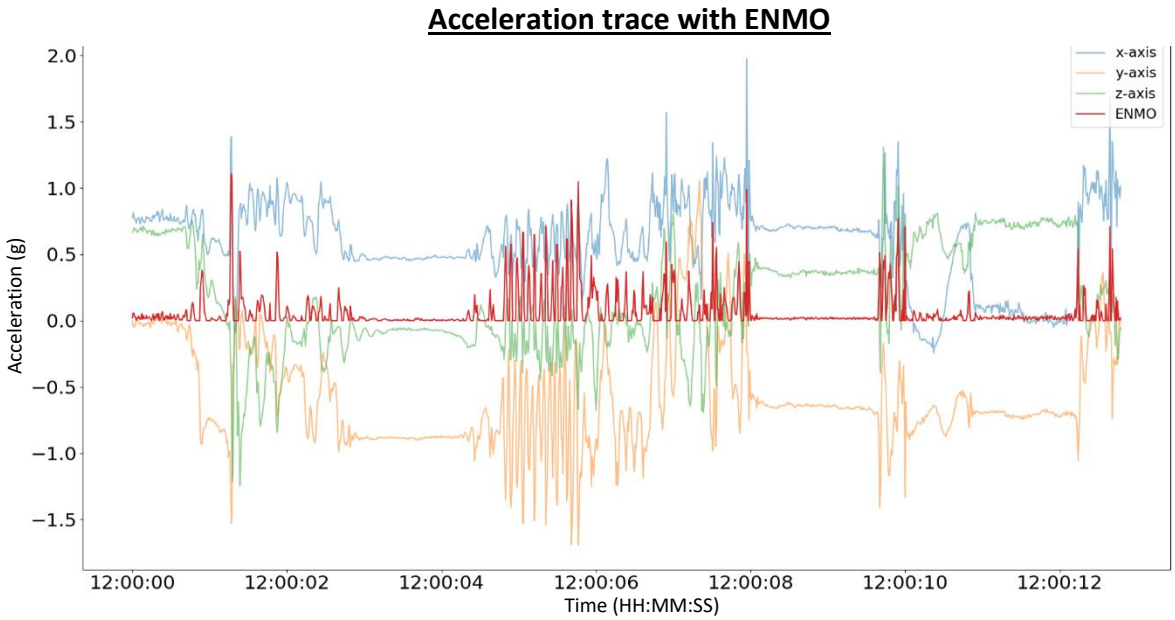

Figure 5: Acceleration trace with ENMO. The faint line representing the unfiltered data, the bold representing the filtered data.

2.10.2 Filtering

Noise represents an issue in activity classification. In this context, noise refers to one of two things: Observational noise: random disturbances in the signal caused by the device (typically 'Gaussian noise'), or additional information in the signal that is not useful for the activity classification. A 45Hz signal generated by riding on a bus in the acceleration is not random but is not useful in activity classification and will disturb the acceleration values, hence is treated as noise.

Typically, in activity classification, the cut-off frequency for a signal being noise ranges from 15-20Hz. Rationalising the choice of these frequencies is the work of Mann et al (100) who determined that 99% of measured body movements are contained within frequency components below 15Hz. However, the Nyquist-Shannon theorem (101) states that for a successful reconstruction, data needs to be sampled with at least twice its highest frequency, which indicates that the cut-off frequency should be at least 30-40Hz. It is important to note however

that Mann et al only accounts for frequencies generated by human movement, not human activities. Riding in a car may generate frequencies in excess of 15 Hz (102) yet at lower frequencies it is indistinguishable from sitting. Therefore, these higher frequency signals may actually benefit the task of activity classification.

A common pre-processing method in any data concerning time series is filtering. This refers to creating an approximation of the time series that can capture important patterns but is less affected by noise. There are three common forms of filtering: Low-pass (removing all frequencies higher than a threshold), high-pass (removing all frequencies lower than a threshold) and band-pass (a combination of high and low, keeping only the frequencies between two thresholds). In this field, low-pass filtering is typically used. Several different algorithms for filtering exist but there are two forms that are commonly used in activity classification; using a moving average (103) and using a Butterworth filter (88,103,104).

2.10.2.1 *Moving Average*

A moving average is ‘dynamic average calculated across successive segments of data’ (typically of constant size and overlapping) of a series of values.’ (105). In this case the series of values represents the acceleration time series. Typically, the average used is the mean, although the median has also shown some success (103). Typically, the segment sizes, referred to as $'n'$, range from two (containing only two points) to 100 (equivalent to 1 second at 100Hz). Generally, activity classification studies do not provide reports on classification performance with and without using the filtering, so it is difficult to gauge the effectiveness of such a method without additional study. Additionally, it is not clear which value of $'n'$ is optimal and whether this differs for different activities/participants.

Mathematically a moving average is calculated by:

$MA(w_M) = \frac{1}{j}\sum_{i=0}^{j-1} w_{M-i}$, this is computed separately for X, Y, Z.

Equation 2: Moving Average.

A moving average filter is a Low-Pass Finite Impulse Response filter. For removing simple observational noise (the noise from measurement error), such a filter is mathematically optimal (no other filter can do better) (106). However, for removing higher frequencies data (such as data generated by a 45Hz signal), this filtering method is poor, for more detail see 5.3.

Figure 6 shows the effect of a moving average filter, with an $'n'$ of 11. This filter drastically smooths the acceleration values, making it less varied.

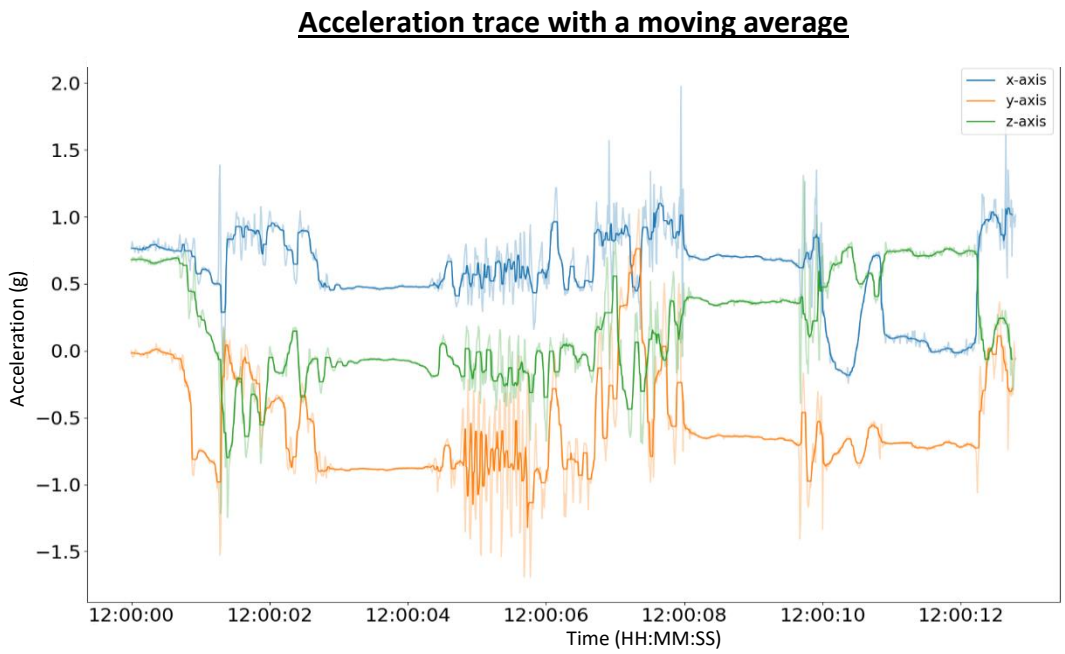

Figure 6: Acceleration trace with a moving average filter. The faint line representing the unfiltered data, the bold representing the filtered data.

2.10.2.2 *Butterworth Filter*

A low-pass Butterworth filter is a filtering method that can be used for the attenuation of high-frequency data from a time series. Like the use of moving

averages, no studies showing the activity classification performance with and without this filtering were found by the author, so it is not known if it is actually beneficial despite its widespread usage.

A Butterworth filter is mathematically optimal for removing the higher frequencies without affecting the lower frequencies in the data (106), in that the frequencies below the threshold are unchanged. In Figure 7, the effect of applying a Butterworth filter to acceleration data can be seen (attenuating all frequencies above 20Hz). As can be seen, there is very little change in the signal after the filtering. This suggests that the majority of the acceleration is not from signals with a high frequency.

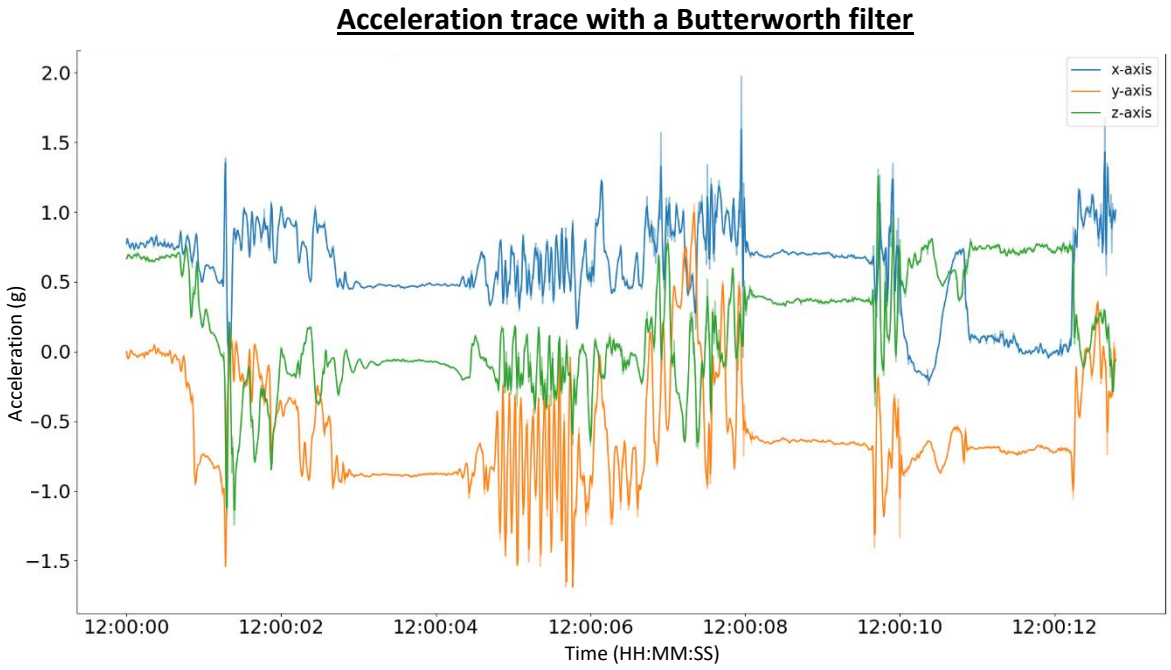

Figure 7: Acceleration trace with a Butterworth filter of 20Hz. The faint line representing the unfiltered data, the bold representing the filtered data.

2.10.3 Orientation Invariance

Orientation-invariant transformations are another method for achieving orientation invariance (73) that do not require the use of aggregation. These are transformations that can be applied to the acceleration time series that result in

an orientation invariant time series. The study that reported this method showed a performance reduction of 15.5% when classifying randomly rotated data against the reference case (no rotation) compared to the 21.2% performance reduction when not using any transformation. However, using ENMO yields a performance reduction of 13.5%, suggesting that ENMO is the superior method (73,107). Much like ENMO, creating extra time series to allow for orientation invariance can be considered either a method of feature extraction or a pre-processing method.

A method related to orientation invariance is inclination correction. This deals with the scenario where the sensor is still broadly at the correct orientation but may have moved slightly. In the work of Fida et al (108) inclination correction is described as "*each data channel value was removed from the average value obtained when standing still for 5 s before starting the activity path*.".

In Fida's study, inclination correction did not allow for a significant increase in classification performance. However, in their activity protocol, the accelerometers were placed by the researchers and it is not indicated if they required inclination correction.

2.10.4 Class Imbalances

The proportion of the 'classes' in the training data can affect the overall classification performance. Classes, in this case, refer to the different activities that the classifier is identifying. Having imbalanced classes (more of one kind of activity than others) tends to decrease the performance, especially if the imbalances differ between the training and test set (69). This is particularly an issue in PA data-sets, due to their generally small size. Additionally, different activities have different occurrence rates; naturally, it would be expected to see a greater amount of time spent sitting or sleeping than running. Reducing this class imbalance should allow for higher inter-protocol performance. The two most common of which in this field are over and under-sampling. Oversampling

refers to generating synthetic data from the under-populated classes in order to make the number of examples from each class equal. Whereas under-sampling refers to removing samples from the well-populated classes until all populations are equal. As under-sampling reduces the amount of data in the training set it is typically not preferred. The majority of oversampling methods are not created for time series data, due to the high level of inter variable correlation where sequential points are related to one another. Specific methods must be created: Cao et al (75) explore the use of Structure-Preserving Oversampling in activity classification in order to correct for a class imbalance. They reported a 5.3% performance increase in performance compared to not correcting the class imbalances.

2.10.5 Conclusion

Many different methods of pre-processing have been identified in this review and most of them have identified that they result in a performance increase in the classification process.

However, all the pre-processing methods have been tested in the context of intra-protocol performance, instead of inter-protocol; testing the performance of the methods when the training and testing data are drawn from the same protocol. Methods that improve intra-protocol performance do not always improve inter-protocol performance. As the aim of this thesis is to maximise the inter-protocol performance (while maintaining intra-protocol performance), these pre-processing methods must be investigated to see their effect on this.

Specifically, the pre-processing methods that will be investigated in Chapter 5 are:

- Filtering
    - Using a moving average
    - Using a Butterworth filter

- Using inclination correction
- Fixing class imbalances
- Aggregating the data via ENMO
- Using orientation invariant transforms

2.11 Segmenting into Windows

Windows refer to the segments of the activity data that undergo the feature extraction and classification process. They typically range from 0.8 - 60 seconds (63,87,109) and can overlap with each other. See Figure 8 for an example of windows used in classification.

2.11.1 Window Size

Larger windows have several advantages over smaller windows. The larger a window, the greater the information available, which should result in improved classification performance. Additionally, noise has a lessened effect on larger windows as a single erroneous point represents less of the total available data. The disadvantages of larger windows are the increased likelihood of activity transitions (changing from one activity to another) occurring in a window. Activity transitions negatively impact the classification process as there does not exist a correct activity label for a window if it contains a transition. Larger windows also represent a greater computational cost to extract features. Additionally, an increase in the window size represents a decrease of the size of the training data, since the acceleration data forms fewer segments.

2.11.2 Multiple Length Windows

Using multiple length windows in the classification process allows for the advantages of small and large windows while mitigating many of the disadvantages. Majority voting is a method making use of multiple windows,

whereupon windows of different sizes are created, all centred on the same data segment (the centre of the smallest window). Each of the windows are then classified, resulting in multiple labels. The majority label (the most common) is then chosen for this segment.

Using multiple windows has shown an increase in performance (110) in other bio-medical time series domains. The only multiple windows technique used in activity classification reported no performance increase (111). However only windows up to 10 seconds were utilised. Using multiple windows increases the required computation, a factor which must be weighed against any potential increase in performance. It may be possible to utilise multiple windows in a more hierarchical fashion, making use of the fact that different window sizes are better at classifying different activities but currently no such studies exist. Another possible method making use of multiple windows is to extract features from all of the windows centred on a single point and concatenate the features into one large feature vector. This can then be used in the classification

Figure 8, shows the process of majority voting over multiple windows. Each window is classified, and the majority classification is then taken, e.g. walking.

Multiple window classification procedure

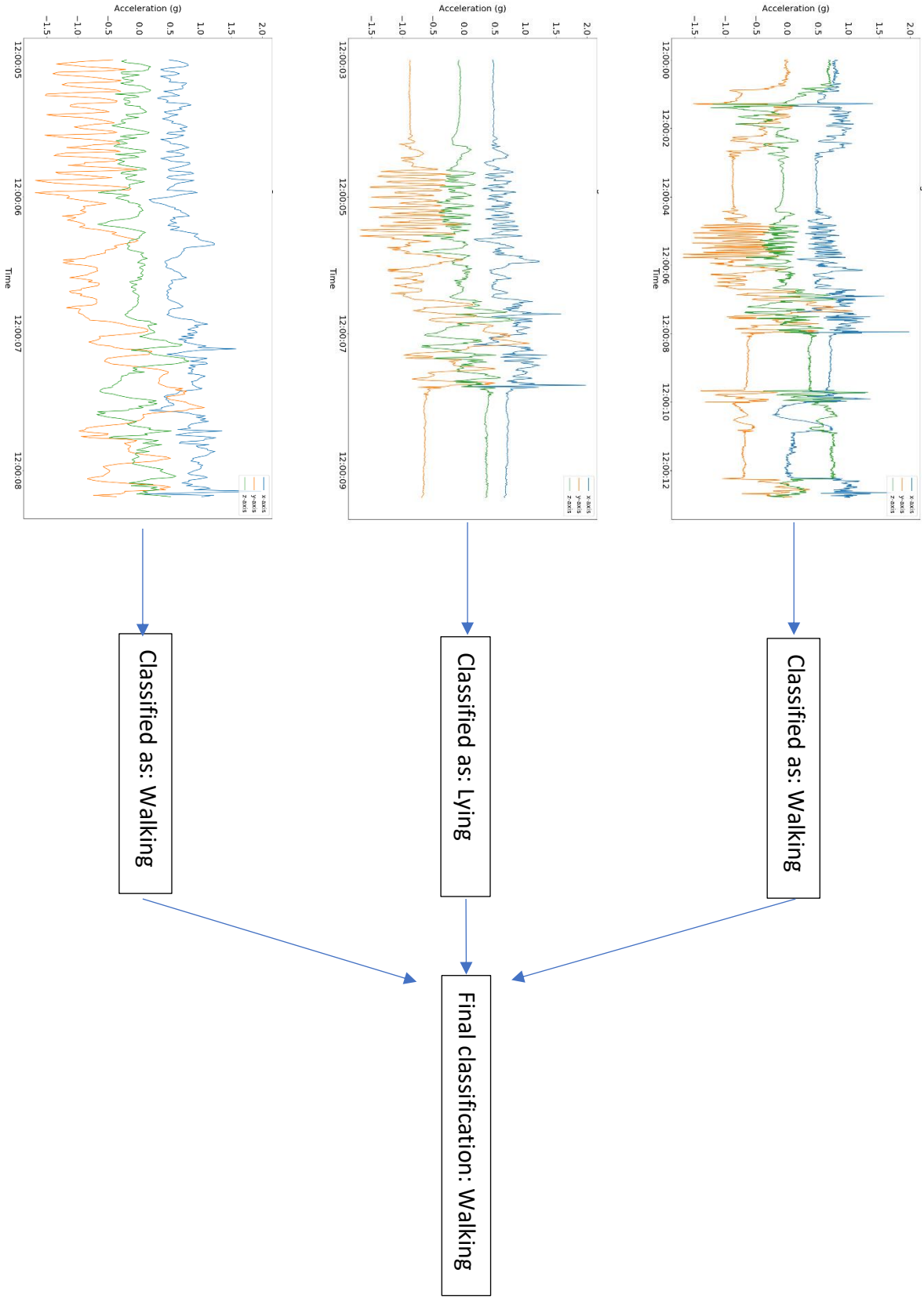

Figure 8:A multiple window classification, making use of a 12.8, 6.4 and 3.2 second window.

2.11.3 Auto-Segmentation

Auto-segmentation describes a technique for creating windows of acceleration data without making use of fixed window sizes. This technique attempts to identify windows based on where the activities change/transitions occur, detecting events and event changes. Typically, methods making use of auto-segmentation outperform fixed window approaches (112). Auto-segmentation approaches increase computational requirement and complexity, precluding them from common usage.

Auto-segmentation produces variable length windows. As optimal window size changes with the activity being classified, this variability is advantageous (63). An additional benefit of auto-segmentation is that activity transitions are unlikely to occur in the windows. This is because the point of the transitions is typically also a point where the data automatically segments. The main disadvantage of auto-segmentation is that it requires a method for transition detection, which is highly complex to create. In Chapter 6 a method of automatic segmentation via transition detection will be investigated.

2.11.4 Overlap

Overlap is a modification to windowing approaches where the sequential segments used to create the windows are not separate but instead share a portion of their data (they overlap). This overlap is most commonly 50% although other proportions are used (74).

The main advantage of this technique is that it increases the amount of data available in the training data, which improves classification performance. This is especially useful when used in conjunction with larger windows as this mitigates their main disadvantage. Additionally, it allows for a better time resolution. The main disadvantage of using overlap is it increases the interdependence of the

windows. Most classification procedures assume that the data is independent; violating this assumption may decrease the performance (113).

2.11.5 Conclusion

In conclusion, the choice of window duration is always a compromise. Short duration windows give good temporal resolution, but it is difficult to determine the PA type because of the limited data in each window and multiple windows need to be combined to understand PA events (63). Longer windows are efficient for data processing and if the window covers a single PA event, longer windows help to accurately identify it. However, longer windows have an increased probability of containing an activity transition, leading to a difficulty in assigning a single label to the entire window (63). Automatically segmenting the acceleration data allows for the creation of variable duration windows without the same compromises as fixed duration windows (112).

The challenge, therefore, is the creation of a method for automatic segmentation that can allow for automatic creation of variable length windows of acceleration data. This will be addressed in Chapter 6.

2.12 Extracting Features

Features characterize the accelerometer windows. The features used for activity classification typically fall into one of three categories, although it is worth noting that any distinction is somewhat artificial:

- Statistical and frequency aggregative features: these are features that attempt to represent the data with a single aggregated value, such as the mean or skewness of the acceleration.
- Morphology-based features are based on the shape of the acceleration trace (the morphology), as opposed to statistical features that describe them (114).

- Automatically derived features: features that are derived from the data itself with respect to either the classification problem or in a more general sense.

2.12.1 Statistical and Frequency Aggregative Features

Statistical features are based on aggregate statistics of the windows, such as mean, skewness and kurtosis. These features represent some of the highest performing (intra-protocol) and most used features in activity classification (54). This is partly due to their simplicity and their use in Bao et al's (54) seminal work on activity classification. Statistical features allow for a lot of information to be extracted from the window and can allow for a high activity classification performance. However, methods using these features typically show a poor inter-protocol performance which represents a major disadvantage (76).

Frequency-based features are features based on the Fourier analysis (a way to approximate functions/signals by sums of trigonometric functions/signals) of the acceleration signal in the window, such as entropy of the spectrum and the dominant frequency (54). These features are typically paired with statistical-based features and show high performance for a variety of activities. Much like statistical features, these show high intra-protocol performance but low inter-protocol performance (76).

Table 3 shows one of the most commonly used feature sets, which uses 39 statistical and frequency features (54) (12 features for each axis, and cross-correlation on each pair of axes, 3). This feature set has been shown to have a high performance on a range of different data protocols.

Statistical Features	*Frequency Features*
Mean	Kurtosis
Standard deviation	Number of zero crossing
Minimum	Energy of acceleration signal
Maximum	Principal frequency of acceleration signal
Variance	Magnitude of principal frequency
Median	Cross-correlation
Skewness	

Table 3: A selection of commonly extracted statistical and frequency features for activity classification.

One potential issue of statistical features for classification is they may be too representative of the acceleration data. A hypothesised reason that statistical features have a high intra-protocol performance, but low inter-protocol performance is that they can represent the data so well they begin to overfit, thus reducing their ability to generalise to unseen data.

2.12.2 Morphology-Based Approaches

Morphology-based approaches generate features that are based on the shape of the acceleration trace (the morphology), as opposed to statistical features that describe them. Typically, these approaches are based on the acceleration trace itself, matching the trace to known examples. Alternatively, much like Fourier analysis, the data may be decomposed to identify features. However, information about the decomposition, not statistical features of the decomposed functions are used (94).

Typically, features extracted from morphology-based approaches show a higher inter-protocol performance at the cost of a lower intra-protocol performance (76). One advantage of morphology-based approaches is they can generate features using unlabelled data (94). This potential use of unlabelled data means that the data from the test set can be used in the feature extraction. This

ensures that features will describe the test data and may improve inter-protocol performance (as features can be modified for each new participant/protocol).

The main disadvantage of morphology-based features is that they typically have a lower intra-protocol performance when compared to statistical features. The lower performance may be because morphology-based features are not as representative as statistical-based features, explaining both the lower intra-protocol performance and the higher inter-protocol performance. However, as they are less representative, the features will not correspond as closely to the training data as statistical features will, decreasing overfitting.

2.12.2.1 Movelets

Movelets are created on a single labelled acceleration segment by segmenting the data into overlapping smaller sub-segments, each with the label of the complete segment. This process is completed for all segments of acceleration data in the training set, resulting in a 'dictionary' of sub-segments (called movelets) with associated labels. When given unlabelled data, it is segmented and the label of the closest 'matching' movelet is assigned to each sub-segment. The most common label for all of these sub-segments is then assigned (combining both feature extraction and classification) (115,116).

2.12.2.2 Template Matching

Template matching is a feature extraction method similar to movelets (115), in that both methods create a representation of each activity. Unlike movelets, template matching creates a single 'template' for each activity. These templates can then be compared to an unseen acceleration window and the most 'similar' template can be used to predict the labels of the activity (76) or the distance from each known label can be computed and used as features.

2.12.2.3 Sparse Dictionary Encoding

Dictionary encoding creates features by decomposing the acceleration segments into simple filters, such that the initial data can be recreated via linear combinations of the filtered signals. Sparsity refers to limiting the number of filters that need to be 'activated' to recreate any one signal. The sparsity helps to eliminate noise and ensures that the signals do not simply perform a Fourier decomposition. Features are created by identifying the activations of the individual filters required to recreate the data (94,117). This will be discussed further in Chapter 8.

2.12.2.4 Recurrent Quantification Analysis

Recurrent Quantification Analysis creates an image based on the recurrent structure of the acceleration and then computes aggregative statistics based on this image, combining both morphology and statistical-based methods. Recurrent Quantification Analysis has shown considerable success in accelerometry gait analysis (118), a field of study similar to that of activity classification. This will be investigated in Chapter 7.

2.12.3 Automatic Feature Extraction

Automatic feature extraction refers to methods of identifying features automatically (119–121). Typically, these methods work in conjunction with the training stage of the classifier, modifying the features extracted to improve performance on the training data. The features are constructed through mathematical operations performed on weighted sums of the input data. In most cases, the mathematical operations are defined by hand and the optimal weighting for the sums are found by the algorithm. While these features can be described as statistically based, their formulation is dependent on the algorithm and the training data and therefore they are thought of as automatically extracted.

Typically, these methods achieve very high performance (109) (higher than other feature methods). Additionally, the features extracted represent the training data well. It may also be possible to use automatic feature extraction methods on testing data in order to create features that represent the test data highly, thus increasing the inter-protocol performance. This is discussed further in Chapter 8.

The main limitation is that they typically require a large amount of data to create representative features. Additionally, as the features are not defined by the researcher, they are not interpretable into meaningful metrics.

2.12.4 Conclusion

In conclusion, there is a consistent trade-off between the intra and inter-protocol performance of features. Statistical-based features and automatically extracted features have high intra-protocol performance but low inter-protocol, while for morphology-based features the opposite holds true.

The challenge, therefore, is to identify features that allow for high inter-protocol performance without reducing the intra-performance. This is addressed in Chapters 7 and 8.

2.13 Creating the Classifier

As discussed previously, the classifier can be thought of as a function that maps input data to the desired output, specifically mapping the acceleration data to the PA labels.

Many different classifier algorithms exist, each with their own advantages and disadvantages, these can be broken into two categories: Discriminative and Generative classifiers (122).

When given observable variables $D, T = \{d_n, t_n\}_{n=1}^{N}$, with d_n and t_n being the n^{th} data point with the corresponding label, a Generative classifier attempts to model the joint probability distribution $P(D, T)$. By comparison, a Discriminative classifier attempts to model the conditional probability of T given D or $P(D|T)$. Classifiers that do not make use of any probability - mapping the inputs directly to the outputs - are also referred to as Discriminative.

When attempting to maximise classification performance, the use of Discriminative classifiers is typically recommended,
"*one should solve the [classification] problem directly and never solve a more general problem as an intermediate step*" (123)

However, when the amount of data is limited, Generative classifiers have been shown to be preferred (122), reaching a high level of performance with logarithmically lower amounts of training data. As the cost of gathering training data for activity classification is so high, this is naturally advantageous.

The issue of inter-protocol performance has been raised many times in this thesis, relating to how well a classifier can predict data from a different protocol. Although this hasn't been tested on activity data, as far as the author knows, Generative classifiers typically have a lower performance on data from highly similar data-sets but a higher performance on data that is slightly different from the training data (122). It may, therefore, be the case that Generative classifiers have a higher inter-protocol performance, at the cost of a lower intra-protocol performance; although this has never been tested (as far as the author knows). This will, therefore, be addressed in Chapter 99, by identifying six classifiers (3 Generative, 3 Discriminative) and comparing their inter and intra-protocol performances.

2.14 Post-Processing

Post-processing refers to the modification of the predicted labels after the classification process in order to improve performance. Most post-processing approaches use the sequential nature of activity data to improve performance, making use of the fact that consecutive segments are likely to be the same activity.

2.14.1 Smoothing

An issue that arises in activity classification is isolated misclassifications embedded in longer sequences of correct classifications, for example, a single segment (typically around 10 seconds in length) identified as running embedded in a 10-minute sequence of sleep segments. Such a classification is generally an effect of imperfect class separation within the classifiers, causing a misclassification. The fact that activities tend to be part of a behavioural sequence, therefore last longer than the length of one window can be used to filter out these misclassifications in a method very similar to a moving average (as discussed in 0), called smoothing.

A moving average is used on the identified activity labels, with the average being the mode. This has the effect of filtering out any isolated misclassifications. Studies have identified that this post-processing typically results in a performance increase, with increases in accuracy ranging from +3.7% in (75) to 0.3% in (124). This is a commonly used method, although the performance difference between the smoothed and unsmoothed versions is often not reported.

The advantages of this method are: it has a very low computational requirement and can easily be used in real-time classification systems (albeit with a slight buffer). It is conceptually simple and makes minimal assumptions about the underlying data (other than assuming very short events are unlikely). A

limitation is that as far as the author is aware no investigation on how window size affects its performance has been undertaken. Similarly, this method is unlikely to perform well with automatic segmentation methods, as sequential windows are unlikely to share labels. Furthermore, this method assumes that low duration events do not occur (an event the duration of a single window would likely be identified as an error) even though this may not be the case when investigating some forms of PA.

2.14.2 Hidden Markov Model Smoothing

Another form of post-processing that attempts to remove isolated misclassifications is the use of Hidden Markov Models.

Hidden Markov Models are statistical models that can be used to describe the creation of an observable time series, making use of internal factors that are not directly observable. The data in the time series are referred to as symbols, and the internal factors are referred to as states. A Hidden Markov Model consists of two probabilistic processes, an invisible series of hidden states (a Markov chain) and a visible series of observable symbols. It is assumed that the observable symbols are probabilistically dependent on these hidden states (the value of the states affects the probability of the value of the symbols). It is also assumed that the hidden states form a time series where each successive state is probabilistically dependent on only the value of the prior state and no other states (the Markov assumption) (69).

The time series of classified activities are the observable symbols, and a hypothetical time series of ‘correct’ classifications are the hidden states. Using information about the probability of transitioning from one activity to another (found from the observable symbols), the probability distribution of the different activities (found from the training data) and the observable symbols themselves, it is possible to recreate these hidden states and effectively recreate the time series of ‘correct classifications’ from the given data (69).

This method typically improves the performance when applied (68). The disadvantage of method is that it typically requires the transition probabilities to be learned from the training data. In the event that the training and testing data are from two different protocols the transition probabilities may radically differ, impacting performance.

2.14.3 Participant Adaptation via Iterative Relearning

A post-processing technique that has seen considerable success is Participant Adaptation via Iterative Relearning (125).

This post-processing method attempts to use the participant's own data to retrain the classifier and improve the classification. To begin with, the participant's data is classified in the usual way, resulting in a time series of accelerations with predicted activity labels. The accelerations and predicted activity labels are then used to retrain the activity classifier, making it specialised towards that participant. This process can be repeated multiple times, adapting to the individual participant more and more.

The technique can lead to a performance increase ranging from 5.2% to 28% after one iteration and after four iterations a minimum increase of 16% has been reported (125).

The disadvantage of this is that this method only works if the original classification is reasonably accurate. Errors in the original classification propagate through the iterations and can cause performance to decrease.

2.14.4 Null Classes

Most classification methods allow for an estimation of the probability of accuracy by the classification process. A low probability of accuracy implies that the label identified may be wrong. A potential post-processing method is to assign all labels that have a probability of accuracy below a threshold to a

separate ‘Null’ class (126). A classifier then attempts to label the Null class data with broader activity labels (such as simply Active or Inactive) that make use of the fact that broader labels typically allow for higher performance.

Observation of this Null class by the researcher will allow for the identification of weakness in the classification model that may then be addressed.

2.14.5 Conclusion

Many different methods of post-processing exist, with the majority increasing the intra-protocol classification performance, however as far as the author is aware, all of them have been tested for intra-protocol performance instead of inter-protocol. Methods that improve intra-protocol performance do not always improve inter-protocol performance.

As the aim of this thesis is to maximise the inter-protocol performance (while maintaining intra-protocol performance), these post-processing methods must be investigated to see their effect on inter-protocol performance.

Specifically, the post-processing methods that will be investigated in Chapter 5 are:

- Using Smoothing
- Using a Hidden Markov Models smoother
- Using Participant Adaptation via Iterative relearning

Using Null classes will not be tested in this work, as the aim of this work is to identify the specific activity labels.

2.15 Methodological challenges and gaps in current research

This review has highlighted many different methodological challenges of the classification of PA type from acceleration data. These include:

- The challenge of identifying methods that allow for invariance to sensor placement location (especially on the wrist), investigated in Chapter 4.
- The challenge of identifying the effects of pre and post-processing on the intra and inter-protocol performance, investigated in Chapter 5.
- The challenge of identifying optimal window length, investigated in Chapter 6.
- The challenge of identifying features that allow for high intra-and inter-protocol performance, investigated in Chapter 7 and 8.
- The challenge of identifying the optimal classifier that allows for high intra-and inter-protocol performance, investigated in Chapter 9.

2.16 Conclusion

In conclusion, there are several different challenges in the research that need to be addressed. The following chapters will attempt to develop methods for mitigating the challenges mentioned in this review, allowing for the creation of a classification pipeline that has high inter and intra-protocol performance. The next chapter will describe the data-sets used in the following chapters.

3. Data-sets and Base Classifier

3.1 Introduction

The purpose of this chapter is to describe the data-sets used in the rest of this thesis, specifically for training the classifier, optimising the hyperparameters of the classification pipeline and appraising the final classification performance of the activity classification pipeline created in this work. The activity protocols used to gather the data-sets are described, as well as several summary statistics relating to the data itself. Additionally, the methods used for assessing the performance of the classifier will be set out. Finally, a Base activity classification pipeline will be created from current research, this will follow the pipeline as defined in section 2.2, with the specific methods discussed in section 3.5. This Base Classifier will act as the criterion approach used in the rest of this thesis with the performance of all new methods developed being compared to this Base Classifier. The data-sets defined in this chapter represent the data-sets used in the training and testing of the methods developed in the rest of this thesis.

3.2 Data-Sets

For training the classifier, two data-sets were used (referred to as classification data-sets). The classification data is comprised of *labelled* activity/acceleration data. These represent the data-sets that the training, testing and validation data are drawn from. These data-sets are named after the activity protocol used to gather the data, Lab-Based and Free-Living respectively. A third *unlabelled* data-set is used for the overall appraisal of the optimised classification pipeline (referred to as the assessment data). The assessment data is comprised of *unlabelled* acceleration data.

When examining the classification data-sets, four factors are important to note:

- The activity protocol: this describes how the activities in the study were performed, hence how the acceleration and the corresponding labels were gathered. It also describes the process of identifying the activity labels. The protocol is important to identify because it allows for comparisons between the protocols. It also shows how 'realistic' (reflective of everyday living) the activity protocols are.
- Participant characteristics: these are essential for evaluating how closely the participants resemble various populations. The participants of the training and testing data can have a considerable effect on the performance of the classifier. For instance, data gathered from normal-weight participants fails to accurately represent data from overweight participants, decreasing inter-protocol performance (when trained on normal weight and tested on overweight) (76). Differences in the participant characteristics between data-sets represent a change in the underlying mapping function from accelerations to activities, hence a potential decrease in performance. In addition to bodyweight, other factors such as age, height, functional capacity and disease status may alter the acceleration values gathered. In essence, both this and the activity protocols are trying to evaluate the external validity of the data, identifying how much the external factors influencing the data have changed between protocols and participants.
- Class balance: this refers to the proportion of the different classes (activity labels) in the data-set. The class balance affects the classification performance (75), with a classifier trained on imbalanced classes being more likely to predict the majority classes; decreasing performance on data-sets where the imbalance no longer holds. Unlike the other points identified here, class imbalances can be alleviated via pre-processing (75), resulting in balanced data-sets that can potentially improve performance.

- Activity transition probabilities: these represent the chances of transitioning from one activity to another in the activity protocol. These probabilities are used in post-processing methods to improve performance (68). These probabilities are computed by first segmenting the data into 12.8 seconds blocks and then computing the transition probabilities (the choice of 12.8 seconds is explained in section 3.5).

When examining the assessment data-set, an additional factor must be identified, as well as the participant characteristics and the activity protocol:

- The inclusion and exclusion criteria: this refers to the criteria that the participants must fulfil to be included in this study. This is to ensure that the participant's PA will not be affected by major health issues (e.g. only having one leg). This also ensures that the only participants who remain in the study are the ones who have worn the accelerometers for a minimum time span. This is so that enough data is gathered to estimate their habitual PA (127).
- Activity transition probabilities and class balances require known activity labels, as the assessment data is unlabelled, these cannot be calculated.

3.2.1 Lab-Based Data-Set

This is one of two classification data-sets used in the creation of the classifier. This is comprised of labelled data, as the name suggests this data was gathered from a Lab-Based protocol.

3.2.1.1 Activity Protocol

During a single visit to a laboratory at the University of Exeter, 16 participants were given an ordered list of physical activities to complete (Table 4). This is the activity protocol that the data was gathered under.

Labelling 1	*Duration*	*Labelling 2*	*Sedentary-Stand-Active Labelling*
Lying	30 minutes	Lying	Sedentary
Transition period	5 minutes	Transition period	Transition period
Watching TV	5 minutes	Sitting	Sedentary
Transition period	1 minutes	Transition period	Transition period
Working at desk	5 minutes	Sitting	Sedentary
Transition period	1 minutes	Transition period	Transition period
Standing still	5 minutes	Standing	Standing
Transition period	1 minutes	Transition period	Transition period
Vacuuming	5 minutes	Household	Standing
Transition period	1 minutes	Transition period	Transition period
Washing Dishes	5 minutes	Household	Standing
Transition period	1 minutes	Transition period	Transition period
Folding Laundry	5 minutes	Household	Standing
Transition period	1 minutes	Transition period	Transition period
Slow walking	5 minutes	Walking	Active
Transition period	1 minutes	Transition period	Transition period
Moderate walking	5 minutes	Walking	Active
Transition period	1 minutes	Transition period	Transition period
Fast walking	5 minutes	Walking	Active
Transition period	1 minutes	Transition period	Transition period
Stair climbing	5 minutes	Walking	Active
Sit to stand transitions	5 minutes	Transition period	Transition period
Sit to walking transitions	5 minutes	Transition period	Transition period

Table 4: Activity protocol for the Lab-Based data with different labelling schemas.

Three different labelling schemes were used to classify each of the twelve activities performed (Table 4). The first scheme uses a description of the specific activity for the label, this represents the most specific labelling. The second labelling groups the data into broad activity classes (Household, Walking, Sitting, Lying and Standing. The third labelling scheme assigns all activities to Sedentary, Standing or Active classes. This third scheme is required in order to match that of the Free-Living data which is restricted to these labels.

During the protocol, participants wore two wrist-mounted tri-axial GENEActivs (61) (one on each wrist), with sampling frequencies set to 100Hz. The accelerometers were placed by the researchers to ensure consistent orientation (see Chapter 4). All activity labels were identified via direct observation. Intensities were self-derived.

The University of Exeter ethics committee approved this (20/4/2017) and informed consent was obtained before participation.

All data that was given the transitioning label was removed before any further data processing. This is because it is impossible to correctly assign an activity label to a window containing an activity transition, due to there being two activities. If these incorrectly labelled windows are included in the training of the classifier, it reduces the performance (54).

3.2.1.2 Transition data

The Sit-to-Stand and Sit-to-Walk transition data were collected in order to have data containing known transitions which could be used for evaluating transition detection methods. Each participant was requested to transition from Sitting to Standing or Walking and back 10 times (transitioning from one activity to another was considered one transition) in the 5-minute interval. The time of these transitions was recorded via direct observation. For all activity

classification, this data was removed with the other transitions, and was used solely for transition detection in Chapter 6.

3.2.1.3 Participant Statistics

Characteristics of the 16 participants are displayed in Table 5. Ages ranged from 18-47. No participants reported health issues that may have affected their movements. All participants abstained from caffeine and alcohol for 8 hours before the data was gathered. Caffeine, alcohol and health issues can all affect movement, therefore not controlling for this may have meant that the movements performed in this data-set were not representative of a 'normal' participant. Body Mass Index (BMI) ranged from 17.3-24.3, hence no participants were overweight.

	Mean	*Range*
Age (years)	25.3	18 - 47
Height (metres)	1.68	1.55 - 1.84
Weight (kgs)	72	56 – 93

Table 5: Lab-Based data-set statistics.

3.2.1.4 Class Balances

As can be seen in Figure 9a, with 'labelling 1' classes were reasonably balanced, with all but one class being of equal size. The largest class was Lying Supine which had a duration of 30 mins per participant. All other classes had an equal duration of 5 minutes. In 'labelling 2' (Figure 9b) the classes were less balanced with 38% of the entire data-set being comprised of a single activity (Lying) and the least common activity only contributing 5% of the total (Standing). In 'labelling 3' (Figure 9c), Sedentary-Stand-Active, the most common class was Sedentary, while Active and Standing classes had equal portions of 25%. This means that in this data an equal amount of time was spent either Active or Standing, which unlikely to be realistic of Free-living data.

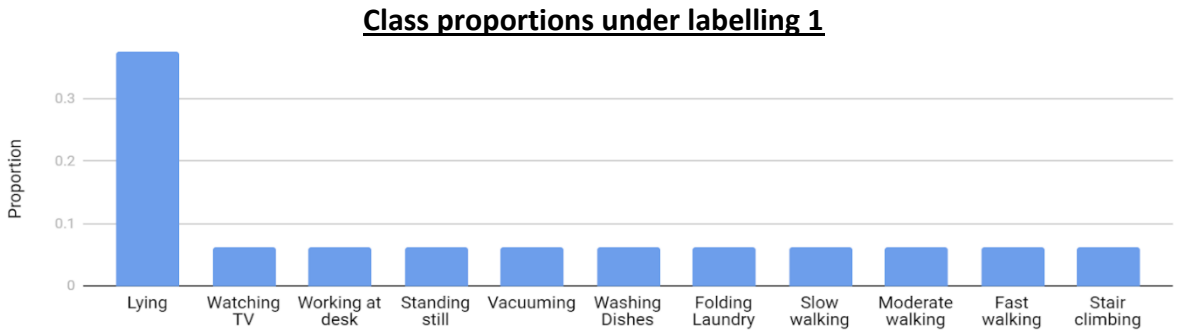

Figure 9a: Class proportions under labelling 1.

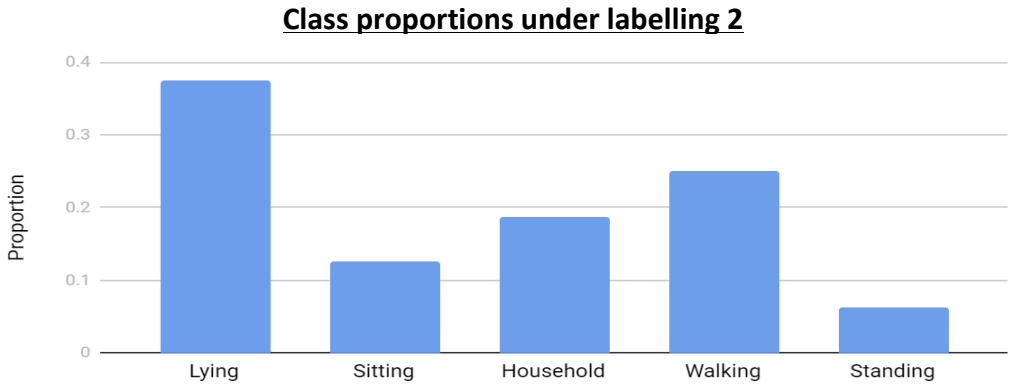

Figure 9b: Class proportions under labelling 2.

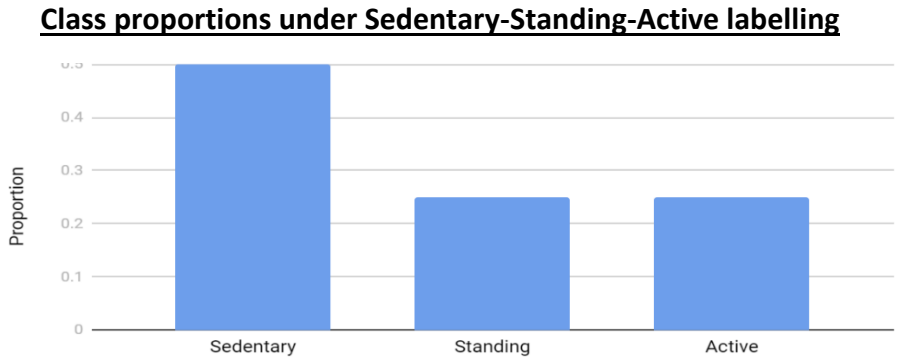

Figure 9c: Class proportions under Sedentary-Stand-Active labelling.

3.2.1.5 Transition Probabilities

As can be seen in Table 6, the probability of transitioning from one activity to another is very low. While the self-transition probability (probability of remaining

in a given activity) is very high. This is due to the activity protocol, assuring that all activities are done in continuous blocks of 5 minutes. This is very different to data that is gathered in a Free-Living scenario. Such a clear divergence from natural behaviour may be one of the major reasons that Lab-Based studies are so poor at classifying data from Free-Living studies (77,128).

	Sedentary	*Standing*	*Active*
Sedentary	0.98	0.02	0
Standing	0.02	0.95	0.03
Active	0	0.08	0.92

Table 6: Transition probabilities for the Lab-Based data under Sedentary-Stand-Active labelling, rows represent activity being transitioned from, columns represent activity being transitioned to.

3.2.2 Free-Living Data-Set

This the second of two classification data-sets used in the creation of the classifier. This is comprised of labelled data, as the name suggests this data was gathered from a Free-Living protocol.

3.2.2.1 Activity Protocol

The participants were able to undertake Free-Living with no fixed activity plan, without observation (hence Free-Living data-set). The data-gathering phase lasted seven days, allowing for 168 hours of acceleration data with paired labels to be gathered. The participants wore a GENEActiv (61) on the right wrist, with a sampling frequency of 100Hz. Additionally, the participants had a thigh-mounted ActivPal accelerometer (129,130), measuring at 20Hz. The accelerometers were placed by the researchers to ensure consistent orientation. As direct observation was not possible, labels for performed activity were generated from the ActivPal. The ActivPal can assign one of three potential labels (Sedentary, Standing or Active) with a high degree of validity and are therefore assumed to be fully correct (129,130). The University of

Exeter ethics committee approved the study (20/4/2017) and informed consent was obtained before participation.

3.2.2.2 *Participants Statistics*

This data-set was comprised of 49 participants with ages ranging from 18-53 (Table 7). No participants reported health issues that may have affected their movements. All participants abstained from caffeine and alcohol for 8 hours before the data was gathered. BMI ranged from 18.1-28.3, hence no participants were obese, although 23% were overweight.

	Mean	*Range*
Age (years)	24.3	18 – 53
Height (metres)	1.69	1.54 - 1.91
Weight (kgs)	74.6	48 – 102

Table 7: Characteristics of participants in the Free-Living data-set.

3.2.2.3 Class Balances

This data-set was less balanced than the Lab-Based protocol with 8% of activities being ‘Active’ and 80% being Sedentary (Figure 10). This level of activity is equivalent to approximately two hours of physical activity per day, levels that coincide with previously reported values (15).

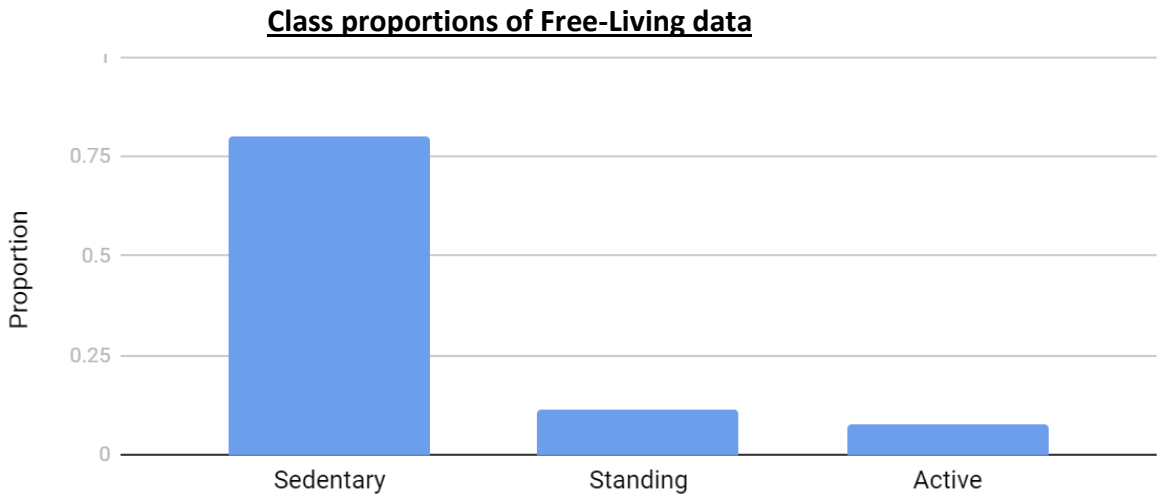

Figure 10: Class proportions in Free-Living data-set.

3.2.2.4 Transition Probabilities

The transition probabilities between activities were much higher in this data-set (Table 8), compared to the Lab-Based data-set. This is most likely because the participants were allowed to decide when to change activities (as no researcher guidance was given).

	Sedentary	*Standing*	*Active*
Sedentary	0.82	0.17	0.01
Standing	0.15	0.62	0.23
Active	0.02	0.34	0.64

Table 8: Transition probabilities for the Free-Living data-set, rows represent activity being transitioned from, columns represent activity being transitioned to.

This difference in class-balance and transition probabilities represents one of the major differences between Lab and Free-Living based data-sets. Although gathering the data in a Free-Living scenario allows for more realistic data, it does have disadvantages. As direct observation is not possible, an ActivPal was used to label the data. As an ActivPAL is a thigh-mounted device it can only identify postural labels (Active, Sedentary, and Standing) (129,130). This means that it cannot identify when a participant transitions between two

activities with similar postures (walking/running). This was found to be a problem when attempting to evaluate methods of activity transition detection in Chapter 6.

3.2.3 Assessment Data-Set

The assessment data-set is from the REtirement in ACTion (REACT) (131), a randomised controlled trial of a PA intervention comprised of acceleration records and health metrics.

3.2.3.1 Activity Protocol

Each participant wore a wrist-mounted GENEActiv (61) accelerometer for 7 days, with a sampling frequency of 30Hz. The accelerometer was placed by the researchers to ensure proper orientation and attachment on the non-dominant wrist.

3.2.3.2 Participants Statistics

The data-set comprises 712 adults with ages ranging from 65 - 98, with a mean of 77.6 (Table 9). Height ranged from 1.35m to 1.97m and BMI ranged from 17.2 to 51.1. 37% of the participants were obese and 37% overweight.

	Mean	*Range*
Age (years)	77.6	65 – 98
Height (metres)	1.63	1.35 - 1.97
Weight (kgs)	78.1	42 - 147

Table 9: Characteristics of participants in the assessment data-set.

3.2.3.3 Inclusion and Exclusion Criteria

Inclusion criteria:

- Aged 65 years or older and not in full-time employment
- Planning to reside in the target area (Bath/Bristol, Devon, Birmingham) for at least 24 months
- A score between 4 and 9 (inclusive) on the Short Physical Performance Battery (132)

Exclusion criteria:

- Self-reported inability to walk across a room without a walker or the help of another person
- Existing major mobility limitation (defined as Short Physical Performance Battery of 3 or less, or unable to complete the 4-m walk component of Short Physical Performance Battery)
- Living in residential or nursing care
- Inability to attend the REACT PA sessions as scheduled
- A documented or patient-reported medical condition that would preclude participation, including:
    - Arthritis so severe it would prevent participation in PA
    - Parkinson's disease or diagnosed dementia
    - Any terminal illness
    - Lung disease requiring the use of orally administered corticosteroids or supplemental oxygen
    - Severe kidney disease requiring dialysis
    - Severe heart disease that would prevent participation in PA (e.g. chest pain when walking 100 or 200 yards or up a flight of stairs)
    - Implanted cardiac defibrillator
    - Cardiac arrest which required resuscitation
    - Severe uncontrolled psychiatric illness

    - Currently receiving radiation therapy or chemotherapy treatment for cancer
    - Awaiting knee or hip surgery
    - Major heart surgery (including valve replacement or bypass surgery) in the last 6 months
    - Unstable heart condition (e.g. uncontrolled arrhythmia, angina, heart failure or hypertension)
    - Spinal surgery in the last 6 months
- Any other clinical condition that the person's GP or clinician considers would make them unsuitable for participation in a PA rehabilitation programme to prevent the decline of lower-limb functioning
- Heart attack (or myocardial infarction), stroke, spinal surgery, hip fracture, hip or knee replacement within the previous 6 months
- Currently receiving physical therapy on legs or enrolled in another PA research or intervention study
- Less than 7 days of consecutive wear time with the accelerometer.

3.3 Evaluation

Chapter 2 identified four comparisons for assessment of the performance of a classifier:

- Intra-protocol-intra-subject
- Intra-protocol-inter-subject
- Inter-protocol-intra-subject
- Inter-protocol-inter-subject

As discussed in Chapter 2, the aim of this thesis is to optimise both the inter-protocol performance (inter-protocol-inter-subject) and the intra-protocol (intra-protocol-inter-subject) performance. This is done by identifying

hyperparameters that allow for a high level of inter and intra-protocol performance.

The intra-protocol performance is calculated by training and testing the classifier on the same protocol (either the Free-Living data-set or the Lab-Based data-set), with different participants. The classifier is trained on participants from the given data-set, and then used to predict the activity labels of other participants (from the same data-set) from their acceleration data. These predicted labels are then compared to the known correct labels so that the performance of the classifier can be determined. Specifically, Leave One Subject Out Cross Validation is used, a common method used in activity classification literature (54). Leave One Subject Out Cross Validation works by training the classifier on all but one participant, and then evaluating the performance on the remaining participant. This procedure is repeated for all participants and the averaged evaluation metric is reported (although the individual performances are retained for statistical testing). This gives an idea of the performance of the classifier over each participant. In this work, Leave One Subject Out Cross Validation performance is computed for both the Lab-Based data and the Free-Living data, giving a performance score for each data-set, referred to as the LabCV and FreeCV scores. These are measures of the intra-protocol performance.

The inter-protocol performance is identified by testing the classifier on data from a different protocol than it was trained on. Specifically, in this work, this is done by assigning the Lab-Based data as either the training or testing data, with the Free-Living data being the converse. In this case, the performance is evaluated for each participant in the test data-set separately, as this allows for paired statistical testing of the participants (133). As the specifics of the statistical testing vary between chapters, more detail will be provided in the Analysis sections of each chapter. The averaged evaluation metric is reported, giving an inter-protocol performance score for each data-set, referred to as the Lab-Free

(trained on the Lab, tested on the Free) and Free-Lab (trained on the Free, tested on the Lab) score.

3.4 Evaluation Metric

The evaluation metric used in this thesis is the F1-score (134), the harmonic mean of precision and recall. This metric was chosen instead of simpler metrics such as accuracy because the F1-score is typically more robust to class imbalances. This metric has seen much use in activity classification (55). The range of this metric is 0-1, with 1 indicating perfect recall and precision.

F1-score is originally a binary classification metric, as such a minor modification is used to allow this to function in a multi-class setting.

A one versus all approach is taken, where the classification labels are transformed into binary labels (either label L or not label L) for each potential label. The number of true positives (data that was classified as label L correctly, TP_L), False Positives (data that was incorrectly classified as label L, FP_L) and False Negatives (data that was incorrectly classified as not label L, FN_L) are computed as standard and then used to solve the precision and recall for Label L ($Precision_L, Recall_L$), resulting in precision and recall measures for each potential L. The weighted average of these precision and recall measures are then used to compute the overall precision and recall. The overall F1-score is then computed as the harmonic mean of the overall precision and recall.

$$Precision_L = \frac{TP_L}{TP_L + FP_L}, \qquad Recall = \frac{TP_L}{TP_L + FN_L}$$

$$Precision(All) = \frac{\sum Precision(L)|All = L|}{|All|}, Recall(All) = \frac{\sum Recall(L)|All = L|}{|All|},$$

$$F1(All) = 2 \times \frac{Precision(All) \times Recall(All)}{Precision(All) + Recall(All)}$$

Equation 3: F1-Score.

3.5 Base Classifier

The pipeline for the Base classifier is:

1. Determination of data type
2. Pre-processing: None
3. Windowing: Using 12.8-second windows
4. Feature extraction: Using 39 features based on the statistical aggregate features and frequency statistics.
5. Building the classification model: Using a Random Forest classifier with 50 separate trees.
6. Post-processing: None

Much of this thesis focuses on identifying and fixing methodological concerns in activity classification. Fixing these methodical concerns involves modifying hyperparameters of the classification pipeline and adding extra-steps in the pipeline. In order to ensure that each 'fix' is successful, this must be compared to a criterion approach. The criterion approach, referred to as the Base classifier, is a classification pipeline based on the work of Chowbury et al (55). This is a classification pipeline that has been validated on both Lab and Free-Living data-sets. As mentioned in Chapter 2, it is difficult to directly compare classification pipelines due to different performance measures and differing data-sets (54,55), so it may not be the case that this is the highest performing of all available methods. However, this classification pipeline was used for a variety of reasons. The lack of pre and post-processing allowed for a simpler examination of the impact of the pre and post-processing methods reviewed in this work, assuring that there was no emergent behaviour that arose due to interaction between processing methods. A Random Forest is a classification model that typically avoids overfitting and can therefore be expected to generalise to unseen data well (135). Additionally, as mentioned above this classification pipeline has been validated on both Lab and Free-living data. Thus, this classification model has a relatively high ability to generalise both in

theory and practice. Furthermore, the features used in this pipeline represent some of the most widely used and validated features in activity classification (54). For these reasons, this classification is used as the 'state of the art' that all methods in this work will be compared to.

3.5.1 Features

The 39 features used in the classification pipeline are those used in the work of Bao et al (54). For each axis, 12 features are computed, resulting in 36 features overall. For each pair of axes, the cross-correlation is also computed, resulting in three extra features. Overall, 39 features are used. These features can be seen in Table 10, these will be referred to as the Base features in the rest of this thesis.

Statistical features	*Frequency Features*
Mean	Kurtosis
Standard deviation	Number of zero crossings
Minimum	Energy of acceleration signal
Maximum	Principal frequency of acceleration signal
Variance	Magnitude of principal frequency
Median	Cross-correlation
Skewness	

Table 10: The classification features used in the Base classifier.

3.5.2 Random Forest and Decision Trees

Decision trees are classifiers that use the values of features to split the data into partitions. The training process corresponds to identifying the features and the values that best allows for splitting the data into the given classes (different labels). The predictive step involves using the feature values to assign the data to a partition and outputting the majority labelling in that partition. Decision trees are a very popular classification method, *"because it is invariant under scaling and various other transformations of feature values, is robust to the inclusion of*

irrelevant features, and produces inspectable models. However, they are seldom accurate" (136). An additional issue with decision trees, is they tend to model the training data too well, preventing them from generalising the predictions to the test data. This overfitting represents the largest issue with decision trees.

Random Forests are an extension to decision trees that can generalise performance to unseen data (prevent overfitting). Random Forests are a combination of multiple decision trees (an ensemble) trained on subsets of the training data. The output of a Random Forest is the majority predicted classification. During training, each tree is only exposed to a subset of the data; this is known as bagging, and each tree can only make use of a subset of the features, known as feature bagging. Bagging is done because individual decision trees are highly sensitive to noise in the training data, but the average of many trees is not sensitive as long as the trees are not correlated. The feature bagging is used to make sure that a single very effective feature does not dominate the training process. This would result in the trees being highly correlated with each other and would decrease the effectiveness of the initial bagging (135).

Decision trees and Random Forests also allow for implicit identification of feature importance. By identifying how well a feature can partition the data, the importance of that feature can be computed. Specifically, the importance of a given feature can be computed by identifying the *“total decrease in node impurity (weighted by the probability of reaching that node (which is approximated by the proportion of samples reaching that node)) averaged over all trees of the ensemble.”* (135), where node impurity is a measure of how mixed the partitions created by the node are.

3.5.3 Performance

The Base classifier achieves a high intra-protocol performance (0.898, 0.765) but a low inter-protocol performance (0.352, 0.415), as shown in Table 11. This result demonstrates the exact problem identified in Chapter 2, that the majority of current approaches are built by optimising the intra-protocol performance, thus reducing the inter-protocol performance, resulting in an inability to accurately classify data from a different protocol (77,128).

	LabCV	*FreeCV*	*Lab-Free*	*Free-Lab*
Averaged F1-Score	0.898 (0.103)	0.765 (0.214)	0.352 (0.132)	0.415 (0.0978)

Table 11: Inter and intra-protocol performance of the Base classifier. Figures in brackets indicate standard deviations.

3.6 Conclusion

This chapter discussed the data-sets used in this thesis. Additionally, the creation of the Base classifier was addressed. This gives the ability to analyse if changes to the classification pipeline are successful at dealing with the methodological concerns identified in Chapter 2.

4. Accelerometer Placement Location

4.1 Introduction

As discussed in Chapter 2, PA classifiers are typically trained on data obtained from sensors at a set orientation relative to the wrist or limb. Changes in this orientation (such as being on a different wrist) result in performance degradation. A method to obtain orientation invariance for classification of wrist-mounted acceleration data is therefore essential for ensuring high-performance classification when the orientation of the sensor cannot be guaranteed.

A possible solution is to make use of features derived from the raw acceleration that are orientation invariant, meaning that the features will be identical regardless of the orientation of the sensor (107); for example, the magnitude of the acceleration values. This prevents the performance reduction but limits the available features that can be used in the classification, which in turn may compromise performance.

Another more effective approach by Gjoreski et al (89) trained the classifier with data from both wrists, resulting in wrist invariance. However, data from the opposing wrist may confuse classification, reducing the performance. An additional limitation of this method is that data from both wrists must be collected in the training stage, increasing both the cost of the data gathering process and the burden to the participant.

A technique known as Domain Adaptation allows for wrist invariance without requiring data from both wrists and the creation of orientation invariant features. Standard machine learning aims to create a classifier based on labelled data from a training data-set that can correctly classify unlabelled data collected in the same way. The training data is said to come from the *source domain*, while the unlabelled data to be classified is said to come from the *target domain*; in

standard machine learning the source and target domains are identical. When the target data comes from a different domain, the classification performance generally drops due to the classifier only being suited to the source domain. Domain Adaptation methods seek to adapt data from the target domain to the source domain so that good performance is achieved (137). Clearly, the source and target domains must be related for Domain Adaptation to be successful.

This chapter investigates whether Domain Adaptation can be used to adapt data measured from the "wrong" wrist (the target domain) to the training data in order to achieve good classification performance regardless of the wrist on which the accelerometer is worn. Domain Adaptation approaches only require acceleration data from one wrist, as well as having no limitations on the features used, which is a substantial advantage over other approaches.

As mentioned above, Domain Adaptation involves allowing a classifier created on a source domain to be applied to a different (but related) target domain without suffering performance reduction. Specifically, it deals with techniques for moderating the performance reduction when classifying over different distributions; making it well suited to attenuating the performance drop from using differing wrists. Domain Adaptation has seen some use in activity classification and accelerometry studies (138), however, no work has been found that allowed for location/orientation invariance using Domain Adaptation. The similarity between visual data and time series data (their local correlations and innate structure) allows for Domain Adaptation algorithms designed for visual applications to function well on time series data. Although it is strongly linked with another field known as Transfer Learning, and most literature uses these terms interchangeably, Domain Adaptation will be used in this work (137).

This chapter will examine Domain Adaptation as a possible solution for achieving location/orientation invariance in activity classification via accelerometry.

4.2 Method

4.2.1 Data

This study made use of the Lab-Based classification data-set (3.2.1) as this utilized data from both wrists. Labelling 2 was used, meaning the acceleration data was assigned one of five activity labels: Lying supine, Sitting, Household tasks, Walking and Standing. These are described in 3.2.1.1.

4.2.2 Procedure

In order to evaluate the effect of Domain Adaptation on the performance reduction observed when applying activity classifiers to the 'wrong wrist', a series of comparisons were made between different Domain Adaptation and Non-Domain Adaptation approaches. Each of the approaches is described below and summarised in Table 12.

Five Domain Adaptation and Non-Domain Adaptation approaches were evaluated:

- The Criterion approach: this refers to creating and testing the classifiers on the same wrist. This approach serves as the gold standard for performance.
- The Domain Adaptation approach: in the approach, a classifier was trained on data of one wrist. Domain Adaptation was used to adapt the target data collected from the opposite wrist data to the source domain so that the trained classifier could be applied to the adapted target data.
- The Non-Domain Adaptation approach: here a classifier was trained from the data of one wrist. The resultant classifier was then used to classify the data from the opposite wrist with no modification. This method served as the control.

- The Not Applicable approach: here a classifier was trained from the data of one wrist. Domain Adaption was then used with the same wrist data serving as the target domain. The resultant domain adapted classifier was then used to classify the data from the same wrist. This method served to investigate the effect of using Domain Adaption when it is not required, in circumstances where the wrist placement of the accelerometer is unknown.
- The Amalgam approach: here a classifier was trained on data from both wrists, similarly to Gjoreski et al (89). The resultant classifier was then used to classify the data from just one wrist with no modification, to examine if using data from both wrists allowed for wrist invariance without the need for Domain Adaptation.

4.2.3 Analysis

For each approach in this chapter the LabCV performance (F1-score) was computed and compared to other approaches. As only one data-set was used in this chapter (because only Lab-Based data has data form both wrists) only the LabCV was computed; this refers to performing cross-validation in the Lab-Data to obtain a notion of the intra-protocol performance.

The comparative performance of the different approaches across different subjects was tested for statistical significance using the Wilcoxon Signed-Rank test, which tests the null hypothesis that two related paired (by participant ID) samples come from the same distribution. A low p-value ($p < 0.05$) indicates that the results are statistically significantly different from one another with high confidence. Due to the fact the multiple hypotheses were evaluated on the same data set, the likelihood of a Type I error is increased. This was compensated for by using Bonferroni corrections. This entails testing each individual hypothesis at a significance level of $\frac{\alpha}{\mathrm{m}}$, where α is the overall hypothesis level (in this case 0.05) and m is the number of hypotheses.

Approach	*Description*	*Training wrist data*	*Testing wrist data*	*Uses Domain Adaption*
Right Criterion	Single, same wrist Non-Domain Adaptation	Right	Right	No
Left Criterion	Single, same wrist Non-Domain Adaptation	Left	Left	No
Right Domain Adaptation	Single wrist Domain Adaptation	Left	Right	Yes
Left Domain Adaptation	Single wrist Domain Adaptation	Right	Left	Yes
Right Non-Domain Adaptation	Single wrist Non-Domain Adaptation	Left	Right	No
Left Non-Domain Adaptation	Single wrist Non-Domain Adaptation	Right	Left	No
Right Not Applicable	Single, same wrist Domain Adaptation	Right	Right	Yes
Left Non-Applicable	Single, same wrist Domain Adaptation	Left	Left	Yes
Right Amalgam	Both wrists Non-Domain Adaptation	Left+Right	Right	No
Left Amalgam	Both wrists Non-Domain Adaptation	Left+Right	Left	No

Table 12: Summary of classification methods using Domain Adaptation and alternatives. Left and Right refers to the testing wrist.

4.3 Classification Procedure

The activity classifier in this chapter follows the Base classification pipeline discussed in the preceding chapter, with three additional steps before the training stage (creation of the classifier). These steps are the Normalization step, Feature Reduction and the Domain Adaptation step.

1. Determination of data type
2. Pre-processing

3. Segmenting into windows.
4. Extracting features
   4.1. Normalization
   4.2. Feature reduction
   4.3. Domain Adaptation
5. Creating the classifier
6. Post-processing

All approaches tested in this work (Domain Adaptation, Non-Domain Adaptation, Not Applicable, Amalgam and Criterion) make use of steps 1-4.2, and 5-6. Only the Not Applicable and Domain Adaptation approaches make use of the Domain Adaptation step (4.3).

Feature Reduction is concerned with reducing the number of features used in the classifier. The Domain Adaptation method used in this work makes use of a Feature Reduction stage and an adaptive stage. In order to ensure comparability between all methods the same Feature Reduction stage was performed regardless of whether Domain Adaptation was utilised. This ensures that the effect of the adaptation stage is not masked by Feature Reduction. The specific form of Feature Reduction used in this work is Principal Component Analysis. It works by projecting high dimensional data (in this case 39) into a smaller number of dimensions - a subspace - while preserving as much variance as possible. The resulting low-dimensional features are linear combinations of the original, high-dimensional features. This technique has commonly been used in activity classification and a more detailed explanation can be found, for example, in the work of Lever et al (139). Principal Component Analysis requires a parameter (k) to be chosen, which is the dimension of the lower-dimensional subspace. In this work, k is chosen to be 12 (as determined with cross-validation) unless stated otherwise, although the performances for all values of k were evaluated.

Normalization is a procedure used to ensure that all features have similar variance, meaning that all features have an equal weighting in the data-set. This procedure is often used in activity classification work, although is not required in most cases. However, the feature reduction step requires normalization to occur. This is because Principal Component Analysis attempts to maximise the variance of the data during its projection to a lower-dimensional subspace. If the values are not normalized then this projection is typically dominated by the features with the largest values, instead of the features that have the largest normalized variances.

4.4 Domain Adaptation

The Domain Adaptation method used in this work is a straightforward modification of the Subspace Alignment algorithm (140). Subspace Alignment was selected because it does not require target labels, does not drastically increase computational load and it is insensitive to the precise value of the single parameter that must be chosen. The underlying idea of the Subspace Alignment algorithm introduced in (140) is to rotate the source data so that it best aligns with the target data; a classifier is then trained on the aligned source data in order to be able make accurate predictions on the target/test data. The dimensionality of the features in both source and target sets is reduced in a feature reduction step prior to the alignment.

In this form Subspace Alignment requires the source data to be rotated to align with each new set of target data, which may represent a substantial computational burden because the classifier must be retrained for each newly-aligned set of training data. Since training the classifier is computationally expensive compared with the cost of classifying new examples with the trained classifier, in this chapter the *target* data is rotated to align with the training/source data, as illustrated in Figure 11. This means that the expensive training of the classifier is done once (using unrotated source data), after which

each new target set for classification is aligned with the source data (a computationally cheap step) allowing it to be classified.

Algorithm 1: Subspace Alignment

Input: Source features F_s, target features F_t, subspace dimension k
1: $P_s \leftarrow$ Principal component analysis (F_s, k) // Generate k principal components of F_s
2: $F_s^a = F_s P_s$ // Reduce dimension of F_s
3: Train classifier using F_s^a and corresponding labels
4: Collect target features F_t
5: $P_t \leftarrow$ Principal component analysis (F_t, k) // Generate k principal components of F_t
6: $F_t^a = F_t P_t P_t^T P_s$ // Reduce dimension of F_t and align with source
7: Use classifier to predict labels for aligned features F_t^a

F_s and F_t denote the feature matrices of the source and target data respectively; each row represents an observation and each column one of the M = 39 features. P_s and P_t respectively denote the M by k (orthonormal) matrices of principal components of the source and target feature matrices, and P^T denotes the transpose of matrix P.

Example of Subspace Alignment

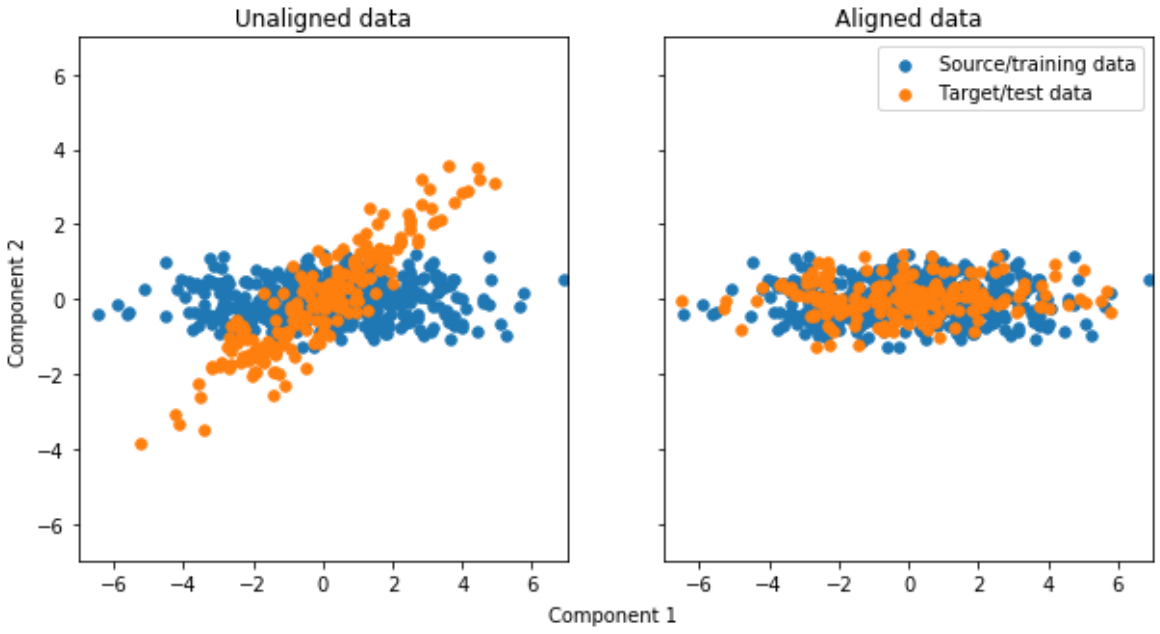

Figure 11: Example data-set, before and after Subspace Alignment. First, the data is reduced to a two-dimensional subspace (k = 2), in which the principal directions of the source data are aligned with the coordinate axes (left panel), then the data-sets are aligned by rotating the target data (right panel).

Algorithm 1 summarises the main steps in classification using Subspace Alignment. Prior to alignment, the source and target data are each projected into a subspace defined by a smaller dimension subspace defined by their principal components. This often has the beneficial effect of discarding noise, improving classification performance (70), and identifies the principal directions in the data (P_s and P_t in Algorithm 1) that should be aligned by rotation. Alignment of the dimension reduced target features is then accomplished by multiplication by the matrix $P_t^T P_s$ in step 6, after which the trained classifier may be used to predicted labels for the target features that have been rotated into alignment.

4.5 Results

Approach	*Description*	*LabCV score*
Right Criterion	Single, same wrist Non-Domain Adaptation	0.84 (0.17)
Left Criterion	Single, same wrist Non-Domain Adaptation	0.82 (0.15)
Right Domain Adaptation	Single wrist Domain Adaptation	0.81 (0.1)
Left Domain Adaptation	Single wrist Domain Adaptation	0.83 (0.13)
Right Non-Domain Adaptation	Single wrist Non-Domain Adaptation	0.72 (0.11)
Left Non-Domain Adaptation	Single wrist Non-Domain Adaptation	0.68 (0.14)
Right Not Applicable	Single, same wrist Domain Adaptation	0.81 (0.15)
Left Not Applicable	Single, same wrist Domain Adaptation	0.81 (0.15)
Left Amalgam	Both wrists Non-Domain Adaptation	0.80 (0.14)
Right Amalgam	Both wrists Non-Domain Adaptation	0.81 (0.14)

Table 13: Performance results (F1-score) of classification approaches using Domain Adaptation and alternatives for each participant. Figures in brackets indicate standard deviations.

Table 13 shows the average LabCV F1-score over 16 participants for all methods, over both wrists. Right and Left, indicates an approach that was tested on the right and left wrist data respectively. The Criterion approaches achieved the highest performance, as expected, because it involves creating

and testing the classifiers using data collected from the same wrist and is the gold standard for classification. Training on one wrist and classifying data from the other wrist (Non-Domain Adaptation) resulted in an average performance reduction of 12% compared to the Criterion approaches. In contrast Domain Adaption reduced this performance drop to an average of 1%. This shows that using Domain Adaptation allowed PA classification without significant reduction in performance regardless of which wrist the accelerometer was worn on. The Amalgam approach, similarly, did not have a significant performance reduction but had a slightly worse performance than the Domain Adaptation method.

Figure 12 shows the performance of the Right Domain Adaptation approach for varying dimensions of the dimensional subspace (represented by k). As can be seen the performance begins low, at 0.44 for one dimension and then increases with the introduction of more dimensions until it reaches, 0.83, at 7 dimensions. The F1-score remains relatively stable for all values of k, until 34 dimensions, where the performance starts dropping. This shows that the choice of the k parameter does not have a great effect on the performance of the approaches if it is in the range 7-34. Hence the number of subspace dimensions was chosen to be 12 (as determined with cross-validation).

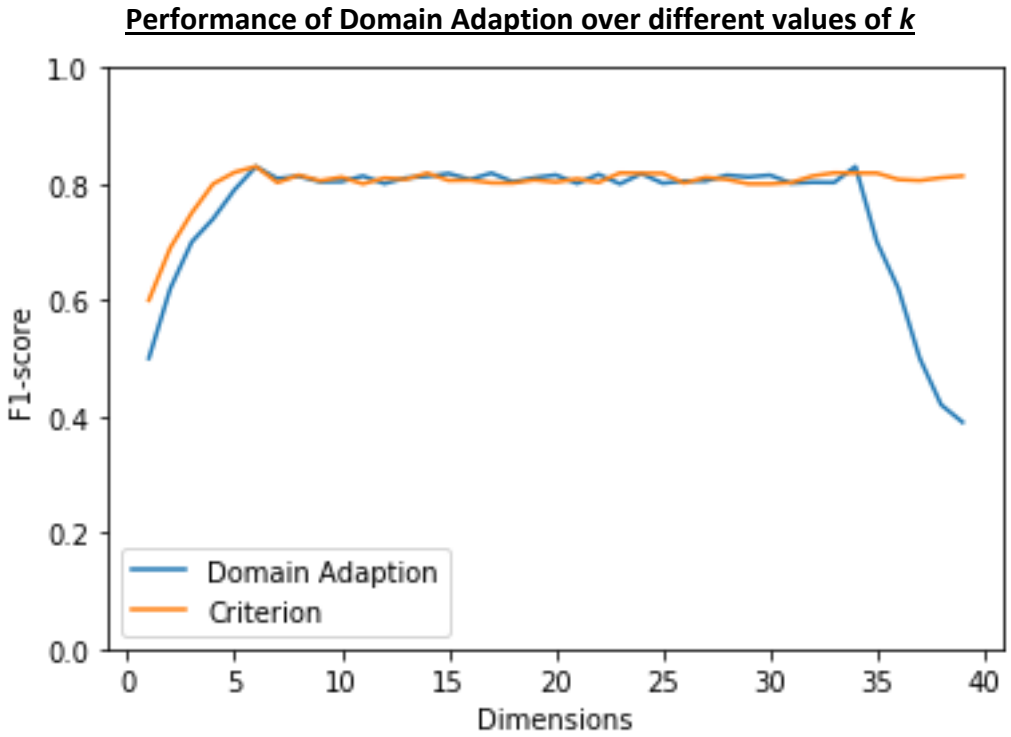

Figure 12: Performance (F1-score) of the Right Domain Adaptation approach versus subspace dimension, k.

When using Domain Adaptation approaches there was no significant difference between the scores compared to the Criterion approaches. When comparing Criterion approaches to Non-Domain Adaptation approaches (Right Non-Domain Adaptation & Left Non-Domain Adaptation in Table 14), significantly different scores were observed, highlighting that the Domain Adaptation has a significant effect on performance reduction. The significance threshold is 0.05, as 10 tests are being performed m equals 10. So, with the Bonferroni corrections a value of $p < 0.005$ implies that the performance change is significant.

Comparison, Wilcoxon Signed-Rank	P
Left Criterion, Left Domain Adaptation	0.07
Right Criterion to Right Domain Adaptation	0.5
Left Criterion to Left Non-Domain Adaptation	0.001
Right Criterion to Right Non-Domain Adaptation	0.000
Left Criterion to Left Non-Domain Adaptation	0.001
Criterion Right Domain Adaptation to Right Non-Domain Adaptation	0.003
Left Criterion, Left Not Applicable	0.8
Right Criterion to Right Not Applicable	0.6
Left Criterion to Left Amalgam	0.05
Right Criterion to Right Amalgam	0.04

Table 14: Results of a Wilcoxon Signed Rank test, comparing the performances of five different methods for statistical significance.

4.6 Discussion

This chapter set out to evaluate five approaches to achieving wrist/orientation invariance in activity classification via accelerometry and specifically examined the efficacy of Domain Adaptation through Subspace Alignment.

The results showed that without Domain Adaptation, classifying data with a classifier trained on the opposing wrist leads to an average performance drop of 12% compared to using a classifier trained and evaluated on the same wrist. However, when using Domain Adaptation, the performance drop was reduced to a statistically insignificant level (Table 14). The performance of Domain Adaptation approaches was not statistically different from the Criterion approaches, whereas there was a difference in performance between Criterion approach and approaches using Non-Domain Adaptation. Additionally, Domain Adaptation approaches outperformed the Non-Domain Adaptation approaches.

Furthermore, the Domain Adaptation approaches did not cause a reduction in the performance of the classifier. Gjoreski et al (89) found that the Amalgamated method outperformed even the Criterion approaches, although that was not the case in this study. This may have been due to the higher amount of activities with asymmetric hand movements in this work (Washing-up, Desk work, see Labelling 1, Section 3.2.1.1). The Domain Adaptation and Amalgamated approaches were equally effective at attenuating the performance reduction associated with Non-Domain Adaptation approaches. However, Domain Adaptation is preferable because it only requires acceleration data from one wrist. Additionally, it has a low cost and participant burden, as well as not requiring extensive computation.

The results of this work are consistent with the work of Montoye et al (81), who found that by making use of features that were invariant to orientation it was possible to reduce the performance drop to a negligible amount (<2%). However, their method limits the features available to the classifier. Specifically, their work only made use of the ENMO feature. As mentioned in Chapter 2, this feature is often poorly performing.

4.6.1 Strengths and Limitations

Principal Component Analysis was utilised both as a feature reduction method and as part of the Subspace Alignment algorithm. Use of Principal Component Analysis with and without using Subspace Alignment ensured that any decrease in the performance drop could be attributed to the Subspace Alignment and not the Principal Component Analysis. The number of dimensions in the projected subspace (k) can be an important parameter with respect to the overall performance. In this chapter, when k was low or high there were clear effects on the performance; however, the effect of k over a wide range between these limits was negligible. Although there are methods for automatically identifying an optimal k value, e.g. (140), it was not necessary, as the aim of this work is to

evaluate the effectiveness of Domain Adaptation, not Principal Component Analysis and performance was unaffected across a wide range of values.

A major strength of using Subspace Alignment is its simplicity. The Domain Adaptation part of the classification pipeline amounts to only a few lines of code expressing standard linear algebra operations. It is not dependent upon a classifier and existing classification schemes are easily augmented with it. The fact that it does not decrease the performance if alignment is not required, means that if there is any uncertainty about the location and orientation of an accelerometer, this technique should be used. Moreover, unlike other techniques, Subspace Alignment allows for all data to be used in the classification as opposed to just data from one wrist.

Furthermore, unlike other approaches, there are no restrictions on the potential features that can be used, linking well with methods of automatic feature extraction (141) where it may be impossible to ensure that the extracted features are rotation invariant.

Some potential weaknesses of this work are as follows: only a single data-set was used in the classification. It would have been preferable to test the use of Domain Adaptation with multiple data-sets so that the inter-protocol performance could be identified, however only the Lab-Based data-set uses accelerometers on both wrists, therefore it was only possible to compute the intra-protocol performance. While the impact of domain adaption on the inter-protocol performance was not computed, it is apparent that training and testing on different wrists will decrease performance regardless of if the data was gathered in Lab or Free-Living scenarios. Therefore, by ensuring that this is no longer an issue (due to achieving wrist invariance) this is removes one potential source of performance degradation and should be included in the classification pipeline.

The modification of the Subspace Alignment procedure to align the target data with source data obviates the expensive retraining of the classifier for each new set of target data. This greatly decreases the computational and data storage burden compared with the original Subspace Alignment algorithm and enables rapid classification of new data. However, enough target data must be collected to characterise the principal directions before the rotation that best aligns it with the source data can be identified. This means that the algorithm cannot be used for online PA classification in its present form. It is envisaged that online classification could be achieved after data is collected to characterise the rotation; a further enhancement would be to track and update the necessary rotation through a non-stationary version of Principal Component Analysis.

This work deals with the issue of training/testing on left/right hands instead of focussing on dominant and non-dominant hands. The activities performed in the activity protocol are not ones where dominance would have much effect (excepting desk work). Achieving invariance between dominant and non-dominant hands would represent a separate and potentially more complex piece of work, as Subspace Alignment requires that all the training data comes from the same wrist (as dominance is not consistent among people this would be problematic).

4.7 Conclusion

Most PA classifiers utilising wrist-worn accelerometers experience performance degradation when they are applied to data extracted from accelerometers located on a different wrist or differently oriented from which the classifier was trained. Domain Adaptation, specifically Subspace Alignment, overcomes this problem as it allows for wrist/orientation invariance. The method is simple and can easily be incorporated in existing classification schemas with no loss in performance even if Domain Adaptation is not required.

As such this method will be utilised in the final classification pipeline developed in this work.

At this moment the classification pipeline is (red text indicates the additions from this chapter):

1. Determination of data type
2. Pre-processing
3. Segmenting into windows
4. Extracting features
   - 4.1. Normalization
   - 4.2. Feature reduction
   - 4.3. Domain Adaptation
5. Creating the classifier
6. Post-processing

5. Pre and Post-Processing

5.1 Introduction

This chapter addresses another of the challenges identified in Chapter 2, that of understanding how pre and post-processing methods affect the inter and intra-protocol performance.

Pre-processing refers to preliminary processing of the acceleration data before any classification steps are carried out. This has a range of uses such as: allowing for rotational invariance (73), addressing any class imbalances in the data (75) and the removal of noise (74). Post-processing refers to the modification of the predicted labels after the classification, typically making use of the time-dependent nature of activity data in a way that classifiers of non-sequential data cannot (68). A wide variety of pre and post-processing methods are used in the literature, but to my knowledge, the work reported here is the first systematic investigation of the efficacy of pre and post-processing methods especially relating to their impact on inter and intra-protocol performance.

Chapter 2 identified two main points with respect to pre and post-processing: how all methods report higher intra-protocol performances than control methods and that the effect on inter-protocol performance is rarely discussed.

In this chapter, different methods of pre and post-processing will be examined and their ability to improve performance will be investigated. It will also look at their impact on the ability to generalise. Each of these methods will be separately incorporated into the classification pipeline and the performance of the resulting system will be compared to the Base classification pipeline identified in Chapter 3. The aim of this chapter is to identify the pre and post-processing methods that result in a statistically significant intra-protocol performance and inter-protocol performance increases. These will then be used

in the classification pipeline developed in this thesis. It is important to note that this chapter is a review of current methods and does not attempt to introduce or modify any methods.

5.1.1 Data

Both the Free-Living and the Lab-Based data, as described in 3.2 were used in this chapter. Sedentary-Standing-Active labelling (as identified in 3.2.1) was used in the Lab-Based data to ensure comparability between data-sets, this meant that both data-sets used the labels: Sedentary, Standing or Active.

5.1.2 Analysis

As Sedentary-Stand-Active Labelling was used, it was possible to compute: LabCV, FreeCV, Lab-Free and Free-Lab. LabCV and FreeCV give an indication of the intra-protocol performance (how well the classification performs on data from the same protocol), while Lab-Free and Free-Lab give an idea of inter-protocol performance (how well the classification performs on data from a different protocol). To determine if there was a significant difference between the Base classification pipeline and the pipeline making use of the pre/post-processing method, a Wilcoxon signed-rank (133) test was used to determine whether the performances were significantly different. In order to ensure a large enough sample size, the comparisons were paired on each participant's performance, instead of the average. Due to the fact the multiple hypotheses were evaluated on the same data set, the likelihood of a Type I error is increased. This was compensated for by using Bonferroni corrections. This entails testing each individual hypothesis at a significance level of $\frac{\alpha}{m}$, where α is the overall hypothesis level (in this case 0.05) and m is the number of hypotheses. A p-value under $\frac{\alpha}{m}$ indicates that the results are statistically significantly different from one another with high confidence.

5.2 Data Aggregation Via Euclidean Norm Minus One

5.2.1 Method

Aggregating the acceleration data into a single stream is beneficial for multiple reasons: it decreases the number of features that must be computed, which decreases the chance of overfitting. Additionally, some forms of aggregation allow for orientation invariance. An additional approach is to aggregate the data from the acceleration streams but retain the separate streams. This gives access to the information contained in the aggregated stream without losing information from the initial three.

In actigraphy, the main aggregate metric used is ENMO (90). This is an aggregate metric that gives an indication of the magnitude of the accelerations created by the participant. ENMO is very vulnerable to calibration errors and typically requires a device-specific calibration protocol (96). This protocol is computationally expensive but does not require labelled data so can be done without requiring extra information.

The calibration protocol attempts to identify periods of non-movement (10-second windows with standard deviations of between 10-13 milli-g). The deviation between these non-movement points and a 1-g sphere (the theoretical ideal for non-movement) is computed. A transformation to match the non-movement data with the ideal is created and used to transform the acceleration data (96).

In this work, the ENMO will be extracted from the acceleration data and the features identified in section 3.5.1 will be extracted from the ENMO time series, resulting in 12 features.

Using ENMO in addition to the measured acceleration will also be tested, based in the work of (97). This increases the number of features to 54 (4*12 + 6 cross-

correlations). Increasing features may increase the chance of overfitting (69). As such, results using feature reduction to decrease the number of features to 39 will also be shown, allowing comparability to the Base classifier which also has 39 features. Feature reduction will use Principal Component Analysis (139). Figure 13, shows an acceleration trace with ENMO (Also Figure 5).

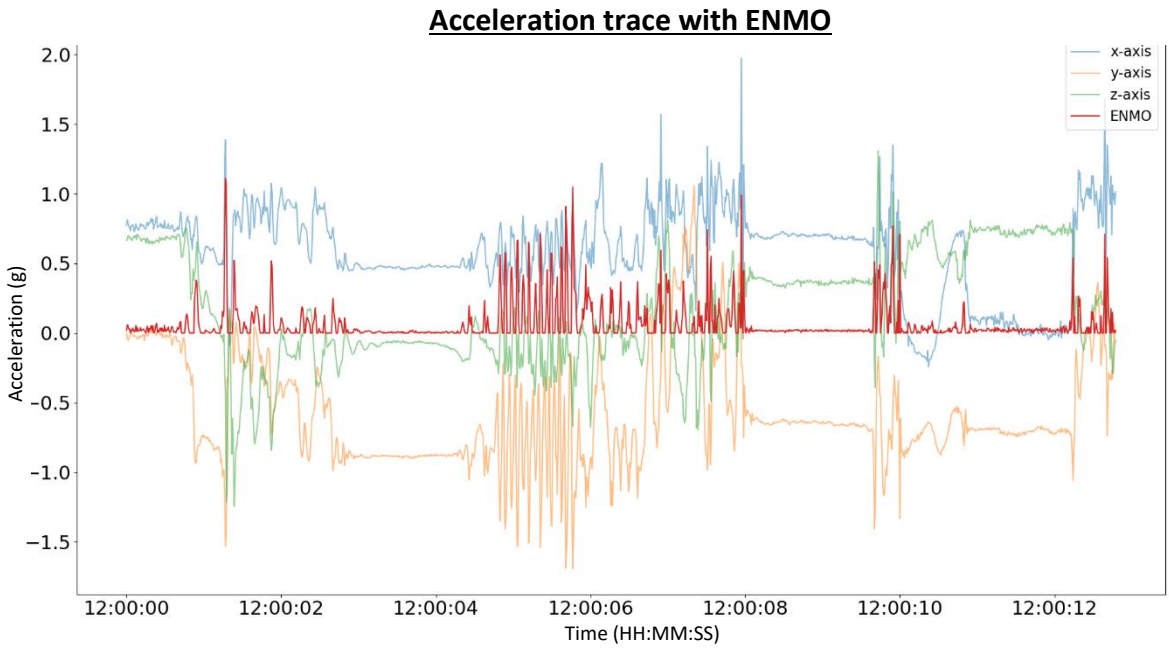

Figure 13: Acceleration trace with ENMO. The faint line representing the unfiltered data, the bold representing the filtered data.

5.2.2 Results

Using just ENMO (without the X, Y, Z data streams) significantly decreased the intra-protocol performance of the classification (0.898, 0.765 versus 0.837, 0.715) but increased the inter-protocol performance (0.352, 0.415 versus 0.399, 0.496). Using ENMO with the X, Y, Z data streams allowed for a significantly higher intra-protocol performance (0.898, 0.765 versus 0.910, 0.802) as well as a higher inter-protocol performance (0.352, 0.415 versus 0.394, 0.534). Using feature reduction outperformed the Base method but did not allow for a higher intra or inter-protocol performance than using ENMO (0.902, 0.803, 0.392, and

0.521 versus 0.910, 0.802, 0.394, and 0.534). These results are shown in Table 15.

	LabCV	*FreeCV*	*Lab-Free*	*Free-Lab*
Base	0.898 (0.103)	0.765 (0.214)	0.352 (0.132)	0.415 (0.0978)
$ENMO$	0.837* (0.098)	0.715* (0.164)	0.399* (0.152)	0.496* (0.100)
$ENMO, X, Y, Z$	0.910* (0.109)	0.802* (0.201)	0.394* (0.127)	0.534* (0.0981)
$ENMO, X, Y, Z$, with feature reduction	0.902* (0.098)	0.803* (0.174)	0.392* (0.109)	0.521* (0.102)

*Table 15: LabCV, FreeCV, Lab-Free and Free-Lab performance when using ENMO. * Indicates significant differences from the Base classification pipeline. Figures in brackets indicate standard deviations.*

5.2.3 Discussion

Using just ENMO significantly decreased the intra-protocol performance but increased the inter-protocol performance. This indicated that when using the Base classifier there may have been some degree of overfitting. Using ENMO decreased the number of features to 12 (from 39) so this may have influenced the overfitting. It may also be that ENMO allows for a greater ability to generalize because of its orientation invariance, meaning that rotated accelerometers are not affecting the performance.

Using ENMO in conjunction with the X, Y, Z acceleration streams increased both the intra and inter-protocol performance. This indicates that the inclusion of the ENMO may be the cause for the increased inter-protocol performance (and not using fewer features) as this method has more features than the Base (54 compared to 39). This hypothesis is strengthened by the result from the feature reduction, in that reducing the number of features does not increase the inter-protocol performance compared to using ENMO with all features. This suggests

that the increased number of features aren't causing overfitting, which is a potential problem when using ENMO and the X, Y, Z streams due to inputting the same information twice.

It is worth noting however that only one feature set has been used, and it may not be the case the using ENMO allows for a high performance for all feature sets.

5.2.4 Conclusion

Using ENMO in conjunction with the X, Y, Z acceleration streams allows for a greater intra and inter-protocol performance, with a minimal increase in computational overhead. As such this is a beneficial pre-processing method and will be included in the final pre-processing methods.

5.3 Filtering

One of the largest issues in activity classification is the presence of noise in the acceleration data. In this context, noise refers to one of two things: random disturbances in the signal caused by the device (referred to as observational noise), or additional information in the signal that is not useful for the activity classification (referred to as frequency noise). A 45Hz signal in the acceleration may not be generated by random disturbances but is not useful in activity classification and will disturb the acceleration values, hence is treated as noise. The aim of filtering is the removal of these sources of noise. As discussed in Chapter 2, the frequency used as the threshold for noise in the literature is 15-20Hz. This choice is rationalised by the work of Mann et al (100) who determined that 99% of measured body movements are contained within frequency components below 15Hz.

Any signal can be decomposed into a spectrum of frequencies over a continuous range according to Fourier analysis. This can be used to investigate how much of the signal is generated signals of each spectrum.

The response of a filtering method on the frequency spectrum can be analysed via a Bode magnitude plot. This is an image that shows the amplitude effect that the filtering has on signals of each frequency. This refers to how much the power of the frequencies are changed by. Figure 14, shows the ideal Bode plot for removing all noise above 15Hz. The passband (the frequencies that are to be kept, below the 15Hz) remains completely unchanged. The stopband (the frequencies that are to be removed, above 15Hz) all have their amplitude decreased to 0 (effectively removing them). Actual filtering methods rarely have Bode plots that are this perfect, either affecting the frequencies in the passband or not decreasing the amplitude of frequencies in the stopband consistently (typically ones close to the threshold).

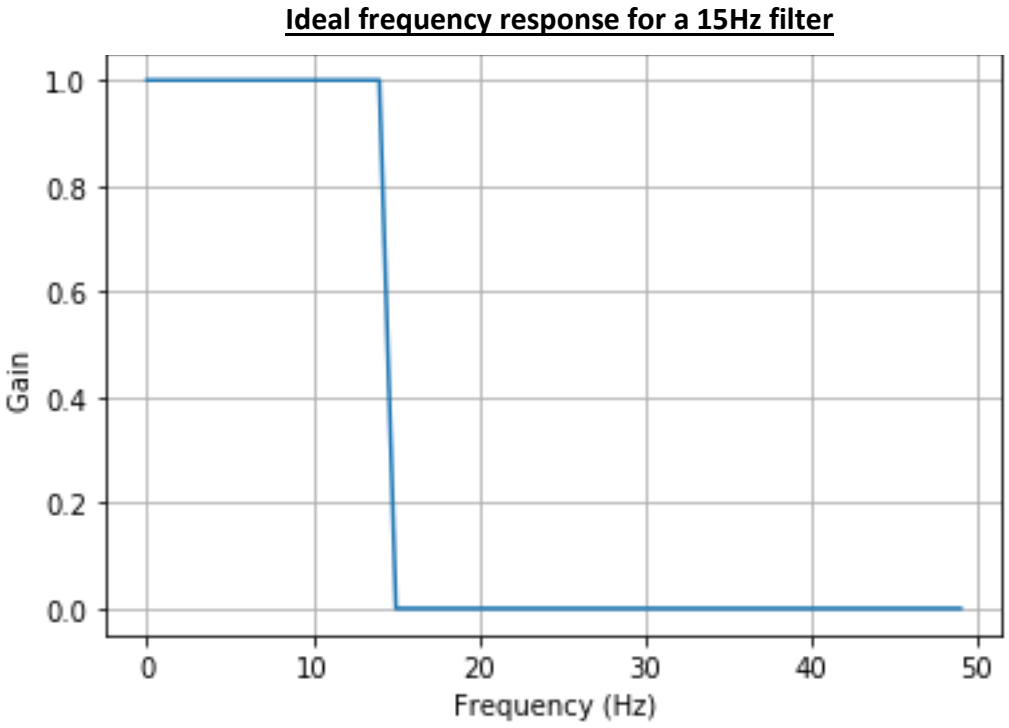

Figure 14: An ideal frequency response of a filter at 15Hz.

It is worth noting that despite Mann et al (100) identifying that 99% of body movements are contained in frequencies of 15Hz or lower, the Nyquist-Shannon theorem (101) states that for a successful reconstruction data needs to be sampled with at least twice its highest frequency, which indicates that the cut-off frequency should be between 30-40Hz, and this is the minimum required, often 3 or 4 times the highest frequency is preferable (106).

5.3.1 Moving Average

A moving average is '*a succession of averages derived from successive segments (typically of constant size and overlapping) of a series of values.*' (105). A moving average can be used to partially remove higher frequency noise from signal data. This is a filtering method that is primarily used to remove measurement noise.

The series of values, in this case, are the acceleration data streams. The size of the segments is represented by n. This value of n is generally between 2-100 for a sampling rate of 100Hz. See Equation 2 for the formulation.

Typically, the average refers to the mean or the median (103). The three values of n that will be investigated in this work are 7 (70), 11, and 15 (142). For removing measurement noise, such a filter is mathematically optimal (no other filter can do better) (106). However, for removing higher frequencies, this filtering method is poor as it has little ability to separate frequencies.

The Bode plot of using a moving average filter on 100Hz data with $n = 11$ can be seen in Figure 15. As can be seem all frequencies in the signal are affected by the filtering, not just the ones in the stopband. Additionally, the effects of the filter on the stopband are not consistent.

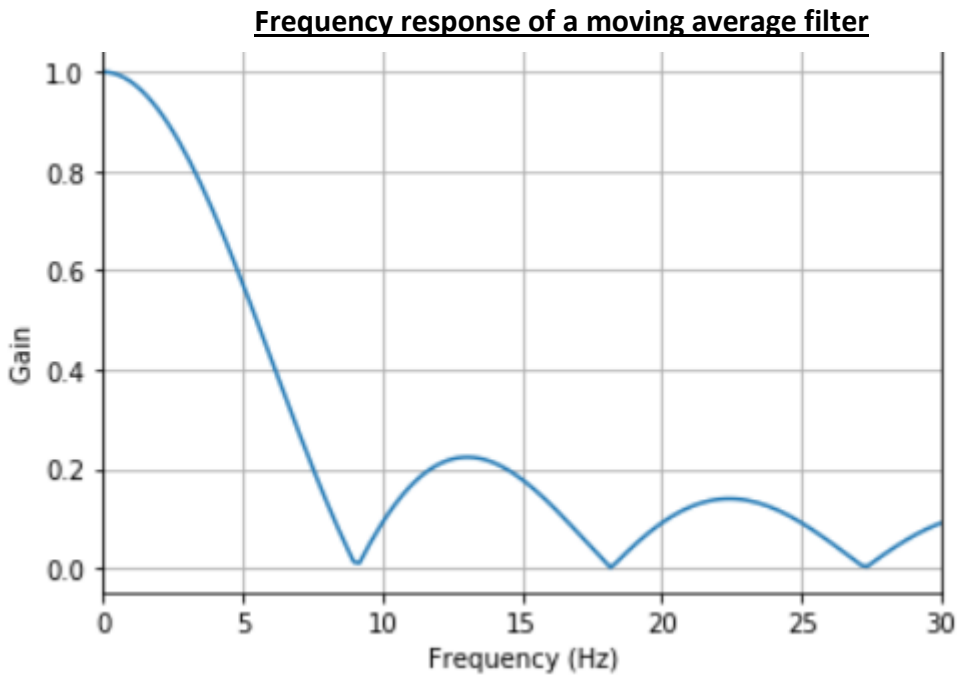

Figure 15: Frequency response of a Moving average filter, with n=11.

Figure 16 shows the effect of a median moving average with a value of 11, using a median average smooths the acceleration signal greatly. It can also be seen that individual values affect the data much less, with ‘spikes’ of acceleration data being removed.

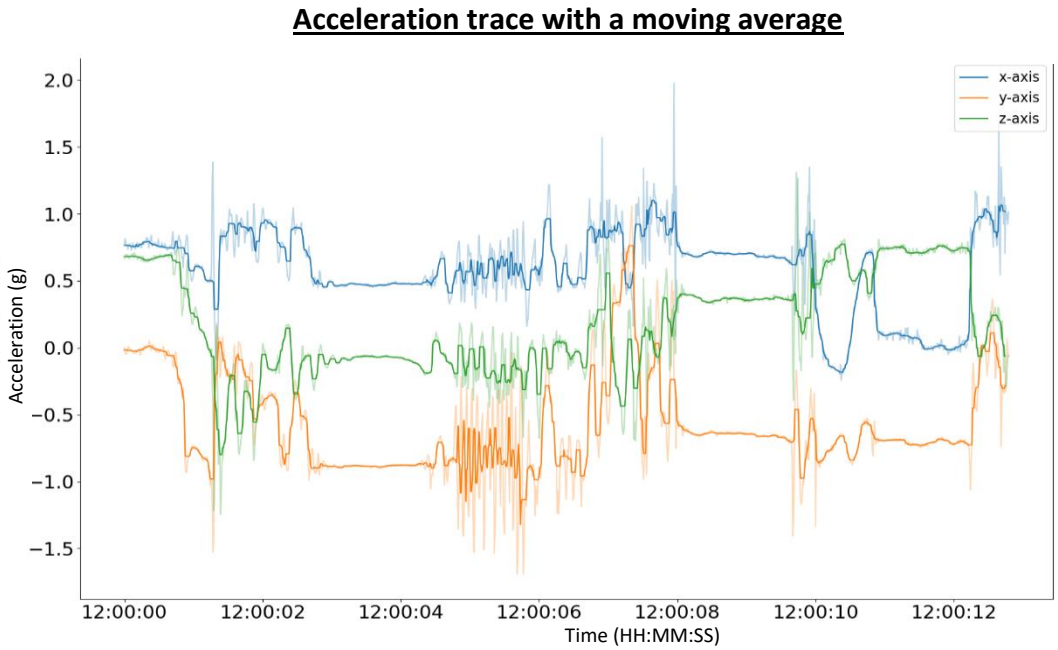

Figure 16: Acceleration trace with a moving average filter. The faint line representing the unfiltered data, the bold representing the filtered data.

5.3.1.1 Results

Filtering the data via a moving average has no significant effect on the intra-protocol performance for any value of n used in this work.

Using a median filter, $n = 7$ had a significant negative effect on the inter-protocol performance decreasing the scores to 0.338 and 0.385. Using a mean filter with $n = 7$ has no significant effects. Using a median filter with $n = 11$, caused a significant drop in the inter-protocol performance, but not the intra-protocol performance. These results are shown in Table 16.

	LabCV	*FreeCV*	*Lab-Free*	*Free-Lab*
Base	0.898 (0.103)	0.765 (0.214)	0.352 (0.132)	0.415 (0.0978)
Mean, $n = 7$	0.894 (0.101)	0.767 (0.214)	0.341 (0.124)	0.390 (0.130)
Median, $n = 7$	0.899 (0.0984)	0.761 (0.198)	0.338* (0.148)	0.385* (0.108)
Mean, $n = 11$	0.895 (0.106)	0.761 (0.231)	0.344 (0.135)	0.369* (0.0981)
Median, $n = 11$	0.896 (0.103)	0.762 (0.212)	0.337* (0.0972)	0.389* (0.100)
Mean, $n = 15$	0.891 (0.110)	0.770 (0.209)	0.321* (0.131)	0.371* (0.102)
Median, $n = 15$	0.892 (0.100)	0.765 (0.214)	0.343 (0.109)	0.362* (0.105)

*Table 16: LabCV, FreeCV, Lab-Free and Free-Lab performance when using a moving average filter, for a variety of n values. * Indicates significant differences from the Base classification pipeline. Figures in brackets indicate standard deviations.*

5.3.1.2 Discussion

Using a moving average filter had no significant effect on the intra-protocol performance for any value of n. All methods, except for using the mean average with $n = 7$, resulted in a significant drop in the inter-protocol performance.

Using $n = 11$, resulted in the worse overall performance. It may be that using larger segments ‘muddies’ the data more, allowing for a lower ability to characterize it with the features in the classification pipeline, hence a lower performance.

As can be seen in Figure 15, using $n = 11$ affects all frequencies, including those under 10Hz. From the work of Mann et al, 98% of all movements are

contained within this. This may be why filtering decreases inter-protocol performance.

5.3.1.3 Conclusion

The effects on the performance for all moving average methods is minimal. As such this method will not be used in the final pre-processing methods. It may be the case that different values of n have differing effects on intra and inter-protocol performance but that is beyond the scope of this thesis.

5.3.2 Butterworth Filtering

5.3.2.1 Method

A Butterworth filter is a smoothing method that allows for the removal of high-frequency data from a time series. Specifically, a Butterworth filter intends to reduce the effect of the stopband frequencies without modifying the passband frequencies (143).

The frequency response of the Butterworth filter is maximally flat (i.e. has no ripples) in the passband and rolls off towards zero in the stopband. Other methods of signal filtering may have faster roll-offs (the response of the stopband decreases faster), but this comes with ripples in the passband that may artificially increase the response of frequencies in the passband. As discussed in Chapter 2, the frequency used as the threshold in the literature is 15-20Hz. However, due to the slow roll-off of Butterworth filters, this will mean that much of the components from 21Hz, will still be available. The Bode plot of a Butterworth filter, filtering at 20Hz can be seen in Figure 17.

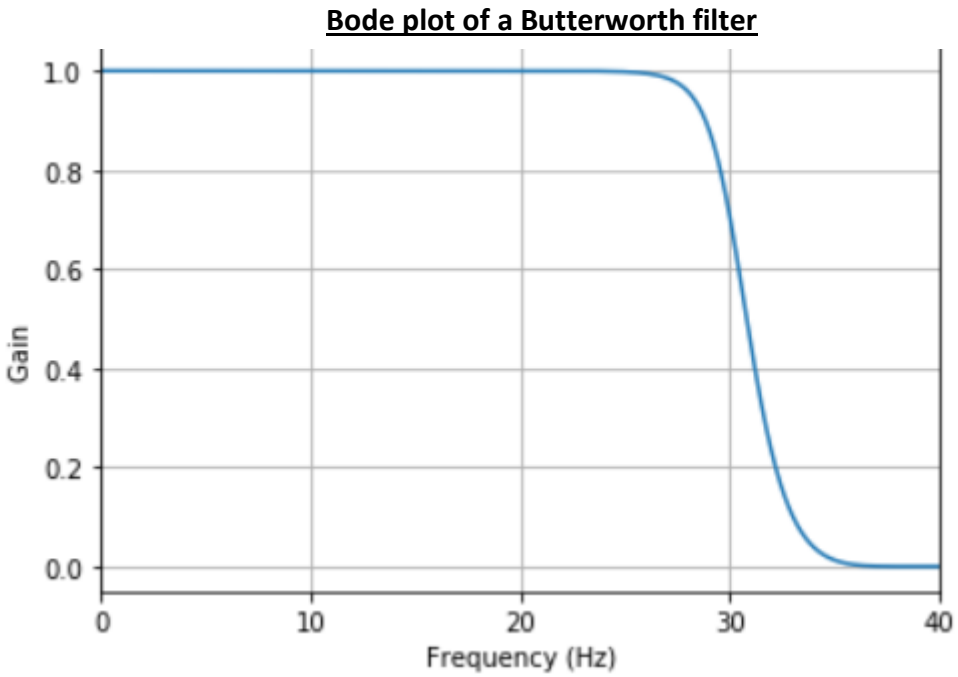

Figure 17: Bode plot of Butterworth filter at 20 Hz.

The equation used to compute a Butterworth filter is:

$$G(\omega) = \frac{1}{\sqrt{1 + \zeta^2 \left(\frac{\omega}{\omega_c}\right)^{2o}}}$$

Equation 4: Butterworth Filter.

With ω is the frequency, o is the order of the filter, ζ is the maximum passband gain and ω_c is the cut-off frequency.

The steepness of the slope in the stopband is affected by the order of the Butterworth filter. In activity classification, a second-order filter is generally used, with a cut-off frequency of 20Hz (70,104).

Figure 18, shows the effect of applying a second order Butterworth filter (cut-off 20Hz) to acceleration data. As can be seen, there is very little change between the two acceleration traces. The only identifiable changes are in the dark trace (after processing) there is a reduction of some of the ‘spikes’ of acceleration

that occur for a single measurement compared to the light trace (before processing).

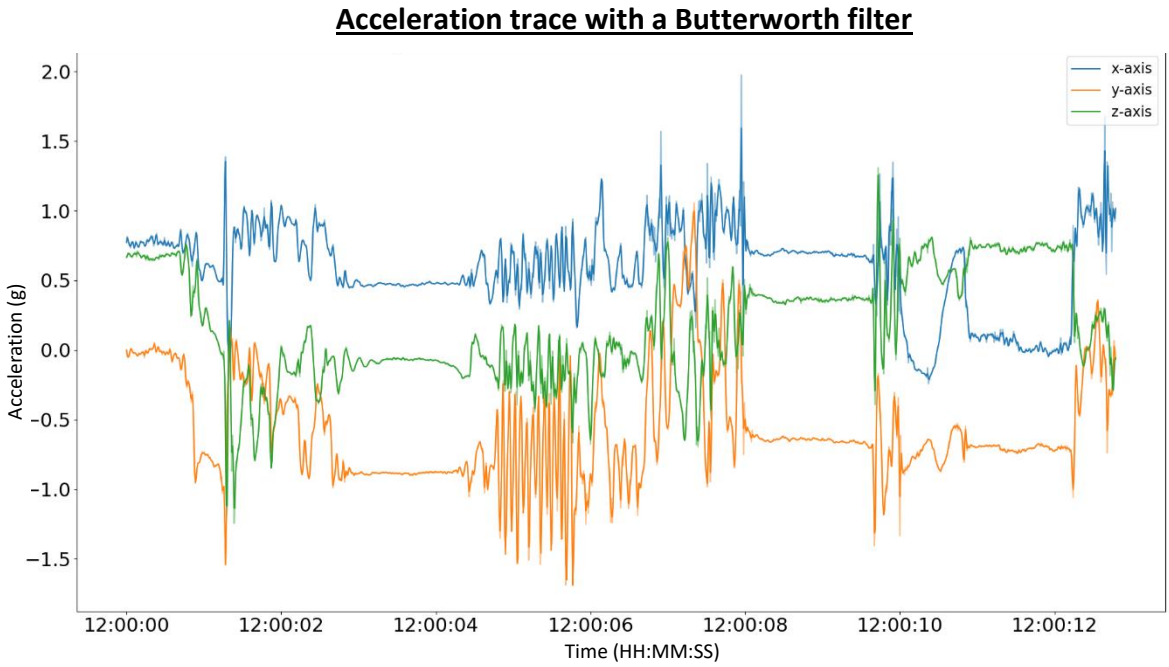

Figure 18: Acceleration trace with a Butterworth filter of 20Hz. The faint line representing the unfiltered data, the bold representing the filtered data.

5.3.2.2 Results

Using the Butterworth filter did not significantly decrease the intra-protocol performance but did significantly degrade the inter-protocol performance (0.352, 0.415 versus 0.342, 0.382). These results are shown in Table 17.

	LabCV	*FreeCV*	*Lab-Free*	*Free-Lab*
Base	0.898 (0.103)	0.765 (0.214)	0.352 (0.132)	0.415 (0.0978)
After filtering	0.894 (0.105)	0.761 (0.187)	0.342 (0.126)	0.382* (0.109)

*Table 17: LabCV, FreeCV, Lab-Free and Free-Lab performance when using Butterworth filter, compared to the Base classifier. * Indicates significant differences from the Base classification pipeline. Figures in brackets indicate standard deviations.*

5.3.2.3 Discussion

Using the Butterworth filter did not significantly impact intra-protocol performance. It did significantly alter the inter-protocol performance (with the Free-Lab scores being significantly different). This may be because different data-sets have different frequency distributions. This lack of an intra-protocol performance change may be because the data-sets used do not have any high-frequency activities (cycling, driving), therefore there is little high-frequency data for the filter to remove.

In future work, it may be worth investigating the effects of using different orders of filter or using different frequency thresholds. Additionally, it may be that the features used are not affected by high-frequency data and therefore the filtering has no great effect. As such, despite the common usage of this method in the literature (88,104), the effects on the overall performance appear to be minimal.

It is also worth noting that studies tend to use cut-off frequencies of 15-20Hz, citing Mann et al (100), however as noted above the Nyquist-Shannon theory (101) states this cut-off is too low. An extension to this work would be to use higher cut-off frequencies.

5.3.2.4 Conclusion

As there are significantly detrimental effects on the overall performance, this method will not be used in the final pre-processing methods.

5.4 Orientation Invariance

5.4.1 Method

The previous chapter dealt with orientation invariance in the context of incorrect wrist placement but not in the correction of the orientation of the device upon the wrist, such as wearing it upside down.

In most of the work on activity classification it is assumed that the sensors are always placed correctly and remain in a fixed orientation. This assumption is unlikely to hold, especially in wrist-based data. Large population studies typically require that the sensor is placed by the participant themselves; this can lead to incorrect orientation or a loose fit that then allows for device slippage. An incorrect orientation can dramatically decrease the performance of a classification model (up to 21.2% drop (107)).

A way to combat this performance decrease is to transform the acceleration data so that it becomes orientation invariant. The majority of these transformations lose information (typically orientation-based information) by aggregating the acceleration streams into one (90).

A method that allows for orientation invariance without this data loss is Heuristic Orientation-Invariant Transformation. This entails transforming the 3-D acceleration data into 9-D orientation invariant data. These data streams are provably invariant to rotation (107). The first 6 data streams are simply orientation invariant streams that are well representative of the data, while data streams 7-9 represent information about the rotational movements in 3-D space. After the transformation of the acceleration data into these 9 orientation invariant streams, features are extracted for them. Due to retaining this orientation-based information, this method is suggested to be able to outperform using ENMO (107).

5.4.2 Heuristic Orientation-Invariant Transformation Data Points

Let $W_S = \big((x_1, y_1, z_1), (x_2, y_2, z_2), \dots, (x_S, y_S, z_S)\big)$, be the S length vector of the X, Y, Z data. Typically S is equal to the window size. The first and second-order time differences are defined as $\Delta W_S = W_{S+1} - W_S$, and $\nabla W_S = \Delta W_{S+1} - \Delta W_S$.

$\|W_S\|$ represents the Euclidean norm of W_S, π represents the angle between two vectors. The 9-dimension data extracted when using this method is comprised of:

1. $\|W_S\|$
2. $\|W_S\|$
3. $\|\nabla W_S\|$
4. $\pi(W_S, W_{S+1})$
5. $\pi(\Delta W_S, \Delta W_{S+1})$
6. $\pi(\nabla W_S, \nabla W_{S+1})$
7. $\pi((W_S \times W_{S+1}), (W_{S+1} \times W_{S+2}))$
8. $\pi((\Delta W_S \times \Delta W_{S+1}), (\Delta W_{S+1} \times \Delta W_{S+2}))$
9. $\pi((\nabla W_s \times \nabla W_{S+1}), (\nabla W_{S+1} \times \nabla W_{S+2}))$

The classification pipeline in this work uses 12 features for each data stream, as well as 1 feature for each pair of data streams, thus resulting in 144 (12*9 + 36) features. Transforming the 3 acceleration streams into the 9 data streams allows for orientation invariance, but also increases the number of features that are created. Due to this, a feature reduction method is used to decrease the number of features back down to 39 allowing for a more direct comparison to the Base classifier (which also has 39 features). The method used is Principal Component Analysis (139).

Unlike the other methods described in this work, Heuristic Orientation-Invariant Transformation is not a method to improve performance in all cases, but instead a tool to mitigate a performance decrease when the orientation of the accelerometer is suspect. As such, this method was also evaluated when the

axes of the accelerometer were scrambled (artificially changing the orientation of the sensor). This was be compared to using an orientation invariant feature set, derived from ENMO.

5.4.3 Results

Using the Heuristic Orientation-Invariant Transformation features (shown in Table 18) decreases the intra-protocol performance (0.898, 0.765 versus 0.823, 0.726) as well as the inter-protocol performance (0.352, 0.415 versus 0.312, 0.350). In all cases these changes are significant. When using Principal Component Analysis to reduce the number of features to 39, the intra-protocol performance decreases further (0.823, 0.726 versus 0.812, 0.725) but allows for a greater inter-protocol performance (0.312, 0.350 versus 0.334, 0.401). However, even with Principal Component Analysis, using the Heuristic Orientation-Invariant Transformation features are significantly worse than the Base classification.

	LabCV	*FreeCV*	*Lab-Free*	*Free-Lab*
Base	0.898 (0.103)	0.765 (0.214)	0.352 (0.132)	0.415 (0.0978)
Heuristic Orientation-Invariant Transformation	0.823* (0.150)	0.726* (0.261)	0.312* (0.201)	0.350* (0.103)
Heuristic Orientation-Invariant Transformation (with Principal component analysis)	0.812* (0.121)	0.725* (0.193)	0.334* (0.153)	0.401* (0.108)

*Table 18:LabCV, FreeCV, Lab-Free and Free-Lab performance when using Heuristic Orientation Invariant Transformation, compared to the Base classifier. * Indicates the scores which are statistically significantly different from the Base classifier. Figures in brackets indicate standard deviations.*

When investigating the intra-protocol performance on data with the axis labels shuffled (artificially changing the orientation of the sensor, Table 19) using the Heuristic Orientation-Invariant Transformation features dramatically increases the intra-protocol performance (0.436, 0.512 versus 0.812, 0.725) as well as

increasing the inter-protocol performance (0.306, 0.401 versus 0.334, 0.401). However, Heuristic Orientation-Invariant Transformation features do not outperform ENMO (another orientation invariant feature set).

The ENMO scores are statistically different from the Heuristic Orientation-Invariant Transformation scores, indicating the ENMO is the most effective method of achieving orientation invariance.

	LabCV	*FreeCV*	*Lab-Free*	*Free-Lab*
Base	0.436 (0.254)	0.512 (0.164)	0.306 (0.143)	0.401 (0.109)
Heuristic Orientation-Invariant Transformation	0.812* (0.121)	0.725* (0.193)	0.334* (0.153)	0.401* (0.108)
$ENMO$	0.837*- (0.098)	0.715*- (0.164)	0.399*- (0.152)	0.496*- (0.100)

*Table 19: LabCV, FreeCV, Lab-Free and Free-Lab performance when using Heuristic Orientation-Invariant Transformation on axes scrambled data, compared to the Base classifier. * Indicates statistically significant differences from the Base classifier. - Indicates statistically significant differences from Heuristic orientation-invariant transformation. Figures in brackets indicate standard deviations.*

5.4.4 Discussion

Heuristic Orientation-Invariant Transformation features decrease the intra-protocol performance and the inter-protocol performance when the axes are not shuffled. This result is in agreement with Yurtman et al (73,107), who found an average intra-protocol performance decrease of 21.2% compared to no transformation. When the axes are shuffled, Heuristic Orientation-Invariant Transformation allows for greater intra-protocol performance and inter-protocol performance. This suggests that in the case where the orientation of the sensor is suspect, such a transformation may be beneficial to use, but not if the orientation is known to be consistent. However, ENMO outperforms using

Heuristic Orientation-Invariant Transformation, again in agreement with the work of Yurtman et al.

5.4.5 Conclusion

When the axes are not shuffled Heuristic Orientation-Invariant Transformation performance is worse than the Base classifier, hence there does not appear to be a reason to use this method in the final pre-processing methods. When the axes are shuffled, ENMO was more beneficial than using Heuristic Orientation-Invariant Transformation. Hence ENMO will be used in the final classification, not Heuristic Orientation-Invariant Transformation.

5.5 Inclination Correction

Another method to obtain some degree of orientation invariance is inclination correction (108). Instead of trying to achieve full invariance to orientation, this method attempts to mitigate the performance reduction caused by the accelerometer shifting on the wrist. Unlike a full rotation, this will not invert any axes but may cause a slight performance drop.

5.5.1 Method

Inclination correction works by ‘centring’ the acceleration streams. This is done by removing the average value obtained when standing still for 5s before starting the activity. This has the effect of transforming the data so that it may have been obtained from a correctly inclined accelerometer.

The plot of transformed acceleration can be seen in Figure 19. As can be seen, the inclination correction has the effect of shifting the acceleration values.

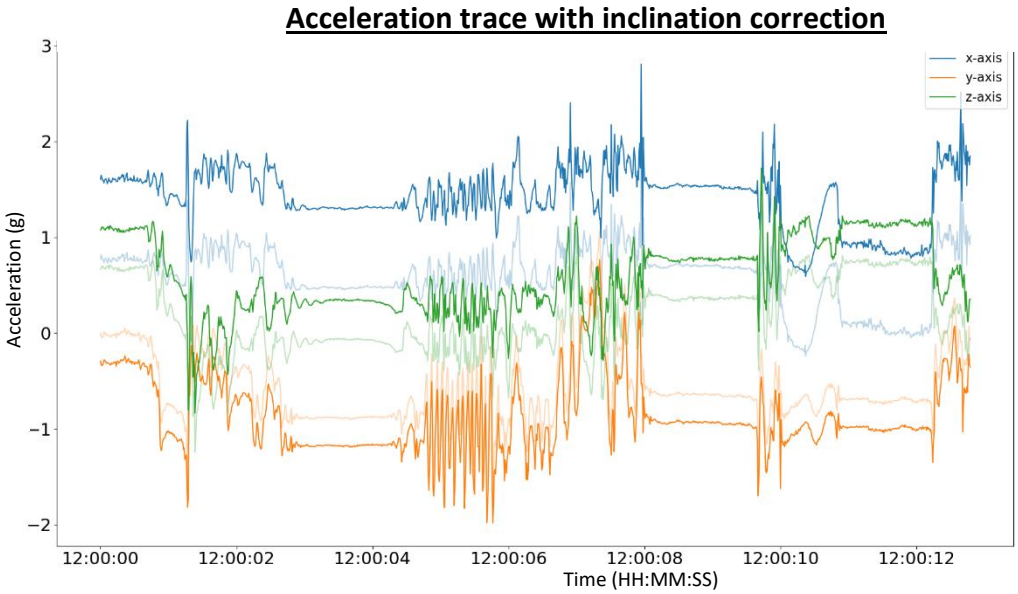

Figure 19: Acceleration trace with inclination correction. The faint line representing the unfiltered data, the bold representing the filtered data.

5.5.2 Results

Using inclination correction did not significantly affect intra-protocol performance. The inter-protocol performance was significantly affected, with the Base outperforming the inclination correction (0.352, 0.415 versus 0.336, 0.410). Only the reduction in the Lab-Free scores was significant. Shown in Table 20.

	LabCV	*FreeCV*	*Lab-Free*	*Free-Lab*
Base	0.898 (0.103)	0.765 (0.214)	0.352 (0.132)	0.415 (0.0978)
After filtering	0.893 (0.121)	0.774 (0.198)	0.336* (0.143)	0.410 (0.100)

*Table 20: LabCV, FreeCV, Lab-Free and Free-Lab performance when using inclination correction, compared to the Base classifier. * Indicates significant differences from the Base classification pipeline. Figures in brackets indicate standard deviations.*

5.5.3 Discussion

Using inclination correction showed no significant effect on intra-protocol performance. This is in agreement with the work of Fida et al (108), who also

identified that inclination correction showed no significant intra-protocol performance increases. Much like Heuristic Orientation-Invariant Transformations, inclination correction is a method for when the orientation of the accelerometer is suspect, and it is believed that the device may have slipped slightly. In the case of Lab-Based data, the researchers attached the accelerometer and observed the protocol, hence could ensure that no slippage occurred. In the case of the Free-Living protocol, the researchers attached the devices but could not observe to ensure that no slippage occurred. It is important to note however, that inclination correction does not significantly decrease the intra-protocol performance, and may prove useful for participant-mounted devices, although this assumption has not been tested.

5.5.4 Conclusion

Due to the lack of a performance increase, this method will not be included in the final pre-processing methods.

5.6 Structure-Preserving Oversampling

5.6.1 Method

As discussed in Chapter 2, imbalanced classes may greatly decrease the classification performance (69). Oversampling is often used to mitigate this, creating synthetic data from the minority classes. Due to the high level of the interrelatedness of time series data, a specific method must be created in order to allow for oversampling without causing issues, such as weakening correlation structures for time series data. An oversampling method for time series data is ‘Structure-Preserving Oversampling’ (75). This is an oversampling technique that generates synthetic samples while preserving the covariance structure of the data (therefore not weakening the correlation structures).

First, the covariance matrix of the time series of raw acceleration from the minority class is computed, the eigenvector decomposition of this covariance matrix is then identified. The eigenvectors are then split into ‘reliable’ and ‘unreliable’ subspaces. Eigenvectors are reliable if the eigenvalues are consistent among subsamples of the data. The unreliable eigenvectors are then regularized to ensure smoothness of the eigen spectrum. These newly regularised eigenvectors are used to transform random Gaussian data, thus creating synthetic data that maintains the covariance structure of the data. An additional step to ensure that the random synthetic data created does not decrease the separation of the classes is to not allow the addition of synthetic data that reduces the minimum distance between two data points in differing classes.

5.6.1.1 Structure-Preserving Oversampling Algorithm:

Inputs: $Class_{min}$ = minority class, $Class_{maj}$ = majority class

1. Generate the covariance matrix of $Class_{min}$ to obtain Cov_{min}.

2. Compute the eigenvalue decomposition of Cov_{min} to obtain $Decomp = e_1, e_2, \dots, e_N$

3. Divide into 'reliable' and 'unreliable' eigenvectors such that $e_1, e_2, \dots, e_B$ are reliable and $e_{B+1}, e_{B+2}, \dots, e_N$ are unreliable.

B (the index of the last reliable eigenvector) is found by using cross-validation for progressively increasing values potential values of B. One partition computes eigenvalues, then projects the other partition. This is compared to the true projection. When the accuracy drops, the optimal value of B has been found.

4. New regularised eigenvectors are then created using the formulation:

$$\widehat{e_j}\begin{cases} e_j & j \le B \\ \gamma(j) & j > B \end{cases}, \quad \gamma(j) = \frac{e_1 e_B (B-1)}{e_1 - e_B} \div \left(\frac{B e_b - e_1}{e_1 - e_b} + j\right)$$

5. A new vector (V) is generated from a Gaussian distribution $N(0,1)$ and then transformed into a synthetic member of class $Class_{min}$ by transforming through V

6. V is checked to make sure that it doesn't decrease the margin, if not then $Class_{min} = Class_{min} \cup V$.

7. Steps 5-6 are repeated until $|Class_{min}| = |Class_{maj}|$.

5.6.2 Results

Using Structure-Preserving Oversampling had no significant effect on the intra-protocol performance (0.898, 0.765 versus 0.897, 0.765) but increased the inter-protocol performance (0.352, 0.415 versus 0.473, 0.498) significantly. This is shown in Table 21.

	LabCV	*FreeCV*	*Lab-Free*	*Free-Lab*
Base	0.898 (0.103)	0.765 (0.214)	0.352 (0.132)	0.415 (0.0978)
Structure Preserving Oversampling	0.897 (0.143)	0.765 (0.213)	0.473* (0.106)	0.498* (0.0843)

*Table 21: LabCV, FreeCV, Lab-Free and Free-Lab performance when using Structure-Preserving Oversampling compared to the Base classifier. * Indicates significant differences from the Base classification pipeline. Figures in brackets indicate standard deviations.*

5.6.3 Discussion

Chapter 3 identified the imbalances in both classification data-sets, showing that both data-sets are highly imbalanced, with a large skew towards sedentary activities. Although it is likely that this imbalance accurately reflects the true distribution of activities (with a high amount of sedentary activities and low amount of active PA), it is still beneficial for data-sets to be balanced for the purposes of classification. Despite this, including Structure-Preserving Oversampling in the classification pipeline does not cause a significant intra-protocol performance increase. It is thought that this lack of an intra-protocol performance increase is for two reasons. Random Forests, as used in this work, are relatively robust in handling imbalanced data (135) and the data imbalances are the same for both the training and the testing sets (when investigating intra-protocol performance). When investigating the effect on the inter-protocol performance of including Structure-Preserving Oversampling in the classification pipeline it is clear that it is beneficial, with a performance increase of up to 0.12. This is in agreement with the work of Cao et al who note a 5.3% increase in performance (75).

5.6.4 Conclusion

As this method does improve the inter-protocol performance, at minimal cost to the intra-protocol performance, this method will be included in the final pre-processing methods.

5.7 Smoothing

5.7.1 Method

Smoothing is a post-processing method that refers to applying a modal filter to the predicted labels, such that each predicted label is replaced by the most common label of the n closest labels (inclusive of itself) (75,124). Typically, this method increases performance, although by variable amounts. Multiple studies have used this method to increase performance, however, a consistent value of n has not been used. A simple modal filter has been used in all studies. The two values of n that will be investigated are 3 and 11.

5.7.2 Results

Using a smoother with $n = 11$ allowed for the greatest average increase in intra-protocol performance (0.902, 0.775) and inter-protocol performance (0.356, 0.419). These differences were significant only in the inter-protocol performance. In most cases the smoothing resulted in non-significant changes, as shown in Table 22.

	LabCV	*FreeCV*	*Lab-Free*	*Free-Lab*
Base	0.898 (0.103)	0.765 (0.214)	0.352 (0.132)	0.415 (0.0978)
$n = 3$	0.905 (0.106)	0.770 (0.209)	0.354 (0.135)	0.415 (0.0987)
$n = 11$	0.902 (0.100)	0.775 (0.231)	0.356* (0.128)	0.419* (0.0953)

*Table 22: LabCV, FreeCV, Lab-Free and Free-Lab performance when using a range of smoothing filters, compared to the Base classifier. * Indicates significant differences from the Base classification pipeline. Figures in brackets indicate standard deviations.*

5.7.3 Discussion

Smoothing allowed for higher inter-protocol performance for $n = 11$, but not $n = 3$. This is most likely because it meant erroneous classifications has less impact.

5.7.4 Conclusion

Using smoothing with a $n = 11$ significantly improves inter-protocol performance with no cost to intra-protocol performance. As such, this is a beneficial post-processing method and will be included in the final post-processing.

5.8 Hidden Markov Models

5.8.1 Methods

A Hidden Markov model can be used for removing isolated misclassifications in the predicted labels. Hidden Markov models use the prevalence of classes, as well as transition probabilities in the training data to 'smooth' the predicted classifications. As discussed in Chapter 2, Hidden Markov models describe the creation of an observable time series, making use of internal factors (hidden states) that are not directly observable. The predicted activities time series are the observable symbols, and a hypothetical time series of 'correct' classifications are the hidden states. Combing these with the transition probabilities and prevalence of different classes in the training data, the Hidden

Markov models can 'predict' the sequence of hidden states that generated the predicted activities. This sequence of hidden states represents a smoothed time series of activity classes that should more closely match the true activities performed (69). In this work, the transition probabilities and class prevalence's are identified in Chapter 3.

5.8.2 Results

Using a Hidden Markov models had a positive effect on the intra-protocol performance (0.898, 0.765 versus 0.925, 0.801) it also increased the inter-protocol performance in the Free-Lab data (0.352, 0.415 versus 0.353, 0.591). This is shown in Table 23. All scores were significantly different from the Base classifier except for the Lab-Free score.

	LabCV	*FreeCV*	*Lab-Free*	*Free-Lab*
Base	0.898 (0.103)	0.765 (0.214)	0.352 (0.132)	0.415 (0.0978)
Hidden Markov model	0.925* (0.0931)	0.801* (0.104)	0.353 (0.251)	0.591* (0.200)

*Table 23: LabCV, FreeCV, Lab-Free and Free-Lab performance when using a Hidden Markov model smoother, compared to the Base classifier. * Indicates significant differences from the Base classification pipeline. Figures in brackets indicate standard deviations.*

5.8.3 Discussion

The Hidden Markov model increased the intra-protocol performance, this was expected and agrees with the work of (68). Hidden Markov models will consistently increase the intra-protocol performance when using data from the same protocol (LabCV and FreeCV). This is because they make use of implicit information about the activity protocol: the average length of each activity and which activities are done next. If the activity protocol remains consistent among participants (typical for a Lab study) then this intra-protocol performance increase will be seen.

Despite the fact that the activity protocol changes, thereby changing the transition probabilities, using Hidden Markov models improves the inter-protocol performance. This is likely because they still have a smoothing effect, removing brief incorrect events, even if they cannot achieve the optimal smoothing; due to the changes in probability.

5.8.4 Conclusion

Using the Hidden Markov model smother improves the intra-protocol performance and the inter-protocol performance. As such, this is a beneficial post processing method and will therefore be included in the final post-processing methods.

5.9 Participant Adaptation via Iterative Relearning

5.9.1 Method

This is a form of post-processing that attempts to adapt the classifier to an individual participant via iteratively retraining it on the participant specific data (125).

Specifically, Participant Adaptation via Iterative Relearning attempts to make use of the Base classifier and then adapts it to specific participants via self-training.

Self-training is a form of semi-supervised learning. First a supervised learning algorithm is trained on the labelled data only (this is the standard classification training step). This classifier is then used to predict labels for unlabelled data. The most confident of these labelling's are then used to retrain the classifier (replacing the original data) allowing for a greater amount of training data, which is specific to the participant. As the unlabelled data used in the self-training comes from the participant themselves, the new classification model is more

closely aligned to their data and becomes more specialised (increasing the intra-protocol performance on the data, while decreasing inter-protocol performance). In essence this method is attempting to alter the problem from inter-subject-inter-protocol performance to intra-subject-intra-protocol performance which typically has a higher performance.

The Participant Adaptation via Iterative Relearning algorithm for a single participant H is:

1. Create general classifier (person non-specific) using labelled data from training set
2. Classify the unlabelled data from H
3. Identify the most confident classification (in this work 5%), and create a classifier based on this
4. Repeat step 3 until no improvement, or for a set number of iterations (in this case 5)

5.9.2 Results

Using Participant Adaptation via Iterative Relearning had a positive effect on the intra-protocol performance (0.898, 0.765 versus 0.916, 0.823) it also increased the inter-protocol performance in the Free-Lab data and the Lab-Free data (0.352, 0.415 versus 0.532, 0.512). Results presented in Table 24. All differences in scores were statistically significant.

	LabCV	*FreeCV*	*Lab-Free*	*Free-Lab*
Base	0.898 (0.103)	0.765 (0.214)	0.352 (0.132)	0.415 (0.0978)
Participant Adaptation via Iterative Relearning	0.916* (0.119)	0.823* (0.198)	0.532* (0.102)	0.512* (0.139)

*Table 24: LabCV, FreeCV, Lab-Free and Free-Lab performance when using Participant Adaptation via Iterative Relearning, compared to the Base classifier. * Indicates significant differences from the Base classification pipeline. Figures in brackets indicate standard deviations.*

5.9.3 Discussion

For all cases Participant Adaptation via Iterative Relearning improved the intra-protocol performance. This is in agreement with the work of (125), who reported an intra-protocol performance increase ranging from 5.2% to 28% after one iteration and after four iterations the minimal increase was 16% (125). The reported intra-protocol performance increase in (125) is similar to the increase shown here.

The increased intra-protocol performance may be a factor of the increased amount of training data, or the specific participant adaption. As FreeCV improves more than LabCV, and FreeCV has substantially more data than LabCV, this suggests that the intra-protocol performance increase is a product of the participant adaption rather than the increased number of training samples.

5.9.4 Conclusion

Using Participant Adaptation via Iterative Relearning improves the intra-protocol performance and the inter-protocol performance. As such this is a beneficial post processing method and will be included in the final post-processing methods.

5.10 Combination of All Methods

5.10.1 Methods

The pre-processing methods that improved performance in this work were:

- Using Structure Preserving Oversampling to rebalance the classes
- Using ENMO with X, Y, Z acceleration streams

The post-processing methods that improved performance in this work were:

- Using Participant Adaptation via Iterative Relearning

- Using a Hidden Markov model
- Using a smoother with $n = 11$

These methods were combined into a single pre/post-processing approach and the effect on the intra and inter-protocol performance was tested. This method is referred to as Pre-Post-Combined approach.

The combination took the following order.

1. Determination of data type
2. Structure Preserving Over-sampling was used to fix any class imbalances
3. Windowing: Using 12.8-second windows
4. Feature extraction: Using the statistical features on both the X, Y, Z streams and the ENMO
5. Building the classification model: Using a Random Forest classifier with 50 separate trees.
6. The classifier then underwent Participant Adaption Via Iterative Re-training, identifying the most confident classification (in this work 5%), and re-training the classifier based on this.
7. After the final classification, the labels were then smoothed using the HMM, followed by the smoother with $n = 11$.

5.10.2 Results

Using the Pre-Post-Combined allowed for the intra-protocol performance to be greater than the Base classifier (0.898, 0.765 versus 0.912, 0.817) and the inter-protocol performance was also higher (0.352, 0.415 versus 0.485, 0.556). The results of the Pre-Post-Combined and all methods used within it are presented in Table 25.

	LabCV	*FreeCV*	*Lab-Free*	*Free-Lab*
Base	0.898 (0.103)	0.765 (0.214)	0.352 (0.132)	0.415 (0.0978)
$ENMO, X, Y, Z$	0.910 (0.109)	0.802 (0.201)	0.394 (0.127)	0.534 (0.0981)
Structure Preserving Oversampling	0.897 (0.143)	0.765 (0.213)	0.473 (0.106)	0.498 (0.0843)
Participant Adaptation via Iterative Relearning	0.916 (0.119)	0.823 (0.198)	0.532 (0.102)	0.512 (0.139)
Hidden Markov Modelling	0.925 (0.0931)	0.801 (0.104)	0.353 (0.251)	0.591 (0.200)
Median, $n = 11$	0.902 (0.100)	0.775 (0.231)	0.356 (0.128)	0.419 (0.0953)
Pre-Post-Combined	0.912* (0.127)	0.817* (0.199)	0.485* (0.154)	0.556* (0.109)

*Table 25: LabCV, FreeCV, Lab-Free and Free-Lab performance when using final pre and post-processing method, compared to the Base classifier. * Indicates significant differences from the Base classification pipeline and the Pre-Post-Combined method. Figures in brackets indicate standard deviations.*

5.10.3 Discussion

The Pre-Post-Combined approach significantly increases both intra and inter-protocol performance compared to the Base classifier. While the average performance of the Participant Adaptation via Iterative Relearning is higher that that of the Pre-Post-Combined approach, the difference is not significant. Therefore, the choice was made to make use of the Pre-Post-Combined approach, as this was not worse than Participant Adaptation via Iterative Relearning.

It is worth noting a potential limitation in this method of combining the pre-processing methods. Due to the order in which they are performed, they do not interact with each other. For instance, combining the Participant Adaption Via

Iterative Re-training and the smoothing would be a potential extension, instead of re-training the classifier on the most confident classifications, the smoothing could be used to determine windows that were incorrectly classified (windows where the label was changed due to smoothing), thus training the classifier on data where it is likely to misclassify.

5.10.4 Conclusion

Pre-Post-Combined approach will be used for the final classification pipeline because it allows for a consistently high increase in intra-protocol performance and inter-protocol performance across all data-sets.

5.11 Conclusion

This chapter set out to examine different methods of pre and post-processing methods and their ability to improve intra-protocol performance and inter-protocol performance. The pre and post-processing methods that improved the intra and inter-protocol performance were identified. This was then used to create a combination of pre-and post-processing methods that will be used in the rest of this thesis.

At this moment the classification pipeline is (red text indicates the additions from this chapter):

1. Determination of data type
2. Pre-processing
   2.1. ENMO extraction
   2.2. Structure Preserving Oversampling
3. Segmenting into windows.
4. Extracting features
   4.1. Normalization
   4.2. Feature reduction

4.3. Domain Adaptation
5. Creating the classification model
6. Post processing
   6.1. Participant Adaption via Iterative Relearning
   6.2. Hidden Markov Modelling
   6.3. Smoothing

6. Segmentation of Acceleration into Windows

6.1 Introduction

This chapter seeks to address a challenge identified in Chapter 2, namely that of developing a method for automatic segmentation of acceleration data. This will allow for variable length windows to be used in the activity classification pipeline, instead of fixed width windows. As discussed previously, variable length windows have advantages over fixed width: different activities have different optimal window sizes, and activity transitions can be avoided, which can decrease performance.

This chapter sets out a methodology for automatic segmentation by combining wrist-worn accelerometry data with change point detection (CPD). CPD is a data driven method for detecting if the underlying process generating time series data changes. If the acceleration data is generated via participant activities, a detected change in the acceleration data is representative of a change in the activity that generates the acceleration (a transition). Hence, using CPD on acceleration data will allow for the detection of changes in the activity, namely transitions.

The aim of this chapter is to create an activity transition detection method and to investigate the effects of varying parameters and data choices on performance of this method. An assumption integral to the majority of transition detection methods, is that all transitions are instantaneous (112). This assumption was tested, found to be false and methods to overcome it were developed.

The performance of the new method created in this work was compared to other activity transition detection methods. After the optimal method for

transition detection was developed, the automatic segmentation allowed through this method was compared to fixed window methods in the context of activity classification.

6.2 Change Point Detection

Change point detection (CPD) refers to the identification of times when the probability distribution governing a stochastic process or time series changes (144). This makes it well suited to segmenting acceleration data based on changes in the acceleration data, as it has been shown that the probability distribution of a data window is linked to the activity performed in the data window (65).

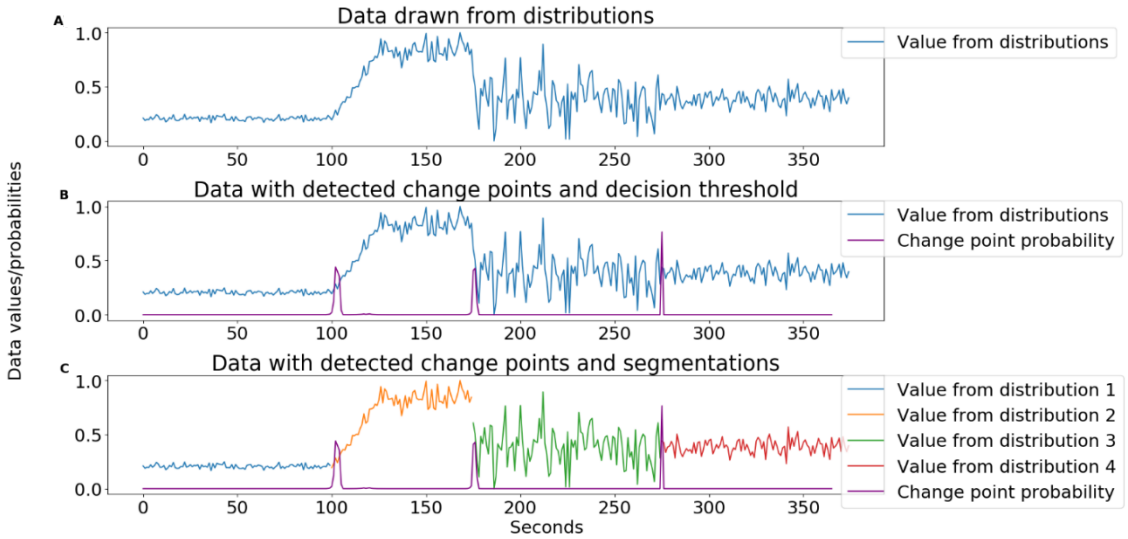

Figure 20: Showing change point detection separating data from four distributions. A shows the created data, B shows the CPD applied to the data, and C shows the correct separation.

As an example of CPD, Figure 20A shows data created by concatenating samples from four normal distributions $N(0.23, 0.02), N(0.82, 0.08), N(0.44\ 0.11)$, $N(0.44, 0.06)$. Between data points 100 and 125 the data is gradually transitioned from being drawn from $N(0.82, 0.08)$ rather than $N(0.23, 0.02)$ to represent a gradual transition.

Figure 20B shows the probability estimated by the Online Bayesian CPD (OBCPD) (145) method that the data distribution has changed. A threshold value can then be used to decide if a point is a change point from the estimated probability. This value can be raised to decrease the number of segments into which the data is partitioned. In most cases an optimal threshold is identified from the data, as is the case in this work.

Figure 20C shows the final segmentation of the data, when using a threshold of 0.4 (chosen as an example). As can be seen, gradual transitions (seconds 100-125) still allow for changes to be detected but not as precisely (assigning high probability to multiple points surrounding the true change). Note also that the transition in variance at data point 275 has been detected even though the means of distributions before and after are identical.

Many different algorithms exist for detecting changes in time series data, each of which has advantages and disadvantages (144). The algorithm chosen for this work is OBCPD, as created by Adams and McKay (145). OBCPD, works by "*estimating the posterior distribution over the current 'run length', or time since the last change point, given the data so far observed*". This means that when the change points are computed, both the probability that each successive point does not belong to the same distribution as previous points and length of runs are estimated. In very simple terms, this method works by identifying the probability that the next point it "sees" is generated by the same probability distribution as the previous points. This form of CPD was chosen for many reasons:

- It is not necessary to know the number of changes *a priori*. In the majority of cases (such as segmenting data from unseen participants) this information would not be known.
- OBCPD allows for the use of multivariate signals when detecting change points. The acceleration data is gathered from a triaxial accelerometer, meaning that three data streams are available. The current methods for

transition detection either only use one axis (146) or aggregate all axes into one data stream (112).

- The OBCPD algorithm reports the probability or score indicating the likelihood of points being a change point instead of giving a hard classification of locations. Using probabilities allows for adjusting the severity of the segmentations by adjusting the threshold value. This allows the user to bias the algorithm towards under or over-segmentation depending on the task at hand.
- Its online nature represents an advantage over retrospective methods, especially with the increasing interest in real-time online activity classification.
- The running time of OBCPD scales linearly with the length of the data. Since accelerometry data can consist of 100Hz acceleration data gathered for 1 week, such a highly scalable algorithm is essential.

OBCPD was used to identify the location of activity transitions by using CPD on the acceleration data and assuming that transitions in PA correspond to detected change points.

6.3 Refractory Period

During the process of investigating the transition detection methods, a limitation of current methodologies (including OBCPD) was identified. It was found that multiple transitions, in close proximity to one another (within 1 second), were predicted when only a single true transition occurred. It appears that these multiple transitions were identified because of the (incorrect) assumption that the transitions are instantaneous (or occur < 1 second). Hence a post-processing method to supress these additional detected transitions was developed.

The post-processing method adds a refractory period in the change point detection (and other transition detection methods) that suppresses the detection of additional transitions within a set interval of other transitions. The interval used was 2.95 seconds, as Kozina et al (112) identified this as the average length of a transition. The time of the 'detected' transition was then identified as the mean of all detected transitions in the 2.95 second window.

This detection of multiple transitions during a single true transition is illustrated in Figure 21. At 57 and 168 seconds a true transition occurs (panel A), but multiple transitions are detected (panel B). Using the refractory period, indicated by the yellow shading, sets the location of the 'detected' transition to the centre of the detected transitions and supresses all other detected transitions within 2.95 seconds of the first detected transition (panel C). The margin is represented in panel D by a grey shaded area around the true transitions. A transition must be within this bounded area to be considered correct.

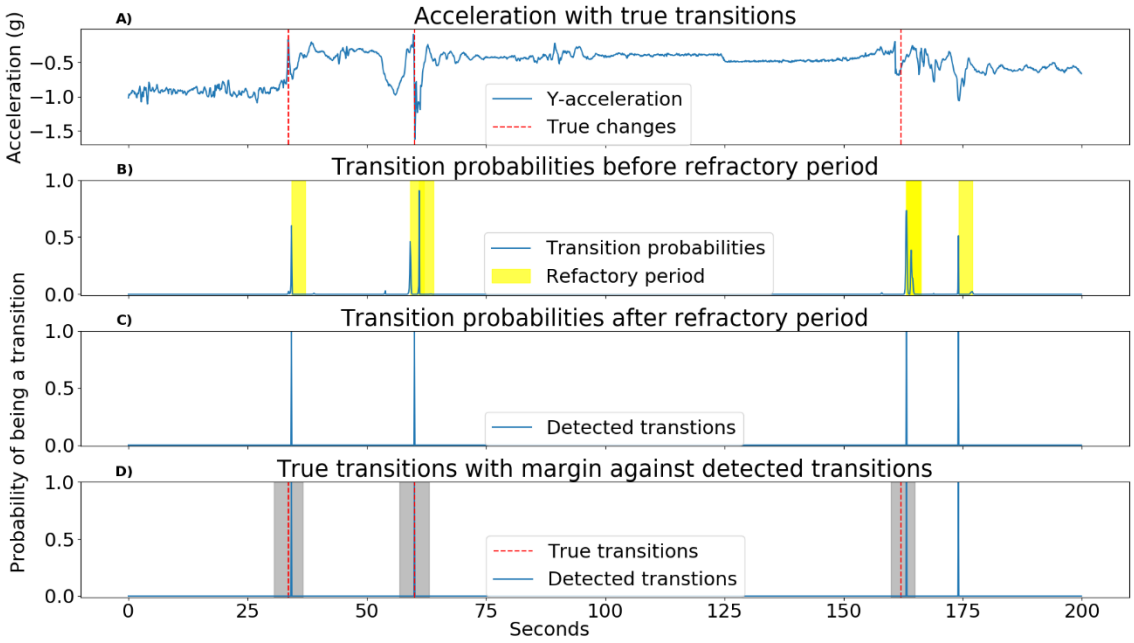

Figure 21: A: the acceleration with the identified true transition locations. B: the estimated probabilities that each point is a transition as found by OBCPD, along with the 2.95 seconds refractory period following the first of each group of detections indicated by yellow shading. C: the transitions are then joined into a single location for each distinct transition, and (D) compared to the true location with margins of acceptance shown as grey shading.

6.4 Method

6.4.1 Data

This study made use of both the Lab-based transition data (3.2.1.2) and the Free-Living data discussed in Chapter 3. The acceleration data was reduced to 10 Hz by averaging (mean) for two reasons: reducing the frequency of the data made the activity transitions more abrupt in the data and increased the performance of all methods; additionally, reducing the amount of data dramatically decreased the computational time required for all methods.

A goal of this work was to determine which combination of the X, Y and Z axes accelerations, recorded by a triaxial accelerometer, yields the most accurate transition detection. The performance of the transition detection was evaluated for all combinations of axes. While the OBCPD is able to make use of multivariate data, the other methods are not. For the alternative methods, the data was aggregated into a single data stream by computing the vector magnitude of the acceleration axes. This represents the standard way of combining multiple acceleration axes for transition detection (112). When making use of multivariate data it is possible to weight each axis to assign greater importance to a specific axis, but this was not done in this work.

The performance of 7 different combinations of acceleration data were evaluated:

1. X-axis
2. Y-axis
3. Z-axis
4. Multivariate X and Y-axes
5. Multivariate X and Z-axes
6. Multivariate Y and Z-axes
7. Multivariate X, Y and Z-axes

6.4.2 Metrics

The performance of the transition detection was evaluated with multiple metrics in order to provide a more comprehensive view than any single metric may provide. The Lab-Based transition data was evaluated with the following metrics, as full information about the location of the true transitions was available (having been observed directly).

- Root mean squared error (RMSE): this is computed by calculating square root of the mean squared time difference between each detected transition and the closest true transition. This evaluation informs how close detected transitions are to the true locations.
- Matthews Correlation Coefficient (MCC): a correlation coefficient between the observed and predicted binary classification of a transition that takes into account true and false, positives and negatives and is generally regarded as a balanced metric which can be used even if the class sizes are very different (147). The formula for computation is:

$$MCC = \frac{(TN \times TP) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

  where TP, FP, TN and FN stand for the numbers of True Positives, False Positives, True Negatives and False Negatives respectively.

Due to the imbalanced nature of the data-sets, which were comprised of mostly negatives (no transitions), the specificity (the proportion of correct negatives) and the accuracy were not used, as an accuracy of 99% is possible by never detecting any transitions.

The Free-Living data does not contain full information about the location of the true transitions, just postural labels. This means that transitions between activities that have the same postural label (sitting to lying) are identified as transitions in the wrist acceleration data but not in the ActivPal (postural) labels. Therefore, different metrics that do not require the full information about the

location of the true transitions are required. These two metrics attempt to provide performance metrics that do not penalise False Positives (as changes in the wrist data that do not correspond to changes in the labels will evaluate as FP's), without the bias toward over-segmentation that typically accompanies this lack of penalisation.

- Sensitivity: sensitivity or true positive rate is defined as $\frac{TP}{TP+FN}$. This metric is used to determine how often the transitions are detected correctly.
- Mean Minimum Distance to transitions (MMD): the mean minimum distance is the mean distance between a true transition and the closest detected transition.

Both metrics reward over-segmentation, having no penalty for False Positives. In order to correct this bias towards over-segmentation, a naïve transition detection method was created that identified the same number of transitions as the evaluated method (over-segmenting to exactly the same extent) and distributed them uniformly through the data. The ratio of the metrics computed for the methods being evaluated and the naïve detection thus gives an evaluation metric that is not influenced by over-segmentation:

$$Ratio\ of\ Sensitivity\ (RoS) = \frac{Sensitivity(Detected\ transitions)}{Sensitivity(Naive\ transition\ detection)}$$

$$Ratio\ of\ MMD(RMMD) = \frac{MMD(Detected\ transitions)}{MMD(Naive\ transition\ detection)}$$

Equation 5: Ratio of Sensitivity and Mean Minimum Distance.

The RoS ranges between 0 and infinity, with larger numbers representing greater levels of success. A value of 1 means that the detected transitions are as sensitive as a naïve transition detection.

The RMMD ranges between 0 and infinity, smaller values indicating detected transitions are closer to true transitions without over segmentation.

It is important to note that transitions are not instantaneous, and therefore a precision at the sampling rate (10Hz) is unfeasible. Hence, a transition is deemed to have been correctly detected (a true positive) if it is within a specified temporal “margin” of the labelled true transition between activities (see Figure 22). Kozina et al (112) estimated that the average transition between activities lasts 2.95 seconds, and Salcic (146) made use of a 3 second fixed window for activity transition detection. Consequently, a 3 second margin was used here, meaning that the detected transition must be within ±3 seconds of a true transition to be considered correct (allowing for the detected transition to occur 3 seconds before or after the labelled transition). The MCC and RoS made use of the margin whereas the RMSE and RMMD do not because they are distance-based metrics.

When using the margin, it is possible for multiple detected transitions to be assigned to a single true transition; to mitigate this only a single detected transition was allowed to match to a true transition. For example, in Figure 22, a very low threshold would detect two transitions (at times 16 and 17), but only the closest {16} is counted as matching the true transition at time 15.

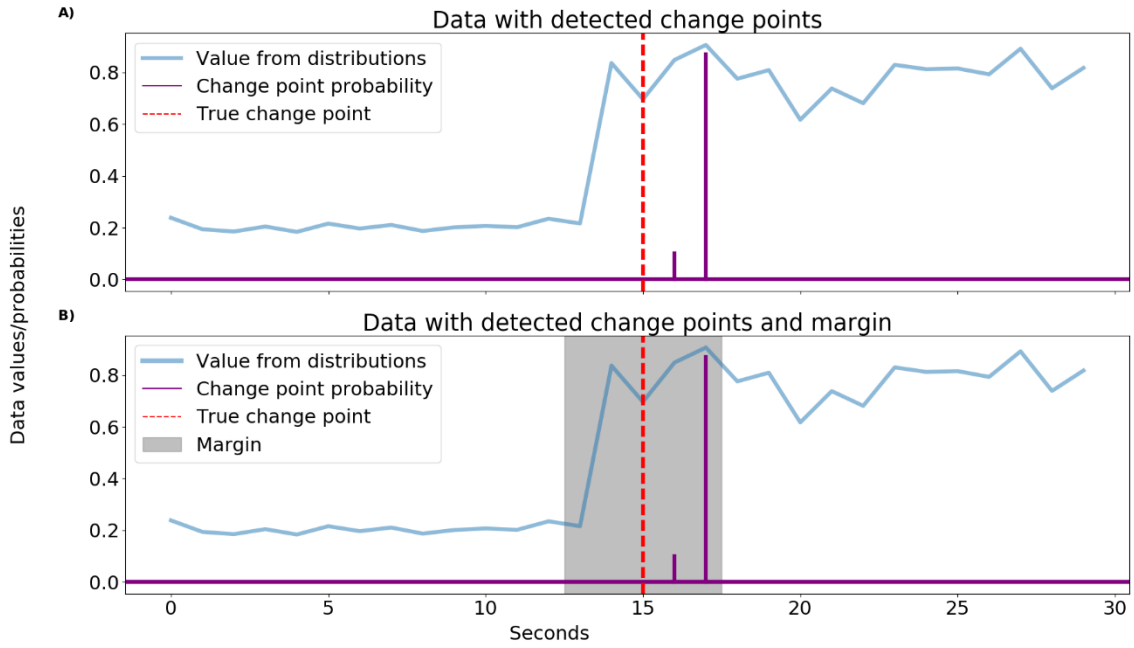

Figure 22: A) Showing detected transition/change point along with location of true change point. B) Showing true change point along with the acceptable 'margin' of error. A detected change point within this margin is deemed correct. Here either one of points 16 or 17 would be counted as correct, but to avoid double counting only one detected transition within the margin is counted.

The performance of the OBCPD method was compared to three other methods:

- Kozina's method (112): this is a data driven approach that identifies points where there is a "significant change between consecutive data samples and divides the data into intervals at that point. The significant change is defined as a sequence of consecutive data samples where the values are in descending order, and the difference between the maximum and the minimum element in the sequence is larger than a threshold". The threshold mentioned here is referred to as the 'derived constant' to differentiate it from the decision threshold used elsewhere in this work. The derived constant is a data derived value obtained from the maximum and minimum values of all of the participant's available data. Like the OBCPD method, a score that any given instant is a transition is provided, and a decision threshold must then be chosen.

- Lyden's method (148): this is a data driven method that identifies instances of rapid acceleration/deceleration and divides the data at those points. Rapid accelerations are defined as "instances where the absolute difference between adjacent counts from the second-by-second vertical acceleration signal is greater than the acceleration cut-off". In the original work the cut-off value was 30 counts. However, since there is no direct conversion from counts to raw acceleration values, this approach was modified to use raw acceleration data. The second-by-second acceleration values were computed (subsampling the data to 1Hz by averaging), and the deciles of the absolute differences between consecutive samples were computed. This gave 10 possible threshold values that could be used to determine activity transitions.
- Salcic's method (146): this is a classification-based approach that creates a classifier trained on some labelled training data to identify whether a three second moving window contains a transition. It does this by forming a decision tree based on the absolute mean difference of the acceleration in the three second window.

All methods require training, either to create and train a classifier or to identify the optimal decision threshold/acceleration cut-off value. To guard against overfitting, testing and training used Leave One Subject Out Cross Validation. In this case the optimal threshold was computed over the training participants and then used to identify the transitions in the held-out participant. For all threshold-based methods, the optimal thresholds are identified by linear search of the values $[0.01, 0.02, \ldots, 1]$. The value which optimises the performance on the left-out data is used as the threshold. It is important to note that the multivariate methods only require a single threshold.

The optimal parameters were identified in the Lab-Based transition data-set, after which their final performance using these parameters was evaluated on the Free-Living data. This gives a more accurate representation of their ability to

generalise to unseen data which will typically be gathered in a Free-Living scenario. The optimal thresholds are found by averaging the thresholds found for each 'fold' of the Leave One Subject Out Cross Validation.

To determine if two methods have significance differences the data was paired by participant ID and a Wilcoxon signed-rank test (133) was used to compare the methods. Due to the fact the multiple hypotheses were evaluated on the same data set, the likelihood of a Type I error is increased. This was compensated for by using Bonferroni corrections. This entails testing each individual hypothesis at a significance level of $\frac{\alpha}{m}$, where α is the overall hypothesis level (in this case 0.05) and *m* is the number of hypotheses. A p-value under $\frac{\alpha}{m}$ indicates that the results are statistically significantly different from one another with high confidence.

6.5 Classification

After the optimal transition detection method was identified (OBCPD) it was used to automatically segment the Lab-Based activity data (distinct from the Lab-Based transition data) and Free-Living acceleration data. This segmented data was then used in the classification pipeline, which was then compared to using fixed window sizes in activity classification.

The Base classifier used in this thesis uses the following pipeline for the creation, training and testing of the classifier. The method tested here replaces using fixed 12.8 second windows with using detected transitions to automatically segment the data.

1. Determination of data type
2. Pre-processing: none
3. Windowing:
    a. Using 12.8 second windows

b. Using automatic segmentation

4. Feature extraction: using 39 features based on the statistical aggregate features and frequency statistics
5. Building the classification model: using a Random Forest classifier with 50 separate trees.
6. Post-processing: none

For the segmentation, the data was reduced to 10Hz, but not for the classification step. Sedentary-Standing-Active labelling in the Lab-Based activity data was used, this means that the Lab-Based activity data and the Free-Living data had comparable labels, hence it was possible to compute LabCV, FreeCV (to compute the intra-protocol performance), Lab-Free and Free-Lab (to compute the inter-protocol performance).

The window sizes automatically generated were also examined.

6.6 Results

6.6.1 Lab-Based Transition Data

Table 26 shows a comparison of the MCC scores for the OBCPD method and the three other representative algorithms drawn from the literature. Each method was evaluated on all combinations of accelerometer axes.

The level of agreement between observed and predicted activity transitions (as measured by the MCC) was highest for the OBCPD (MCC 0.767). The best results were obtained from detecting transitions from the Y-axis only and making use of the refractory period (Table 26). For all single axes, the OBCPD achieved highest MCC values, outperforming the other three methods. The Salcic method had the lowest MCC values for all combinations of axes. The difference between the performance of the Y-axis OBCPD and the next best method (Lyden Y-axis) was significant.

The lowest RMSE values were obtained from the OBCPD method (Table 27), for all single axis methods, but for multiple axes Lyden's method performed best. The lowest RMSE value (3.17 seconds) was once again obtained from detecting transitions in the Y-axis (vertical accelerations) only and making use of the refractory period.

	OBCPD *(145)*		Kozina *(112)*		Lyden *(148)*		Salcic *(146)*
		Ref		Ref		Ref	
X	0.636 (0.173)	**<u>0.726</u>** (0.0998)	0.384 (0.104)	0.528 (0.146)	0.477 (0.122)	0.566 (0.183)	0.232 (0.270)
Y	0.686 (0.244)	**<u>0.763</u>** (0.108)	0.362 (0.149)	0.415 (0.154)	0.475 (0.126)	0.582 (0.199)	0.209 (0.199)
Z	0.592 (0.260)	**<u>0.648</u>** (0.265)	0.320 (0.127)	0.395 (0.178)	0.437 (0.225)	0.524 (0.175)	0.183 (0.210)
XY	0.452 (0.199)	**<u>0.508</u>** (0.133)	0.261 (0.152)	0.431 (0.163)	0.353 (0.104)	0.432 (0.159)	0.246 (0.308)
XZ	0.419 (0.171)	0.489 (0.221)	0.426 (0.131)	0.487 (0.125)	0.474 (0.138)	**0.580** (0.160)	0.233 (0.174)
YZ	0.451 (0.186)	0.501 (0.211)	0.267 (0.105)	0.449 (0.200)	0.461 (0.216)	**0.568** (0.169)	0.234 (0.234)
XYZ	0.381 (0.201)	**0.425** (0.128)	0.252 (0.193)	0.416 (0.244)	0.306 (0.225)	0.403 (0.166)	0.253 (0.201)

Table 26: Matthews Correlation Coefficient of the transition's detection methods for all combinations of axes using Lab-Based transition data. Columns with labelled "Ref" refer to calculations using the refractory period. Bold indicates the highest value in that row; bold and underlined indicates a significant difference between that value and the next highest. The Salcic method made use of fixed length windows for transition detection therefore the refractory period was not applicable. Figures in brackets indicate standard deviations.

For single axis measurements, the X (horizontal right-left) and Y-axis (vertical) were consistently better than the Z-axis (horizontal front-back) for all methods. The performance of axes combinations was dependent on the method used, but

for the OBCPD method it was never the case than any combination of axes outperformed any single axis approach.

In all cases the refractory period increased the MCC of the transition detection, typically increasing the performance by around 25%. However, the effect of the refractory period was significantly reduced when the detection threshold exceeded 0.5. As Table 27 shows, the refractory period tended to increase the RMSE for all algorithms, although only by 7%.

Single axis measurements are preferable for the OBCPD method, but pairs of axes and all axes are beneficial for the Lyden method. Nonetheless in terms of location accuracy, the OBCPD method using any single axis is superior to Lyden's method. In common with the MCC metric, the Y-axis yields the most accurate transition detections. For all methods except Salcic, all axes reported a bias (the sum of all errors) of less than ±1.4 seconds, additionally the errors were normally distributed, meaning the methods were neither consistently late nor early in identifying transitions.

	OBCPD *(145)*		Kozina *(112)*		Lyden *(148)*	
		Ref		Ref		Ref
X	**3.63**	3.88	5.61	6.02	4.51	4.81
	(1.62)	(1.80)	(2.33)	(3.49)	(1.14)	(1.57)
Y	**3.17**	3.38	6.02	6.20	4.42	4.58
	(1.53)	(1.18)	(3.70)	(2.09)	(1.34)	(2.15)
Z	**4.78**	5.04	5.89	7.43	5.10	5.28
	(3.12)	(2.22)	(2.75)	(2.16)	(2.29)	(2.34)
XY	6.73	7.17	7.34	7.29	**5.53**	6.41
	(2.59)	(1.01)	(2.85)	(2.22)	(1.51)	(1.99)
XZ	6.90	7.30	5.30	5.62	**4.86**	5.08
	(3.83)	(4.42)	(3.52)	(2.29)	(2.26)	(2.00)
YZ	6.64	7.09	6.64	7.25	**4.56**	4.75
	(4.00)	(3.18)	(3.86)	(3.22)	(3.11)	(3.24)
XYZ	7.69	7.99	7.38	7.34	**6.15**	6.83
	(4.56)	(4.42)	(3.92)	(1.99)	(3.84)	(2.02)

Table 27: Root Mean Squared Error (seconds) of the transition detection methods for all combinations of axes using Lab-Based transition data. Columns with labelled "Ref" refer to calculations using the refractory period. Bold indicates the lowest value in that row; bold and underlined indicates a significant difference between that value and the next lowest. The Salcic method made use of fixed windows for transition detection so the RMSE is not meaningful. Figures in brackets indicate standard deviations.

The optimal threshold for the OBCPD method as identified by averaging all the thresholds found in the Leave One Subject Out Cross Validation was 0.128.

6.6.2 Free-Living

It is not meaningful to create a naive classifier using a fixed window approach, so the Salcic method is not used with the Free-Living data.

As shown in Table 28 of the Free-Living data, the highest performing method was OBCPD obtaining a sensitivity of 0.802 and a Ratio of Sensitivity (RoS) of 2.4. This means that the OBCPD method is 2.4 times more sensitive than a

naïve segmentation with the same number of detected transitions. The RoS value was determined to be significantly higher than other methods, but it was noted that all methods performed better than the naïve, uniformly spaced segmentation. The Lyden and Kozina methods achieved roughly equal RoS, but different sensitivity values. Specifically, the lower sensitivity values from Kozina with equivalent RoS values to Lyden's method indicate that Lyden's method detects a greater number of segmentations (over-segmenting) in the Free-Living data.

The OBCPD obtained a mean minimum distance (MMD) score of 3.67 seconds indicating that a transition was detected an average of 3.67 seconds from a true transition. This is about 0.7 seconds longer than the 2.95 seconds Kozina (112) estimates for duration of a transition and indicates that the OBCPD method is effectively detecting transitions. The RMMD is 3.21, which is comparable to the Kozina and Lyden methods, although these have different MMD scores, indicating a different level of segmentation. The accuracy of detection by these methods is significantly worse than the OBCPD method (5.16 seconds and 8.06 seconds versus 3.67 seconds).

The use of the refractory period decreased the sensitivity and MMD scores for all methods but significantly increased the RoS and RMMD for all methods. This indicates that while the refractory period decreases the sensitivity of methods, it also reduces the number of detected transitions. This reduction in detected transitions outweighs the reduced sensitivity when computing the RoS and RMMD.

	OBCPD *(145)*		Kozina *(112)*		Lyden *(148)*	
		Ref		Ref		Ref
Sens	0.831 (0.283)	0.802 (0.268)	0.675 (0.209)	0.632 (0.224)	0.872 (0.342)	0.839 (0.415)
RoS	2.37 (1.10)	*2.40 (0.750)*	1.74 (0.721)	1.85 (0.739)	1.86 (0.945)	1.91 (0.927)
MMD	3.54 (1.47)	3.67 (0.624)	8.73 (1.67)	8.06 (1.74)	5.16 (0.873)	5.16 (0.824)
RMMD	3.16 (0.914)	3.21 (1.27)	2.98 (2.35)	3.02 (1.31)	2.71 (0.936)	2.88 (1.27)

Table 28: Reporting the average Sensitivity (Sens), Ratio of Sensitivity (RoS), Mean Minimum distance (MMD) and Ratio of MMD (RMMD) of the transition detection methods for each method in the Free-Living data over each person, figures in brackets represent standard deviations.

6.6.3 Window Sizes

The window sizes generated for the Lab-Based activity data under labelling 2 and Sedentary-Standing-Active labelling, as well as the Free-Living data, are shown below in Table 29, Table 30 and Table 31.

Lab Label 2	*Min*	*Median*	*Max*	*Mean*	*Standard deviation*
Desk	1	14	284	42	62
Standing	1	28	297	77	96
Walking	1	7	300	20	45
Household	1	6	123	6	4
Lying	3	15	1800	150	380

Table 29: Statistics about the window distribution in the Lab-Based activity data using automatic segmentation under labelling 2. All units are seconds (s).

Sedentary-Standing-Active labelling	*Min*	*Median*	*Max*	*Mean*	*Standard deviation*
Sedentary	1	15	1800	90	262
Standing	1	6	297	8	20
Active	1	7	300	20	45

Table 30: Statistics about the window distribution in the Lab-Based activity data using automatic segmentation under Sedentary-Standing-Active labelling. All units are seconds (s).

Free-Living label	*Min*	*Median*	*Max*	*Mean*	*Standard deviation*
Sedentary	1	8	8175	28	122
Standing	1	6	3229	9	21
Active	3	6	626	8	11

Table 31: Statistics about the window distribution in the Free-Living data using automatic segmentation. All units are seconds (s).

The window sizes identified through automatic segmentation had a very broad distribution with a very large tail. Under labelling 2, the median window sizes range between 6-28 and the maximum sizes range between 123-1800. The median values themselves aren't enough to distinguish between labels (as the Free-Living Standing and Active are very close). However, it does seem that different activities have different window sizes or have more varied window sizes. The median window sizes of the Lab-Based activity data under Sedentary-Standing-Active labelling, range between 6 and 15. Thus suggesting that the 12.8 second window used in the Base classifier is a good choice. However, the median window sizes are smaller in the Free-Living data, 6-8 seconds

6.6.4 Classificiation Performance

As seen in Table 32 the classifier making use of the automatic window sizes, as determined by transition detection has a significantly higher inter-protocol performance (0.352, 0.415 versus 0.529, 0.675), without a significant decrease in the intra-protocol performance (0.898, 0.765 versus 0.885, 0.837).

	LabCV	*FreeCV*	*Lab-Free*	*Free-Lab*
Base	0.898 (0.103)	0.765 (0.214)	0.352 (0.132)	0.415 (0.0978)
Auto-segmentation (OBCPD)	0.885 (0.117)	0.837* (0.180)	0.529* (0.0809)	0.675* (0.0860)

*Table 32: Classification performance using automatically segmented data compared to the Base classifier. * Indicates significant differences from the Base classification pipeline. Figures in brackets indicate standard deviations.*

6.7 Discussion

The OBCPD method developed in this study outperformed the three other methods it was compared against in the Lab-Based transition data and the Free-Living data. It achieved the highest MCC and Ratio of Sensitivity (RoS) and lowest RMSE and Ratio of Mean Minimum Distance (RMMD). Best performance was obtained through using the Y-axis alone. A limitation in activity transition detection methods was the assumption that transitions between activities are instantaneous. This assumption leads to the identification of multiple transitions when only one actually occurred. This limitation was overcome by introducing a refractory period of 2.95 seconds. The introduction of the refractory period improved MCC, RoS and RMMD values across all axes for the new OBCPD method as well as the other methods examined.

The ability of this OBCPD to successfully identify transitions, validates the a-priori assumption made in this work that ‘a change in the acceleration data is representative of a change in the activity that generates the acceleration’, as changes in the acceleration data are representing transitions in activity.

The higher performance from the Y-axis is likely to arise from two factors: the specific activities in these data-sets are concerned with transitions from sitting to standing activities, with each causing an orientation change in the wrist. This change in orientation is likely to be captured by the Y-axis due to its position on

the body. Other combinations of axes might be more effective for detecting transitions between a broader range of activities. Additionally, using multiple axes together increased the prevalence of false positives. As a false positive in any axis will trigger the detection of a transition this would explain why all single axis measures outperformed any multi-axes measure in the OBCPD. This argument is strengthened by the fact that the other transition detection methods did not experience this reduction in performance when using multi-axis method, as they aggregate all multi-axis methods into a single data stream.

The method of Lyden et al (148) was a high performing method for transition detection in the Lab-Based transition data, obtaining high MCC values and low RMSE values compared to the other non OBCPD methods. The low RMSE may be because the transitions were worked out on a second by second resolution unlike the Kozina method which used a 10Hz resolution.

Kozina's (112) method was the poorest performing data driven method on the Lab-Based transition data, likely because it required two different values to be identified: the probability threshold use to identify if a given point was a transition and the derived constant used in the identification of the transitions. As the derived constant was based on the maximum and minimum values of the entire acceleration stream, noise or extreme values may have contributed to this lower performance.

Kozina and Lyden achieved comparable RoS and RMMD values but had different sensitivity and MMD values. The higher sensitivity and lower MMD values of Lyden's method with the equivalent RoS and RMMD values indicate that it segmented the data more than Kozinas' method (because the naïve sensitivity would have to be higher to obtain the same RoS). This over-segmentation may be due to using 1 Hz data as opposed to the 10Hz Kozina data.

The bias values of ±1.4 seconds or under as reported by the OBCPD, Kozina and Lyden methods, suggest the main source of error is the gradual nature of the transitions. It has been noted that transitions typically take 2.95 seconds (112), the values of ±1.4 suggest that total error is roughly half that of the length of the transitions. Such errors would be expected when the transition label occurs in the centre of the transition and the detection occurs near the end or the beginning. As the errors are normally distributed, it is not the case that the transitions are consistently detected at specifically the end or the beginning.

The method developed by Salcic (146) consistently achieved the poorest results, likely due to its status as a classification method rather than to a data driven approach. This means that it has the disadvantages of using fixed windows. It may have also been affected by the relatively small amounts of training data. The method involved creating fixed windows of 3 second duration resulting in only 1600 windows and of these only 10% contained transitions. The paucity of data combined with the imbalance of the data-set (containing 90% non-transitions) is likely to have contributed to the lowered performance.

The refractory period added to all the methods in this study led to increases in the MCC values for all methods and all axes combinations. This suggests that the problem of identifying multiple transitions for each true transition can be overcome by the addition of the refractory period. The refractory period has less effect on the performance when the decision threshold is increased, suggesting that most of the extra transitions detected have a low probability. The refractory period does not always decrease the RMSE, which may be because the refractory period methods places the detected transition at the central point of all detected transitions in the 2.95 second window. However, placing the detected transition at the end or the beginning of the detected transitions gives comparative results to placing it at the centre.

In the Free-Living data the refractory period increases the RoS, while decreasing the sensitivity. Using the refractory period reduces the number of

transitions identified, which then decreases the sensitivity of the naïve classifier, offsetting the decrease in sensitivity of the non-naïve classifier. Similarly, use of the refractory period increased the MMD but decreased the RMMD for all methods, most likely for the same reason.

An important facet of this data was observed in the Free-Living data. Transitions in the acceleration data (GENEActiv) did not always correspond to transitions in the postural data (ActivPal), probably because some activities have the same postural labelling. An example of this is shown in Figure 23: the acceleration values noticeably change but there is no recorded change in the postural labels. A transition is detected in the acceleration values of Figure 23 at 30.4 seconds, which is noticeably a valid change. However, as the postural labels don't change this would be reported as a false positive. No modification to any of the algorithms used or parameter manipulations can stop these changes being identified; as the changes are there. However, because these changes are not present in the 'known transitions' (obtained from the ActivPal labelling) they are treated as incorrect by the evaluation metrics. For this reason, evaluation metrics that do not penalise False Positives were used.

Evaluation methods that do not use False Positives are biased towards over-segmentation. To mitigate this the performance of a naïve transition detection with identical amounts of over-segmentation (having equal amounts of segments) was computed in order to compare against the evaluated transition detection system.

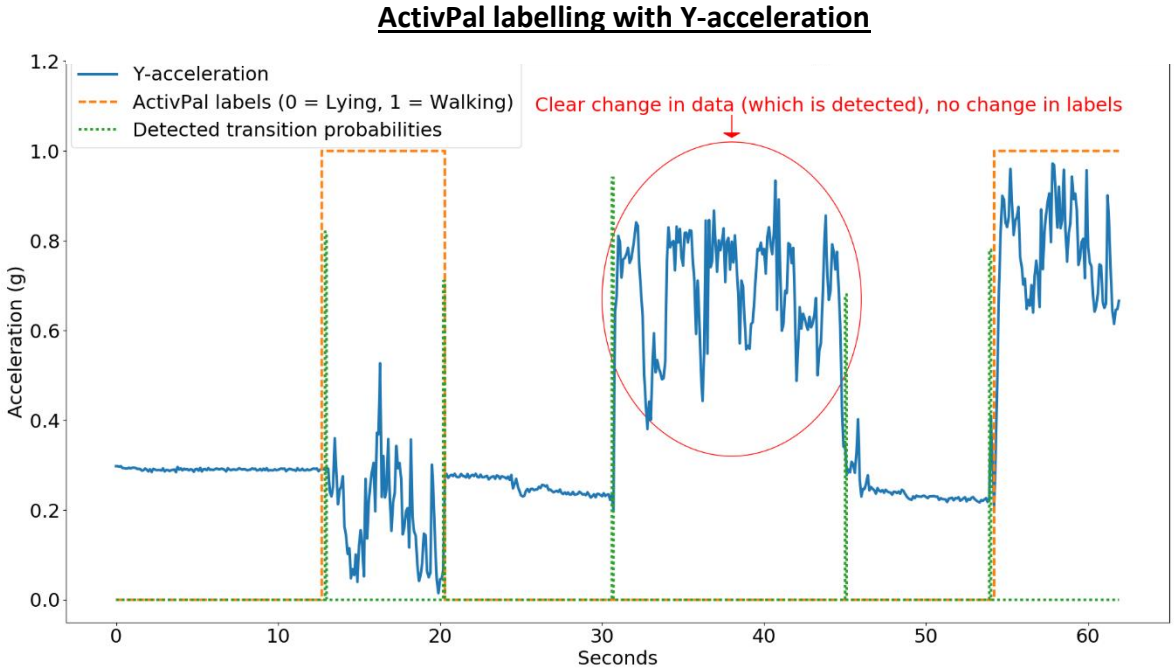

Figure 23: Acceleration value with labels, showing a change in acceleration that does not correspond to a change in the identified activity. The detection of a change in acceleration that does not correspond to a change in the label.

Making use of variable length windows, as defined by activity transition detection allowed for a significantly higher inter-protocol performance without a decrease in the intra-protocol performance. This indicates that using the variable length windows allows for a greater ability to generalise to unseen data than fixed windows.

The median window sizes are slightly lower in the Free-Living data than the Lab-Based activity data, this is likely due to the fact that the Lab-Based activity data is gathered under a strict protocol with uninterrupted segments of specific activities.

6.7.1 Strengths And Limitations

A strength of this work comes from the use of Free-Living data as using Free-Living data allows for validation, which is not possible with only Lab-Based data.

The majority of accelerometry studies make use of only Lab-Based data (55), which typically do not generalise well to Free-Living (77). This inability to generalise means that methods that perform well in Lab-Based data may have limited value as the majority of population studies use accelerometers in Free-Living environments (58).

A further strength of this work was the introduction of the refractory period. This improved the performance for all axis combinations and all methods. It is a simple modification that can be made to existing activity transition methods (either standalone or as part of classification systems) that can be expected to increase the performance with minimal extra computation.

An additional strength of the work is that the OBCPD algorithm used permits online computation of the change points (and therefore the activity transitions) and segmentation of accelerometry data. With the increased adoption of wearable devices, such as Fitbit™ and Apple™ watches, real time *online* wrist-worn activity evaluation is now becoming more prevalent, giving online segmentation algorithms a great advantage.

A limitation of this work is that the activities used are fairly simple. It is possible that when multiple activities are involved, especially more moderate and vigorous ones, the performance of the method may vary. The current conclusions may or may not be valid under more free form activities.

Another limitation of this work was the uncertainty in the locations of the true transitions and the use of the RMSE. The locations of the true transitions were only known with a resolution of 1 second and, more fundamentally, this and all similar work has to confront the difficulty of assigning the location of a transition taking, perhaps, 3 seconds to a single instant. Here the use of the margin around a single instant appears to reduce the precision of the RMSE.

Additionally, when computing the MCC it was possible to have multiple detected transitions assigned to one true transition; to mitigate this only a single detected

transition was matched. Only allowing a single detected transition to match to a true transition was not possible with the RMSE as ‘unmatched’ transitions had an infinite error (due to not having any correct transitions to match to). Due to this, it was decided to let multiple detected transitions match to true transitions when computing the RMSE.

An additional weakness was the limited labelling available in the Free-Living data, meaning that additional metrics had to be identified in order to evaluate the performance of the activity transition detection methods. In future work, this could be addressed by making use of Free-Living data gathered in such a way that allows for more detailed labelling, such as participant-mounted cameras (149).

6.8 Conclusion

The performance of the new Online Bayesian Change Point Detection (OBCPD) method was equal to or better than existing methods depending on the metrics used for evaluation and the data used. In high quality Lab-Based data and Free-Living data the OBCPD method outperformed all others. As such the new OBCPD method is a useful addition to existing activity transition methods and has the advantage of online computation.

The development of this transition detection method allowed for the automatic segmentation of acceleration data into variable length windows. The classification pipeline making use of these windows, outperformed the fixed window approach of the Base classifier with respect to inter-protocol performance. As such this method will be utilised in the final classification pipeline developed in this work.

At the moment the classification pipeline is (red text indicates the addition from this chapter):

1. Determination of data type
2. Pre-processing
   2.1. ENMO extraction
   2.2. Structure Preserving Oversampling
3. Automatically segmenting acceleration data
4. Extracting features
   4.1. Normalization
   4.2. Feature reduction
   4.3. Domain Adaptation
5. Creating the classification model
6. Post processing
   6.1. Participant Adaption via Iterative Relearning
   6.2. Hidden Markov Modelling
   6.3. Smoothing

7. Recurrence Quantification Analysis

7.1 Introduction

This chapter examines the use of Recurrence Quantification Analysis (RQA) features for improving inter-protocol performance.

As mentioned in Chapter 2, a common issue in activity classification is the poor ability of a classifier to generalise to data different to its training data. Often this manifests as lower classification performance in Free-Living data than in the Lab-Based data the classifier was trained on (intra-subject-inter-protocol), although it can also be seen when attempting to classify activities of different populations, such as classifying the activity of overweight participants when trained on normal weight participants (inter-subject-intra-protocol). Several methods for mitigating this performance decrease have been investigated: individual participant adaption (125), the use of specific classification models (77) and the use of specific classification features that are hypothesized to have a greater inter-protocol/inter-subject performance than standard statistical features (76).

Classification features with high inter-protocol performance typically report lower intra-protocol performances than standard statistical features (76). The challenge therefore is to identify features that allow for high inter-protocol performance without reducing the intra-protocol/intra-subject performance. This work has identified RQA as a potential method for this. RQA creates an image based on the recurrent structure of the acceleration and then computes aggregative statistics based on this image, combining both morphological (structural) and statistical methods; methods based on the structure of the signal and the statistical distribution of the signal values, respectively. RQA has shown considerable success in accelerometry gait analysis (118), a field of study similar to that of activity classification.

RQA is a method of statistically analysing data generated by dynamical systems; specifically, it is a way of analysing recurrence plots of a dynamical system (150). Recurrence plots identify the states at which a system approximately repeats a previous state. These recurrence plots characterise the structure of the dynamical system: simple dynamical systems, such as a limit cycle, have simple recurrence plots with few points of recurrence, while complex dynamical systems will have many points of approximate recurrence. The features extracted through RQA describe this recurrence plot and hence the structure of the acceleration data.

The principal benefits of RQA compared to other morphology-based methods are that it requires no filtering before analysis, and it can provide useful information when using data of short durations; whereas some statistical features need large amounts of data before they become meaningful; especially if the data is non-normal.

No research has been identified using RQA in PA classification. This work investigates the use of RQA features for classification, comparing them with the features used in the Base classifier and concentrating on their inter-protocol-inter-subject performance. Additionally, this work challenges the *a priori* assumption that using all axes of the accelerometer for feature extraction results in improved performance by investigating the performance for all combinations of axes.

7.2 Method

7.2.1 Data

Both the Free-Living and the Lab-Based data, as described in 3.2 were used in this chapter. Sedentary-Standing-Active labelling (as identified in 3.2.1) was used in the Lab-Based data to ensure comparability between data-sets, this meant that both data-sets used the labels: Sedentary, Standing or Active. The

data was down sampled to 50Hz because the computation of the RQA features was very computationally expensive; decreasing the frequency of the data reduces the size of the signals to be computed and therefore the computational load.

7.2.2 Analysis

As Sedentary-Stand-Active Labelling was used it was possible to compute: LabCV, FreeCV, Lab-Free and Free-Lab. LabCV and FreeCV give an indication of the intra-protocol performance (how well the classification performs on data from the same protocol), while Lab-Free and Free-Lab give an idea of inter-protocol performance (how well the classification performs on data from a different protocol).To determine if there was a significant difference between the Base classification pipeline and the pipeline making use of the RQA features, a Wilcoxon signed-rank (133) test was used determine whether the performances were significantly different. In order to ensure a large enough sample size, the comparisons were paired on each participant's performance, instead of the average. Due to the fact the multiple hypotheses were evaluated on the same data set, the likelihood of a Type I error is increased. This was compensated for by using Bonferroni corrections. This entails testing each individual hypothesis at a significance level of $\frac{\alpha}{\mathrm{m}}$, where α is the overall hypothesis level (in this case 0.05) and *m* is the number of hypotheses. A p-value under $\frac{\alpha}{\mathrm{m}}$ indicates that the results are statistically significantly different from one another with high confidence.

7.2.3 Recurrence Quantification Analysis Feature Extraction:

RQA feature extraction over a signal (F) works by first computing the distance matrix (DM) of the signal and then extracting features corresponding to the matrix. DM is a matrix, such that in position (i, j) its value is the distance between vertices i and j.

$$DM(i,j) = \|F(i) - F(j)\|$$

Equation 6: Distance matrix.

For RQA, a threshold (ε) is then applied to DM in order to create a binary recurrence matrix R. If the distance between points i, j is less than the threshold ε, then point $R(i,j)$ is 1, else it is 0.

$$R(i,j) = \begin{cases} 1 \; if \; DM(i,j) < \varepsilon \\ 0 \; if \; DM(i,j) \geq \varepsilon \end{cases}$$

Equation 7: Recurrence matrix.

Once the recurrence matrix has been created, features describing R are extracted. The features generally focus on the frequencies of contiguous 'lines of ones', referencing segments of the signal that are similar. An example of R can be seen in Figure 24 as well as the signal F that was used in its creation. Two similar segments can be seen highlighted in F, creating a 'line of ones' in the R.

Four types of feature are usually extracted from the recurrence matrix (150), some of which consider the entire matrix and others that focus only on the distribution of the 'lines of ones'. Typically for these distribution-focussed features, indicated with a * below, three variants are computed; once for diagonal lines of '1's, once for vertical lines of '1's and once for vertical lines of '0's. These are described fully on Table 33.

- Recurrence rate is the density of the recurrence points in the matrix. This corresponds with the probability that any given state will recur. This is computed once over the entire matrix.
- Determinism* measures the predictability of the dynamical system modelled by the recurrence matrix. A random process will have almost only single dots and no lines in the recurrence matrix, whereas a deterministic process will have mostly lines in the recurrence matrix.

- Divergence* estimates the maximal Lyapunov exponent of the system which is a measure of the predictability of the system (151).
- Entropy* measures the complexity of the system. Entropy is the average rate at which information is produced by a stochastic process. In the case of a signal, this is a measure of the signal's complexity (152).

Acceleration trace with corresponding recurrence matrix

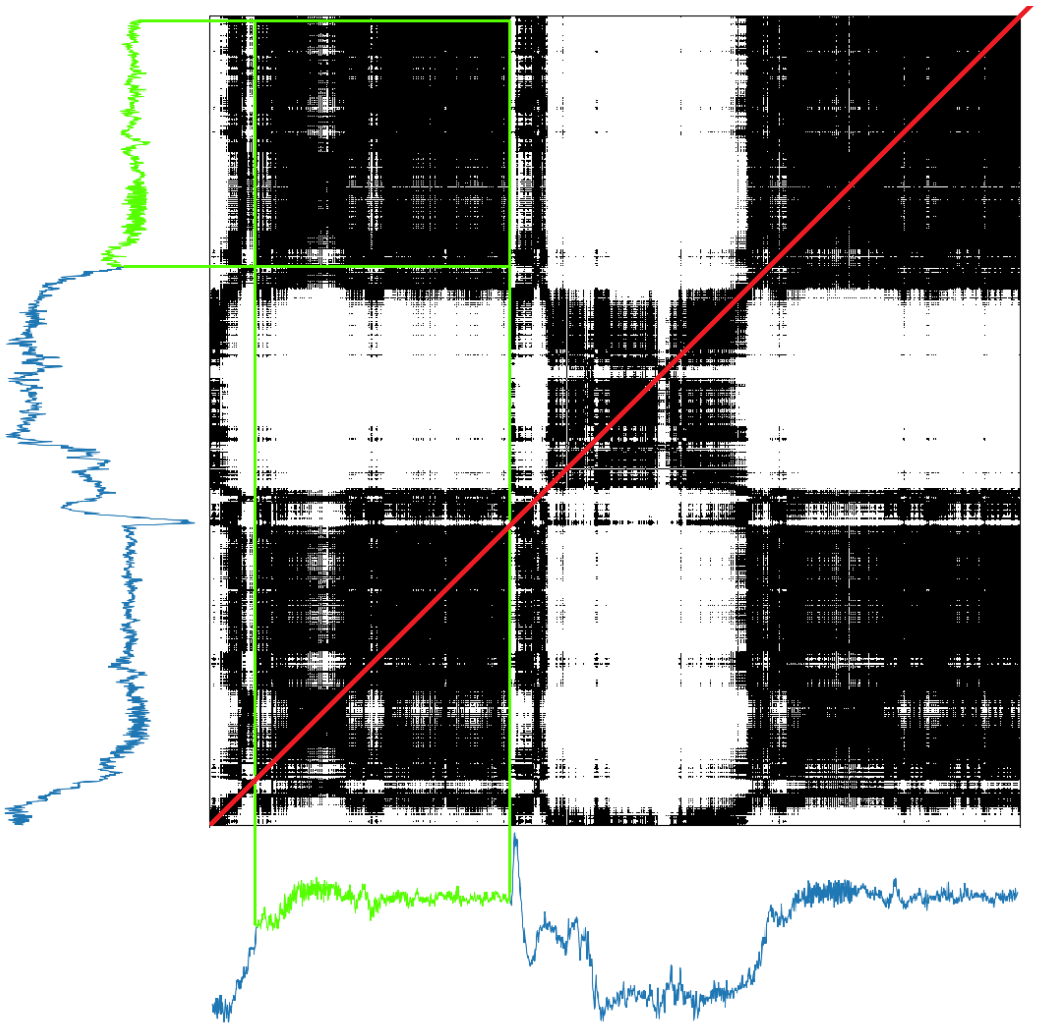

Figure 24: This illustrates an accelerometer trace, and the corresponding recurrence matrix (R) created. Black indicates a value of 1, white indicates a value of 0. This matrix is then used for feature extraction. The signal (F) used to generate the matrix is in blue. The red line identifies where i=j. The green sections identify two segments of the signal that show high similarity with one another and generate a large black patch. This identifies a high amount of recurrence for this segment of the signal.

Feature Number	*Feature symbol*	*Feature Name*	*Formulae*
1	RR	Recurrence Rate	$\frac{1}{N^2}\sum_{i,j=1}^{N} R(i,j)$
2	DET_d	Diagonal determinism	$\frac{\sum_{l=l_{min}}^{N} lP(l)}{\sum_{l=1}^{N} lP(l)}$
3	μ_d	Average diagonal line length	$\frac{\sum_{l=l_{min}}^{N} lP(l)}{\sum_{l=1}^{N} P(l)}$
4	Max_d	Maximum diagonal line length	$\max(l)$
5	DIV_d	Diagonal Divergence	$\frac{1}{\max(l)}$
6	$ENTR_d$	Diagonal Entropy	$\sum_{l=l_{min}}^{N} p(l)\ln(p(l))$
7	DET_h	Horizontal determinism	$\frac{\sum_{v=v_{min}}^{N} vP(v)}{\sum_{v=1}^{N} vP(v)}$
8	μ_h	Average Horizontal line length	$\frac{\sum_{v=v_{min}}^{N} vP(v)}{\sum_{v=1}^{N} P(v)}$
9	Max_h	Maximum Horizontal line length	$\max(v)$
10	$ENTR_h$	Horizontal Entropy	$\sum_{v=v_{min}}^{N} p(v)\ln(p(v))$
11	μ_w	Average white horizontal line length	$\frac{\sum_{w=w_{min}}^{N} wP(w)}{\sum_{w=1}^{N} P(w)}$
12	Max_w	Maximum white horizontal line length	$\max(w)$
13	$ENTR_w$	White horizontal Entropy	$\sum_{w=w_{min}}^{N} p(w)\ln(p(w))$
14	RR/DET_h		$\frac{1}{N^2}\sum_{i,j=1}^{N} R(i,j) \div \frac{\sum_{v=v_{min}}^{N} vP(v)}{\sum_{v=1}^{N} vP(v)}$
15	RR/DET_d		$\frac{1}{N^2}\sum_{i,j=1}^{N} R(i,j) \div \frac{\sum_{l=l_{min}}^{N} lP(l)}{\sum_{l=1}^{N} lP(l)}$

Table 33: Feature symbols, names and formulations for all features used in RQA. Here $P(v)$, $P(l)$ are the frequency distributions of the vertical and horizontal lines respectively. $p(v)$, $p(l)$ are the probabilities that a vertical line/horizontal line has length v/l . v_{min} and l_{min} are the minimum length vertical and horizontal lines considered. N is the number of vertical/horizontal lines.

The majority of activity classification methods make use of all three axes of the accelerometer (55), however it was not known *a priori* if using all three axes was best for RQA features. Due to this, features were extracted for all combinations of the X, Y, Z axes.

- X-axis
- Y-axis
- Z-axis
- Multivariate X and Y-axes, $\overline{XY}$
- Multivariate X and Z-axes, $\overline{XZ}$
- Multivariate Y and Z-axes, $\overline{YZ}$
- Multivariate X, Y and Z-axes, $\overline{XYZ}$

The multivariate combination of the axes was created via averaging the axes at each point to make one aggregate signal.

7.2.4 Takens' Theorem

For activity classification it is possible to extract the RQA features from the acceleration signals. However, an extension to this method exists. According to the work of Hekler et al (153), a participant's movements can be thought of as a chaotic dynamical system. This means that the initial starting condition of PA can greatly change the values throughout time, and how much the current point depends on previous points changes with the value of the current point.

Takens' theorem (154) states that a dynamical system (DS) can be reconstructed from a sequence of observations (o) and the state of the system using a time delay τ and an embedding dimension em, such that:

$$DS(i) = \big(o(i), o(i+\tau), o(i+2\tau), \dots, o(i+(em-1)\tau)\big)$$

Equation 8: Reconstruction of dynamical system via Takens' theorem.

In order to be a faithful reconstruction (a diffeomorphism between the original system and the reconstruction) the embedding dimension must be greater than twice the intrinsic dimension of the original system.

In the context of this work, the participant's PA is the dynamical system and the sequence of observations are the acceleration values observed. For different axis combinations these observed values can be 1-3 dimensional (X, Y, Z). According to Takens theorem, it is possible to recreate the dynamical system (the participants PA) from these observed values (the accelerations), the recreated dynamical system can then used for extracting RQA features instead of just using the acceleration values.

The choice of delay time and embedding dimension are dependent on the acceleration signal and are therefore are affected by the choice of accelerometer axes used.

The optimal value for the time delay is often near the earliest occurring minimum in the autocorrelation of the signal. The optimal value of the embedding dimension can be computed by constructing the phase space according to Takens' theorem and then using Principal Component Analysis to identify how much information (expressed as variance explained) each dimension contributes toward the phase space reconstruction. The minimum number of dimensions to retain most of the information is often near the optimal embedding dimension (155).

Figure 25 identifies the optimal time delays for each axis combination, this information is also expressed in Table 34. This shows that the optimal delay for each axis is reasonably consistent, with the exception of the Z axis, being between 11 and 15. At 50Hz these delays correspond to around 0.24 seconds.

Axis	X	Y	Z	$\overline{XY}$	$\overline{XZ}$	$\overline{YZ}$	$\overline{XYZ}$
Optimal delay (s)	11	12	24	14	11	12	15

Table 34: Table showing the optimal delay as computed via the autocorrelation for each axis/aggregated axis.

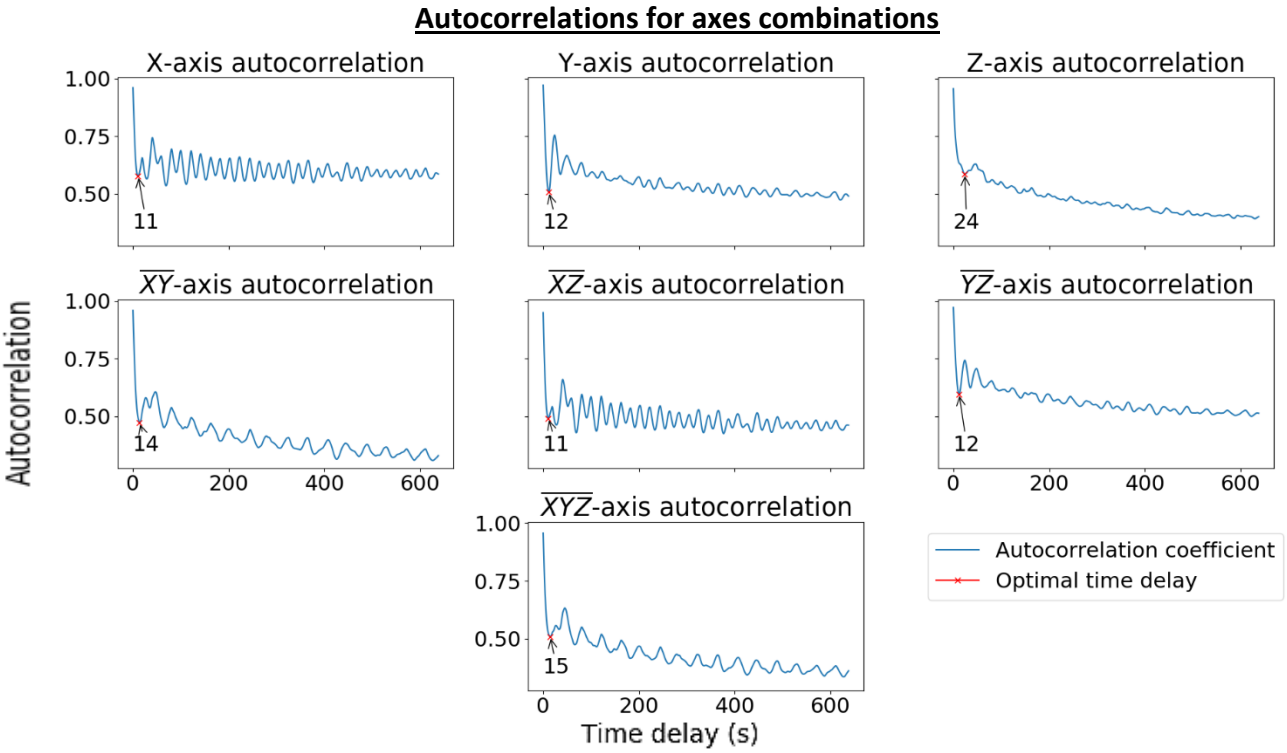

Figure 25: This image shows the autocorrelation values for time delays, for a variety of axis combinations.

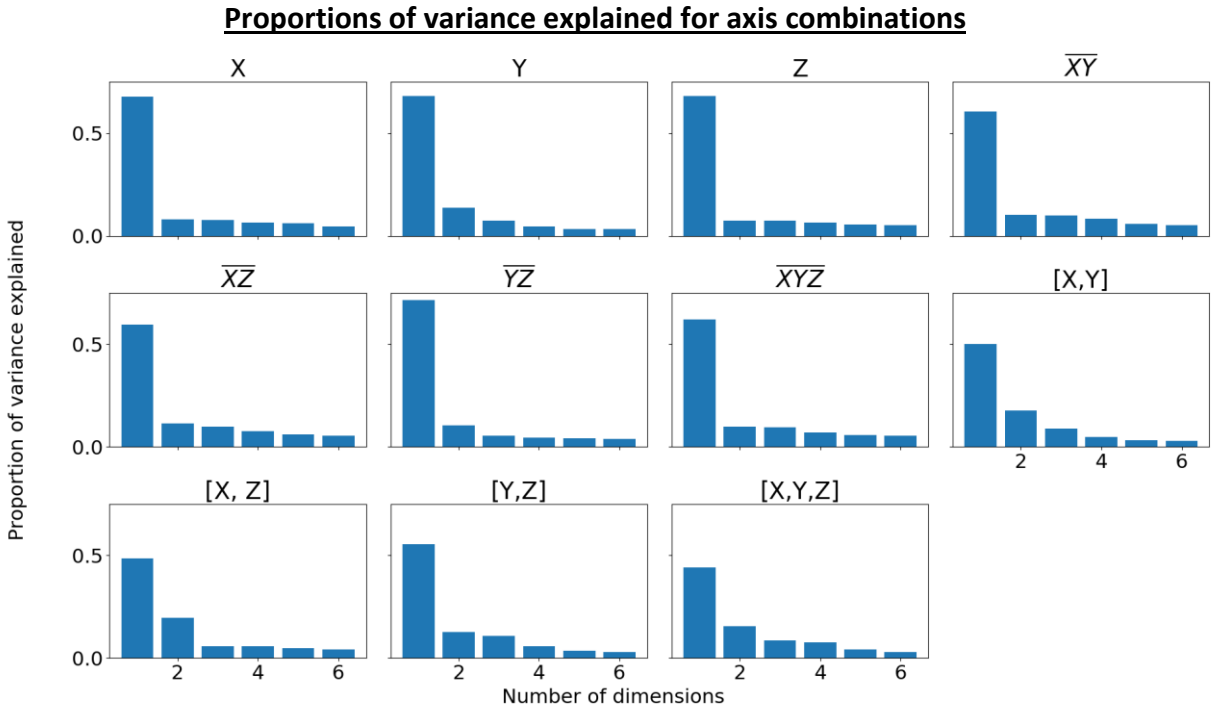

Figure 26: Image showing the proportion of variance explained for each principal component in a Principal Component Analysis of a reconstructed phase space. Where [a, b] represents using both a and b at once two-dimensional data such that $[a, b]_1 = (a_1, b_1)$.

As can be seen from Figure 26, the first dimension contains most of the variance/information for all axis combinations. Due to this, for all axis combinations, the value of the embedding dimension (em) was set to 1. Takens' theorem works if em is greater than twice the intrinsic dimension of the dynamical system generating the motion. For a person walking, swinging their arms, the intrinsic dimension is at least 3 or 4, implying that an embedding dimension of at least 7 or 9 is needed. However, it was found that the optimal embedding dimension was only one. A potential reason for this is that this method for finding the optimal dimension is only an approximation and may not identify the true optimum. It may also be that case that while the intrinsic dimension of swinging arms is 3 or 4, the limitations applied to participants in the Lab-Based activity protocol mean less of the potential dimensions are

achieved (for example, walking was done on a treadmill, so no lateral motion occurred).

7.2.5 Parameter Identification

For the activity classification, parameters such as the thresholding value (ε) used in the creation of the recurrence plot must also be identified. In order to be robust to changes in scale, this was determined as a proportion of the standard deviation of the data (C, representing the proportion of the standard deviation σ).

The optimal axis combinations and threshold values were identified by optimising them on the LabCV score, as defined in Chapter 3. The results can be seen in Table 35.

7.2.6 Comparison of RQA Features

Both RQA and Base features (those used in the Base classifier) were used for computing LabCV, FreeCV, Lab-to-Free and Free-to-Lab. Additionally, combining the RQA features and the Base into one larger set of features was also evaluated. This feature set is referred to as the Combination feature set.

7.2.7 Creating, Training And Evaluating The Classifier

The activity classifier in this chapter follows the Base classification pipeline discussed in preceding chapters, with step 4, simply modifying the features extracted from the Base (see 3.5.1) to either RQA or Combination features.

1. Determination of data type
2. Pre-processing
3. Segmenting into windows
4. Extracting features
    i) Using Base features

   ii) Using RQA features
   iii) Using Combination features
5. Creating the classification model
6. Post processing

7.3 Results

Table 35 investigates how varying the axes used and the values of C impact the performance on the Lab-Based data set. There is a low variation over all axes and values of C, suggesting that the axes used, and the values of C do not impact the performance greatly. In most cases a C value of 0.05 leads to a lower performance, except for using the $[X, Y]$ axis, where the best performance was reported with this value. No single C value achieves consistently higher results, although higher values of C to tend toward higher performances, possibly in response to the relatively high amounts of noise in acceleration data. The X axis is the poorest performing single axis, with a maximum value of 0.86, compared to the maximal single axis score of 0.88. The Y axis is the highest performing single axis with a maximal value of 0.88. Using combinations of single axes $[X, Y]$, $[X, Z]$, $[Y, Z]$ and $[X, Y, Z]$ achieved lower performances than single axis techniques with a maximal value of 0.87. Using combinations of axes by averaging achieved the highest performances with $\overline{XYZ}$ and $\overline{XY}$ reporting F1-scores of 0.89 for a range of C values.

Axis	*Threshold as a proportion of the standard deviation of the embedded signal, C:*									
	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
X	0.81	0.83	0.84	0.85	0.85	0.85	0.85	0.85	0.85	0.85
Y	0.86	0.86	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87
Z	0.81	0.82	0.83	0.84	0.84	0.84	0.85	0.85	0.85	0.85
$\overline{XY}$	0.85	0.86	0.87	0.87	0.88	0.88	0.88	0.89	0.89	0.88
$\overline{XZ}$	0.82	0.84	0.85	0.86	0.86	0.87	0.87	0.87	0.87	0.87
$\overline{YZ}$	0.84	0.85	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86
$\overline{XYZ}$	0.84	0.86	0.88	0.88	0.88	0.88	0.88	0.89	0.88	0.88
$[X, Y]$	0.87	0.86	0.86	0.86	0.86	0.85	0.84	0.84	0.83	0.83
$[X, Z]$	0.77	0.81	0.83	0.83	0.83	0.83	0.83	0.83	0.84	0.83
$[Y, Z]$	0.85	0.85	0.85	0.85	0.85	0.86	0.86	0.85	0.84	0.84
$[X, Y, Z]$	0.86	0.87	0.86	0.86	0.86	0.86	0.86	0.85	0.85	0.84

Table 35: the LabCV score for a range of threshold values and combinations of axes when classifying the Lab-Based data.

It was found that using the mean of the X, Y accelerations, denoted by $\overline{XY}$, with $\tau = 14, em = 1, C = 0.4$ achieved the highest LabCV performance. However, this performance was not significantly different to $\overline{XY}$ with a range (0.10-0.50) of other C values, or from $\overline{XYZ}$ for a range of C values (0.15-0.50).

The parameters that obtained the highest performance (C = 0.4, $\overline{XY}$), were then used in the computation of the features in the Free-Living data-set. This permitted for the computation of FreeCV, Lab-Free and Free-Lab scores, allowing comparison of the RQA and Combination features against the Base features with respect to their inter-protocol and intra-protocol performance.

Features	LabCV	FreeCV	Lab to Free-Living	Free-Living to Lab
Base	0.90 (0.10)	0.77 (0.21)	0.35 (0.13)	0.42 (0.098)
RQA	0.89 (0.076)	0.74 (0.088)	0.69 (0.086) *	0.84 (0.035) *^
Combination features	0.90 (0.13)	0.80 (0.18) *^	0.71 (0.16)*	0.73 (0.12) *

*Table 36: the performance of all features sets on all data-sets, with the standard deviation of CV performances in brackets. * indicates the performance is statistically different from the Base features. ^ * indicates the performance is statistically different from the RQA features.*

As can be seen in Table 36, the Combination features and the RQA features do not have significantly higher LabCV performances than the Base features.

In the FreeCV performances, the Base and RQA features are not significantly different. The Combination features outperform both the RQA and the Base features.

The RQA (0.69) and Combination (0.71) features significantly outperform the Base (0.35) features when evaluating the Lab-to-Free-Living data, indicating that they have a greater inter-protocol performance than the Base features.

The RQA (0.84) features significantly outperform both the Combination features (0.73) and the Base (0.42) features when evaluating the Free-Living-to-Lab data. This indicates that the RQA features have greater inter-protocol performance than the Base features.

7.4 Discussion

This chapter set out to investigate the use of RQA features for PA classification, comparing them with Base features and concentrating on their inter-protocol performance. These features were compared to current state of the art features (Base features) when classifying activity from two different data-sets, evaluating their intra and inter-protocol performances.

There was no difference in the intra-protocol performance of Base features and RQA features when trained and tested on the Free-Living data or the Lab-Based data.

RQA features outperformed Base features when trained and tested on inter-protocol data-sets (Lab-to-Free-Living / Free-Living-to-Lab). This indicates that the RQA features have a greater ability to generalise than the Base features. The ability to generalise from laboratory derived classification models to unlabelled data from differing populations is highly desirable as most acceleration studies utilise unlabelled data from a wide range of populations. Classification of activity type on unseen data based on laboratory models is associated with a reduction in performance (128), a high level of generalisability in the features should mitigate this decrease. This increased generalisation can also be seen in the lower standard deviation scores obtained with the RQA features, indicating a lower level of inter-participant variation.

When investigating the intra-protocol performance (LabCV, FreeCV), the highest performing feature set was Combination. This result is unsurprising as increasing the number of features will almost always increase performance if overfitting doesn't occur. As the classification model used in the study was a Random Forest, overfitting was unlikely (135).

When investigating the inter-protocol performance (Lab-to-Free-Living / Free-Living-to-Lab), the Combination features never significantly outperformed the RQA features. When evaluating the Free-Living to Lab-Based data, the RQA features outperformed the Combination features. As the Combination features outperformed the Base features but not the RQA features, it can be inferred that the inclusion of the Base features in the Combination features are the cause of this reduced ability to generalise.

The Lab-Based protocol only had a narrow range of activities, and these activities were typically performed in a constrained manner. Consequently,

creating a model based on this data may be expected to perform poorly on less constrained activities and a broader range of activities, regardless of features used. This is most likely why all feature sets suffered a performance drop when trained on the Lab-Based data and evaluated on the Free-Living data.

It would be expected that any activities performed in the Lab-Based data would be present in the Free-Living data, however this is not true for the converse. Therefore, any reduction in performance when evaluating Free-Living-to-Lab data is most likely due to an inability to generalise to the differing populations, rather than unknown activities. The RQA features achieve a much higher performance that the Base features in this case (0.84 against 0.42), once again suggesting they have greater ability to generalise to unseen/different participants.

A potential reason for the increased ability to generalise may be because the Base features focus on the aggregating information about the actual acceleration values in the window (mean, skewness, etc.). Whereas the RQA features represent features extracted about the morphology of the acceleration in the window. Focussing on morphology-based features as opposed to aggregate statistical features typically under-performs in other work (76). However, the main strength of morphology-based approaches appears to be the ability to generalise to unseen participants, especially when the population is different. Margarito et al (76), show that template matching (a morphology-based approach) had significantly better generalisation performance than aggregate statistical features (trained on healthy weight participants, tested on overweight participants). This result, combined with the ones presented in this work suggest that although the acceleration values may differ amongst participants, the pattern in which the values are gathered does not.

This argument is further strengthened when investigating the feature importance of the RQA features. Random Forests allow for a notion of how important each feature is in the classification model. For the Base features, the

Lab-Based and Free-Living based feature importance's correlate with $r = 0.602$, whereas the RQA feature importance's correlate with $r = 0.929$. This shows that the individual importance of Base features varies with differing populations but that the RQA feature importance remains consistent. This consistency suggests that in different populations the relationships between the RQA features are similar, hence models based on them still apply. This low correlation also holds true when using a combination of RQA and Base features.

7.4.1 Strengths and Limitations

The major limitation of this work is that the labels used in the Free-Living data only have three separate classes. The limited number of classes identified was due to gathering the true labels via an ActivPAL on the thigh which only identifies three classes. However, there does not exist a way to gather large amounts of accurately labelled Free-Living data other than direct observation, which is very time consuming and unfeasible. An additional limitation of this work was the assumption that the ActivPAL labelling was correct. While the device has been validated against direct observation (129,130), it does not achieve perfect accuracy. The high levels of agreement between the Lab-Based classifiers and the Free-Living labels do suggest some levels of correctness. Nevertheless, additional work making use of Free-Living data labelled via direct observation should be carried out to ensure the correctness of the labelling procedures for both thigh-mounted ActivPal and the RQA features derived from wrist-mounted GENEActiv.

A minor limitation comes from the fact that there are 39 different features in the Base features, whereas the RQA features only have 15. This means that the increased ability to generalise of the RQA features may just be a consequence of having less features and therefore being less likely to overfit. This argument is flawed for two reasons: firstly, the Combination features (comprising of both

Base and RQA) have more features that the Base but show a greater ability to generalise. Furthermore, it is unlikely that the classifiers are overfitting, due to the Random Forests resilience to overfitting (135).

The main strength of this work comes from the use of Free-Living data. The majority of accelerometry studies make use of only Lab-Based data (54,75), which typically does not generalise well to Free-Living (77). This inability to generalise means that methods that perform well in Lab-Based data may not perform as well in Free-Living data, therefore the effectiveness of any studies using solely Lab-Based data are questionable at best.

An additional strength of this study is that the RQA features are focussed on the morphology of the acceleration data as opposed to the actual values. As noted above this is most likely why they generalise better that the Base features, but an additional benefit is that the features are invariant to simple transformations of the acceleration data. The most common such transformation occurs if a participant wears the wrist-worn accelerometer upside down or on the opposite wrist, which has the effect of inverting all accelerations along one or more of the axes. Naturally this inversion decreases the classification performance when using Base features, but the extracted RQA features remain unchanged.

Recently, methods of automatic feature extraction have been performed on the recurrence matrices created as part of the RQA computation (156). Since methods of automatic feature extraction can achieve higher performance on activity classification (124) than traditional features, this could represent a potential extension of this work.

A natural and required extension to this work would be to repeat the experiments with more classes and a more valid labelling schema (direct observation) in order to strengthen the conclusions of this study.

7.5 Conclusion

RQA based features are simple to understand and can easily be computed from acceleration data. Classification when using RQA features is comparable to current state of the art (Base) features when tested on data similar to the training data. RQA based features showed a far greater inter-protocol performance than current state of the art features, possibly due to their focus on the underlying morphology of the acceleration as opposed to its values.

As such this method will be utilised in the final classification pipeline developed in this work. At this moment the classification pipeline is (red text indicates the addition from this chapter):

1. Determination of data type
2. Pre-processing
    2.1. ENMO extraction
    2.2. Structure Preserving Oversampling
3. Automatically segmenting acceleration data
4. Extracting RQA features
    4.1. Normalization
    4.2. Feature reduction
    4.3. Domain Adaptation
5. Creating the classification model
6. Post processing
    6.1. Participant Adaption via Iterative Relearning
    6.2. Hidden Markov Modelling
    6.3. Smoothing

8. Sparse Features

8.1 Introduction

Chapter 2 reviewed the various strategies of constructing features, highlighting that automatically extracted features may have a greater ability to generalise to unseen data. Unlabelled automatic feature extraction generates features from the data without making use of data labels, this means it can be used with unlabelled data. This means that these features can be generated on the test data as opposed to the training data.

Several methods of automatic feature extraction for activity classification have been used before. Bhattacharya (94) used Sparse Feature Encoding (SFE) to generate features from accelerometer data that could distinguish between six states (standing still, walking and travelling by tram, metro, bus and train), while Vollmer (117) utilised a similar approach with some success. Both automated approaches outperformed statistically-based features on their respective data-sets and Bhattachayra's features did not experience any substantial loss in accuracy upon the addition of previously unseen activity.

No work has investigated the effect on the inter-protocol performance of these features. It may be the case that generating the features on the test data allows for a greater inter-protocol performance that other methods.

8.2 Method

8.2.1 Data

Both the Free-Living and the Lab-Based data, as described in 3.2 were used in this chapter. Sedentary-Standing-Active labelling (as identified in 3.2.1) was

used in the Lab-Based data to ensure comparability between data-sets, this meant that both data-sets used the labels: Sedentary, Standing or Active .

8.2.2 Analysis

As Sedentary-Stand-Active Labelling was used it was possible to compute: LabCV, FreeCV, Lab-Free and Free-Lab. LabCV and FreeCV give an indication of the intra-protocol performance (how well the classification performs on data from the same protocol), while Lab-Free and Free-Lab give an idea of inter-protocol performance (how well the classification performs on data from a different protocol).To determine if there was a significant difference between the Base classification pipeline and the Sparse Features, a Wilcoxon signed-rank (133) test was used determine whether the performances were significantly different. In order to ensure a large enough sample size, the comparisons were paired on each participant's performance, instead of the average. Due to the fact the multiple hypotheses were evaluated on the same data set, the likelihood of a Type I error is increased. This was compensated for by using Bonferroni corrections. This entails testing each individual hypothesis at a significance level of $\frac{\alpha}{m}$, where α is the overall hypothesis level (in this case 0.05) and *m* is the number of hypotheses. A p-value under $\frac{\alpha}{m}$ indicates that the results are statistically significantly different from one another with high confidence.

8.2.3 Sparse Feature Encoding

Sparse Feature Encoding is a method of automatically extracting features from accelerometry that operates in the following way:

8.2.3.1 Input Data

Sparse Feature Encoding takes as input data the raw acceleration, segmented into windows. Unlike statistical and morphology-based features, Sparse Feature Encoding cannot be run on each window as it occurs but instead requires all the training acceleration data at once, thus precluding it from online training but not online classification. This data-set will be referred to as Acc.

8.2.3.2 Sparse Basis Vectors

Sparse Feature Encoding works by identifying basis vectors of the acceleration data-set, Acc. Basis vectors are vectors that can recreate a data-set through linear combinations, (adding multiples of the vectors together). For example, $[0,1]$ and $[1,0]$ are basis vectors of $[\text{Ь}, \text{Ч}]\ \forall\ \text{Ь}, \text{Ч}\ \in \mathbb{R}^2$.

In this case, basis vectors can be thought of as simple acceleration time series that can be combined to recreate the acceleration times series from the data.

Sparsity refers to ensuring that for all signals that must be recreated, only a few of the basis vectors need to be activated (used) in the reconstruction.

Sparse Basis Vectors are computed from Acc using the formula:

$$\min_{\text{Ѧ},\Omega} \left(recon(\text{Ѧ}, \Omega, Acc) + ₪\left|\left|a_i\right|\right| \right)$$

$$recon(\text{Ѧ}, \Omega, Acc) = \sum_{\text{i}=1}^{\text{fs}} \left\| \sum_{j=1}^{fs} \Omega_j \ * \ a_j^i - \text{Acc}_\text{i} \right\|_2^2$$

$$\text{subject to } \left|\left|\Omega_\text{j}\right|\right|_2 < 1\ \forall \text{j}$$

Equation 9: Sparse Basis Vector computation.

Where Ω, is the dictionary of filters, $\mathbb{A} = (a_1, \dots a_{fs})$ represents the activations of the filters. fs refers to the number of basis vectors used when recreating the data. ₪ is the sparsity coefficient.

Typically, the value of fs is the same size as the dimension of the data (in this case 1280 (100*12.8)), however it is possible to use more than the dimension of the data-set. This is referred to as an over-complete dictionary. Using an over-complete dictionary allows for increased resistance to noise (94). Bhattacharya (94) identifies the value of fs in his work by computing how the value of fs impacts the reconstruction error of the filters. The value of fs that has the lowest reconstruction error is used. This method was also used in this work, see Figure 27. It was found that using 640 filters allowed for the lowest reconstruction error.

$recon(\mathbb{A}, \Omega, Acc)$, is a function that computed how well the identified basis vectors can recreate the data, this ends up being a trade-off between maintaining the sparsity of the vectors and allowing for a lower reconstruction error (the difference between the original data set and the reconstructed data set from the dictionary). Due to the constraint of sparsity it may not be possible that the basis vectors can recreate the signal perfectly, see Figure 27.

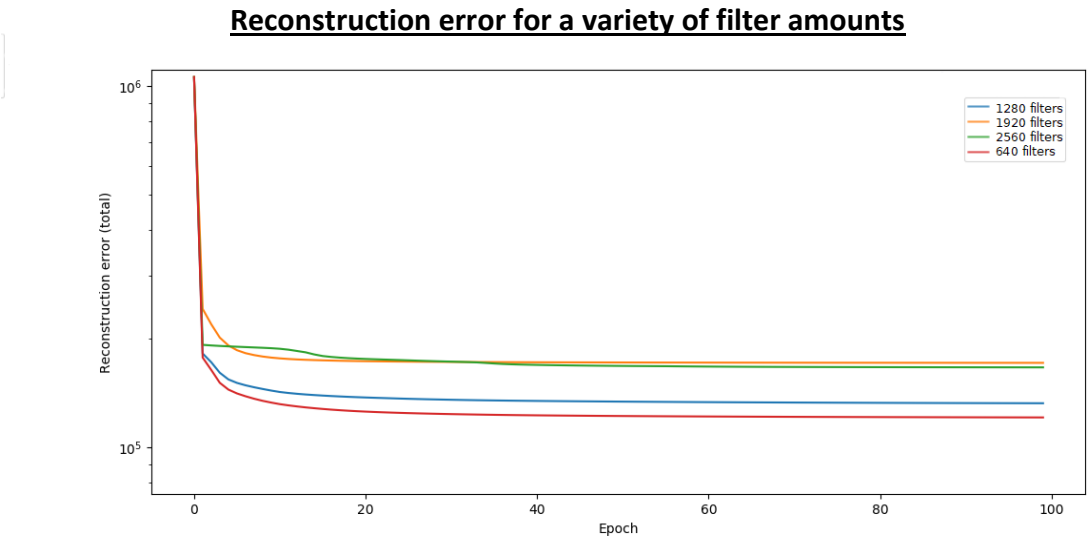

Figure 27: Reconstruction error of varying amounts of filters.

The value of ₪ determines the importance of sparsity in the reconstruction. Setting ₪ to 0 allows for a reconstruction error of 0, whereas a higher value will prioritise sparse activations over ones that can recreate the data with high fidelity.

8.2.3.3 Feature Extraction

The features generated through Sparse Feature Encoding correspond to which filters must be activated in order to recreate the test data. For each test data window tw_I the optimal activation vector $\hat{a}^i$ is computed with the following equation.

$$\hat{a}^i = \arg\min_{a_i} \left\| tw_i - \sum_{j=1}^{fs} \Omega_j * a_j^i \right\|_2^2 + ₪\|a_i\|$$

Equation 10: Sparse Basis Vector Activation.

The vector $\hat{a}^i$ is the feature vector used in the classification process to identify the activity that tw_i corresponds to.

8.2.4 Parameter Identification

When using Sparse Encoding features, it is required to identify the value of ₪.

The value of ₪ is found through optimising the LabCV score.

8.2.5 Comparison Sparse Feature Encoding Features

In order to evaluate the effect of using Sparse Feature Encoding features in classification, their performance and ability to generalise were compared to the Base classifier (Chapter 3), using state of the art features (Base features). Sparse Feature Encoding is an unsupervised method of automatic feature extraction; therefore, it does not require the labels in order to generate the features. As the labels are not required, the features can be generated on the test data and the training data separately. Both methods of generating the features will be tested, Sparse Feature Encoding on the training data is denoted as (train_SFE), this refers to using Sparse Feature Encoding on the training data to identify features, these features are then extracted on the training data and used to create a classifier which attempts to classify the test data. Sparse Feature Encoding of the test data (test_SFE) refers to generating the features from the testing data, then extracting these features from the training data and using this to create a classifier. This classifier then attempts to classify the test data. In both cases the classifiers are trained on the training data first, but the features used are identified from the training or testing data respectively for train_SFE and test_SFE.

This method will also be tested against RQA features.

8.2.6 Creating, Training And Evaluating The Classifier

The activity classifier in this chapter follows the Base classifier pipeline discussed in preceding chapters, with step 4, simply modifying the features (See 3.5.1) to Sparse Feature Encoding.

1. Determination of data type
2. Pre-processing
3. Segmenting into windows
4. Extracting features
    4.1 Using Base features (Base method)
    4.2 Train_SFE
    4.3 Test_SFE
    4.4 RQA
5. Creating the classification model
6. Post-processing

8.3 Results

Table 37 shows the LabCV performance of the test_SFE for various values of ₪. The highest performance is from an ₪ of 0.001. The different values of ₪ do not greatly change the performance; with a range of performance from 0.765 to 0.802. The Wilcoxon-signed-rank test identifies that only the value of $₪ = 1$ results in a statistically different performance to the other values. The highest performing value of ₪ is 0.001 (although this is non-significant), this value was used to compute features in both the Lab and Free-Living data-sets using both train_SFE and test_SFE in order to analyse the FreeCV, LabCV, Free-to-Lab and Lab-to-Free performances for all feature sets.

Value of ₪	*LabCV F1-Score*
1.000000	0.765* (0.102)
0.100000	0.801 (0.122)
0.010000	0.800 (0.124)
0.001000	0.802 (0.122)
0.000100	0.801 (0.113)
0.000010	0.789 (0.149)

*Table 37: LabCV Performance of test_SFE for values of ₪, * indicates significance in the Wilcoxon-signed rank test. Figures in brackets indicate standard deviations.*

Features	*LabCV*	*FreeCV*	*Lab to Free-Living*	*Free-Living to lab*
Base	0.898 (0.103)	0.765 (0.214)	0.352 (0.132)	0.415 (0.0978)
Train_SFE	0.844 (0.127)	0.772 (0.273)	0.601* (0.310)	0.421 (0.164)
Test_SFE	0.802* (0.149)	0.721* (0.190)	0.667* (0.287)	0.645* (0.219)
RQA	0.890* (0.0759)	0.743 (0.0882)	0.691^ (0.0857)	0.844^ (0.0354)

*Table 38: the performance of all feature sets on all data-sets. * indicates significantly different classification performance than the Base features. ^ indicates significantly different classification performance than the highest performing non-RQA method. Figures in brackets indicate standard deviations.*

Table 38 shows the performance of all features over all data-sets. The test_SFE features have a significantly lower intra-protocol performance (LabCV, FreeCV) than the Base features. The inter-protocol performance of test_SFE features is significantly greater than the Base features and the train_SFE features.

The inter-protocol performance using the train_SFE features is only significantly higher than the inter-protocol performance of the Base features for the Lab to Free-Living.

The RQA features outperform both the test_SFE and train_SFE features with respect to inter-protocol performance and are not significantly worse with respect to intra-protocol performance.

8.4 Discussion

This chapter set out to investigate the use of Sparse Feature Encoding features for PA classification, comparing them to state of the art features (see section 3.5.1) and concentrating on inter-protocol performance.

Train_SFE outperformed the test_SFE with respect to the intra-protocol performance. This is most likely because the test_SFE extracted features on data from only person (the left-out participant), whereas the train_SFE had more data (all participants except the left out) to extract features from.

The test_SFE showed a significantly greater inter-protocol performance than the Base features. This is unsurprising as the features were extracted over the test data, therefore allowing for a higher performance of that data-set. However, the increased inter-protocol performance is at the cost of a reduced intra-protocol performance.

The increased ability to generalise at the expense of the intra-protocol performance is typical with morphology-based features (Sparse Feature Encoding features and RQA features). It appears to be from a reduced ability to represent the data. The lowered representative power reduces the performance but also decreases any overfitting that may have occurred, thereby allowing for a greater ability to generalise.

It was found that the inter and intra-protocol performance is affected by the data used to generate the Sparse Feature Encoding features. Sparse Feature Encoding does not require labels in order to generate features, because of this the test data could be used to generate the features instead of the training data (test_SFE). This allowed for a greater inter-protocol performance at the cost of the intra-protocol performance.

The first three filters extracted from performing Sparse Feature Encoding on the Free-Living data are completely straight lines (see Figure 28), which

corresponds to multiplying the accelerations by a set factor across the window. Curiously in the work by Bhattacharya (94), all but one of these would have been removed due to their high temporal correlation. Filters 1-3 are interesting due to both their fixed values and the fact that for each of the filters, one axis from X, Y or Z is set to 0. This means that through a linear combination of these three vectors, any constant windows (constant in X, Y, Z respectively) may be created. The fact that they act as constants that can be applied to the signals may explain their prevalence in the data. The next two filters (4-5) are reminiscent of sine waves at different frequencies. This would correspond to a Fourier decomposition of the sequence, much akin to some of the statistical features used by Bao et al (54).

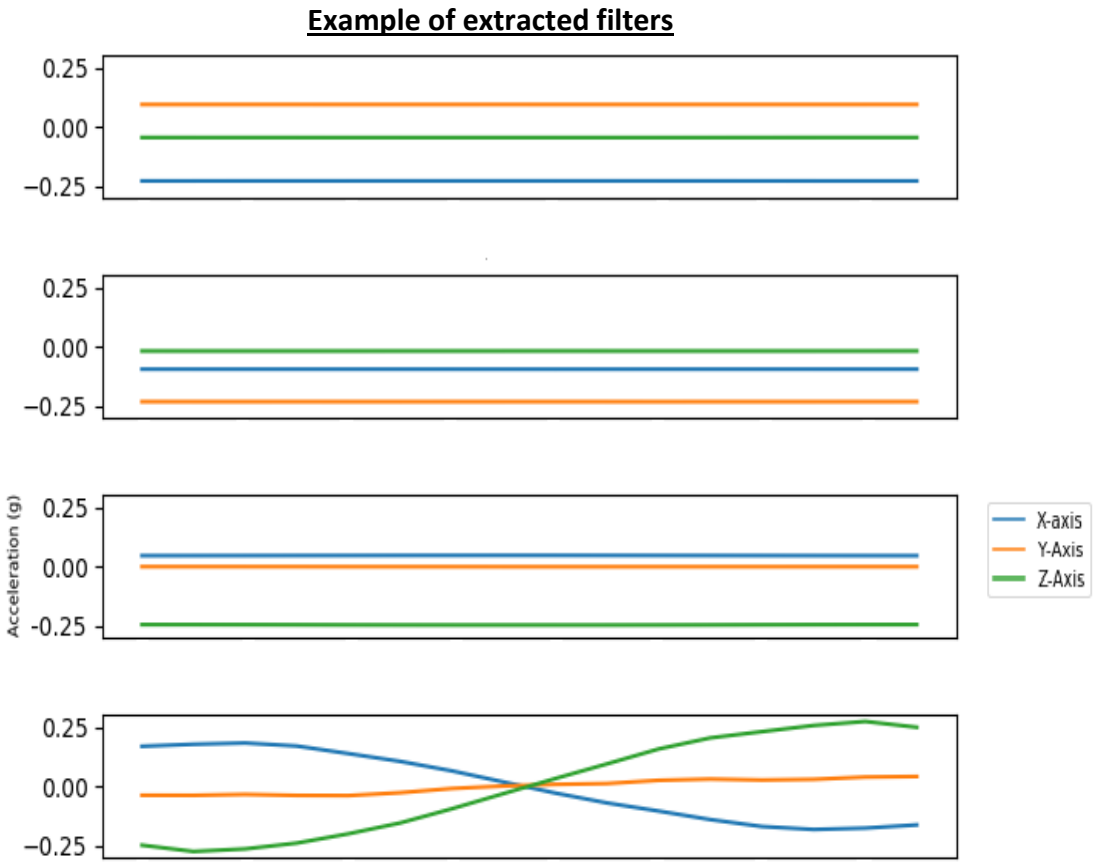

Figure 28: Filters extracted from data.

8.4.1 Strength and Limitations

A major limitation of this work is the computational cost of generating the dictionary of filters. In other work this is less of an issue as the features are created on the training data, therefore they are only created once and can be created prior to the classification. This work however created the features on the test data, to allow for a higher inter-protocol performance. Therefore, the features must be recalculated for every new participant.

A more applicable approach may be to make use of the work of Coates and Ng (157), who use a KMeans model to identify features in a way very similar to Sparse Feature Encoding, Although, they do state, *"Empirically though, sparse coding appears to be a better performer in many applications. The advantage however of the KMeans approach is twofold: it requires no parameter tuning, unlike Sparse Feature Encoding and it is substantially faster and more scalable."*

The main strength of this work is the use of Free-Living data and the focus on inter-protocol performance. Most approaches focus solely on Lab-Based data which typically do not produce results that generalise to Free-Living scenarios.

An extension of this work would be to combine the Free-Living and Lab-Based data-sets together and use Sparse Feature Encoding on the both data-sets at once as opposed to only one at a time. This was not done due to the extreme computational costs this would have evoked.

8.5 Conclusion

This chapter focused on automatically generated features which may improve the classification performance with respect to both the inter and intra-protocol performance. While Sparse Feature Encoding features has greater inter-

protocol performance than state of the art, this was mitigated by the lower intra-protocol performances.

RQA features discussed in Chapter 7 outperformed Base features and Sparse Feature Encoding features. Hence the Sparse Feature Encoding features will not be used in this thesis.

The classification pipeline remains as:

1. Determination of data type
2. Pre-processing
    2.1. ENMO extraction
    2.2. Structure Preserving Oversampling
3. Automatically segmenting acceleration data
4. Extracting RQA features
    4.1. Normalization
    4.2. Feature reduction
    4.3. Domain Adaptation
5. Creating the classification model
6. Post processing
    6.1. Participant Adaption via Iterative Relearning
    6.2. Hidden Markov Modelling
    6.3. Smoothing

9. Classifiers Used

9.1 Introduction

This chapter focuses on evaluation of classifiers in the pursuit of high inter-protocol performance.

In Chapter 2 it was noted that there is no consensus about the optimal form of classifiers used in activity classification pipelines. It was observed that although some studies have tested multiple classifiers against each other, no such test of the ability to generalise have been carried out. Furthermore, it was identified that two major classes of classifiers exist in the activity classification literature: Generative and Discriminative.

When given observable variables D (input data, acceleration features) and target variable T (output labels, activity labels), a classifier that models the joint distribution is a Generative classifier (122); one that models the conditional distribution $p(T \mid D)$ is Discriminative (69). However, a Discriminative classifier may also just produce a score or a class label without actually modelling or producing a probability.

When attempting to maximise classification performance, the use of Discriminative classifiers is typically recommended, "*one should solve the [classification] problem directly and never solve a more general problem as an intermediate step*" (123). However, when the amount of data is limited, Generative classifiers have been shown to be preferred. Generative classifiers reach a high level of performance with logarithmically lower amounts of training data (122). As the cost of gathering training data for activity classification is so high, this is naturally advantageous.

The issue of inter-protocol performance has been addressed many times in this thesis, being related to a well-known concept of the “bias-variance trade-off”. This refers to attempting to minimise two sources of error in supervised learning. The bias refers to the ability of the model to capture the relationship between the training data and labels. Whereas the variance refers to the fluctuation in the relationship caused by minor changes in the training data. A low bias but high variance will allow for a high intra-protocol performance at the cost of a low inter-protocol performance, whereas a high variance and low bias will have a high inter-protocol performance but low intra-protocol performance. Obviously, both a low bias and a low variance is desired. It is known that Generative classifiers typically have a lower variance than Discriminative counterparts (122). As the focus of this thesis is to ensure a high inter-protocol performance while maintaining a high intra-protocol performance, investigating the effects of Discriminative and Generative classifiers on activity classification may be useful.

As such, this chapter will identify the effects on the intra and inter-protocol performance of classification pipelines using different (Discriminative/Generative) classifiers.

9.2 Method

9.2.1 Data

Both the Free-Living and the Lab-Based data, as described in 3.2 were used in this chapter. Sedentary-Standing-Active labelling (as identified in 3.2.1) was used in the Lab-Based data to ensure comparability between data-sets, this meant that both data-sets used the labels: Sedentary, Standing or Active.

9.2.2 Analysis

As Sedentary-Stand-Active Labelling was used it was possible to compute: LabCV, FreeCV, Lab-Free and Free-Lab. LabCV and FreeCV give an indication of the intra-protocol performance (how well the classification performs on data from the same protocol), while Lab-Free and Free-Lab give an idea of inter-protocol performance (how well the classification performs on data from a different protocol).To determine if there was a significant difference between the Base classification pipeline and the pipeline making use of the different classifier, a Wilcoxon signed-rank (133) test was used determine whether the performances were significantly different. In order to ensure a large enough sample size, the comparisons were paired on each participant's performance, instead of the average. Due to the fact the multiple hypotheses were evaluated on the same data set, the likelihood of a Type I error is increased. This was compensated for by using Bonferroni corrections. This entails testing each individual hypothesis at a significance level of $\frac{\alpha}{m}$, where α is the overall hypothesis level (in this case 0.05) and *m* is the number of hypotheses. A p-value under $\frac{\alpha}{m}$ indicates that the results are statistically significantly different from one another with high confidence.

9.2.3 Classification Procedure

The activity classifier in this chapter follows the Base classification pipeline discussed in preceding chapters, replacing the classifier in stage 5, with one of six classifiers (3 Discriminative, 3 Generative)

1. Determination of data type
2. Pre-processing
3. Segmenting into windows
4. Extracting features
5. Creating the classifier

5.1. Naive Bayes (Generative)
5.2. Logistic regression (Discriminative)
5.3. Random Forest (Discriminative)
5.4. Quadratic discriminant analysis (Generative)
5.5. Neural Network (Discriminative)
5.6. Generative adversarial model (Generative)

6. Post processing

The Random Forest classifier refers to the Base classifier as defined in Chapter 3. Additionally, the average (mean) of the Generative classifiers and Discriminative classifiers will be computed, referred to as the Generative mean and Discriminative mean respectively.

9.2.4 Generative-Discriminative pairs

It is the case that classifiers can exist in Generative and Discriminative 'forms' when the underlying model is the same; these are referred to as 'Generative-Discriminative pairs'. Naïve Bayes and Logistic regression use linear separation of the data for classification, with methodology being Generative and one being Discriminative, hence they represent a Generative-Discriminative pair. This pair will be closely examined as it represents a way to identify how Generative/Discriminative classifiers effect the inter and intra-protocol performance when the underlying function is consistent, therefore not affecting performance.

9.3 Classifiers

Six classifiers will be tested in this work, three Discriminative and three Generative. The performance and ability to generalise will be compared against the Discriminative and Generative models.

9.3.1 Naive Bayes

9.3.1.1 Methodology

Naive Bayes (69,158,159) classifiers make use of conditional probability. When given an input $D = (d_1, d_2, \dots, d_n)$, the instance probability of this input coming from each potential class is computed and the class with the highest probability is chosen. This is known as the maximum a posteriori (MAP) rule.

This method has seen much use in this field, although never reaching high levels of performance (65,85).

9.3.1.2 Parameters

When creating the Naïve Bayes classifier, the prior probabilities of the classes are used. These are derived from the class proportions reported in Chapter 3.

9.3.2 Logistic Regression

9.3.2.1 Methodology

Logistic regression can be thought of as attempting to model the conditional probability $p(T|d_1, d_2, \dots, d_n)$ directly, by making use of linear regression. This method has seen much use in this field, although never reaching high levels of performance (160–162).

9.3.2.2 Parameters

A penalty value is chosen which identifies the cost of errors in the classification process. This value was set to one, as this is the standard default value.

9.3.3 Quadratic Discriminant Analysis

9.3.3.1 Methodology

Quadratic discriminant analysis is a Generative classification method. Quadratic discriminant analysis simply models the class conditional distribution of the data. This method assumes that all features are normally distributed, which greatly decreases the number of parameters required to be learned in the classification training.

As with Naïve Bayes, this method attempts to maximise the probability over the conditional

$$\underset{cl\epsilon\{1,\ldots,Class\}}{argmax}\ (p(T_{cl}|D)).$$

Equation 11: Probability maximisation in QDA.

This method has seen some use in activity classification and measuring energy expenditure (163,164).

9.3.3.2 Parameters

Like Naïve Bayes, this method makes use of the class probabilities, as computed in Chapter 3.

9.3.4 Neural Networks

9.3.4.1 Methodology

Neural Networks are a Discriminative classifier that attempts to directly model the decision boundary between classes. A Neural Network can be viewed as a network of nodes, with an input layer, an output layer and hidden layer(s). Each node has an input, output and an activation function.

In the input layer, the inputs correspond to the features used in the classification process. In the other layers (hidden or output) the inputs are weighted sums of the outputs of connected nodes.

When trained a Neural Network identifies weights for each neuron that best maps the inputs (acceleration features) to the activity classes. The weights of the neurons are continuously adjusted over many epochs, gradually improving the ability to correctly classify the data. Neural Networks represent one of the highest performing methods of activity classification in the literature, often outperforming other models (85,124,165).

9.3.4.2 Parameters

The structure of the Neural Network of this work was a Multi-Layered Perceptron making use of:

- Input layer: 39 Inputs (corresponding to features)
- Hidden layer with 25 nodes
- Hidden layer with 12 nodes
- Output layer with 1 output (probability of sample being from each class)

The training process ran for 300 iterations, with a batch size of 512, enough to ensure that the performance of the network had stabilised.

The Activation function (the function applied to the weighted sum of the node inputs to determine the output of a node) used was the rectifier function, $Rect(\omega) = \max(0, \omega)$. This is a function that has seen success in many domains. The Learning rate was 0.01, with a decay and momentum of 0.01 and 1.0 respectively. These parameters relate to the speed at which the node weights are updated. The algorithm used to determine the optimal values for the node weights was the Adam optimisation algorithm, a popular algorithm that

has reported success in many domains. These values were chosen through optimisation on previously gathered acceleration data distinct from the data used in this thesis.

9.3.5 Generative Adversarial Networks

9.3.5.1 Methodology

Generative Adversarial Networks are an extension of Neural Networks. Unlike Quadratic Discriminative Analysis and Naïve Bayes, this method is not a true Generative method, as it only seeks to discriminate between the classes and does not compute the joint probabilities. However, this method shares much in common with Generative approaches and represents one of the highest performing methods in Machine Learning (166).

Generative Adversarial Networks are classifiers that consists of two Neural Networks, a generator and a discriminator. The generator learns to create synthetic data that is indistinguishable from real data (the training data); the discriminator attempts to identify if data is real or synthetic. By playing the two networks off against each other (training the generator on the error function of the discriminator) it greatly improves the ability to the discriminator to identify if an input is real, compared to just using the training data.

For each class, a Generative Adversarial Network can be trained, resulting in a discriminator for each class that can output a predicted probability that a sample belongs to the class. The class that has the highest probability is then chosen.

9.3.5.2 Parameters

The parameters used by the Generative Adversarial Network are the same as the Neural Network discussed above, the Generative Network has the input and output layers swapped, however.

9.4 Results

As can be seen in Table 39, the Generative Adversarial Network classifier achieves the highest score for all methods. For the LabCV the Generative Adversarial Network achieves an F1-score of 0.902, although this is statistically equivalent to both the Neural Network and the Base classifier scores. The other classifiers achieve scores that are significantly lower.

Model	*LabCV*	*FreeCV*	*Lab-Free*	*Free-Lab*
Naïve Bayes	0.823*	0.736	0.379	0.424
	(0.108)	(0.204)	(0.109)	(0.141)
Logistic regression	0.858*	0.784	0.398	0.402
	(0.111)	(0.220)	(0.128)	(0.113)
Quadratic discriminant	0.871*	0.756	0.339	0.331
analysis	(0.0908)	(0.216)	(0.153)	(0.107)
Random Forest	0.898	0.765	0.352	0.415
	(0.103)	(0.214)	(0.132)	(0.0978)
Neural Network	0.885	0.798*	0.415*	0.217*
	(0.0999)	(0.198)	(0.167)	(0.211)
Generative Adversarial	0.902	0.831*	0.507*	0.628*
Networks	(0.109)	(0.163)	(0.175)	(0.202)
Generative mean	0.865	0.774	0.409	0.461
	(0.100)	(0.202)	(0.149)	(0.103)
Discriminative mean	0.880	0.782	0.388	0.345
	(0.102)	(0.210)	(0.145)	(0.122)

*Table 39: Inter and intra-protocol performance of 6 classifiers. * indicates the score is significantly different from the Base classifier. Figures in brackets indicate standard deviations.*

The Generative Adversarial Network achieves the highest FreeCV score (0.831), this is significantly different from the Base classifier (0.765); it is also significantly different from the next highest achieving method (Neural Network). However, no other methods are significantly different from the Base classifier.

The highest achieving classifier on the Lab-Free data is the Generative Adversarial Network (0.507) compared to the Base classifier (0.352). The next highest score is from the Neural Network with a score of 0.415. Both the Generative Adversarial Network and the Neural Network achieve scores that are significantly higher than the Base classifier.

The highest Free-Lab score is from the Generative Adversarial Network (0.628), this is far higher than the Base classifier (0.415). This score is significantly higher than all other classifiers. The Neural Network achieves a score of 0.217 in the Free-Lab domain. This is significantly lower than all other classifiers.

The Generative classifiers have higher inter-protocol performances (0.409, 0.461 vs 0.388, 0.345) than the Discriminative classifiers, but lower intra-protocol performances (0.865, 0.774 vs 0.880, 0.782). However only the difference in the Free-Lab score is significantly different.

9.5 Discussion

The aim of this chapter was to investigate different classifiers and their effects on the intra and inter-protocol performance for activity classification. It also tested whether Generative classifiers had a lower difference between the intra-protocol performance (LabCV and FreeCV) and the inter-protocol performance (Lab-to-Free, Free-to-Lab) due to their hypothesised greater ability to generalise to different protocols.

The results show that different classifiers do have significantly different inter and intra-protocol performances.

The Generative Adversarial Network achieved the highest scores in all four domains, also achieving the lowest inter-intra protocol difference. This indicates that this is the best classifier to use in this work. The LabCV was not significantly higher than the Base classifier, but all other scores were

significantly greater. Generative Adversarial Networks are one of the best performing methods in a variety of domains (166), so this high performance is consistent with current results. However, this is the first time a Generative Adversarial Network has been used in activity classification, so no direct comparisons are available.

The Neural Network showed a high level of intra-protocol performance, but a significantly lower inter-protocol performance. This indicates that the classifier may be overfitting to the individual protocols. This is a common issue in Neural Networks (85). As the Generative Adversarial Network uses the same structure for the discriminator as the Neural Network approach, it is likely that the Generative Adversarial Network avoidance of overfitting is due to its generative nature rather than the model being able to avoid overfitting. However, the Naive Bayes classifier and the Logistic regression classifier which represent a Generative-Discriminative pair did not achieve significantly different results in any domain. This suggests that neither a generative or discriminative nature is particularly beneficial in this work and that any increase in performance may be due to the greater representational power of the specific model rather than its underlying class.

The variance of intra-protocol (LabCV and FreeCV) scores, across all classifiers, are much lower (0.027, 0.031) than the variance of inter-protocol (Lab-Free and Free-lab) scores (0.055, 0.12). This may be because the intra-protocol performances are reaching the maximum performance possible, therefore different classifiers are unlikely to have much effect.

9.5.1 Strengths and Limitations

Only six classifiers were used in this work; there exists a surfeit of classifiers that have been used in activity classification, with many additional classifiers in other domains. It was not possible to test every possible classifier. However, it may have been that the classifiers tested in this work were not optimal for

activity classification. As such, a potential extension for this work would be to explore different classifiers.

A limitation of this work is that only two data-sets were used, gathered over two protocols. It may not be the case that the performances remain consistent amongst different protocols. An additional limitation was only one choice of parameter was made for the activity classifiers. For example, a Neural Network has many different parameters that can affect its performance (number of hidden layers, number of training epochs, optimiser function and training rate). In this work only one set of parameters were used for each model. It may be the case that modifying these parameters may allow for a different level of intra and inter-protocol performance for each method, thus modifying the findings of this work.

A potential extension of this work would be to use an ensemble classifier. It is possible to aggregate multiple classifiers into one ensemble classifier that has the advantages of all methods. Typically, these ensemble classifiers can outperform any single classifier. While Random Forests are already ensembles on decision trees, it is possible to have ensembles of ensembles.

One of the potential benefits of generative classifiers is that they reach a high level of performance with logarithmically lower amounts of training data. As such, varying the amounts of training data would be an interesting addition to this work. This would help to identify if the current amount of training data is enough for achieving optimal performance. Also testing performances with decreased amount of training data would have informed whether generative classifiers do perform better with lower amount of training data in activity classification, as hypothesised. This would help inform additional studies of the optimal classifier to use when training data is less available.

The Neural Network and Generative Adversarial Network used in this study, made use of the same features as Base classifier. One of the main strengths of

Neural Networks is that they can also perform automatic feature extraction, using Convolutional Layers. A very high level of performance in a variety of domains (including activity classification) has been found when making use of these layers (141). Therefore, a potential next step would be to repeat this process making use of automatic feature extraction methods, instead of using statistically derived features. This was not done in this work, because the aim was to identify which classifier was best when using the same features.

9.6 Conclusion

In conclusion, the classifier that allows of the highest level of inter and intra-protocol performance is a Generative Adversarial Network. As such, this will be used in the classification pipeline. The classification pipeline is now (red text indicates the addition from this chapter):

1. Determination of data type
2. Pre-processing
    2.1. ENMO extraction
    2.2. Structure Preserving Oversampling
3. Automatically segmenting acceleration data
4. Extracting RQA features
    4.1. Normalization
    4.2. Feature reduction
    4.3. Domain Adaptation
5. Creating the classification model, a Generative Adversarial Network
6. Post processing
    6.1. Participant Adaption via Iterative Relearning
    6.2. Hidden Markov Modelling
    6.3. Smoothing

This represents the final classification pipeline used in this work.

10. The Final Classification

10.1 Introduction

In the previous chapters, various issues relating to the maintenance of a high inter-protocol performance without decreasing the intra-protocol performance have been dealt with, resulting in an activity classification pipeline with a high intra AND inter-protocol performance. In Chapter 3 a data-set was identified that consisted of unlabelled acceleration data, called the assessment data. In this chapter the activity classifier developed in this thesis is used to predict PA from this assessment data as an example of how the classifier may be used.

10.2 The Final Classification Pipeline

Chapter 3 identified a classification pipeline that was based on the current state of the art research, this is referred to as the Base classifier:

1. Determination of data type
2. Pre-processing: None
3. Windowing: Using 12.8-second windows
4. Feature extraction: Using 39 features based on the statistical aggregate features and frequency statistics.
5. Building the classification model: Using a Random Forest classifier with 50 separate trees.
6. Post-processing: None

This classification pipeline had high intra-protocol performance, but a limited ability to classify acceleration data gathered from a different protocol (inter-protocol performance). This inability to perform well on data from a different

protocol is one of the largest flaws with activity classification (68) and the overall aim of this thesis was to develop methodologies to mitigate this.

Chapter 2 identified various methodological gaps in the research that resulted in a lowered inter-protocol performance. Chapters 4-10 developed various methodologies that deal with these gaps, resulting in a modified classification pipeline that had a greater ability to perform on data from different protocols. The modified classification pipeline can be seen below, henceforth this is referred to as the final classification pipeline, or the final classifier.

1. Determination of data type
2. Pre-processing
    2.1. ENMO extraction
    2.2. Structure Preserving Oversampling
3. Automatically segmenting acceleration data
4. Extracting RQA features
    4.1. Normalization
    4.2. Feature reduction
    4.3. Domain Adaptation
5. Creating the classification model, a Generative Adversarial Network
6. Post processing
    6.1. Participant Adaption via Iterative Relearning
    6.2. Hidden Markov Modelling
    6.3. Smoothing

10.2.1 Data

Both the Free-Living and the Lab-Based data, as described in 3.2 were used in this chapter. Sedentary-Standing-Active labelling (as identified in 3.2.1) was used in the Lab-Based data to ensure comparability between data-sets, this meant that both data-sets used the labels: Sedentary, Standing or Active.

10.2.2 Analysis

As Sedentary-Stand-Active Labelling was used it was possible to compute: LabCV, FreeCV, Lab-Free and Free-Lab. LabCV and FreeCV give an indication of the intra-protocol performance (how well the classification performs on data from the same protocol), while Lab-Free and Free-Lab give an idea of inter-protocol performance (how well the classification performs on data from a different protocol). To determine if there was a significant difference between the Base classification pipeline and final classification pipeline, a Wilcoxon signed-rank (133) test was used determine whether the performances were significantly different. In order to ensure a large enough sample size, the comparisons were paired on each participant's performance, instead of the average.

10.3 Results

As can be seen in Table 40, there is no statistical difference between the two classifiers with respect to their intra-protocol performance. However, the final classifier has a significantly better inter-protocol performance than the Base classifier. This means that the final classifier is more likely to correctly classify acceleration data when it is gathered from a different protocol than it was trained with.

	LabCV	*FreeCV*	*Lab-Free*	*Free-Lab*
Base classifier	0.898 (0.103)	0.765 (0.214)	0.352 (0.132)	0.415 (0.0978)
Final classifier	0.916 (0.0980)	0.826 (0.154)	0.759* (0.100)	0.897* (0.106)

*Table 40: The F1-score for the Base classifier and the final classifier, over four different domains, testing both the intra-protocol performance and the inter-protocol performance. A * indicates that the result was significantly different. Figures in brackets indicate standard deviations.*

10.4 Assessment Data Investigation

The final classifier was created making use of both the Free-Living and the Lab-Based data and the final classification pipeline. The classifier was then applied to the assessment data, allowing for the creation of a time series of PA events for each participant. These will now be examined.

10.4.1 Proportion of Time Spent in Each Activity

Figure 29 shows the proportion of time spent in each activity: 71.9% of the time is spent in Sedentary activities, 20.3 % is spent in Standing activities and 7.8% is spent in Active.

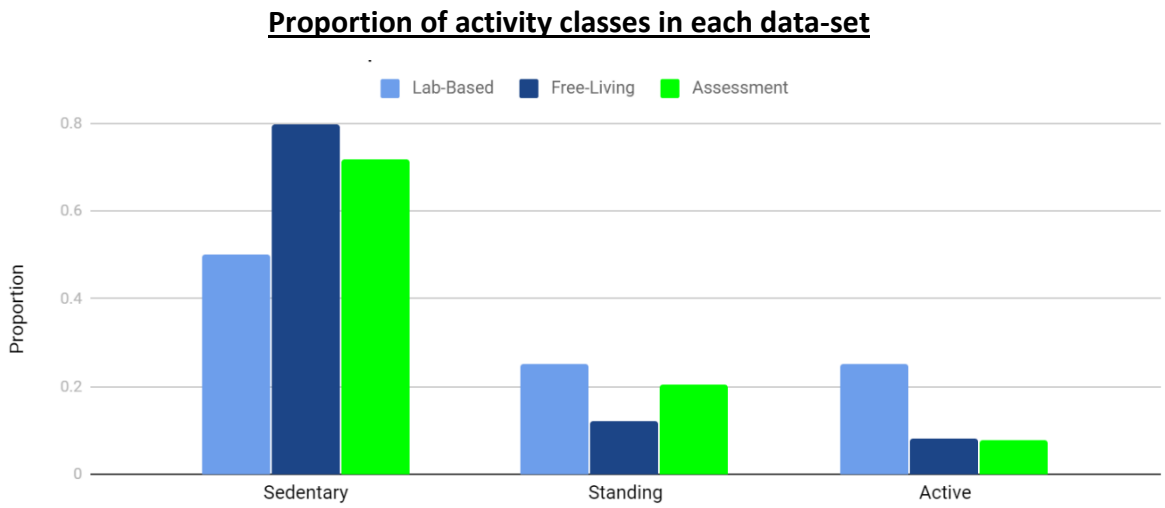

Figure 29: Proportion of each class in assessment data.

The Free-Living data identified in Chapter 3, has very similar proportions; as 8% of activities are Active, 80% are Sedentary and 12% are Standing. The assessment data has almost equal proportions of activities labelled as Active as the Free-Living data-set. However, the assessment data has more Standing activities, replacing Sedentary activities. This is strange as the participants in the assessment data are much older than that of the Free-Living data and it

might be expected that the amount of PA would decrease with age. Therefore, it might be expected that the assessment data would have a larger proportion of time spent in Sedentary with less activity. However, this only considers type of activity, not volume.

The majority of studies that identify PA stratified by age make use of intensity estimates from acceleration thresholds, whereas this work does not. It may be that older participants perform the same duration of activities labelled as Active, but at a lower intensity. This would result in the findings seen here. The lowered intensity would mean that intensity-based thresholds do not characterise the activities as Active, therefore the PA is under-reported. Whereas by making use of type instead of intensity, the PA can be identified.

10.4.2 Transition Probabilities

Table 41 shows the transition probabilities as identified in the assessment data. The self-transition probabilities are high for all labels, this means that the participants are likely to stay in the activity once they have begun. The chance of transitioning to or remaining in Sedentary activities is highest, with Standing being the next most likely.

	Sedentary	*Standing*	*Active*
Sedentary	0.96	0.03	0.01
Standing	0.10	0.86	0.04
Active	0.13	0.11	0.76

Table 41: Activity transition probability in assessment data, rows represent activity being transitioned from, columns represent activity being transitioned to.

The probability of transitioning from Active to Sedentary is 0.13, whereas the probability of transitioning from Active to Standing is 0.11. This means it is more likely to go from Active activities immediately to Sedentary than to first transition to Standing. This is more likely to be an artefact of the transition from Active-Standing-Sedentary being too fast, rather than the participants omitting the

Standing step. In Chapter 6, it was identified that the average durations of activity transitions were 2.95 seconds. In order to compensate for this non-instantaneous transition, a refractory period not allowing for more than one transition to be detected in a 2.95 second window was used. This may be affecting the transition probabilities, if the standing activity lasts for less than 2.95 seconds, then it will not be detected. The work of Kozina (112) only identified the average duration of transitions, this was not stratified by the type of activity that was being transitioned to and from. It may be that Standing activities have a lower duration transition time. Additionally, the variances of the transition durations were not identified, it may be the case that the transitions durations are not normally distributed, and durations of less than 2.95 may be more common than previously thought. Additionally, this duration was only tested for one population, it may not be consistent among all populations (as is often the case in activity classification literature). As such, further work investigating the distribution of transition durations for a range of different activities and populations is recommended.

The Free-Living data which is created via 12.8 second fixed windows and is therefore not affected by the automatic segmentation has a much higher probability of transitioning from Active to Standing (0.34), as can be seen in Table 42.

	Sedentary	*Standing*	*Active*
Sedentary	0.82	0.17	0.01
Standing	0.15	0.62	0.23
Active	0.02	0.34	0.64

Table 42: Transition probabilities for the Free-Living data-set, rows represent activity being transitioned from, columns represent activity being transitioned to.

The assessment data has higher self-transition probabilities for all activities than the Free-Living data, suggesting that once a ‘bout’ of the activity has begun the participants are less likely to change activity in the assessment data. This may be an effect of the post-processing methods used in the final classification

pipeline. Smoothing with a modal filter and using a Hidden Markov model to smooth the data decreases the chance of changing activities, thereby increasing the self-transition probability.

The obvious next step in this work is to attempt to link these PA metrics with health outcomes. Additionally, the classification pipeline developed should be used on multiple data-sets with known activity labels in order to determine its inter-protocol performance over a wider range of data.

10.5 Conclusion

In conclusion, the classification pipeline developed in this work has a significantly increased (over 100%) inter-protocol performance compared to the Base Classifier but is not significantly different with respect to intra-protocol performance. This means that compared to the Base classifier the final classifier is more likely to correctly classify acceleration data when it is gathered from a different protocol than it was trained with. Hence, the aim of this thesis; to maximise the inter-protocol performance (while maintaining intra-protocol performance); has been met.

11. Conclusion

The opening chapter in this thesis drew attention to uncertainties that exist in our understanding of the relationship between physical activity and health (3), the prevalence of inactivity in the population (15), and the effectiveness of interventions that are at least partly due to a reliance on subjective measures of physical activity (22). The importance of more precise, objective measurement of the frequency (28), intensity (32), time (36) and type (37) of physical activity was argued. It was claimed that one of the main drawbacks of the ubiquity of accelerometers was the inability to capture the type of physical activity being undertaken (61). Therefore, the main aim of this thesis was to create an activity classification pipeline that could accurately characterise type of physical activity from raw acceleration data. The goal was to develop an activity classification pipeline that had the ability to perform as well on data from different protocols as it did on data from the same protocol that it was trained on. Overall this aim has been achieved, the classification pipeline in this work achieves almost equal intra and inter-protocol performance and is significantly better than a current state of art approach (55).

Chapter 2 reviewed the current research in activity classification. The methodology behind activity classification was described, being a specific case of supervised learning. Activity classification pipelines are created from training data, comprised of activity labels and accelerations, in order to accurately classify unseen data. Additionally, an activity classification pipeline was synthesised from the current literature (54,55), breaking the classification process into a sequence of data processing steps. The methods for testing the performance of activity classification pipelines and attempting to predict their performance on unseen acceleration data were also discussed. Four different types of performance (65,76,77) were identified in this domain, with the most commonly reported performance being intra-protocol-inter-subject; where the

test data has different participants from the same protocol as the training data. This performance is generally high but does not accurately predict the performance of the classification pipelines on unseen data from a different protocol (54). While the majority of classification pipelines report performances of over 90% accuracy (65), it is unknown if any are able to actually classify activity from different protocols with a similar performance. The major limitation in activity classification research was identified as an over-reliance on optimising this intra-protocol performance. This over-reliance leads to a high ability to classify data from the training data but an inability to classify data from a different protocol (66,77). However, it was also noted that classification methods that attempt to maintain a high level of generalisability tend to have consistently lower performances (bias-variance trade-off) (76). Focusing on maximising the inter-subject-inter-protocol performance, where the test data is from different protocols with different participants, tends to lower the intra-protocol-inter-subject performance. Therefore, in this thesis it was decided to attempt to maximise both inter and intra-protocol performances in order to create a classification pipeline that maintains high performance on both training and unseen data. Aside from the lack of generalisability, other issues were identified relating to how the individual steps of the classification pipeline effect the performance, specifically: the lack of ability to maintain classification performance across different wrists (89), the lack of information about how pre and post-processing methods effect the inter-protocol performance (103), the lack of consensus about which classification features to be used (94), the lack of accord about optimal windows sizes (63) and the lack of agreement about the optimal classification algorithms to be used (65). The chapter concluded that in order to achieve both high inter and intra-protocol performance, these methodological limitations/uncertainties needed to be dealt with.

Chapter 3 described the data-sets used in the thesis. Several summary statistics relating to the data-sets were described. The activity protocols and participant characteristics were discussed in order to give an idea of the

external validity of the data. Additionally, statistics required for the classification pipeline, transition probabilities and class balances, were evaluated. The evaluation metric for this thesis was discussed and justified (134). Finally, a criterion classification pipeline was created to serve as a control method. This pipeline was based on the work of Chowdhury et al (55) which has been validated on both Lab and Free-living data, having been shown to generalise to unseen data better than other approaches reviewed. The work of Chowdhury et al was chosen because its ability to generalise was supported both practically (being validated on Lab and Free-living data) and theoretically (utilising methodologies that typically allow for high generalisability). Also, its relative simplicity allowed for examination into how changes in the pipeline affected the overall performance. The classification pipeline had a high intra-protocol performance of 0.898 and 0.765, however the inter-protocol performance was much lower at 0.352 and 0.415.

Chapter 4 dealt with the first of the methodological challenges identified in Chapter 2, the lack of ability to maintain classification performance across different wrists. Activity classification pipelines are typically trained on data obtained from sensors at a set orientation (107). Changes in this orientation (such as being on a different wrist) result in performance degradation. In order to mitigate this degradation a method for allowing wrist invariance via Domain Adaption (137) was identified. Domain Adaptation was then tested against another method of achieving this invariance, using both wrists in the same classification (89) and found to be effective. The performance drop when the classification pipeline was tested on data from the opposing wrist to which the training data was collected on was entirely mitigated when using Domain Adaption. The use of Domain Adaption was thought to have the advantage over other methods of achieving wrist invariance due to its simplicity and the fact that the use of Domain Adaption when not required has no impact on the performance. There are a number of advantages to having an activity classification pipeline that is invariant to wrist orientation. In some large-scale

studies (58) the accelerometer is sent in the post with written instructions for placement. In these situations, there is no means of knowing whether or not the accelerometer was worn as instructed. Even in studies where the accelerometer is fitted by a researcher (167) there is still no guarantee that the accelerometer was not moved to the other wrist for comfort or convenience. It is even possible that during the observation period the accelerometer was worn on both wrists for different periods prior to being returned. Consequently, activity classification pipelines that have been developed in controlled settings when the location of the accelerometer is always observed, are likely to lose performance when applied to unseen data collected in free-living settings where the location of the accelerometer is unknown and could be different from the location on which the classification pipeline was trained (73,87).

Chapter 5 examined the effects of pre and post-processing methods on inter and intra-protocol performance. A variety of pre-processing techniques were reviewed including: data aggregation (97), filtering noise (70), orientation invariant transformations (107), inclination correction (108) and over-sampling (75). The majority of studies do not report the performance both with and with-out the pre-processing methods, so it is difficult to judge if the performance changes reported here are consistent with other work. Data aggregation (using ENMO) was found to improve inter and intra-protocol performance if ENMO was used in conjunction with the separate axes, instead of replacing them. Filtering appeared to show no increase in performance. Using orientation invariant transformations only improved performance when the orientation of the sensor was artificially changed, however using ENMO outperformed the transformation, consistent with the work of Yurtman et al (107), who reported similar findings. Structure preserving oversampling improved the inter-protocol performance, in agreement with Cao et al (75) who reported a 5.3% increase in performance. In contrast to pre-processing methods, studies involving novel post-processing do tend to report performance both with and without the post-processing, thus allowing for comparison with this work. Three methods of post-processing were

reviewed: smoothing (124), hidden Markov models (69) and Participant Adaption via Iterative Re-learning (125). Smoothing improved the inter-protocol performance, a finding not reported elsewhere. Using hidden Markov models improved both the intra and inter-protocol performance, consistent with the work of Ellis et al (69). Participant Adaption via Iterative Re-learning improved both the intra and inter-protocol performance which agrees with the work of Yuan et al (125) who reported a performance increase of at least 16%. Finally, the efficacy of using multiple pre and post-processing methods at once was analysed, making use of: Structure preserving oversampling, Data aggregation, Participant Adaption via Iterative Re-learning, a hidden Markov model and smoothing. This was found to improve the inter-protocol performance significantly. The findings of this chapter provide new evidence for researchers on how processing techniques improve performance in activity classification. This will allow for the reduction of processing time in classification studies, as some common pre-processing methods can be eliminated from classification pipelines (such as Butterworth filtering (74)). Additionally, Chapter 5 identified which methods of pre and post-processing are in further need of research due to continued lack of clarity of their efficacy.

Chapter 6 dealt with the lack of consensus on the optimal window size for activity classification, and the fact that different activities appear to have different optimal window sizes (63). The chapter identified that automatic segmentation of acceleration data would allow for variable length windows, thus avoiding the many issues raised from fixed length windows. A method for automatic segmentation that detected transitions in acceleration data was created. The transition detection method was tested against several other approaches: Kozina (112), Lyden (148) and Salcic (146). The method created in this work showed a significantly greater ability to detect transitions than other methods, potentially due to its data-driven nature and not making use of all three axis, unlike Kozina's (112) and Lyden's (148) methods of transition detection. After the transition detection method was created it was used to

automatically segment acceleration data prior to classification. Making use of the automatic segmentation allowed for an increase in intra and inter-protocol performance. Ni et al (168) also showed that making use of automatic segmentation could achieve a significant performance improvement, although only the intra-protocol performance was remarked on.

Chapter 7 tackled the limitations of features that allow for a high intra-protocol performance typically having a low inter-protocol performance and vice-versa. In Chapter 2 it was discussed that statistical based features allowed for a high intra-protocol performance at the cost of a lowered inter-protocol performance (54) and for morphological features the converse was true (76). The chapter investigated the idea that by combining statistical and morphological based features it may be possible to obtain both a high intra and inter-protocol performance. Recurrence Quantification Analysis was used because its features are created from statistical analysis of the recurrence plots of the data, thus allowing for a combination of statistical and morphological based features (150). Recurrence Quantification Analysis was also chosen due to its high performance in Gait Analysis (118), a domain that is highly similar to activity classification. Recurrence Quantification Analysis allowed for a high level of inter-protocol performance without any degradation in intra-protocol performance. To the best of my knowledge this was the first work that made use of Recurrence Quantification Analysis in this domain.

Chapter 9 tested an alternative method to Recurrence Quantification Analysis for finding features with high inter and intra-protocol performance. Sparse Feature Encoding (94) was used to automatically generate features from the testing data. These features were found to increase the inter-protocol performance but with a corresponding decrease to the intra-protocol performance. Sparse Feature Encoding has been used in activity classification and has shown an ability to adapt to changes in the activity protocol (94), therefore the findings in this work are concurrent with previous results.

However, Sparse Feature Encoding features had a lower performance than Recurrent Quantification Analysis features, therefore they were not included in the final classification pipeline.

Chapter 9 dealt with the final methodological challenge identified in this work, the lack of consensus about which classification algorithm to use for activity classification (65). An investigation into whether Generative or Discriminative classifiers were superior in this domain was performed. It was found that neither Discriminative nor Generative were inherently superior on the data considered. However, it was found that making use of a Generative Adversarial Network allowed for a greater level of classification performance (both inter and intra) than other methods. No work has investigated the effect of Generative or Discriminative classification algorithms on activity classification, although previous work has reviewed a variety of classification algorithms with respect to their performance in classification (165). This previous work only focused on the individual performance of different algorithms, not the underlying methodologies of Generative and Discriminative. Additionally, previous work (165) has focused entirely on the intra-protocol performance with no investigation on the ability to generalise to unseen data. The findings of the previous work were in agreement with this chapter, as it was found that Neural Networks were high performing in both this work and the work of Kwapisz (165). To the best of my knowledge no other studies have been carried out using Generative Adversarial Networks for activity classification. However, in general Generative Adversarial Networks do outperform other classification algorithms in other domains (166), consistent with this work.

Chapters 4-10 all focused on achieving a high inter-protocol performance without a corresponding loss in intra-protocol performance. This resulted in a classification pipeline that has a good ability to classify activity from acceleration data gathered from a range of different protocols. This will allow for the precise, objective measurement of the type of physical activity from accelerometer data.

11.1 Strengths and Limitations

The main strength of this work was the focus on both inter and intra-protocol performance, compared to the majority of research that focuses solely on intra-protocol performance (54,75). The major limitation of existing research that focuses only on intra-protocol performance is that classification pipelines are optimised to improve their intra-protocol performance at the cost of their inter-protocol performance, leading to a lack of an ability to generalise to unseen data (76,77). As the majority of data will be unseen, this is a considerable limitation and prevents accurate estimates of type of physical activity in studies collecting Free-Living data.

Similarly, this work made use of both Lab-Based and Free-Living data. The majority of accelerometry studies make use of only Lab-Based data (54,75), which typically does not generalise well to Free-Living (77). This inability to generalise means that methods that perform well in Lab-Based data may not perform as well in Free-Living data. Again, activity classification pipelines that are only based on Lab-Based data have little practical utility in research settings outside of the laboratory.

An additional strength of this work is simply the reporting of the inter-protocol performance. This means that there is some indication how well the classification pipeline will perform on unseen data from a different protocol. The majority of classification pipelines created do not investigate this. The issues with this can be seen in the work of Doherty et al (170), where a classification pipeline that has not be investigated with respect to its inter-protocol performance is used on a large accelerometry data-set, with the assumption that the classifications are entirely correct. It may be the case that the classification pipeline is failing to correctly characterise type, thus any health associations computed from the type information may be entirely incorrect.

An additional strength of this work is the focus on all facets of the classification pipeline, instead of focusing solely on one aspect such as features or the classification model, a common limitation in other studies (65). Chapters 6 and 8 showed that each facet can change the overall performance significantly. Therefore, a focus on all facets is logical. Additionally, by focusing on each facet of the classification pipeline, this work may help inform other studies on the development of a high-performance classification pipeline where individual steps of the classification pipeline can be easily modified. For example, if a new classification algorithm was being developed, this work gives high performing pre and post-processing steps that can be used regardless of the classification algorithm.

All methodological developments in this thesis are compared to a criterion classification pipeline that uses a current state of the art method. Consequently, it allowed for each development to be tested against the criterion, akin to clinical trial methodology that tests for non-inferiority, equivalence or superiority to current best practice (171). Many studies showcase new developments, but only focus on the performance on their data set without comparison with a criterion method or any other published methods (66,74,94). Although such studies might be able to demonstrate that a development improves the performance of a classification pipeline compared to the authors base classification pipeline, it does not tell us whether the development outperforms other methods or whether it performs consistently well on other data sets.

The major limitation of this work is the limited labels used for classification. The majority of the chapters make use of Sedentary-Stand-Active Labelling, meaning that there were only three separate classes. The limited number of classes identified was due to gathering the true labels via an ActivPAL on the thigh which only identifies three classes. However, there does not exist a gold standard criterion measure that is capable of measuring physical activity type in Free-Living settings. Direct observation may be possible for very short periods

of time in small sample studies, but it is unfeasible in everyday living and in larger studies. An additional limitation of this work was the assumption that the ActivPAL labelling was correct. While the device has been validated against direct observation (129,130), it does not achieve perfect accuracy. Slow walking is known to result in poor recognition of walking when using an ActivPal (129), especially in older populations. However, the participants in the Free-living data set were below 54 years old, so age was unlikely to be a factor. Giving any instruction on walking speed would be against the principle of achieving Free-living, so misclassification of slow walking is a potential limitation. Some work has been carried out in using participant mounted cameras paired with accelerometers, however, use of these cameras raises many privacy issues, labelling is very labour intensive, and typically suffers from large amounts of missing data (46). Additionally, cameras may cause an observer effect bias.

A related issue is that this work did not have a specific Sleep label, merely identifying Sedentary behaviour. It is well documented that Sleep has many health benefits independent of physical activity (172,173) and therefore ideally it should be identified.

Another limitation of this work was the data-sets used. Although both Lab-based and Free-Living data was gathered, only healthy, younger participants were used. It has been shown that changes in the anthropometric measures of participants impacts the performance of activity classification pipelines (76), therefore this classification pipeline may have difficulties generalising to diseased older participants.

The work focused on all facets on the activity classification pipeline, finding the optimal hyperparameters for each stage of the activity classification pipeline. However, each stage was considered independently. How these hyperparameters affected each other was not identified. It may be the case, for example, that the classification model used may impact the optimal combination of features, but this was not considered in this work.

The creation of this pipeline can be likened to an iterative greedy approach, if one of the methods evaluated in this thesis was found to increase the performance it was added to the final pipeline. This resulted in a high performing pipeline but did lead to an increased level of complexity compared to the original pipeline (The Base pipeline). Occam's razor states, 'entities should not be multiplied without necessity'; to paraphrase, a simple model is better than a complex model. This notion of the principle of parsimony is particularly prevalent in Machine Learning as simple models are typically more resistant to overfitting than complex models. It is also the case that some of the steps of the final classification provide overlapping functionality (particularly HMM smoothing and median smoothing) therefore it may not be necessary to have all the steps in the pipeline to obtain such a high level of performance. A potential next step would be to see which (if any) steps can be removed from the pipeline without deleterious impacts on the performance. This would allow for a reduction of the complexity of the model, thus allowing for a simper solution. Some additional work may be required to develop an acceptable performance reduction in exchange for removing a step in the pipeline (thus reducing the complexity), akin to a regularisation constant that is often used in Machine Learning.

This work assumed that the accelerometers were worn constantly, an assumption that typically is not the case in population surveillance. Non-wear detection is a well-researched area and many methodologies exist for its detection (174,175). Activity classification is typically performed on acceleration data after non-wear detection has been performed and any data with too little valid data is removed, so it was felt that it could be assumed non-wear was not an issue in this work.

Another limitation of this work is that the pipeline developed in this work, made use of ActivPal labels as the gold standard. This means that at best the pipeline performs as well as the activPAL itself, which relies on a very simple scheme not requiring all the methodology and computing power developed in this work.

While this is the case, ActivPals are thigh-mounted accelerometers which have higher participant burden and lower compliance than wrist-worn devices with wear time criteria (86). Additionally, wrist worn devices are more commonly used in research than thigh mounted devices.

11.2 Future Work

Future work should investigate how each aspect of the classification pipelines hyperparameters may affect each other, instead of assuming independence. For instance, it is likely that different features are optimal for different window sizes. This was not investigated in this thesis but represents potential future work that should be examined.

Perhaps the most obvious future work would be to repeat the work in this thesis using more activity classes, instead of only the Sedentary-Stand-Active Labelling, specifically a Sleep activity class. The ability to classify a broader range of behaviours, objectively, would allow for a greater understanding of how activity type is independently associated with health outcomes. This is likely to be particularly important in studies involving population sub-groups where certain types of activity are more prevalent (e.g., domestic activities in the elderly) or of specific interest (stair climbing, sit-to-stand transitions in the elderly).

Additionally, future work is to repeat the work in this thesis on larger data-sets with a broader range of demographic groups, especially those with chronic disease and older participants. If the ability to accurately classify activity type could be achieved in a broad range of populations then wrist-worn accelerometers becomes more viable for population surveillance.

Obviously, this would require labelled acceleration data-sets that are typically expensive to compute. A potential avenue for increasing the amount of labelled acceleration data comes from the related field of activity classification from

video data. There exists a large amount of video data with labelled activities, for instance recordings of sporting matches. Some research has investigated the identification and tracking of wrist motions from video data with respect to sign language translation (176). It may be possible to use wrist tracking to generate wrist acceleration data from video data, thereby creating large labelled acceleration data-sets without requiring the use of accelerometers.

Despite this research being restricted to the classification of Sedentary-Stand-Active behaviours it does offer the potential to greatly increase our understanding of how physical activity is related to health. A person's waking day cycles between a series of sedentary, standing, and physically active behavioural events. The work of this thesis means it is now possible to output a time series of these events from raw acceleration data. The active events can be characterised in terms of their: frequency, duration, intensity, type, volume and pattern, permitting much more detailed analysis of the relationship between these characteristics and health. Many unlabelled acceleration data sets with associated health metrics exist, representing a great opportunity for using the classification pipeline to investigate the impact of physical activity type on health (58). This is where the wrist invariance reported in Chapter 4 is especially useful; few if any available acceleration data sets have a record of the location of the accelerometer for each participant, making a wrist invariant classification pipeline essential.

Chapter 1 highlighted that some of the uncertainties that persist in our knowledge about physical activity and health are related to the low resolution of data available from self-reported physical activity. Already, studies that have collected both self-reported and objective measures concurrently, report stronger associations between physical activity and health with objective measures (177,178). Although such studies are advancing our understanding of the magnitude of the relationship between physical activity levels and health,

they do not provide insights into the level and pattern of specific behavioural types.

The availability of a time series of activity types at high resolution, that overcomes recall and social desirability bias, will lead to a more precise understanding of the association between variations in levels and patterns of activities of daily living and specific health and disease outcomes (170). Chinapaw et al (179) created a time series of behavioural data, classifying events by a combination of intensity (sedentary, light, moderate, and vigorous intensity) and duration. They found that clustering of temporal patterns of intensity and duration were associated with body mass index and fitness. Similarly, Carson et al (180) reported that the composition of physical activity at various intensities throughout the day was associated with cardiometabolic biomarkers in children and youth. A limitation of constructing time series data based on thresholds of acceleration to classify the intensity, is that intensity thresholds vary by location of device and left or right wrist for wrist-worn devices (61,181).

Chastin and Granat (182) used a time series of behavioural event-based data to analyse the volume and pattern of sedentary events (lying and sitting). They found that time total time spent sedentary did not discriminate between healthy and unhealthy adults whereas the pattern of sedentary events did. Other studies have also shown that the pattern of behaviours as well as the volume can offer new insights to the health effects of physical activity (183–185). A study by Paraschiv-Ionescu et al (185) used time series data of behavioural events to combine different features of physical activity (type, intensity, duration) in order to define various physical activity states. They found that the temporal sequence of various behavioural states was associated with chronic pain in older people independent of total activity.

The ability to accurately represent an individuals' pattern of behaviour as a series of events has beneficial outcomes for adoption of objective

measurements approaches in both the commercial and research domains. The data content of a stream of data compressed with varying length epochs, to which meta-data (type, intensity etc.) can be reliably attached, is much higher than fixed epoch approaches. This improves both storage and data transmission efficiency giving the potential for battery life and effective recording time extension in wearable devices.

As new knowledge emerges from studies utilising time series data to characterise patterns of specific objectively measured behavioural events, results could eventually translate into more tailored behavioural guidelines rather than the current one size fits all of 150 minutes of at least moderate intensity physical activity (13). For example, the American Diabetes Association recommends interrupting prolonged sedentary events with light intensity activity every 30 minutes to improve blood glucose in adults with type 2 diabetes (186).

A move towards physical activity guidelines that focus on the pattern of specific types of physical activity has implications for population surveillance. Monitoring the prevalence of such guidelines could only be achieved via high resolution accelerometer data that could be translated into a time series of behavioural events. This more granular characterisation of physical activity may also provide new insights into how physical activity varies according to population sub-group – variations that might be masked by characterisation restricted to volume and intensity metrics.

The pipeline described in the thesis builds the first stage of a wider framework for the processing of large amounts of data. Raw acceleration data is difficult to analyse on demand in the way that is required to operationalise many measurement and behaviour change projects where timeliness is of the essence. The ability to convert raw data to a good common format for further processing as and when needed will allow new services to be created and delivered. Lifestyle profiles and digital endpoints for a wide range of health applications, based on the underlying behavioural bouts, will be near-instantly

available once converted. In the future, this pipeline should be extended further to allow for off-the shelf usage. This will allow for a version that researchers and clinicians can use. Thus, it will allow for exporting a time series of behavioural events which can then be mined for health associations without the requirement of recreating the classification pipeline for new data-sets.

Bringing event-based capabilities in the realm of raw data analytics allows immediate application of emerging specifications for the processing of these data types, such as AlPHABET (187, 188, 189), supporting the wider take-up of these advanced techniques through standardised data formats, interfaces and data governance processes.

11.3 Ethical implications

Research into the ethics of PA monitoring have been considered under three categories (190):

1. Autonomy: The right of a participant to self-govern. Typically, relating to whether a participant can withdraw consent or cease monitoring.
2. Privacy: The right of an individual or group to control access to personal information.
3. Harm: Negative consequences, potentially; physical, psychological, economic or sociological.

Autonomy: Autonomy is perhaps the simplest issue with respect to accelerometery as participants can easily remove the accelerometers to cease any information being gathered. Under current General Data Protection Regulation (GDPR) laws, all participants have the right to remove their consent at any time (191). A potential issue is when the participants are not knowingly sharing their data. For instance, participants may be happy to share their acceleration data but not know that this allows for information about their PA and sleep to be extracted (54,170). The work in this thesis represents such an issue, allowing for information about PA to be derived from acceleration data.

This issue can be dealt with by maintaining transparency on all data processing procedures prior to gaining consent. The scope of data that can be identified from acceleration data is constantly growing, potentially meaning that permissions would have to be continuously resought. This could be overcome by careful selection of wording in participant information sheets and consent forms. This is similar to the storage of human tissue samples that may be re-analysed when new discoveries are made in the future.

Privacy: PA classification research and specifically this work impacts privacy by allowing for personal information to be extracted about the participants; activity (in this research) and sleep (54,170). In an attempt to mitigate these privacy issues, acceleration data is treated as biomedical data and bound by privacy laws (193), one consequence of these laws is that acceleration data is anonymised as standard practise when published. Typically published data sets contain pseudonymised acceleration data paired with other medical data. Personal data is stored separately from acceleration data with linkage only possible by named researchers (103). Due to a number of data leaks from commercial smart-watch companies, there is an increasing amount of personally identifying information with matched acceleration profiles (194). Recent research has investigated the ability to use acceleration profiles as a bio-identifier, specifically recognizing if it is possible to identify if two activity profiles come from the same person (195). Therefore, by matching published health data sets with data obtained from smart watch companies, it is theoretically possible to obtain personal information and matched health information, through acceleration data, which represents a major ethical issue. Thus, it may be the case that the current privacy laws relating to acceleration data may need to be updated. The closest parallel to this would be DNA data. There exist many data sets with paired DNA and other medical data, and commercial DNA analysis companies store both DNA and personally identifying information in a pseudonymised format (196).

Harm: The next two points focus on the potential harmful impacts of the activity classification on the participants themselves, specifically under the assumption that they are able to access the recorded PA data. There is sufficient evidence to show that wearing PA monitoring devices causes a temporary increase in PA prevalence (197). However, this is not the only potential impact of such devices. As discussed in Chapter 1, governing bodies have released PA guidance detailing a minimum recommend amount of PA to be undertaken each week (10). A potential side effect of activity monitoring is that participants will find that they have met these guidelines and use this as a rationale for doing no more PA. While this in itself is not harmful, as the classification is not perfect, it may be the case that participants stop because they falsely believe they have met the minimum guidance, thus not meeting the recommended PA due to the sensor misclassifications. An additional potential for harm through activity classification is the relationship between eating disorders and PA tracking (198). While the research is in its infancy, it is clear that there is a link between activity tracking and disordered eating. It is not clear which direction this relationship is in, therefore there exists the potential that the tracking of activity increases the likelihood of disordered eating. This is obviously a relationship that needs to be further investigated to fully understand the potential negative impacts of PA measurement, especially as the risk groups for disordered eating and low PA are overlapping (198). In this work this should not be a problem as the classification pipeline developed is an off-line pipeline, therefore the PA is only identified after the end of the data collection and unlikely to impact the participants behaviour. Further, the acceleration devices used provide no feedback to the wearer, preventing any judgements being made about personal levels of physical activity.

Classification models making use of anthropometric measures (such as acceleration) are at risk of bias, due to homogenous training data (199). One of the most prevalent examples of this bias is that of speech recognition. Speech recognition systems are typically trained on data from white males, thus

resulting in a significantly lower performance (as low as 50%) when attempting to classify data from either females or non-white males (199). In general, classification models trained on a homogenous population may become inaccurate if there are changes to the population, such as race, gender or disease status (69). This is both a limitation of the study (discussed previously) but also a potential ethical issue. One of the aims for this work was to develop a classifier that would allow for investigation into how type of physical activity can impact health. Any associations that are found making use of this classifier will likely be strongest in data similar to the training population and may therefore lead to health messages that are biased towards the training population. This is especially important in this work, as while the participants used to generate the data sets were heterogeneous with respect to gender, they were comprised exclusively of Caucasians and were relatively young. As differences exist in acceleration data with respect to race and age (192), this represents a strong possibility of bias in this work.

The previous points have focused on how this research and activity classification can cause unintentional harm. These points focus on how PA classification and this research may be used to commit intentional anti-ethical behaviours causing harm to participants. With the rising amounts of obesity in the world, proponents of measures such as denying or reducing access to health care if a participant is obese have been increasing (200). A possible extension to this would be treating adherence to PA guidelines in a similar way. Aside from the ethical issues in denying healthcare for any reason, PA adherence has a particular flaw. Unlike obesity which is a simple accurate metric (Body Mass Index), PA classification is not 100% accurate, which may lead to denial of treatment when a participant has met the PA guidelines. For example, if a single acceleration threshold, calibrated in a young sample, is applied to data collected in older people to classify physical activity intensity, it will underestimate the prevalence of activity in the older sample potentially leading to calls for older adults to increase their physical activity (34). Even

without the extremity of denying health care it can still be seen how information about a participants PA may be used to influence medical opinions, in some cases this could obviously be beneficial, specific treatment interventions based on their PA. However, this information could have negative impacts as well, in particular patients may find their adherence to the PA guidelines effecting procurement of limited treatment. A related issue is insurance companies offering discounts based on PA (201) (as measured via activity trackers). While this in itself is not intentionally unethical, it does show that adherence to PA is currently being used to impact the cost/availability of certain services.

My next point focuses on how activity classification gives an opportunity for the participants to commit unethical actions. The previous points of PA resulting in lower insurance and potentially acting as bar from accessing healthcare, give a rationale of why a participant may wish to provide artificially inflated PA data - social desirability bias. This concept of deceptive behaviour within activity classification has seen some prior research (202) with the authors concluding that by iteratively retraining on successfully deceptive behaviour it was possible to accurately disambiguate between real and deceptive PA. It is worth noting that this method of iterative retraining on the deceptive behaviour is conceptually similar to that of a GAN, which was used in this work. Therefore, although this work should be unaffected by such deceptive behaviour, it does represent a wider issue. As mentioned earlier, because the measurement devices in this research provided no feedback on behaviour, social desirability bias is less likely compared with self-reported behaviour.

In conclusion; there are many potential ethical issues with PA monitoring via accelerometery and the work in this thesis. Further research is needed into the potential harms that PA monitoring may cause and whether any benefits outweigh the risks. Additionally, with the increased ubiquity of accelerometers and the increasing ability to extract information from acceleration data, further

discussion incorporating stakeholders, public health officials and researchers is needed to ensure privacy and autonomy of participants is maintained.

11.4 Conclusion

The objective measurement of physical activity with accelerometers is becoming ubiquitous in epidemiological, surveillance, screening, and intervention studies but a major limitation, compared to self-reported measures, has been the inability to estimate the type of physical activity being undertaken. Estimates of physical activity type enable researchers to examine a much broader range of physical activity metrics and how they are related to health, that are not subject to recall or social desirability bias. The body of research in this thesis provides new and important knowledge including; {1} improved methods for the creation of an activity classification pipeline that maximised both intra and inter-protocol performance; {2} a new method for creating an activity classification pipeline that is invariant to accelerometer location; {3} the creation of an activity classification pipeline, with high intra and inter-protocol performance, that can output a time series of activity types.

The main limitations of this work are the limited number of labels used for classification and the limited heterogeneity in the populations studied. Further research is required to test whether the findings of this thesis can be extended to a broader range of activities and populations.

As a result of this work researchers can reduce many of the uncertainties that exist in physical activity research, due to a reliance on self-report measures of physical activity, which in turn should lead to a more precise understanding of the association between variations in levels and patterns of activities of daily living and specific health and disease outcomes.

In conclusion, the goal of this thesis has been achieved, this will allow for a new level of understanding into how type of physical activity can impact health.

References

1. Lee I-M, Shiroma EJ, Lobelo F, Puska P, Blair SN, Katzmarzyk PT, et al. Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. The lancet. 2012;380(9838):219–229.

2. 2018 Physical Activity Guidelines Advisory Committee. 2018 Physical Activity Guidelines Advisory Committee Scientific Report. Washington, DC: U.S. Department of Health and Human Services; 2018.

3. Dowd KP, Szeklicki R, Minetto MA, Murphy MH, Polito A, Ghigo E, et al. A systematic literature review of reviews on techniques for physical activity measurement in adults: a DEDIPAC study. Int J Behav Nutr Phys Act. 2018;15(1):15.

4. Prince SA, Adamo KB, Hamel ME, Hardt J, Gorber SC, Tremblay M. A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. Int J Behav Nutr Phys Act. 2008;5(1):56.

5. Murphy JJ, Murphy MH, MacDonncha C, Murphy N, Nevill AM, Woods CB. Validity and reliability of three self-report instruments for assessing attainment of physical activity guidelines in university students. Meas Phys Educ Exerc Sci. 2017;21(3):134–141.

6. American College of Sports Medicine position statement on the recommended quantity and quality of exercise for developing and maintaining fitness in healthy adults. Med Sci Sports. 1978;10(3):7-10.

7. Pollock ML, Froelicher VF. Position stand of the American College of Sports Medicine: the recommended quantity and quality of exercise for developing and maintaining cardiorespiratory and muscular fitness in healthy adults. J Cardiopulm Rehabil Prev. 1990;10(7):235–245.

8. Heath GW, Parra DC, Sarmiento OL, Andersen LB, Owen N, Goenka S, Montes F, Brownson RC, Lancet Physical Activity Series Working Group. Evidence-based intervention in physical activity: lessons from around the world. The lancet. 2012 Jul 21;380(9838):272-81.

9. Grand View Research, Wearable Technology Market Size, Share & Trends Analysis Report By Product (Wrist-wear, Eye-wear & Head-wear, Foot-wear, Neck-wear, Body-wear), By Application, By Region, And Segment Forecasts, 2020 – 2027, [internet].

10. Physical activity guidelines for Americans: be active, healthy, and happy! US Department of Health and Human Services; 2008.

11. Chief Medical Officers. Start active, stay active: A report on physical activity from the four home countries' Chief Medical Officers. Department of Health and Social Care; 2011.

12. Mück JE, Ünal B, Butt H, Yetisen AK. Market and patent analyses of wearables in medicine. Trends in biotechnology. 2019 Jun 1;37(6):563-6.

13. UK Chief Medical Officer. UK Chief Medical Officers' Physical Activity Guidelines. Department of Health and Social Care; 2019.

14. Gershuny J, Fisher KD. General Household Survey Cross-Year Leisure Activities, 1973-1997. Office for National Statistics; 1999.

15. Mindell J, Biddulph JP, Hirani V, Stamatakis E, Craig R, Nunn S, et al. Cohort profile: the health survey for England. Int J Epidemiol. 2012;41(6):1585–1593.

16. Allied Dunbar Assurance plc, Health Education Authority, Sports Council. Allied Dunbar national fitness survey: main findings: a report on activity patterns and fitness levels commissioned by the Sports Council and Health Education Authority. London: Allied Dunbar (in association with) Health Education Authority (and) Sports Council; 1992.

17. Ahmad S, Harris T, Limb E, Kerry S, Victor C, Ekelund U, et al. Evaluation of reliability and validity of the General Practice Physical Activity Questionnaire (GPPAQ) in 60–74 year old primary care patients. BMC Fam Pract. 2015;16(1):113.

18. Hagströmer M, Oja P, Sjöström M. The International Physical Activity Questionnaire (IPAQ): a study of concurrent and construct validity. Public Health Nutr. 2006;9(6):755–762.

19. Craig R, Mindell J, Hirani V. Health survey for England. Health Soc Care Inf Cent. 2013.

20. National Institute for Clinical Excellence (NICE). Physical activity: brief advice for adults in primary care. London: NICE public health guidance; 2013.

21. Lobelo F, Rohm Young D, Sallis R, Garber MD, Billinger SA, Duperly J, et al. Routine assessment and promotion of physical activity in healthcare settings: a scientific statement from the American Heart Association. Circulation. 2018;137(18):495–522.

22. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. J Chronic Dis. 1986;39(11):897–906.

23. van Poppel MN, Chinapaw MJ, Mokkink LB, Van Mechelen W, Terwee CB. Physical activity questionnaires for adults. Sports Med. 2010;40(7):565–600.

24. Caspersen CJ, Powell KE, Christenson GM, others. Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research. Public Health Rep. 1985;100(2):126–131.

25. Moore SC, Lee I-M, Weiderpass E, Campbell PT, Sampson JN, Kitahara CM, et al. Association of leisure-time physical activity with risk of 26 types of cancer in 1.44 million adults. JAMA Intern Med. 2016;176(6):816–825.

26. Chomistek AK, Henschel B, Eliassen AH, Mukamal KJ, Rimm EB. Frequency, Type, and Volume of Leisure-Time Physical Activity and Risk of Coronary Heart Disease in Young Women. Circulation. 2016;134(4):290–299.

27. Laursen AH, Kristiansen OP, Marott JL, Schnohr P, Prescott E. Intensity versus duration of physical activity: implications for the metabolic syndrome. A prospective cohort study. BMJ Open. 2012;2(5):e001711.

28. Clarke J, Janssen I. Is the frequency of weekly moderate-to-vigorous physical activity associated with the metabolic syndrome in Canadian adults? Appl Physiol Nutr Metab. 2013;38(7):773–778.

29. O'Donovan G, Lee I-M, Hamer M, Stamatakis E. Association of "Weekend Warrior" and Other Leisure Time Physical Activity Patterns With Risks for All-Cause, Cardiovascular Disease, and Cancer Mortality. JAMA Intern Med. 2017;177(3):335-342.

30. Ainsworth BE, Haskell WL, Whitt MC, Irwin ML, Swartz AM, Strath SJ, et al. Compendium of physical activities: an update of activity codes and MET intensities. Med Sci Sports Exerc. 2000;32(1):498–504.

31. Costigan SA, Lubans DR, Lonsdale C, Sanders T, del Pozo Cruz B. Associations between physical activity intensity and well-being in adolescents. Prev Med. 2019;125:55–61.

32. Chastin SFM, De Craemer M, De Cocker K, Powell L, Van Cauwenberg J, Dall P, et al. How does light-intensity physical activity associate with adult cardiometabolic health and mortality? Systematic review with meta-analysis of experimental and observational studies. Br J Sports Med. 2019;53(6):370–376.

33. Amagasa S, Machida M, Fukushima N, Kikuchi H, Takamiya T, Odagiri Y, et al. Is objectively measured light-intensity physical activity associated with health outcomes after adjustment for moderate-to-vigorous physical activity in adults? A systematic review. Int J Behav Nutr Phys Act. 2018;15(1).

34. Dibben GO, Taylor RS, Dalal HM, Hillsdon M. One size does not fit all-application of accelerometer thresholds in chronic disease. Int J Epidemiol. 2019 01;48(4):1380.

35. Tarp J, Child A, White T, Westgate K, Bugge A, Grøntved A, et al. Physical activity intensity, bout-duration, and cardiometabolic risk markers in children and adolescents. Int J Obes. 2018;42(9):1639–1650.

36. Powell KE, King AC, Buchner DM, Campbell WW, DiPietro L, Erickson KI, et al. The scientific foundation for the physical activity guidelines for Americans. J Phys Act Health. 2018;16(1):1–11.

37. Lee D, Brellenthin AG, Thompson PD, Sui X, Lee I-M, Lavie CJ. Running as a key lifestyle medicine for longevity. Prog Cardiovasc Dis. 2017;60(1):45–55.

38. Vlachopoulos D, Barker AR, Ubago-Guisado E, Ortega FB, Krustrup P, Metcalf B, et al. The effect of 12-month participation in osteogenic and non-osteogenic sports on bone development in adolescent male athletes. The PRO-BONE study. J Sci Med Sport. 2018;21(4):404–409.

39. Morris JN, Heady J, Raffle P, Roberts C, Parks J. Coronary heart-disease and physical activity of work. The Lancet. 1953;262(6796):1111–1120.

40. Paffenbarger Jr RS, Hale WE, Brand RJ, Hyde RT. Work-energy level, personal characteristics, and fatal heart attack: a birth-cohort effect. Am J Epidemiol. 1977;105(3):200–213.

41. Epstein L, Miller G, Stitt F, Morris J. Vigorous exercise in leisure time, coronary risk-factors, and resting electrocardiogram in middle-aged male civil servants. Heart. 1976;38(4):403–409.

42. Wilson PW, Paffenbarger Jr RS, Morris JN, Havlik RJ. Assessment methods for physical activity and physical fitness in population studies: report of a NHLBI workshop. Am Heart J. 1986;111(6):1177–1192.

43. Swain DP, Franklin BA. Comparison of cardioprotective benefits of vigorous versus moderate intensity aerobic exercise. Am J Cardiol. 2006;97(1):141–147.

44. Bird ME, Datta GD, Van Hulst A, Kestens Y, Barnett TA. A reliability assessment of a direct-observation park evaluation tool: the Parks, activity and recreation among kids (PARK) tool. BMC Public Health. 2015;15(1):906.

45. Sallis JF. Measuring physical activity: practical approaches for program evaluation in Native American communities. J Public Health Manag Pract JPHMP. 2010;16(5):404-410.

46. Doherty AR, Kelly P, Kerr J, Marshall S, Oliver M, Badland H, et al. Using wearable cameras to categorise type and context of accelerometer-identified episodes of physical activity. Int J Behav Nutr Phys Act. 2013;10(1):22.

47. Doherty AR, Smeaton AF. Automatically Segmenting LifeLog Data into Events. In: 2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services. 2008. p. 20–23.

48. Kelly P, Marshall SJ, Badland H, Kerr J, Oliver M, Doherty AR, et al. An Ethical Framework for Automated, Wearable Cameras in Health Behavior Research. Am J Prev Med. 2013;44(3):314–319.

49. Sylvia LG, Bernstein EE, Hubbard JL, Keating L, Anderson EJ. Practical guide to measuring physical activity. J Acad Nutr Diet. 2014;114(2):199–208.

50. Dontje ML, Groot M de, Lengton RR, Schans CP van der, Krijnen WP. Measuring steps with the Fitbit activity tracker: an inter-device reliability study. J Med Eng Technol. 2015;39(5):286–90.

51. Hallam KT, Bilsborough S, de Courten M. “Happy feet”: evaluating the benefits of a 100-day 10,000 step challenge on mental health and wellbeing. BMC Psychiatry. 2018;18(1):19.

52. Strycker LA, Duncan SC, Chaumeton NR, Duncan TE, Toobert DJ. Reliability of pedometer data in samples of youth and older women. Int J Behav Nutr Phys Act. 2007;8(4):4.

53. Tudor-Locke C, Williams JE, Reis JP, Pluto D. Utility of pedometers for assessing physical activity: convergent validity. Sports Med Auckl NZ. 2002;32(12):795–808.

54. Bao L, Intille SS. Activity Recognition from User-Annotated Acceleration Data. In: Ferscha A, Mattern F. Pervasive Computing. Berlin, Heidelberg: Springer Berlin Heidelberg; 2004. p. 1–17.

55. Chowdhury AK, Tjondronegoro D, Chandran V, Trost SG. Physical Activity Recognition Using Posterior-Adapted Class-Based Fusion of Multiaccelerometer Data. IEEE J Biomed Health Inform. 2018;22(3):678–85.

56. Ruch N, Rumo M, Mäder U. Recognition of activities in children by two uniaxial accelerometers in free-living conditions. Eur J Appl Physiol. 2011;111(8):1917–1927.

57. Twomey N, Diethe T, Fafoutis X, Elsts A, McConville R, Flach P, et al. A comprehensive study of activity recognition using accelerometers. In: Informatics. Multidisciplinary Digital Publishing Institute; 2018;5(2):27-35.

58. Doherty A, Jackson D, Hammerla N, Plötz T, Olivier P, Granat MH, et al. Large Scale Population Assessment of Physical Activity Using Wrist Worn Accelerometers: The UK Biobank Study. PLOS ONE. 2017;12(2):1–14.

59. Zenko Z, Willis EA, White DA. Proportion of adults meeting the 2018 physical activity guidelines for Americans according to accelerometers. Front Public Health. 2019;7:135-152.

60. Silfee VJ, Haughton CF, Jake-Schoffman DE, Lopez-Cepero A, May CN, Sreedhara M, et al. Objective measurement of physical activity outcomes in lifestyle interventions among adults: A systematic review. Prev Med Rep. 2018;11:74–80.

61. Esliger DW, Rowlands AV, Hurst TL, Catt M, Murray P, Eston RG. Validation of the GENEA Accelerometer. Med Sci Sports Exerc. 2011 Jun;43(6):1085–1093.

62. Migueles JH, Cadenas-Sanchez C, Tudor-Locke C, Löf M, Esteban-Cornejo I, Molina-Garcia P, et al. Comparability of published cut-points for the assessment of physical activity: Implications for data harmonization. Scand J Med Sci Sports. 2019;29(4):566–574.

63. Banos O, Galvez J-M, Damas M, Pomares H, Rojas I. Window Size Impact in Human Activity Recognition. Sensors. 2014;14(4):6474–6499.

64. Pirttikangas S, Fujinami K, Nakajima T. Feature Selection and Activity Recognition from Wearable Sensors. In: Youn HY, Kim M, Morikawa H, editors. Ubiquitous Computing Systems. Berlin, Heidelberg: Springer Berlin Heidelberg; 2006. p. 516–27.

65. Ravi N, Dandekar N, Mysore P, Littman ML. Activity recognition from accelerometer data. In: Aaai. 2005. p. 1541–1546.

66. Ellis K, Kerr J, Godbole S, Staudenmayer J, Lanckriet G. Hip and Wrist Accelerometer Algorithms for Free-Living Behavior Classification: Med Sci Sports Exerc. 2016;48(5):933–40.

67. Veltink PH, Bussmann HJ, De Vries W, Martens WJ, Van Lummel RC, others. Detection of static and dynamic activities using uniaxial accelerometers. IEEE Trans Rehabil Eng. 1996;4(4):375–385.

68. Ellis K, Kerr J, Godbole S, Staudenmayer J, Lanckriet G. Hip and Wrist Accelerometer Algorithms for Free-Living Behavior Classification: Med Sci Sports Exerc. 2016;48(5):933–40.

69. Bishop CM. Pattern recognition and machine learning. Springer; 2006.

70. Ortega-Anderez D, Lotfi A, Langensiepen C, Appiah K. A multi-level refinement approach towards the classification of quotidian activities using accelerometer data. J Ambient Intell Humaniz Comput. 2019;10(11):4319–4330.

71. Mannini A, Rosenberger M, Haskell WL, Sabatini AM, Intille SS. Activity Recognition in Youth Using Single Accelerometer Placed at Wrist or Ankle: Med Sci Sports Exerc. 2017;49(4):801–12.

72. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res. 2016;18(12):e323.

73. Aras Yurtman, Billur Barshan. Activity Recognition Invariant to Sensor Orientation with Wearable Motion Sensors. Sensors. 2017;17(8):1838.

74. Nguyen ND, Truong PH, Jeong G-M. Daily wrist activity classification using a smart band. Physiol Meas. 2017;38(9):10–16.

75. Cao H, Nguyen MN, Phua C, Krishnaswamy S, Li X-L. An integrated framework for human activity classification. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing - UbiComp ’12. Pittsburgh, Pennsylvania: ACM Press; 2012. p. 331.

76. Margarito J, Helaoui R, Bianchi AM, Sartor F, Bonomi AG. User-independent recognition of sports activities from a single wrist-worn accelerometer: A template-matching-based approach. IEEE Trans Biomed Eng. 2016;63(4):788–796.

77. Pavey TG, Gilson ND, Gomersall SR, Clark B, Trost SG. Field evaluation of a random forest activity classifier for wrist-worn accelerometer data. J Sci Med Sport. 2017;20(1):75–80.

78. Quionero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND. Dataset shift in machine learning. The MIT Press; 2009.

79. Atallah L, Lo B, King R, Yang G-Z. Sensor positioning for activity recognition using wearable accelerometers. IEEE Trans Biomed Circuits Syst. 2011;5(4):320–329.

80. Mannini A, Intille SS, Rosenberger M, Sabatini AM, Haskell W. Activity recognition using a single accelerometer placed at the wrist or ankle. Med Sci Sports Exerc. 2013;45(11):2193-2203.

81. Montoye AH, Pivarnik JM, Mudd LM, Biswas S, Pfeiffer KA. Wrist-independent energy expenditure prediction models from raw accelerometer data. Physiol Meas. 2016;37(10):1770-1784.

82. Trost SG, Zheng Y, Wong W-K. Machine learning for activity recognition: hip versus wrist data. Physiol Meas. 2014;35(11):2183–2189.

83. Strath SJ, Kate RJ, Keenan KG, Welch WA, Swartz AM. Ngram time series model to predict activity type and energy cost from wrist, hip and ankle accelerometers: implications of age. Physiol Meas. 2015;36(11):2335–2351.

84. Trost SG, Cliff DP, Ahmadi MN, Tuc NV, Hagenbuchner M. Sensor-enabled Activity Class Recognition in Preschoolers: Hip versus Wrist Data. Med Sci Sports Exerc. 2018;50(3):634–41.

85. Ronao CA, Cho S-B. Human activity recognition with smartphone sensors using deep learning neural networks. Expert Syst Appl. 2016;59:235–44.

86. Fairclough SJ, Noonan R, Rowlands AV, Van Hees V, Knowles Z, Boddy LM. Wear Compliance and Activity in Children Wearing Wrist- and Hip-Mounted Accelerometers. Med Sci Sports Exerc. 2016;48(2):245–53.

87. Aguirre PLR, Torres LAB, Lemos AP. Autoregressive Modeling of Wrist Attitude for Feature Enrichment in Human Activity Recognition. Congresso Brasileiro de Inteligência Computacional. 2017.

88. Valarezo E, Rivera P, Park JM, Gi G, Kim TY, Al-Antari MA, et al. Human Activity Recognition Using a Single Wrist IMU Sensor via Deep Learning Convolutional and Recurrent Neural Nets. 2017;5:1-5.

89. Gjoreski M, Gjoreski H, Luštrek M, Gams M. How Accurately Can Your Wrist Device Recognize Daily Activities and Detect Falls? Sensors. 2016;16(6):800.

90. Bakrania K, Yates T, Rowlands AV, Esliger DW, Bunnewell S, Sanders J, et al. Intensity thresholds on raw acceleration data: Euclidean norm minus one (ENMO) and mean amplitude deviation (MAD) approaches. PloS One. 2016;11(10):e0164045.

91. Marcotte RT, Petrucci JG, Cox MF, Freedson PS, Staudenmayer JW, Sirard JR. Estimating Sedentary Time from a Hip-and Wrist-worn Accelerometer. Med Sci Sports Exerc. 2020 Jan;52(1):225-232

92. Ellis K, Kerr J, Godbole S, Lanckriet G, Wing D, Marshall S. A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers. Physiol Meas. 2014;35(11):2191–203.

93. Deng W-Y, Zheng Q-H, Wang Z-M. Cross-person activity recognition using reduced kernel extreme learning machine. Neural Netw. 2014;53:1–7.

94. Bhattacharya S, Nurmi P, Hammerla N, Plötz T. Using unlabeled data in a sparse-coding framework for human activity recognition. Pervasive Mob Comput. 2014;15:242–262.

95. Zhang S, Rowlands AV, Murray P, Hurst TL. Physical Activity Classification Using the GENEA Wrist-Worn Accelerometer: Med Sci Sports Exerc. 2012;44(4):742–8.

96. van Hees VT, Fang Z, Langford J, Assah F, Mohammad A, da Silva IC, et al. Auto-calibration of accelerometer data for free-living physical activity assessment using local gravity and temperature: an evaluation on four continents. Am J Physiol-Heart Circ Physiol. 2014;117(7):738-744.

97. Shoaib M, Scholten H, Havinga PJM. Towards Physical Activity Recognition Using Smartphone Sensors. In: 2013 IEEE 10th International Conference on Ubiquitous Intelligence and Computing and 2013 IEEE 10th International Conference on Autonomic and Trusted Computing. Italy: IEEE; 2013. p. 80–7.

98. Sun L, Zhang D, Li B, Guo B, Li S. Activity Recognition on an Accelerometer Embedded Mobile Phone with Varying Positions and Orientations. In: Yu Z, Liscano R, Chen G, Zhang D, Zhou X, editors. Ubiquitous Intelligence and Computing. Berlin, Heidelberg: Springer Berlin Heidelberg; 2010. p. 548–62.

99. Bai J, Di C, Xiao L, Evenson KR, LaCroix AZ, Crainiceanu CM, et al. An activity index for raw accelerometry data and its comparison with other activity metrics. PloS One. 2016;11(8):e0160644.

100. Antonsson EK, Mann RW. The frequency content of gait. J Biomech. 1985;18(1):39–47.

101. Shannon CE. Communication in the Presence of Noise. Proc IRE. 1949 Jan;37(1):10–21.

102. Song JT, Ahn SJ, Jeong WB, Yoo WS. Subjective absolute discomfort threshold due to idle vibration in passenger vehicles according to sitting posture. Int J Automot Technol. 2017;18(2):293–300.

103. Anguita D, Ghio A, Oneto L, Parra X, Reyes-Ortiz JL. A Public Domain Dataset for Human Activity Recognition Using Smartphones. 21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN) 2013. p. 437-442

104. Ortiz JLR. Smartphone-based human activity recognition. Springer; 2015.

105. Murray JA, Bradley H, Craigie WA, Onions CT. The Oxford english dictionary. Vol. 1. Clarendon Press Oxford; 1933.

106. Smith SW. The scientist and engineer's guide to digital signal processing. California Technical Publishing, San Diego, CA. 1997.

107. Yurtman A, Barshan B, Fidan B. Activity Recognition Invariant to Wearable Sensor Unit Orientation Using Differential Rotational Transformations Represented by Quaternions. Sensors. 2018;18(8):2725.

108. Fida B, Bernabucci I, Bibbo D, Conforto S, Schmid M. Pre-Processing Effect on the Accuracy of Event-Based Activity Segmentation and Classification through Inertial Sensors. Sensors. 2015;15(9):95–109.

109. Naik GR. CNN based approach for activity recognition using a wrist-worn accelerometer. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Jeju Island, South Korea: IEEE; 2017. p. 2438–41.

110. Dutta S, Sarkar S, Ghosh AK. Multi-scale classification using localized spatial depth. J Mach Learn Res. 2016;17(1):7657–7686.

111. Zheng Y, Wong W-K, Guan X, Trost S. Physical activity recognition from accelerometer data using a multi-scale ensemble method. In: Twenty-Fifth IAAI Conference. 2013.

112. Kozina S, Lustrek M, Gams M. Dynamic signal segmentation for activity recognition. In: Proceedings of International Joint Conference on Artificial Intelligence, Barcelona, Spain. 2011. p. 1522.

113. Dundar M, Krishnapuram B, Bi J, Rao RB. Learning Classifiers When the Training Data Is Not IID. In: IJCAI. 2007. p. 756–761.

114. Olszewski RT. Generalized feature extraction for structural pattern recognition in time-series data. Doctoral dissertation, Air Force Research Laboratory.

115. Bai J, Goldsmith J, Caffo B, Glass TA, Crainiceanu CM. Movelets: A dictionary of movement. Electron J Stat. 2012;6:559–578.

116. He B, Bai J, Zipunnikov VV, Koster A, Caserotti P, Lange-Maia B, et al. Predicting Human Movement with Multiple Accelerometers Using Movelets: Med Sci Sports Exerc. 2014;46(9):1859–66.

117. Vollmer C, Gross H-M, Eggert JP. Learning Features for Activity Recognition with Shift-Invariant Sparse Coding. In: Mladenov V, Koprinkova-Hristova P, Palm G, Villa AEP, Appollini B, Kasabov N, editors. Artificial Neural Networks and Machine Learning – ICANN 2013. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013. p. 367–74.

118. Labini FS, Meli A, Ivanenko YP, Tufarelli D. Recurrence quantification analysis of gait in normal and hypovestibular subjects. Gait Posture. 2012;35(1):48–55.

119. Sani S, Wiratunga N, Massie S. Learning Deep Features for kNN-Based Human Activity Recognition. CEUR Workshop Proceedings, 2017.

120. Wang L. Recognition of Human Activities Using Continuous Autoencoders with Wearable Sensors. Sensors. 2016 Feb 4;16(2):189.

121. Xiao L, He B, Koster A, Caserotti P, Lange-Maia B, Glynn NW, et al. Movement prediction using accelerometers in a human population: Movement Prediction Using Accelerometers in a Human Population. Biometrics. 2016;72(2):513–24.

122. Ng AY, Jordan MI. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In: Advances in neural information processing systems. 2002. p. 841–848.

123. Vapnik V. Statistical learning theory. Wiley; 1998.

124. Yang JB, Nguyen MN, San PP, Li XL, Krishnaswamy S. Deep Convolutional Neural Networks on Multichannel Time Series for Human Activity Recognition. In Twenty-Fourth International Joint Conference on Artificial Intelligence. 2015.

125. Smartphone-based Activity Recognition using Hybrid Classifier - Utilizing Cloud Infrastructure for Data Analysis: In: Proceedings of the 4th International Conference on Pervasive and Embedded Computing and Communication Systems. Lisbon, Portugal: SCITEPRESS - Science and and Technology Publications; 2014. p. 14–23.

126. Pirttikangas S, Fujinami K, Nakajima T. Feature Selection and Activity Recognition from Wearable Sensors. In: Youn HY, Kim M, Morikawa H, editors. Ubiquitous Computing Systems. Berlin, Heidelberg: Springer Berlin Heidelberg; 2006. p. 516–27.

127. Dillon CB, Fitzgerald AP, Kearney PM, Perry IJ, Rennie KL, Kozarski R, et al. Number of days required to estimate habitual activity using wrist-worn GENEActiv accelerometer: a cross-sectional study. PloS One. 2016;11(5):e0109913.

128. Sasaki JE, Hickey AM, Staudenmayer JW, John D, Kent JA, Freedson PS. Performance of Activity Classification Algorithms in Free-Living Older Adults: Med Sci Sports Exerc. 2016;48(5):941–950.

129. Kim Y, Barry VW, Kang M. Validation of the ActiGraph GT3X and activPAL accelerometers for the assessment of sedentary behavior. Meas Phys Educ Exerc Sci. 2015;19(3):125–137.

130. Edwardson CL, Winkler EAH, Bodicoat DH, Yates T, Davies MJ, Dunstan DW, et al. Considerations when using the activPAL monitor in field-based research with adult populations. J Sport Health Sci. 2017;6(2):162–78.

131. Stathi A, Withall J, Greaves CJ, Thompson JL, Taylor G, Medina-Lara A, et al. A community-based physical activity intervention to prevent mobility-related disability for retired older people (REtirement in ACTion (REACT)): study protocol for a randomised controlled trial. Trials. 2018;19(1):228.

132. Guralnik JM, Simonsick EM, Ferrucci L, Glynn RJ, Berkman LF, Blazer DG, et al. A short physical performance battery assessing lower extremity function: association with self-reported disability and prediction of mortality and nursing home admission. J Gerontol. 1994;49(2):85–94.

133. Wilcoxon F. Individual comparisons by ranking methods. In: Breakthroughs in statistics. Springer; 1992. p. 196–202.

134. Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies 2011;2:37-63.

135. Breiman L. Random forests. Mach Learn. 2001;45(1):35–32.

136. Hastie T, Tibshirani R, Friedman J, Franklin J. The elements of statistical learning: data mining, inference and prediction. Math Intell. 2005;27(2):83–85.

137. Csurka G. Domain adaptation in computer vision applications. Springer; 2017.

138. Cook D, Feuz KD, Krishnan NC. Transfer learning for activity recognition: A survey. Knowl Inf Syst. 2013;36(3):537–556.

139. Lever J, Krzywinski M, Altman N. Points of significance: Principal component analysis. Nature Publishing Group; 2017.

140. Fernando B, Habrard A, Sebban M, Tuytelaars T. Unsupervised visual domain adaptation using subspace alignment. In: Proceedings of the IEEE international conference on computer vision. 2013. p. 2960–2967.

141. Gjoreski H, Bizjak J, Gjoreski M, Gams M. Comparing Deep and Classical Machine Learning Methods for Human Activity Recognition using Wrist Accelerometer. Proceedings of the IJCAI 2016 Workshop on Deep Learning for Artificial Intelligence. p. 936–43

142. Garcia-Ceja E, Brena R. Long-term activity recognition from accelerometer data. Procedia Technol. 2013;7:248–256.

143. Fubini E, Guillemin E. Minimum insertion loss filters. Proc IRE. 1959;47(1):37–41.

144. Reeves J, Chen J, Wang XL, Lund R, Lu QQ. A review and comparison of changepoint detection techniques for climate data. J Appl Meteorol Climatol. 2007;46(6):900–915.

145. Adams RP, MacKay DJ. Bayesian online changepoint detection. ArXiv Prepr ArXiv07103742. 2007.

146. Noor MHM, Salcic Z, Kevin I, Wang K. Adaptive sliding window segmentation for physical activity recognition using a single tri-axial accelerometer. Pervasive Mob Comput. 2017;38:41–59.

147. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta BBA-Protein Struct. 1975;405(2):442–451.

148. Lyden K, Keadle SK, Staudenmayer J, Freedson PS. A method to estimate free-living active and sedentary behavior from an accelerometer. Med Sci Sports Exerc. 2014;46(2):386-397.

149. Doherty AR, Caprani N, Conaire CÓ, Kalnikaite V, Gurrin C, Smeaton AF, et al. Passively recognising human activities through lifelogging. Comput Hum Behav. 2011;27(5):1948–1958.

150. Webber Jr CL, Marwan N. Recurrence quantification analysis. Theory Best Pract. 2015;

151. Boeing G. Visual analysis of nonlinear dynamical systems: chaos, fractals, self-similarity and the limits of prediction. Systems. 2016;4(4):37.

152. Young L-S. Entropy, Lyapunov exponents, and Hausdorff dimension in differentiable dynamical systems. IEEE Trans Circuits Syst. 1983;30(8):599–607.

153. Hekler EB, Buman MP, Poothakandiyil N, Rivera DE, Dzierzewski JM, Aiken Morgan A, et al. Exploring behavioral markers of long-term physical activity maintenance: a case study of system identification modeling within a behavioral intervention. Health Educ Behav. 2013;40:51–62.

154. Takens F. Detecting strange attractors in turbulence. In: Dynamical systems and turbulence, Warwick 1980. Springer; 1981. p. 366–381.

155. Broomhead DS, King GP. Extracting qualitative dynamics from experimental data. Phys Nonlinear Phenom. 1986;20(2–3):217–236.

156. Hatami N, Gavet Y, Debayle J. Classification of time-series images using deep convolutional neural networks. In: Tenth International Conference on Machine Vision (ICMV 2017). International Society for Optics and Photonics; 2018.

157. Coates A, Ng AY. Learning feature representations with k-means. In: Neural networks: Tricks of the trade. Springer; 2012. p. 561–580.

158. Rish I, others. An empirical study of the naive Bayes classifier. In: IJCAI 2001 workshop on empirical methods in artificial intelligence. 2001. p. 41–46.

159. Hand DJ, Yu K. Idiot's Bayes: Not So Stupid after All? Int Stat Rev. 2001;69(3):385-399.

160. Liu L, Peng Y, Liu M, Huang Z. Sensor-based human activity recognition system with a multilayered model using time series shapelets. Knowl-Based Syst. 2015;90:138–52.

161. Liu L, Peng Y, Wang S, Liu M, Huang Z. Complex activity recognition using time series pattern dictionary learned from ubiquitous sensors. Inf Sci. 2016;340–341:41–57.

162. Lockhart JW, Weiss GM. The Benefits of Personalized Smartphone-Based Activity Recognition Models. In: Zaki M, Obradovic Z, Tan PN, Banerjee A, Kamath C, Parthasarathy S. Proceedings of the 2014 SIAM International Conference on Data Mining. Philadelphia, PA: Society for Industrial and Applied Mathematics; 2014. p. 614–22.

163. Siirtola P, Koskimäki H, Röning J. Personalizing human activity recognition models using incremental learning. ArXiv Prepr ArXiv190512628. 2019.

164. Pober DM, Staudenmayer J, Raphael C, Freedson PS. Development of Novel Techniques to Classify Physical Activity Mode Using Accelerometers: Med Sci Sports Exerc. 2006;38(9):1626–34.

165. Kwapisz JR, Weiss GM, Moore SA. Activity recognition using cell phone accelerometers. ACM SIGKDD Explor Newsl. 2011;12(2):74.

166. Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, et al. Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 4681–4690.

167. Gilbert E, Conolly A, Tietz S, Calderwood L, Rose N. Measuring young people's physical activity using accelerometers in the UK Millennium Cohort Study. Cent Longitud Stud Work Pap. 2017.

168. Ni Q, Patterson T, Cleland I, Nugent C. Dynamic detection of window starting positions and its implementation within an activity recognition framework. J Biomed Inform. 2016;62:171–180.

169. Anderson MM. Physical activity recognition of free-living data using change-point detection algorithms and hidden Markov models. 2013; Doctoral Dissertation, Oregon State University.

170. Willetts M, Hollowell S, Aslett L, Holmes C, Doherty A. Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 UK Biobank participants. Sci Rep. 2018;8(1):1–10.

171. Lesaffre E. Superiority, equivalence, and non-inferiority trials. Bull NYU Hosp Jt Dis. 2008;66(2):150-154.

172. Alvarez GG, Ayas NT. The impact of daily sleep duration on health: a review of the literature. Prog Cardiovasc Nurs. 2004;19(2):56–59.

173. Tremblay MS, Carson V, Chaput J-P, Connor Gorber S, Dinh T, Duggan M, et al. Canadian 24-hour movement guidelines for children and youth:

an integration of physical activity, sedentary behaviour, and sleep. Appl Physiol Nutr Metab. 2016;41(6):311–327.

174. Oliver M, Badland HM, Schofield GM, Shepherd J. Identification of accelerometer nonwear time and sedentary behavior. Res Q Exerc Sport. 2011;82(4):779–783.

175. Zhou S-M, Hill RA, Morgan K, Stratton G, Gravenor MB, Bijlsma G, et al. Classification of accelerometer wear and non-wear events in seconds for monitoring free-living physical activity. BMJ Open. 2015;5(5):e007447.

176. Crasborn O, Sloetjes H, Auer E, Wittenburg P. Combining video and numeric data in the analysis of sign languages with the ELAN annotation software. In: 2nd workshop on the representation and processing of sign languages: lexicographic matters and didactic scenarios. ELRA; 2006. p. 82–87.

177. Tucker JM, Welk GJ, Beyler NK, Kim Y. Associations between physical activity and metabolic syndrome: comparison between self-report and accelerometry. Am J Health Promot. 2016;30(3):155–162.

178. Sabia S, Cogranne P, van Hees VT, Bell JA, Elbaz A, Kivimaki M, et al. Physical activity and adiposity markers at older ages: accelerometer vs questionnaire data. J Am Med Dir Assoc. 2015;16(5):438–447.

179. Chinapaw MJ, Wang X, Andersen LB, Altenburg TM. From total volume to sequence maps: sophisticated accelerometer data analysis. Med Sci Sports Exerc. 2019;51(4):814–820.

180. Carson V, Tremblay MS, Chaput J-P, McGregor D, Chastin S. Compositional analyses of the associations between sedentary time, different intensities of physical activity, and cardiometabolic biomarkers among children and youth from the United States. PloS One. 2019;14(7).

181. Hildebrand M, Hansen BH, van Hees VT, Ekelund U. Evaluation of raw acceleration sedentary thresholds in children and adults. Scand J Med Sci Sports. 2017;27(12):1814–1823.

182. Chastin S, Granat MH. Methods for objective measure, quantification and analysis of sedentary behaviour and inactivity. Gait Posture. 2010;31(1):82–86.

183. Cavanaugh JT, Kochi N, Stergiou N. Nonlinear analysis of ambulatory activity patterns in community-dwelling older adults. J Gerontol Ser Biomed Sci Med Sci. 2010;65(2):197–203.

184. Boerema ST, van Velsen L, Vollenbroek MM, Hermens HJ. Pattern measures of sedentary behaviour in adults: A literature review. Digit Health. 2020.

185. Paraschiv-Ionescu A, Perruchoud C, Buchser E, Aminian K. Barcoding human physical activity to assess chronic pain conditions. PloS One. 2012;7(2).

186. Colberg SR, Sigal RJ, Yardley JE, Riddell MC, Dunstan DW, Dempsey PC, et al. Physical activity/exercise and diabetes: a position statement of the American Diabetes Association. Diabetes Care. 2016;39(11):2065–2079.

187. Tremblay MS, Aubert S, Barnes JD, Saunders TJ, Carson V, Latimer-Cheung AE, et al. Sedentary behavior research network (SBRN)–terminology consensus project process and outcome. International Journal of Behavioral Nutrition and Physical Activity. 2017;14(1):75.

188. Chastin SF, Helbostad JL, Tremblay MS, Ainsworth B, Mork PJ, Rochester L, et al. AlPHABET: Development of A Physical Behaviour Taxonomy with an international open consensus. 2016.

189. OASIS COEL TC. Classification of Everyday Living Version 1.0. OASIS. 2018.

190. Ganyo M, Dunn M, Hope T. Ethical issues in the use of fall detectors. Ageing & Society. 2011 Nov;31(8):1350-67.

191. Voigt P, Von dem Bussche A. The EU general data protection regulation (GDPR). A Practical Guide, 1st Ed., Cham: Springer International Publishing. 2017.

192. Yates T, Henson J, Edwardson C, Bodicoat DH, Davies MJ, Khunti K. Differences in levels of physical activity between White and South Asian populations within a healthcare setting: impact of measurement type in a cross-sectional study. BMJ open. 2015 Jul 1;5(7).

193. Fuller D, Shareck M, Stanley K. Ethical implications of location and accelerometer measurement in health research studies with mobile sensing devices. Social Science & Medicine. 2017 Oct 1;191:84-8.

194. Hern A. Fitness tracking app strava gives away location of secret us army bases, Jan 2018. URL

https://www.theguardian.com/world/2018/jan/28/fitness-tracking-app-gives-away-location-of-secret-us-army-bases.

195. Thang HM, Viet VQ, Thuc ND, Choi D. Gait identification using accelerometer on mobile phone. 2012 International Conference on Control, Automation and Information Sciences (ICCAIS) 2012 Nov 26 (pp. 344-348). IEEE.

196. Ayme S. Data storage and DNA banking for biomedical research: technical, social and ethical issues. European Journal of Human Genetics. 2003 Dec;11(12):906-8.

197. de Vries HJ, Kooiman TJ, van Ittersum MW, van Brussel M, de Groot M. Do activity monitors increase physical activity in adults with overweight or obesity? A systematic review and meta-analysis. Obesity. 2016 Oct;24(10):2078-91.

198. Plateau CR, Bone S, Lanning E, Meyer C. Monitoring eating and activity: links with disordered eating, compulsive exercise, and general wellbeing among young adults. International Journal of Eating Disorders. 2018 Nov;51(11):1270-6.

199. Koenecke A, Nam A, Lake E, Nudell J, Quartey M, Mengesha Z, Toups C, Rickford JR, Jurafsky D, Goel S. Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences. 2020 Apr 7;117(14):7684-9.

200. Eyal N. Denial of treatment to obese patients—the wrong policy on personal responsibility for health. International Journal of Health Policy and Management. 2013 Aug;1(2):107.

201. Henkel M, Heck T, Göretz J. Rewarding Fitness Tracking—The Communication and Promotion of Health Insurers' Bonus Programs and the Use of Self-tracking Data. InInternational Conference on Social Computing and Social Media 2018 Jul 15 (pp. 28-49). Springer, Cham.

202. Saeb S, Körding K, Mohr DC. Making activity recognition robust against deceptive behavior. PloS one. 2015 Dec 11;10(12):e0144795.