COMPARING ACCELEROMETER PROCESSING METRICS FOR PHYSICAL ACTIVITY CLASSIFICATION ACCURACY USING MACHINE LEARNING METHODS

By Musa, Sumayyah Bamidele

A thesis submitted to the School of Graduate Studies in partial fulfillment of the requirements for the degree of

Master of Science in Kinesiology

Memorial University of Newfoundland

September 2022

St. John’s, Newfoundland & Labrador

Table of Contents

ABSTRACT

The purpose of this study was to compare the performance of three accelerometer processing metrics, Euclidean Norm Minus One (ENMO), ActiGraph Counts, and Monitor Independent Movement Summary (MIMS) units, in classifying physical activity using Random Forest (RF) and k-Nearest Neighbors (KNN) machine learning models, as well as to investigate the effect of hyperparameter tuning and feature selection on each processing metric. The dataset was sourced from a laboratory-based protocol involving raw acceleration data from 50 participants who held a smartphone device in their right hand while completing six activities.

Findings indicated that even though the acceleration metrics performed well above 80% accuracy with both RF and KNN, the best performance was achieved with ENMO and the raw data as features. Additional accuracy of between 1% to 5% was achieved when the model hyperparameters were tuned before classification, and there was no difference when other features were included in the classification.

In conclusion, ENMO is the best acceleration metric for classifying PA from accelerometers. Tuning the models and using a few selected features affected the models' accuracy.

Keywords: accelerometer-based physical activity, machine learning classification, ActiGraph counts, MIMS units, ENMO

GENERAL SUMMARY

This study was carried out to compare three methods of processing data retrieved from accelerometers (acceleration metrics) in smartphone devices. Fifty participants held the smartphone in their right hand while performing six laboratory activities for 65 minutes. These activities were sitting, lying, walking on the treadmill at a selected pace, walking at 3 Mets, running at 5 Mets, and running at 7 Mets. The acceleration metrics, ActiGraph counts, ENMO and MIMS units were computed using R programming. The metrics were used along with 58 other extracted variables to classify the activities using random forest, support vector machine and K-nearest neighbour algorithms. Results showed that ENMO is the best acceleration metric to implement when classifying physical activity with machine learning-based methods; utilizing the raw data and acceleration metrics without the other extracted features produced the best classification results.

ACKNOWLEDGEMENTS

First and foremost, I would like to praise Allah the Almighty, the Most Gracious, and the Most Merciful for His blessing during my study and in completing this thesis. May Allah's blessing goes to His final Prophet Muhammad (peace be upon him), his family and his companions.

I would like to express my gratitude and sincere thanks to my research supervisor, Dr. Daniel Fuller, Ph.D., Associate Professor and Canada Research Chair in Population Physical Activity, Memorial University of Newfoundland, for his invaluable guidance, continuous support, advice, and encouragement so I could complete this thesis in time. His dynamism, vision, sincerity, and motivation have deeply inspired me. Working and studying under his guidance was a great privilege and honour. I am extending my heartfelt thanks to his wife and kids for their patience during the several meetings I had with him on research work and thesis preparation. I would also like to thank the rest of my thesis committee, Dr. Kevin Stanley, Dr. Fabien Basset, and Dr. Angela Loucks-Atkinson, for their encouragement, corrections, insightful comments, and suggestions.

My ultimate thanks are dedicated to my beloved parents, Dr. N.A Musa and Hj. R.B Oladimeji, and parents-in-law, N. Sulaiman and R. Adesanwo, for their endless support, love, and prayers. I am incredibly grateful to my husband for his love, patience, sacrifice, understanding, prayers, and continuing support along the long and winding road of completing this thesis. Also, I express my special thanks to my sisters, brothers, sisters-in-law and brothers-in-law for their support and valuable prayers.

Finally, my gratitude goes to all my friends and all persons whose names cannot be mentioned one by one for their help, kindness, and support, directly or indirectly, in completing this thesis.

LISTS OF FIGURES AND TABLES

LISTS OF APPENDICES

LISTS OF ABBREVIATIONS

CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW

1.1 Background to the study

Physical Activity (PA) is any bodily movement produced by skeletal muscles that result in energy expenditure (Siscovick et al., 1985). This term encompasses a range of human movements, including competitive sports, exercise, activities of daily living like occupational, household (e.g., caregiving, domestic cleaning), transport (e.g., walking or cycling to work) and leisure-time activities (e.g., dancing, swimming). Regular PA is a crucial protective factor for preventing and managing non-communicable diseases (NCDs) such as cardiovascular disease, type-2 diabetes, and several cancers (I.-M. Lee et al., 2012). PA also benefits mental health, including preventing cognitive decline and symptoms of depression and anxiety and can contribute to maintaining a healthy weight and general well-being (World Health Organization, 2020).

PA is one of the essential components of successful health promotion and disease prevention for individuals and communities. Physical inactivity is the fourth leading cause of chronic disease mortality, such as cancers, heart disease and stroke, contributing to over three million preventable deaths worldwide (World Health Organization, 2020). Physical behaviours relating to PA during the 24 hours of the day are of significant public health interest due to their well-documented influence on mortality (Ekelund et al., 2019). A recent systematic review of 8 studies representing 36,386 participants showed that physical activity, regardless of intensity, was associated with reduced mortality risk. Compared to the least active group, even people who did a small amount of exercise was 54% less likely to die prematurely. For the most active group, premature death was 73% less likely than in the least active group.

Physical inactivity and sedentary behaviour have been identified as critical issues in Canada and internationally. For example, *A Common Vision for Increasing Physical Activity and Reducing Sedentary Living in Canada: Let's Get Moving* is a policy that strives for a Canada where all

Canadians move more and sit less, more often. A study by Colley et al., (2018) assessed the trend of moderate-to-vigorous physical activity (MVPA) levels in 13,173 Canadian adults aged 18 - 79 years from 2007 - 2017 using the Canadian Health Measures Survey (CHMS) accelerometer data. They found that about 3% of Canadian adults accumulated no MVPA, while nearly 36% did not accumulate any MVPA in bouts of at least 10 minutes. Their findings suggest that fewer than one in five Canadian adults met the Canadian Physical Activity Guidelines. MVPA levels were higher among men compared with women and younger adults compared with older adults.

WHO data also show that one in four adults, and four out of five adolescents, do not get enough PA. Globally this is estimated to cost US$54 billion in direct health care and another US$14 billion in lost productivity (World Health Organization, 2020). Global estimates also indicate that 27.5% of adults (Guthold et al., 2018) and 81% of adolescents (Guthold et al., 2020) do not meet WHO recommendations for PA, with almost no improvements seen during the past decade (World Health Organization, 2010). A systematic review examining the associations between sedentary behaviour and all-cause mortality found that high levels of moderate-intensity PA seem to eliminate the increased risk of death associated with high sitting time (Ekelund et al., 2016). Lee et al. (2012) quantified the effect of physical inactivity on non-communicable diseases. They estimated that more than 1.3 million deaths could be averted globally if physical inactivity was reduced by 25%.

There are several ways to measure physical activity, including using self-report questionnaires and accelerometers. While there is considerable debate in the literature about the similarities and differences between these methods, accelerometers are becoming more commonly used. They are being used in more extensive studies (Kelly et al., 2016). As the use of accelerometers to measure and monitor physical activity (PA) continues to be emphasized by many researchers (Wijndaele et al., 2015), a significant concern is the standardization of raw acceleration data processing techniques. As a result, for this thesis, it is worth reviewing commonly used methods for physical activity measurement.

1.2 Physical Activity Measurements

1.2.1 Self-report Measurements

Accurate and valid measurement of free-living PA is vital to estimate the prevalence of populations meeting the current PA guidelines, assess the success of interventions aiming to increase PA in specific people, and quantify PA's dose-response impact on health (I. M. Lee & Shiroma, 2014). Self-reported physical activity measures have long been used to measure physical activity because they are relatively simple and inexpensive to administer.

The two most common self-report PA measurement tools for adults are the International Physical Activity Questionnaire (IPAQ) and the Global Physical Activity Questionnaire (GPAQ) (Strath et al., 2013). The IPAQ validation study was published in 2003 by Craig et al., 2003. The purpose of the IPAQ is to provide standard instruments that can be used to obtain internationally comparable data on health-related PA. The IPAQ has two different versions, the short and the long. The short version includes nine items that capture four domains of activity intensities, including sedentary time and light, moderate, and vigorous activity over the past week. The extended version of the IPAQ includes 27 items that capture five activity domains (work, leisure, household, transport, and sedentary time) and their intensities over the past week. The IPAQ can be administered over the telephone or self-administered online. The general structure of all IPAQ questions is as follows. First, the participant is asked how many days they did an activity in the last seven days of the week. Second, participants are asked to provide how much time they spent on average doing the activity each day. These two answers allow researchers to calculate the number of minutes an activity was conducted over the past seven days or estimate Metabolic Equivalents (METs) using the Compendium for Physical Activity (Ainsworth et al., 2011).

The GPAQ validation study was published in 2009 by Bull et al., 2009. The World Health Organization developed the GPAQ in 2002 to enhance the IPAQ and develop a standardized tool that enables PA comparisons across culturally diverse populations (Armstrong & Bull, 2006). The original version of the GPAQ included 19 items that capture five activity domains (work, leisure, household, transport, and sedentary time) and their intensities over the past week, like the IPAQ. However, a shorter version was later developed, eliminating three redundant questions, totalling 16 items and four domains (activity at work, travel to and from places, recreational activities, and sedentary behaviour) for the most updated version. The GPAQ was developed for face-to-face interviews by trained researchers.

The IPAQ (Craig et al., 2003) has reasonable concurrent validity compared to accelerometer data. In an international sample collected in 12 countries (Australia, Brazil, Canada, Finland, Guatemala, Netherlands, Japan, Portugal, South Africa, Sweden, United States, and the United Kingdom) at 14 different sites, 2721 participants completed the short and long-form of the IPAQ and wore an accelerometer for seven days. The criterion validity of the IPAQ was a Spearman correlation coefficient of 0.3, which is comparable to other self-report physical activity measures compared to accelerometry. The IPAQ questionnaires produced repeatable data with Spearman's correlation coefficient of around 0.8) and had similar data from short and long forms. The GPAQ (Keating et al., 2019) has poor to fair concurrent validity compared to accelerometer data. In a systematic review of 26 studies that validated the GPAQ in 23 countries (Bangladesh, China, India, Korea, Malaysia, Saudi Arabia, Singapore, Thailand, United Arab Emirates, Vietnam, Belgium, Chile, Spain, Brazil, Ethiopia, Portugal, South Africa, France, Switzerland, Japan, Indonesia, United States, United Kingdom) from 2002 to 2019. The criterion validity of the GPAQ measures was Spearman's correlation coefficient of less than 0.5 across all the studies, which shows that the validity is relatively low when using an accelerometer as the criterion standard. Test-retest reliability of the GPAQ ranged from moderate to very good, with Pearson's correlation coefficient of between 0.58 to 0.89, which shows that the results of the GPAQ are relatively consistent and reproducible when compared to the IPAQ (Bull et al., 2009). Using accelerometers as the criterion measure for self-report questionnaires is somewhat debated in the literature (Kelly et al., 2016); however, there is considerable variability in physical activity measurement when using accelerometer data.

1.2.2 Accelerometer Measurements

The assessment of daily PA in population studies requires valid, cheap, and feasible measurement technology (Corder et al., 2008; Wong et al., 2003). Accelerometers, which quantify the acceleration and deceleration in orthogonal directions of three-dimensional space, have become the preferred and feasible device to assess PA in large-scale studies (P. Freedson et al., 2012). However, their utilization requires several different data processing and aggregation methods, which can considerably impact physical activity estimation using accelerometers (Migueles et al., 2017).

Broadly, four main steps are required to convert raw accelerometer data to measures of physical activity. First, raw accelerometer data must be processed to extract human movement signals. Many different methods are used to process this data (see section 1.3), all of which use some form of signal processing on the raw accelerometer data. Second, the processed data is filtered to include only data where the participant is wearing the device. Third, the filtered data is used to generate estimates of physical activity intensity or type using rule-based cut-point or machine learning approaches. Fourth, physical activity intensity or type estimates are aggregated to represent meaningful activity metrics. **Figure 1** presents a graphical representation of the accelerometer data processing steps. As an example of how research might process accelerometer data, the most used method for processing raw accelerometer data is referred to as "counts" (or

ActiGraph counts), an aggregate measure of the intensity and magnitude of accelerations over a given time epoch (Colley et al., 2011; Hagströmer et al., 2010). These counts are produced via proprietary algorithms developed and patented by the manufacturers of ActiGraph accelerometers, making it difficult to compare results from studies that have employed ActiGraph and non-ActiGraph devices (Marschollek, 2013). A researcher could convert raw data to counts, then in step 2, apply a wearing algorithm; here, the Choi algorithm (Choi et al., 2011) is very commonly used. In step 3, data from different axes are used individually or combined into a single metric, often using the formula presented in equation 1.

$$vm_i = \sqrt{x_i^2 + y_i^2 + z_i^2}$$

Equation 1

Where *vm* is the vector magnitude (or Euclidean norm), x, y, and z are the raw accelerometer data for each axis, and *i* is the lowest data collection frequency.

Then using a cut-point-based approach, a researcher could aggregate the data by summing the counts to the minute level, then use the Troiano cut-points for adults to define ranges of ActiGraph counts for sedentary behaviour (0-99 counts/min), light intensity PA (100 - 2019 counts/min), moderate-intensity PA (2020-5998 counts/min), and vigorous-intensity PA (5999 - max counts/min) for every minute of accelerometer data in a database. In the fourth step, a researcher could calculate the number of moderate to vigorous physical activity minutes per day or week (Troiano et al., 2008).

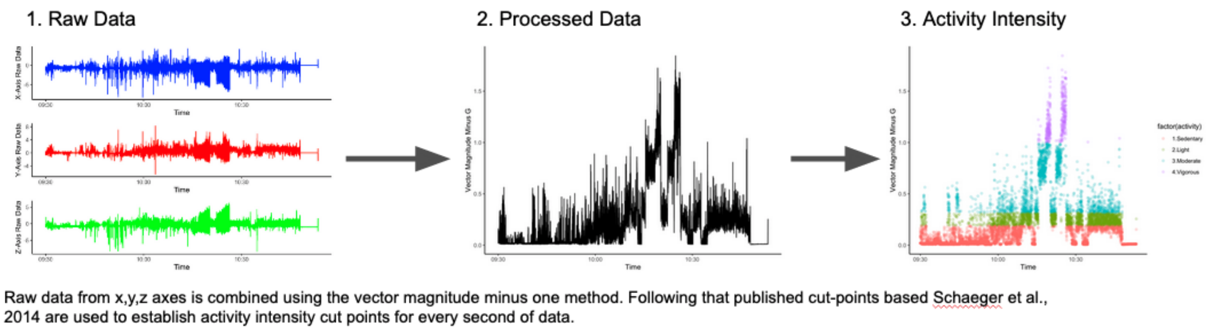


Figure 1: Example of accelerometer data visualized through data processing.

While the above example may seem straightforward, there is considerable debate about the optimal approach for each accelerometer data processing step. Over 100 different methods have been published for cut-point-based approaches using ActiGraph counts (Migueles et al., 2017). Technological advances in storage and computation mean raw data can now be stored at high frequencies, with no need to summarize into proprietary ActiGraph counts (John et al., 2013). Consequently, there is a need to understand better the similarities and differences between raw acceleration data analysis, particularly the data processing and machine learning aspects (Bakrania et al., 2016).

1.3 Comparison between self-report and Accelerometer in Canada

Several studies have shown significant variations between self-reported and accelerometer-measured PA. Dyrstad et al. (2014) compared PA and sedentary time from the self-administered, short version of the IPAQ with data from the ActiGraph accelerometer in 1751 Norwegian adults over seven days. They found that most participants reported less sedentary time, less moderate-intensity, and a higher vigorous-intensity PA level than the accelerometer data. These differences were affected by education level, sex, and age, but not body mass index (BMI). The disparity between self-reported and measured sedentary time and vigorous-intensity PA was most significant among men with less education and men 65 years and older. Their study also showed that men reported 47% more moderate to vigorous physical activity (MVPA) than women, but there were no differences in accelerometer-determined MVPA.

Additionally, they reported weak correlation coefficients of 0.20 and 0.46 between self-reported variables and accelerometer PA measures. Colley et al., (2018) compared the estimates of self-reported PA among 2,372 Canadian adults using a newly developed Canadian questionnaire - Physical Activity Adult Questionnaire (PAAQ), with those obtained objectively using an Actical accelerometer. Their findings showed that, on average, the participants reported more PA than they accumulated when measured with the accelerometer. The correlation they discovered between self-reported data from the new questionnaire module and accelerometer-measured physical activity was Pearson's coefficient of 0.21, which was weak. Generally, when physical activity is measured with self-report, people tend to have higher estimated levels of moderate to vigorous PA and lower levels of sedentary behaviour. Despite the potential differences between self-report and accelerometer data, there can also be considerable variability in physical activity estimation between different accelerometer data processing methods.

1.4 Current challenges with accelerometer measurement

Several factors affect the analyses of raw acceleration signals. These include the management of the vast amount of data which are generated; the requirement to remove the gravitational and noise components incorporated within the signals (van Hees et al., 2013); and the requirement of feasible mathematical and statistical tools to accurately analyze and make valid estimates of physical activity from the data (Bakrania et al., 2016). The following section discusses two primary aspects of accelerometer data analysis that are the focus of this thesis, raw accelerometer data processing and machine learning. It does not focus on wearing algorithms or cut-point-based approaches to estimation.

1.5 Raw Accelerometer Data Processing

Accelerometers measure the acceleration of the body segment to which the monitor is attached. They capture voluntary and involuntary human movements and movements influenced by the environment (e.g., vibrations from being in a car). These signals are then filtered and preprocessed to generate summary measures of human activity, which attempt to capture the intensity and magnitude of the acceleration due to body movements (Migueles et al., 2017). Data processing aims to eliminate the gravitational component and noise from the raw acceleration signal to create an acceleration signal aggregation metric (hence called acceleration metric) that can be used to represent the bodily movement (van Hees et al., 2013). These acceleration metrics are then used to classify the amount and intensity of daily PA in a specific time interval (epoch length) with a set of cut-points, i.e., thresholds for PA intensity classification. ActiGraph counts are generated via patented and proprietary algorithms, making it difficult to compare data and results between studies using different accelerometer brands. Bakrania et al. 2016, have developed various techniques to remove gravitational and noise components and correctly estimate PA from raw acceleration signals. Commonly used procedures for processing the raw acceleration data and separating the movement and gravitational components of the movement signals include ActiGraph counts, Euclidean Norm Minus One (ENMO) metrics, and the Monitor-Independent Movement Summary (MIMS) units.

ActiGraph counts are a proprietary manufacturer-specific metric used for ActiGraph brand accelerometers (Brond & Arvidsson, 2016). ActiGraph counts are unitless measures, computed to 1Hz, of either single or combined tri-axial acceleration. Brond & Arvidsson (2016) attempted to reverse engineer the ActiGraph counts algorithm using several methods, including orbital shakers and human participants. Brond’s work showed that the ActiGraph counts' algorithm involved many data processing steps, including alias filtering, down sampling, and signal rectification. **Figure 2** summarizes the process steps to convert raw accelerometer signals to ActiGraph counts.

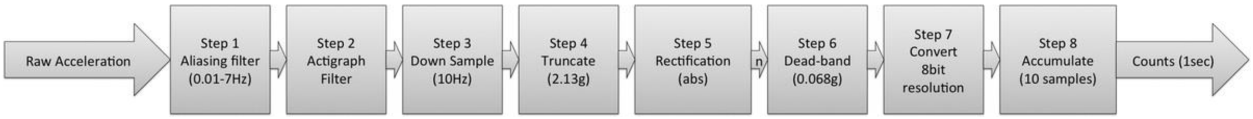


Figure 2: Steps for processing raw accelerometer data and generating ActiGraph counts. Figure adapted from (Brond & Arvidsson, 2016)

ActiGraph accelerometers and ActiGraph counts are the most used accelerometers in physical activity research. As a result, a considerable body of literature is developing cut-points of physical activity estimation among different populations (Hernando et al., 2018). ActiGraph counts were developed for specific accelerometers and are not comparable across devices or models from the same manufacturer (Grydeland et al., 2014). A cross-sectional study of 50 participants was carried out by Sasaki et al., (2011) to compare activity counts from the ActiGraph GT3X to those from the ActiGraph GT1M during treadmill walking/running. Their findings showed that the anterior-posterior activity counts and the vector magnitude of the vertical activity counts were significantly higher in the GT1M monitor than those of the GT3X monitor at different speeds.

Based on the work done by Brond & Arvidsson (2016), Brondeel et al., (2021) have developed open-source codes in MATLAB, Python and R, three different programming languages, to convert raw accelerometer data to activity counts irrespective of the accelerometer brands, models, or devices implemented. The development of open-source code has allowed for the calculation of ActiGraph counts for studies using any accelerometer brand.

Euclidean norm minus one (ENMO) is an open-source method to process raw acceleration data. ENMO does not require the raw data to be filtered to correct gravity since it systematically considers this element within its algorithm (Vähä-Ypyä et al., 2015). It adjusts for gravity by subtracting a fixed offset of one gravitational unit from the Euclidean Norm of the three raw acceleration signals. The Euclidean Norm is obtained by taking the Euclidean norm (sometimes known as vector magnitude in the physical activity literature), i.e., the square root of the sum of the squares of the three raw acceleration signals and subtracting the fixed offset value of 1. See equation 2:

$$Euclidian\ Norm\ Minus\ One\ (ENMO) = r_i - 1000$$

1000 = 1000 Milli gravitational units = 1 gravitational unit

Where,

$$r_i = \sqrt{x_i^2 + y_i^2 + z_i^2} = i^{th}\ vector\ magnitude\ at\ each\ time\ point$$

Equation 2

A comparison of the performances of ENMO and the Mean Amplitude Deviation (MAD) metrics were carried out by Bakrania et al., (2016) in 33 participants that wore four accelerometers in different body segments and performed 16 activities (11 sedentary behaviours and five. light-intensity physical activities) for 5 minutes each. Their study showed that both metrics performed similarly across activities and accelerometer brands. Bai et al., (2016) compared ActiGraph counts to ENMO using a sample of 194 women aged 60-91. They found that ENMO were more sensitive to moderate and vigorous physical activities than ActiGraph counts. Clevenger et al., (2020), in a sample of 54 adult participants, showed that epoch-level data were not identical between ENMO and ActiGraph counts. However, most outcomes were strongly related between models (e.g., ENMO, ActiGraph counts) and similar once aggregated to the number of minutes spent in different activity intensities.

MIMS-unit is abbreviated for Monitor Independent Movement Summary unit. This open-source measurement was developed recently by John et al., (2019) to harmonize accelerometer data processing from different devices. The algorithm was developed using digital signal processing techniques to harmonize raw data from devices with different dynamic ranges and sampling rates and then aggregate the raw data to capture normal human motion. MIMS unit uses raw signal harmonization to eliminate inter-device variability (e.g., dynamic range, sampling rate), bandpass filtering (0.2 – 5.0 Hz) to eliminate non-human movement and signal aggregation to reduce data to simplify visualization. **Figure 3** shows a representation of the MIMS-unit processing steps.

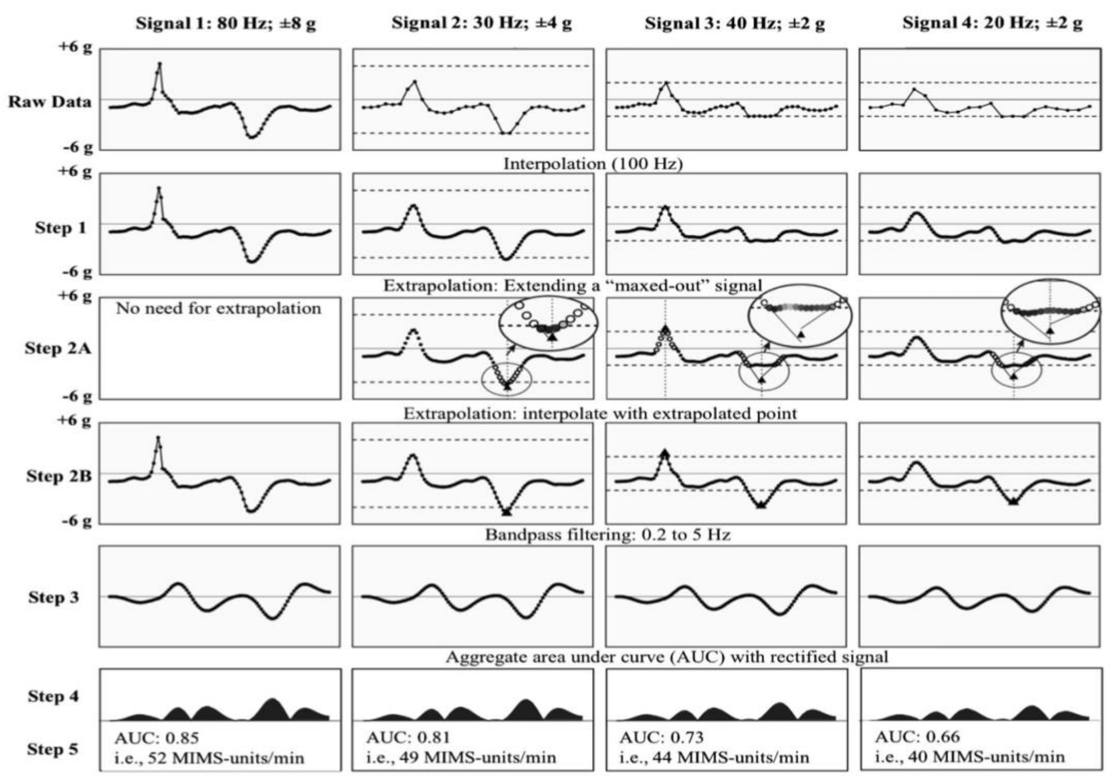


Figure 3: MIMS unit process steps. Figure adapted from John et al., (2019)

The MIMS algorithm was tested on 60 participants by John et al., (2019) to compare the inter-output behaviour patterns among MIMS units, ActiGraph counts, and ENMO during human movement. Their findings showed that ActiGraph counts' inability to detect signals representing sedentary behaviour with minimal movement was overcome by MIMS units and ENMO processing methods. They did not show any differences between MIMS-unit and ENMO for most activities studied, except in 2 cases, playing frisbee and activities involving running faster than 8.8km/h, where a significant intra-location difference was found between the hip and wrist.The wrist-worn accelerometers produced higher values than the hip-worn.

1.6 Physical Activity Classification

Physical activity measurement using accelerometer data relies on processing raw data, as described above, followed by methods to classify activity intensity, activity type, or a combination of activity intensity and activity type. There are two general approaches for PA classification (activity classification will be used to classify activity intensity, type, or a combination of intensity and type); cut-point-based and machine learning-based approaches. Cut-point-based approaches are prevalent in PA research and have been the subject of a systematic review (Bianchim et al., 2019). Cut-point-based methods use a single summary measure of acceleration (e.g., ActiGraph counts or ENMO) and apply thresholds, known as cut-points, to define categories of activity classification. For example, the Freedson cut-points, arguably the most used cut-points in the PA literature, determine PA intensities as sedentary (<99 counts), light (100-759 counts), moderate (760-5724 counts), and vigorous (5725-max counts) intensity (P. S. Freedson et al., 1998). Cut-points for activity classification are device, wear location, and population-specific. There are many different cut-points, and it is often challenging for researchers to select the appropriate cut-point for their specific study population, accelerometer device, and wear location (Kim et al., 2012). Due to these limitations, manufacturers were urged to provide access to raw, unfiltered accelerometer data to promote transparency (John & Freedson, 2012). However, raw accelerometer data are voluminous (usually between 30 and 100 samples per second), making the data challenging to manage and interpret. While cut-point-based approaches continue to be developed, machine learning is increasingly used in PA classification (Narayanan et al., 2020).

Researchers have employed machine learning (ML) approaches to process and analyze raw accelerometer data while capturing all PA components (frequency, intensity, time, and type). Classification of PA with ML algorithms relies on using multiple features (i.e., variables) derived from the raw accelerometer signal (S. Liu et al., 2012). These ML algorithms generate a predictive model by learning how patterns in the accelerometer data features are related to an activity type or intensity. Classification of PA using ML is done by training a model with features of the accelerometer signal (e.g., mean, SD, and correlations) extracted from the raw or processed accelerometer data. Trained models can then classify activity from features extracted from different accelerometer data. This method, called supervised learning, requires a direct observation measure on which the model can be trained (Narayanan et al., 2020).

Three significant aspects related to the performance of machine learning models in the classification of PA are essential for this thesis. First are the acceleration metrics and the features extracted from the raw acceleration signals used as predictors in the machine learning models. The acceleration metrics included in this thesis were Euclidean Norm Minus One (ENMO), ActiGraph counts, and Monitor Independent Movement Summary (MIMS) units. On the other hand, the extracted features included both time and frequency domain features, for example, interquartile range, peak amplitude and zero crossings. While several studies have compared the performances of ENMO with ActiGraph counts (Migueles et al., 2019), no study has methodically evaluated the performance of the MIMS unit's metric with machine learning relating to physical activity.

Second, the ML algorithms utilized in classifying PA. This is the first study to examine the interaction of different ML algorithms with various acceleration metrics. Third, the hyperparameter settings for the machine learning algorithms. A hyperparameter is a parameter whose value can be used to control a machine learning algorithm's training process and behaviour. It is related to how the model learns the patterns based on data. It has been reported byLavesson & Davidsson, (2006) that the settings of a hyperparameter have a significant impact on the classification accuracy of the resulting trained model as it relates to training time and model accuracy. Additionally, (Probst et al., 2018) showed that tuning these hyperparameters, as opposed to using their default values, is often more important than the choice of the machine learning algorithm itself. To my knowledge, no study has reported the hyperparameter settings or tested the potential impact of model tuning in examining the performance of machine learning algorithms concerning acceleration metrics and PA. There is a need to understand better how acceleration metrics combined with different machine learning algorithms influence PA classification.

CHAPTER 2: RESEARCH PROBLEM

2.1 Statement of the Problem

Several methods for processing raw accelerometer data have been developed and used in the literature. Studies have shown that these methods might not be optimal for representing the movement behaviour (John et al., 2019). There are also potential interactions between accelerometer processing methods and the machine learning algorithms used to classify activity. A systematic study has not been undertaken to test these possible interactions using an open-source dataset. Therefore, this study aims to (1) examine the acceleration metrics using dominant hand wear location during different activities; (2) compare and evaluate the performances between the ENMO, ActiGraph counts and MIMS-unit metrics using three machine learning algorithms for PA classification, RF, K-Nearest Neighbour (KNN) and SVM; and (3) examine the impact of hyperparameter tuning on these machine learning algorithms as related to PA.

2.2 Objective of the Study

The purpose of the study is first to compare three commonly used acceleration metrics, ActiGraph counts, ENMO, and MIMS-units, using three different machine learning approaches, and second to examine the effects of hyperparameter tuning on PA classification when machine learning methods are implemented.

2.3 Research Hypothesis

The research hypothesis for objective 1 is that the MIMS-units metric will perform better than ENMO and ActiGraph counts in PA classification using RF, KNN, and SVM classification models. The hypothesis for objective 2 is that model tuning of hyperparameters will influence model performance by between 5-10% for each type of model.

2.4 Importance of the Study

The study results will provide evidence on (1) the consistency and performance of the MIMS-units metric as a standardized acceleration data cleaning and processing technique for PA classification using ML methods and (2) the influence of hyperparameter tuning when using ML for PA classification. To my knowledge, few studies have directly compared MIMS to other accelerometer processing methods. No studies around physical activity intensity or type classification have examined the impact of hyperparameter tuning on model performance.

CHAPTER 3: METHODOLOGY

3.1 Research Design

This is a cross-sectional study to examine the accuracy of PA classification algorithms when using each of these summarized acceleration metrics; MIMS-units, ENMO (Euclidean Norm Minus One) and ActiGraph counts.

3.2 Dataset

This dataset utilized for this study was raw, time-stamped accelerometer data collected from a laboratory-based protocol conducted by the Built Environment and Active Populations (BEAP) Lab. Fuller et al., (2021) give a detailed description of that study. Fifty participants (25 females) aged 18 to 56 wore a SenseDoc hip-worn accelerometer and had Samsung Galaxy S7 devices in three locations - pocket, right hand, and backpack. Each device measured acceleration in three axes. The SenseDoc was programmed to continuously measure at 100Hz, while the Samsung Galaxy S7 phones' data collection frequency was an average of 17Hz, varying between 10-30Hz. In this study, only the hand-held acceleration data was used.

The data collection protocol included these activities (**Table 1**): lying, sitting, self-paced walk, walking at 3 METs, running at 5 METs and running at 7 METs in sequential order for a 65-minute lab-based protocol with 40 minutes on the treadmill and 25 minutes of sedentary time. Participant energy expenditure was measured using the Oxycon Pro metabolic cart (Oxycon Pro, Jaeger, Hochberg, Germany). The Oyxcon Pro is a good and reliable unit for measuring energy expenditure (Ismail et al., 2019), and it was used to determine the Metabolic Equivalents (METs) at which each activity was performed. The metabolic cart was calibrated according to manufacturer specifications every morning of data collection. This study was approved by the Memorial University Interdisciplinary Committee on Ethics in Human Research (ICEHR 20180188-EX).

Table 1: Description of the 65-minute lab-based protocol.

Time (Minutes)	Activity	Detailed activity
0 - 5	Lying	Lying on a cot
5 - 10	Sitting	Sitting on a chair
10 - 20	Self-Pace Walk	Walking on the treadmill at a chosen self-paced speed
20 - 25	Lying	Lying on a cot
25 - 35	Walking at 3 METs	3 METs pace walking on the treadmill
35 - 40	Lying	Lying on a cot
40 - 50	Running 5 METs	5 METs running on the treadmill
50 - 55	Sitting	Sitting on a chair
55 - 65	Running 7 METs	7 METs running on the treadmill

Abbreviation: METs = Metabolic Equivalents

3.3 Data Analysis

The procedure for this study involved a three-step approach to data analysis and machine learning. These steps include data preprocessing (data cleaning, feature selection, feature extraction and scaling), model building and performance evaluation. The model-building step involved data splitting, hyperparameter tuning and model training by implementing the most used algorithms based on the systematic review by Narayanan et al., 2020. In this study, all analysis steps were done using R software version 4.0.4 (The R Foundation) and RStudio version 1.4.1106 (RStudio Inc).

3.3.1 Data Preprocessing

A. Data Cleaning:

In this step, the raw triaxial acceleration signals from the hand wear location were first transformed to generate the acceleration metrics - ActiGraph counts, ENMO and MIMS units over a 1-second epoch. ENMO was calculated by hand using **Equation 2,** as illustrated above. ActiGraph counts and MIMS units were computed using the "activityCounts" and "MIMSunit" packages, respectively. The latest versions of these packages were installed from CRAN (The Comprehensive R Archive Network), a centralized software repository for all R packages and documentation. Observations that contained missing values were replaced with values generated using the linear interpolation method. Linear interpolation is a method of fitting a curve using linear polynomials and creating new data points within the range of the original values for which the linear interpolation is done. The *na.approx()* function from the *zoo* package and the *mutate()* function from the *dplyr* package were used to impute these interpolated values. Data points that fall above and below 1.5 of the interquartile range in each accelerometer metric were omitted in the analysis. Duplicates in the timestamps were also dropped before the next step.

B. Feature Selection and Extraction:

In this step, variables that provided no helpful information for the physical activity classification task were dropped. These variables were participant id, wear location, and height and weight attribute removed after BMI was calculated to avoid multicollinearity in our model. A range of time and frequency domain features were extracted over a 1-second sliding window with no overlap. A 1-second window was chosen since it allows for many periodic movements to be captured for all activities. In total, 58 feature sets were extracted from the accelerometer data proposed by Liu et al., (2012). These feature sets included sum, mean, standard deviation, signal power, coefficient of variation, skewness, interquartile range, peak amplitude, and kurtosis of each axis of the three-axis accelerometer. In addition to these simple time-domain features, frequency domain characteristics such as dominant frequency, zero crossings, and cross-axis correlations were also extracted. Peak amplitude is the maximum value minus the minimum value of the signal at each window. The number of times the acceleration signal changed sign from positive to negative was represented by Zero crossings. A detailed description of extracted features from the raw acceleration signals over a 1-second window is provided in **Table 2**.

Table 2: Detailed description of featured extracted from raw acceleration signals.

Name of extracted features	Description
Signal Power	The sum of the absolute squares of signal at each window
Dominant Frequency	The signal frequency that carries the maximum energy among all frequencies at each window
Peak Amplitude	The maximum value minus the minimum value of signal at each window
Zero Crossings	The number of times the signal crosses its median at each window
Sum Log Energy	The sum of the natural logarithms of the power of signals at each window (sum(log(x^2 +1))
Coefficient of variation	The ratio of standard deviation to the mean of signals at each window
Peak Intensity	The number of signal peak appearances within a certain window
Vector Magnitude	The measure of the distance between one signal and another.
Entropy	The measure of randomness or uncertainty of the signal at each window
Descriptive features	Mean, Sum, Interquartile Range, Standard Deviation, Skewness, Kurtosis and Correlation

C. Feature Scaling:

This step converts categorical variables into factors and numeric binary (0 and 1) variables using the *step_dummy* recipe specification from the *Tidymodels* package. Since all the feature values vary in measurement scales, feature normalization is required so that each extracted feature contributes equally to the classification model. For this study, all numeric features were normalized to a mean of zero and standard deviation of one by using a specification of a recipe step in the *Tidymodels* package, *step_normalize*. Variables with significant correlations with other variables were removed using a correlation-based step on all the predictor features, *step_corr*. A correlation threshold of 0.7 was used, and 32 features (Appendix C) were fed into the selected models. The features chosen for each acceleration metric are provided in Appendix C.

3.3.2 Model Building

A. Model Selection:

In selecting the machine learning models for this study, three of the four most used models based on the systematic review examining machine learning and physical activity classification conducted by Narayanan et al., 2020, were implemented. Appendix A summarizes the papers and methods from Narayanan et al., 2020. According to the review and table, the four most used machine learning methods are Support Vector Machines (SVM), Random Forest (RF), K-Nearest Neighbour (KNN), and Artificial Neural Networks (ANN). These four models represent different mathematical and statistical approaches to physical activity classification. SVM uses a hyperplane approach based on linear algebra that classifies activity types on a multi-dimensional feature space (*Support-Vector Machine*, 2021). RF examines each feature and provides a decision about the cut-off for that feature in predicting different physical activity categories. Each feature is used in the analysis to create a tree of predictions for each feature. RF is a random decision tree and is a subclass of very commonly used ensemble-based methods (*Random Forest*, 2021). RF requires a user to input parameters, for example, the number of trees to create and the depth of trees in the feature space.

KNN is a method based on the probability that each point in a feature space is close to other related points (*K-Nearest Neighbors Algorithm*, 2021). An essential feature of KNN is that the user-defined definition of “close” can have significant consequences for the analysis. ANNs are a collection of connected nodes called artificial neurons (*Artificial Neural Network*, 2021). Each node can transmit a signal to other nodes. An artificial node receives a signal, then processes it, and can signal other nodes connected. The signal at a connection is a feature, and each node's output is a weight representing a nonlinear function for feature inputs. Nodes can be grouped into one or more hidden layers where weights are estimated. As with other machine learning methods, the user must specify the model's number of nodes and layers. **Figure 4** presents an example of each of the four machine learning models. SVM, RF, and KNN were selected for this thesis because these models have 2-3 hyperparameters that can be examined for model tuning. ANNs have many model hyperparameters, and with little to derive from the literature about where to start model tuning, it was decided to investigate only three of the four most common models.

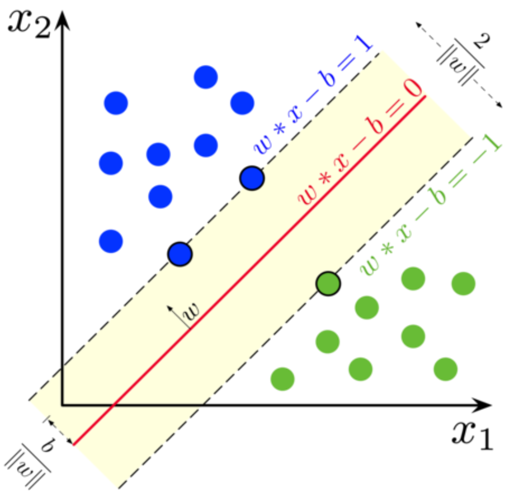
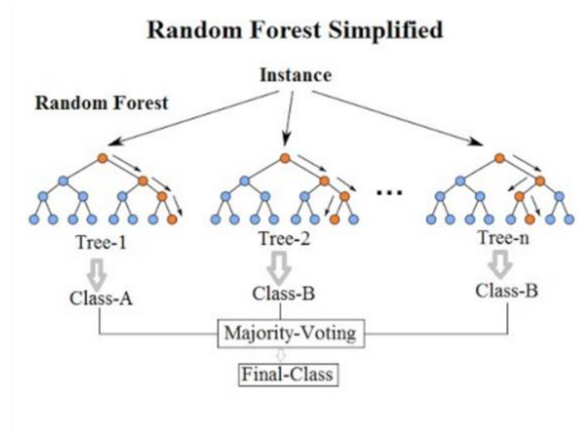
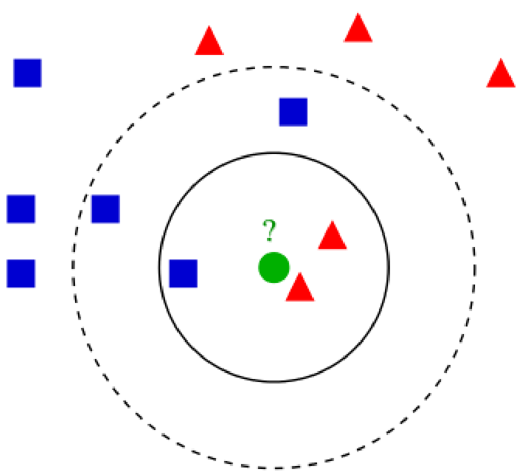
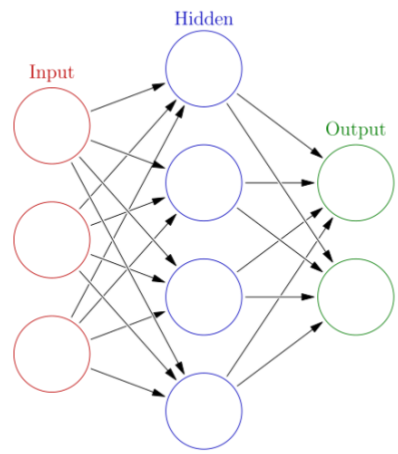
Support Vector Machines	Random Forest
 Larhmam, CC BY-SA 4.0, via Wikimedia Commons	 Venkata Jagannath, CC BY-SA 4.0, via Wikimedia Commons
K-Nearest Neighbour	**Artificial Neural Networks**
 Antti Ajanki AnAj, CC BY-SA 3.0, via Wikimedia Commons	 Glosser.ca, CC BY-SA 3.0, via Wikimedia Commons

Figure 4: A visual example of Support Vector Machines, Random Forest, K-Nearest Neighbour, and Artificial Neural Networks.

B. Data Splitting:

The whole dataset was first split into three using the acceleration metrics. Each subset was divided into training and testing sets comprising 75% and 25% of the observations. The validation set was a set of 10 validation folds obtained from the training sets using a k-fold cross-validation method, with the function *vfold_cv*. The training, validation and testing sets were activity-based; the split was based on a stratified sampling using the six activity classes.

C. Hyperparameter Tuning:

One of the objectives of this study is to determine the effect of model hyperparameters on the performances of each model. Different iterations of hyperparameters were tested for each ML algorithm and accelerometer metric to examine this impact. The hyperparameters (**Table 3**) for each of the three machine learning models were tuned on the cross-validation set to get the optimal settings for the models. This tuning was done using the Random Grid Search method (*tune_grid*) in Tidymodels to generate ten combinations of random hyperparameter settings. The tuning process presents a model for each proposed hyperparameter setting, evaluates the results on the cross-validation set and produces the settings that yield the best performance. The last step is to train a new model on the entire dataset (training and testing sets) under the best hyperparameter setting and evaluate the final model performance.

Table 3: Description of model hyperparameters tuned for RF, KNN and SVM.

ML Algorithm	**Hyperparameter**	**Description**
Random Forest	mtry	The number of predictors that will be randomly sampled at each split when creating tree models.
	min_n	The minimum number of data points in a node that is required for the node to be split further.
	trees	The number of trees contained in a random forest
K Nearest Neighbour	weight_func	The kernel function used to give a weight to the nearest k points
	neighbours	The number of neighbours used for the models

ML Algorithm	Hyperparameter	Description
Support Vector Machines	cost	A positive number for the cost of predicting a sample within or on the wrong side of the margin

D. Model Training and Testing:

Random Forest, Support Vector Machine and K-Nearest Neighbour algorithms were trained for each acceleration summary measure (ENMO, ActiGraph Counts and MIMS units). The number of subjectively selected features was varied to compare the differences between the acceleration metrics. After the preprocessing step, the models were first trained using all extracted features (Full feature set). They were additionally trained with two other subsets of selected feature subsets, one containing the acceleration metrics and the demographic characteristics only (Single feature set) and the other containing the raw acceleration signals and the acceleration metrics (Raw feature set). Eighteen models were trained and evaluated, plus 36 models were used for hyperparameter tuning and testing.

3.3.3 Model Performance Evaluation

After fitting the best-trained models on the testing sets, the performance was evaluated by computing the confusion matrix and the classification accuracy score. A confusion matrix is a summary of prediction results on any classification problem. It is a contingency table that gives insight into the errors made by the models during predictions on the testing data. Classification Accuracy is a metric that summarizes the performance of a classification model as the proportion of correct predictions among the total number of predictions made by the model (*Accuracy and Precision*, 2021). The higher the classification accuracy, the better the classification model.

CHAPTER 4: RESULTS

4.1 Participant Characteristics

Observations for forty-seven participants (25 females, mean age, 29.9 ± 9.1 years, BMI, 24.3 ± 3.4 kg/m^2) comprising hand accelerometer data from the original dataset were analyzed for this thesis. Data for three participants were excluded from the study because they were not correctly classified. **Table 4** shows the demographic characteristics of the final dataset that was analyzed.

Table 4: Demographic characteristics of the participants.

	Female (n = 25)	**Male (n = 22)**	**Total (n = 47)**
Age (years)	31.7 (8.1)	27.9 (9.6)	29.9 (9.1)
Height (cm)	162.2 (7.0)	178.6 (6.8)	169.7 (10.6)
Body mass (kg)	62.1 (10.2)	80.6 (12.7)	70.5 (14.5)
BMI (kg/m^2)	23.5 (3.0)	25.3 (3.7)	24.3 (3.4)

Values are mean (standard deviation). Abbreviation: BMI = Body Mass Index

4.2 Acceleration summary metrics

Descriptive summaries of Actigraph Counts, ENMO, and MIMS units by the physical activity type are provided in **Table 5** and **Figure 5**. The values for ENMO were 8.7 m/s^2 for lying, 8.6 m/s^2 for sitting, 9.3 m/s^2 for self-paced walking, 9.6 m/s^2, 11.2 m/s^2 and 12.6 m/s^2 for running 3, 5 and 7 METs, respectively. ActiGraph counts were 218 counts/minute for lying, 438cpm for sitting, 703cpm for self-paced walking, 921cpm, 1237cpm and 1314cpm for running at 3, 5, and 7 METs, respectively. The MIMS units were 4.8 units/minute for lying, 8.3upm for sitting, 24.1upm for self-paced walking, 30.6upm, 46.3upm, and 53.9upm for running at 3, 5 and 7 METs, respectively. Overall, the acceleration values of MIMS units and ActiGraph Counts produced a clear difference between each activity. For ActiGraph Counts and MIMS, there is a transparent gradient in values as activity intensities increase. Visual inspection suggests that there were only minimal differences between ENMO values for lying and sitting and between self-paced walking and running 3 METs. Compared to ActiGraph Counts and MIMS units, which exhibited substantial variation, ENMO values produced for each activity type had lower variability.

Table 5: Summary of acceleration metrics of the participants.

	ActiGraph (counts/min)	**ENMO (gravity units)**	**MIMS (units/min)**
Lying	218.0 (413.4)	8.7 (0.9)	4.8 (12.7)
Sitting	437.7 (595.7)	8.6 (1.3)	8.3 (15.4)
Self-pace walk	702.7 (433.6)	9.3 (1.3)	24.1 (17.2)
Running 3 METs	920.8 (431.9)	9.6 (1.4)	30.6 (19.2)
Running 5 METs	1236.5 (536.0)	11.2 (2.0)	46.3 (26.9)
Running 7 METs	1313.9 (742.1)	12.6 (4.3)	53.9 (33.9)

Values are mean (standard deviation). Abbreviation: METs = Metabolic Equivalents

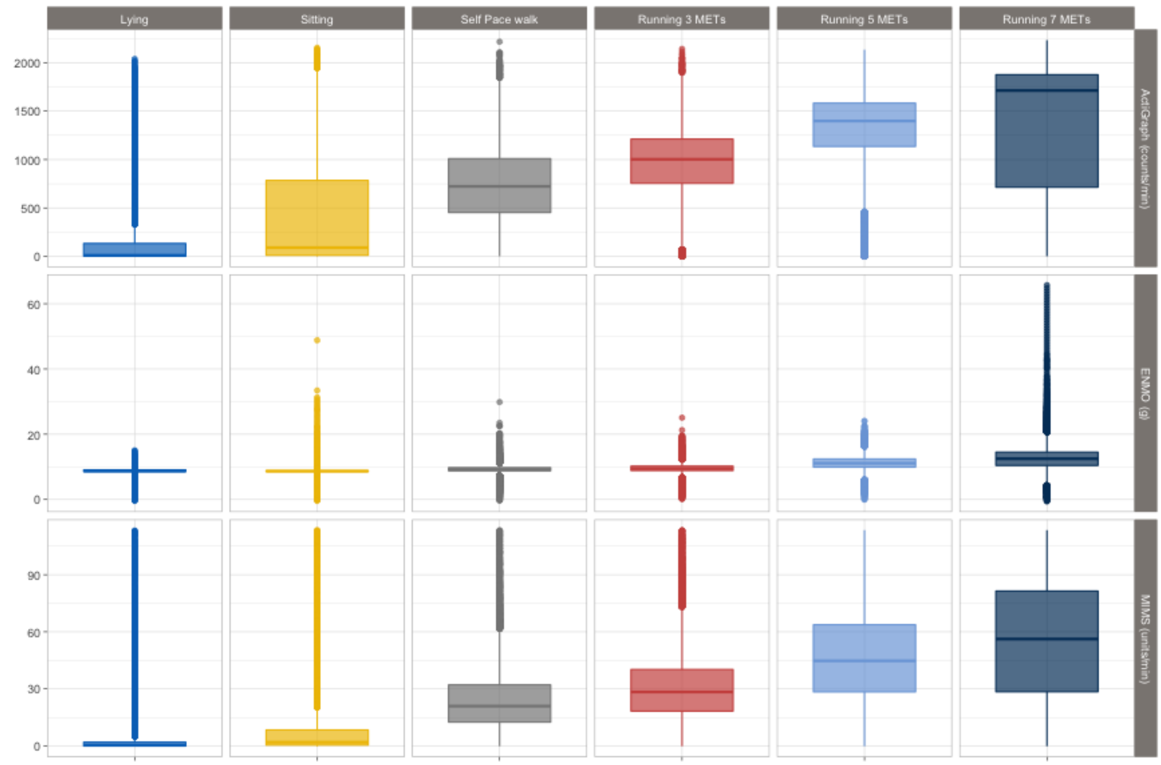


Figure 5: Euclidean Norm Minus One (ENMO), ActiGraph Counts and Monitor Independent Movement Summary (MIMS) units calculated from raw acceleration data

4.3 Hyperparameter Testing Results

To examine the impact of model hyperparameters on classification accuracy, several different hyperparameter specifications were tested for each acceleration metric and ML algorithm. **Table 6** shows each model's lowest and highest classification accuracy for ten iterations of hyperparameter specifications. The models with only a single feature formed the worst, as expected. The results show a relatively high variation in classification accuracy by the ML model, acceleration metric and number of features, ranging from 40.8% to 88.4%. **Figure 6** shows the graphical representation of model accuracy for the RF and KNN models for the acceleration metrics based on different iterations of the model hyperparameters. For models with only one feature, hyperparameter testing had an effect of between 2 and 10 percentage point improvement in model performance. For Actigraph Counts, ENMO, and MIMS with Random Forest models, accuracy ranged from 42.3% to 44.3%, 47.7% to 58.3%, and 40.8% to 42.6% for the worst and best-fitting models, respectively. For Actigraph Counts, ENMO, and MIMS with KNN models, accuracy ranged from 45.9% to 53.2%, 46.1% to 55.8%, and 38.8% to 46.3% for the worst and best-fitting models, respectively.

For the models with all features, Actigraph Counts, ENMO, and MIMS with Random Forest, accuracy ranged from 84.8% to 88.1%, 86.0% to 88.4%, and 86.9% to 88.3%, for the worst and best-fitting models respectively. For Actigraph Counts, ENMO, and MIMS with all features using KNN models, accuracy ranged from 82.1% to 83.3%, 81.6% to 83.3%, and 81.7% to 82.9%, for the worst and best-fitting models respectively. The hyperparameter test for RF models with all features showed that models could improve by between 1 and 4 percentage points depending on the accelerometer summary metric. For KNN models, hyperparameter testing had a negligible impact, with a 1 to 2 percentage point increase with hyperparameter optimization.

Figure 6 shows the graphical representation of the relationships between hyperparameters and model accuracy. Visual inspection of these figures shows that there does not appear to be a clear relationship between hyperparameters and model accuracy. We cannot assume that changing hyperparameter values will improve model accuracy.

Table 6: Hyperparameters and accuracy for best fitting Random Forest and K-Nearest Neighbour models for ActiGraph Counts, Euclidean Norm Minus One, and MIMS unit.

Single Features	**Model Name**	**Minimum Hyperparameter values**	**Accuracy**	**Optimal Hyperparameter Values**	**Accuracy**
ActiGraph Counts	RF	mtry: 3 trees: 78 min_n: 3	42.3%	mtry: 2 trees: 706 min_n: 36	44.3%
	KNN	neighbors: 2 weight_func: cos	45.9%	neighbors: 13 weight_func: epanechnikov	53.2%
Euclidean Norm Minus One	RF	mtry: 4 trees: 456 min_n: 3	47.7%	mtry: 2 trees: 1462 min_n: 24	58.3%
	KNN	neighbors: 2 weight_func: rank	46.1%	neighbors: 15 weight_func: biweight	55.8%
MIMS units	RF	mtry: 4 trees: 1807 min_n: 29	40.8%	mtry: 2 trees: 622 min_n: 39	42.6%
	KNN	neighbors: 2 weight_func: optimal	38.8%	neighbors: 13 weight_func: cos	46.3%
Raw Feature Sets					
ActiGraph Counts	RF	mtry: 1 trees: 703 min_n: 27	85.1%	mtry: 4 trees: 484 min_n: 8	88.4%
	KNN	neighbors: 2 weight_func: rectangular	88.4%	neighbors: 11 weight_func: triangular	89.3%
Euclidean Norm Minus One	RF	mtry: 1 trees: 840 min_n: 16	83.2%	mtry: 2 trees: 718 min_n: 12	87.0%
	KNN	neighbors: 2 weight_func: rank	86.4%	neighbors: 15 weight_func: biweight	87.5%

Single Features	**Model Name**	**Minimum Hyperparameter values**	**Accuracy**	**Optimal Hyperparameter Values**	**Accuracy**
MIMS units	RF	mtry: 6 trees: 251 min_n: 35	84.2%	mtry: 2 trees: 1416 min_n: 7	87.1%
	KNN	neighbors: 2 weight_func: optimal	85.6%	neighbors: 14 weight_func: triweight	86.9%
Full Feature Sets					
ActiGraph Counts	RF	mtry: 2 trees: 1353 min_n: 39	84.8%	mtry: 15 trees: 851 min_n: 12	88.1%
	KNN	neighbors: 2 weight_func: cos	82.1%	neighbors: 13 weight_func: epanechnikov	83.3%
	SVM**				
Euclidean Norm Minus One	RF	mtry: 4 trees: 2000 min_n: 39	86.0%	mtry: 11 trees: 484 min_n: 8	88.4%
	KNN	neighbors: 2 weight_func: rank	81.6%	neighbors: 15 weight_func: biweight	83.3%
	SVM	cost: 0.0179	60.2%	cost: 21.9	60.9%
MIMS units	RF	mtry: 17 trees: 20 min_n: 17	86.9%	mtry: 11 trees: 1881 min_n: 12	88.3%
	KNN	neighbors: 2 weight_func: optimal	81.7%	neighbors: 13 weight_func: cos	82.9%
	SVM	cost: 1.74e-3	61.6%	cost: 0.121	62.4%

RF = Random Forest. KNN = k nearest neighbors, SVM = Support Vector Machines. * All models include the variables of age, gender, and Body Mass Index. ** SVM model convergence failed.

Figure 6: Hyperparameter tuning for RF with raw and full feature sets

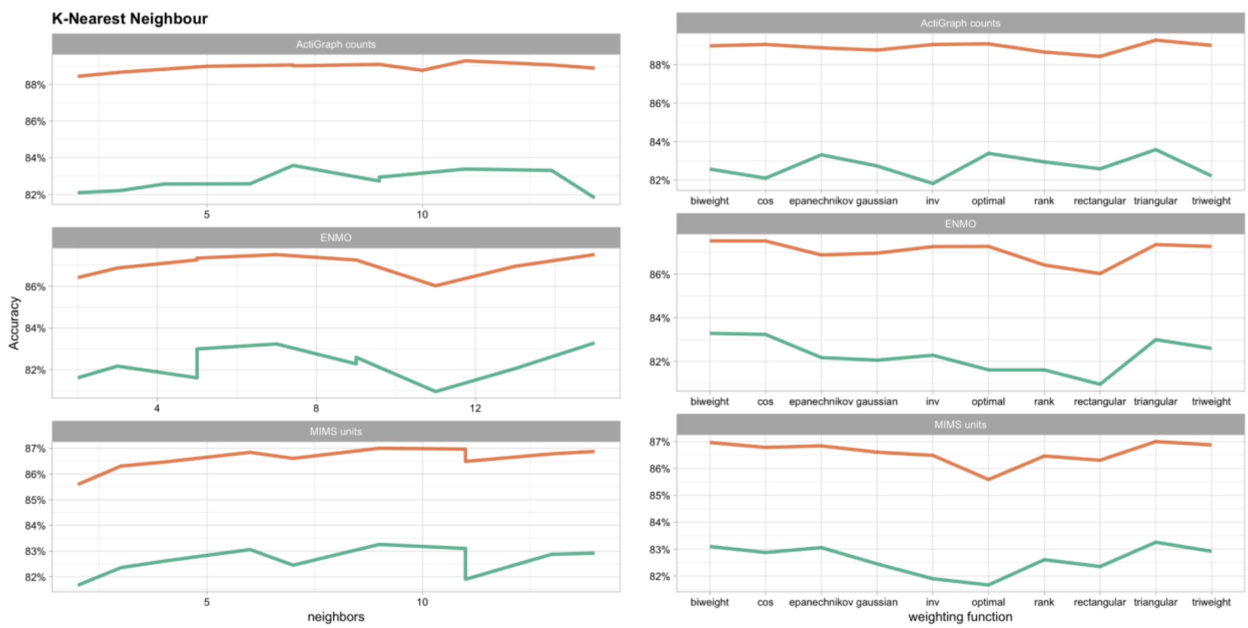


Figure 7: Hyperparameter tuning for KNN with raw and full features

4.4 Machine learning classification

Classification accuracy for the best fitting models for each model type, accelerometer metric and selected feature sets are presented in **Table 7**. Across the acceleration metric and model type, models fed with the acceleration metric, age, sex, and BMI (single feature sets) had lower classification accuracy of between 42.3% and 58.4%. MIMS units had the most insufficient classification accuracy, less than 50%, compared to ENMO and ActiGraph Counts with single feature sets. The models with full feature set obtained higher classification accuracy with RF (88.5% - 88.9%) than KNN (83.4% - 85.4%). Of the three selected feature sets, the raw feature sets provided better performance with ActiGraph Counts, 88.6% with RF and 89.1% with KNN. ENMO performed better with RF using the full feature sets (88.9%) and KNN using the raw feature sets (89.1%). MIMS units performed well with raw and full feature sets, with accuracy scores between 85.1% and 88.6%. Regardless of the accelerometer summary metric, the Random Forest models with all features have very similar accuracy, 88.9% with ENMO, 88.5% with Actigraph Counts, and 88.6% with MIMS units.

Table 7: Classification accuracy of models trained with feature subsets for each acceleration metric.

	Acceleration Metrics	Single features	Raw features	Full features
RF	**ENMO**	58.4%	87.2%	88.9%
	ActiGraph counts	44.3%	88.6%	88.5%
	MIMS units	42.3%	87.5%	88.6%
KNN	**ENMO**	55.9%	87.9%	85.4%
	ActiGraph counts	53.0%	89.1%	83.4%
	MIMS units	47.2%	87.3%	85.1%

Note. Single features include only the acceleration metrics. Raw features include the acceleration metrics and the raw X, Y, and Z acceleration. Full features include the acceleration metric and 58 features.

Figure 8 shows the confusion matrices for each model. Visual inspection of the confusion matrices suggests that the biggest classification challenge for the models was differenciating between self-paced walking and running/walking at 3 METs. This classification challenge is because, for the self-paced walk, participants chose a pace that was, on average, 2.7 METs; as a result, the self-paced walk and 3 METs walks were very similar in intensity. There also appears to be difficulty distinguishing between lying and sitting, which is expected. The confusion matrices also show that the raw feature sets in RF models (Figure 8) provided a better recognition of lying, sitting, self-paced walking and running at 3 METs, while the full feature sets only had better recognition with running at 5 and 7 METs. In contrast, the confusion matrices (Figure 9) of KNN models had consistently better recognition with the raw feature sets.

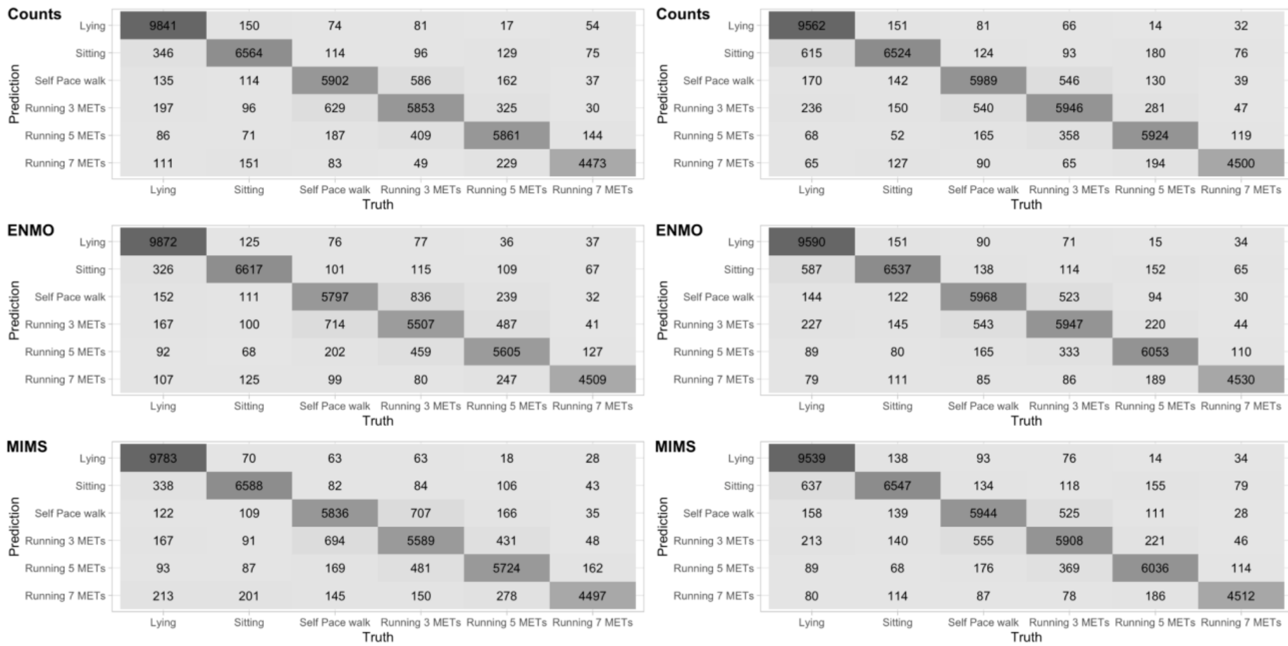


Figure 8: Confusion Matrices for RF models with raw (LEFT) and full (RIGHT) features

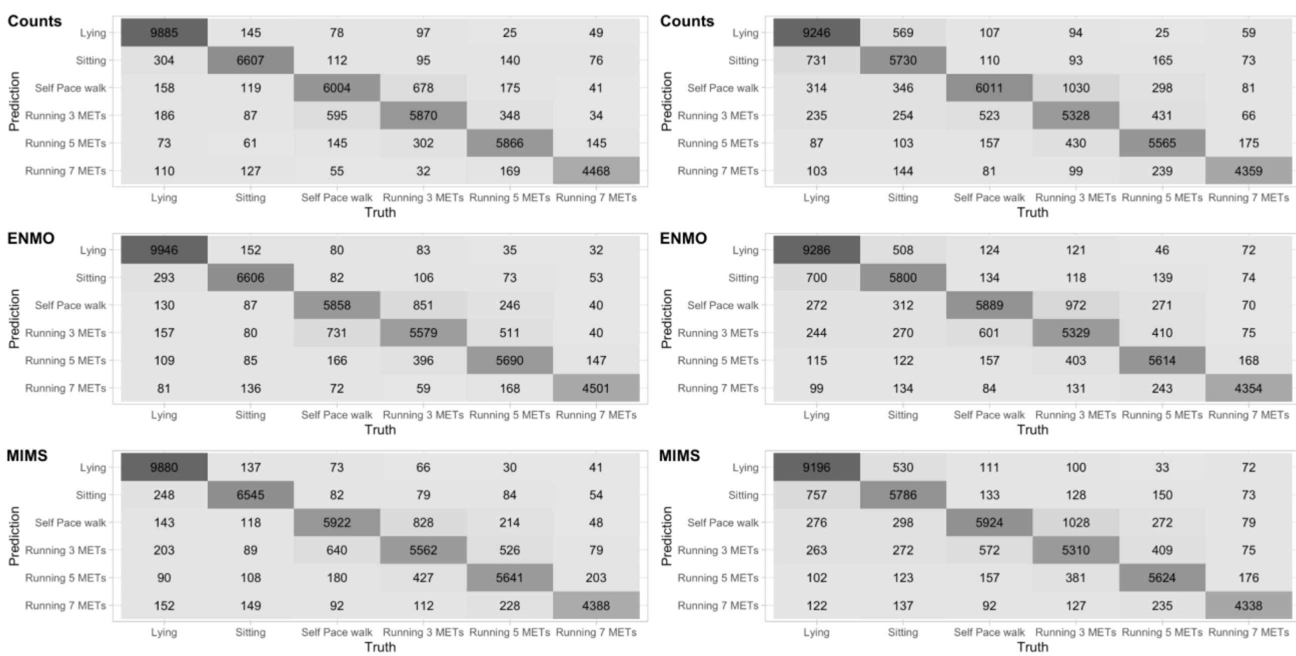


Figure 9: Confusion Matrices for KNN models with raw (LEFT) and full (RIGHT) features

CHAPTER 5: DISCUSSION

After training, this study examined the classification accuracy achieved by RF and KNN with three acceleration summary metrics and three sets of subjectively selected features. The influence of tuning the models' hyperparameters on classification accuracy was also investigated. Across the three feature sets, RF provided consistently higher classification accuracy with both the raw feature and the complete feature sets. KNN had a higher precision with the raw feature set. Of the three acceleration summary metrics examined, ActiGraph counts provided marginally better performance with the raw feature sets, and ENMO performed better with the complete feature sets. Hyperparameter tuning made a difference in accuracy by an average of 4% for RF models and 6% for KNN models.

The primary finding obtained from this study showed that RF models had the highest accuracy scores compared to the performance of KNN models in all three feature sets. These findings are consistent with previous studies showing that RF performed consistently better than many other supervised machine learning methods in classifying physical activity types and intensities. A systematic review by Narayanan et al., (2020) on physical activity classification using machine learning reported that almost half of the studies reviewed achieved the highest accuracy with RF (~85%). Similarly, (Chowdhury et al., 2017) compared the physical activity classification accuracy between different machine learning algorithms, including RF and KNN, using three independent datasets from wrist-worn accelerometers. Their findings showed that the RF classifier provided consistently high classification performance, with F1 scores ranging from 79.6% to 85% across three data sets.

The second finding in this study was that models fed with the raw feature set, i.e., the acceleration metric, the raw signals, and demographics, performed consistently better in the KNN models across all acceleration metrics. In contrast, the full features set performed better in the RF models. These findings are consistent with a study by (Mesanza et al., 2020), which compared the effect of the number of features fed into RF and KNN models on classification accuracy. The results reported from Mesanza's study indicated that a small set of features could be used to design the ML-based PA classifier, as the effect of increasing the number of features was insignificant in the total success rate of the classifier. Their study extracted 176 features but found that classification accuracy of over 90% was achieved using only the best nine features for all the models.

Interestingly, our models show that model performance was very similar between Actigraph counts, ENMO, and MIMS units. The study published on the development of the MIMS unit (John et al., 2019) showed that Actigraph counts tended to be very similar for lying, sitting, and standing. In contrast, ENMO and MIMS values showed variability between these movement types. The original study did not apply machine learning models to the 17 activity types and intensities examined. Differences in results could also be due to the kind of device. This study used a smartphone, while the studies included in John et al., used an Actigraph GTX9 worn on the hip and wrist for data collection. It is possible that holding the phone in your hand could result in more movement than a wrist-worn device. It is also interesting that ENMO outperformed both Actigraph counts and MIMS units in the single feature models. A pre-print using data from a sample of 655 adults wearing a wrist-worn Actigraph GTX9 device showed that the correlation between Actigraph counts and MIMS was 0.99 and between Actigraph counts and ENMO was 0.89, accounting for age, BMI, and gender (Karas et al., 2022). The result of my research and the pre-print suggest that despite some differences, Actigraph counts, ENMO, and MIMS may not perform differently when machine learning models with many features are used. However, the results clearly show that hyperparameter tuning has an essential impact on model performance.

The models trained in this study utilized the best hyperparameter settings for each model to achieve the best classification results. Few studies use machine learning for PA intensity or type prediction that reports or tests model hyperparameters. My results show that hyperparameter tuning can have up to a 5-percentage point increase in model performance, which is a significant performance gain for an ML model. Previous studies have examined the performance of RF and KNN with different acceleration metrics using the default hyperparameter settings of each model. Khataeipour et al., (2022) investigated the performance of RF, with the default hyperparameter settings, on physical activity classification using ActiGraph counts from right-hand accelerometer data as the summary metric. Their findings showed that the performance of RF was 40.3% with single features and 48.5% with full features. These results indicate that accuracy scores did not perform as well as a similar set of models with hyperparameter tuning.

Similarly, a study by Zhao et al., (2013) examined the performance of a model on the classification of six physical activity types using the best hyperparameter values after 44 iterations of hyperparameter settings. Their tuned models achieved more than 90% overall classification accuracy across four activity types. This showed that their model performance with 44 iterations was relatively better than this present study's because we achieved an accuracy score of ~80% by tuning ten iterations of hyperparameter settings for each model.

This current research has several contributions. It is the first study to compare the accuracy of activity classification models, RF and KNN using ENMO, ActiGraph Counts, and MIMS acceleration summary metrics. ENMO provided the best accuracy with all three models; MIMS units performed relatively well. In addition, the study demonstrated that proper feature selection is necessary when developing PA classification models. Adding more variables does not guarantee that the models will become more accurate.

Furthermore, this is the first study to publish its evaluated models' hyperparameters; the results indicate that proper tuning of model hyperparameters greatly influences ML-based PA classification. It should be noted that although the models examined in this study were RF and KNN, Support Vector Machines (SVM) and Artificial Neural Networks (ANN) have been published to be one of the most implemented ML models in PA classification research. ANN is known to have six primary hyperparameters available for tuning: hidden units, penalty, dropout, activation, epochs, and seeds. To get the best SVM model performance, two hyperparameters must be tuned: cost and kernel function. This study examined some SVM models; however, there were model convergence challenges. The models trained with ActiGraph Counts failed to compute results. The hyperparameter tuning process was also computationally costly. It took about a week to get the tuning results for each acceleration metric. Due to this, future research will concentrate on analyzing the performance of SVM and ANN on PA classification with ENMO, ActiGraph Counts, and MIMS, as well as the impact of tuning the hyperparameters as opposed to utilizing the default hyperparameter settings.

5.2 Limitations

There are a few limitations identified in this study. First, the study participants were predominantly younger and middle-aged adults; specifically, no children or adults above 60 years were included. Second, the classification models were computed using accelerometry data from laboratory-based settings. Participants performed a limited number of activity types that do not reflect how they are carried out in real-life situations. When models are trained on free-living accelerometer data, participants might engage in entirely different activities or the same types of activity in completely different ways. ML models are known not to perform well when comparing models derived from lab-based versus free-living study designs (Kerr et al., 2016). Third, the classification models were trained on accelerometer data from one type of smartphone held in the right hand. The study did not investigate the performance of the acceleration metrics based on other accelerometer devices, smartphone types or wear locations, i.e., the results might not translate to accelerometer data retrieved from smartphones in the left hand, pocket, or backpack. Fourth, each classification model obtained the hyperparameter tuning results using ten iterations because the tuning process can be computationally intense. It can take days to tune the hyperparameters of any model on a regular computer. The performance of the classification models could be much better if more iterations of hyperparameter specifications were implemented. Finally, the findings from this study are conditional on the chosen acceleration metrics and machine learning models which were selected because there are either limited studies on them (e.g., MIMS units) or they are commonly used, prospective studies using other types of ML models or acceleration summary metrics should use caution when using the hyperparameter settings from this study.

5.3 Conclusion

The comparison of ENMO, Actigraph Counts and MIMS units in this study has shown their performance with RF and KNN to enhance comparison for present and future studies. The effectiveness of tuning model hyperparameters and the importance of feature selection in PA classification using ML methods have been emphasized in this study. Prospective researchers are encouraged to implement and report on hyperparameter tuning. For this study, we used the "Tidymodels" package in R and feature importance with the "vip" package when training classification models from other smartphone devices, accelerometer sensors and placements.

REFERENCES

Accuracy and precision. (2021, November 9). In Wikipedia. https://en.wikipedia.org/wiki/Accuracy_and_precision

Ainsworth, B. E., Haskell, W. L., Herrmann, S. D., Meckes, N., Bassett, D. R., Tudor-Locke, C., Greer, J. L., Vezina, J., Whitt-Glover, M. C., & Leon, A. S. (2011). 2011 Compendium of Physical Activities: a second update of codes and MET values. *Medicine and Science in Sports and Exercise*, *43*(8), 1575–1581. https://doi.org/10.1249/MSS.0B013E31821ECE12

Andreu-Perez, J., Garcia-Gancedo, L., & McKinnell, J. (2017). Developing fine-grained actigraphies for rheumatoid arthritis patients from a single accelerometer using machine learning. *Sensors*, *17*(9), 2113. https://doi.org/10.3390/s17092113

Arif, M., & Kattan, A. (2015). Physical activities monitoring using wearable acceleration sensors attached to the body. *PLoS One*, *10*(7), e0130851. https://doi.org/10.1371/journal.pone.0130851

Armstrong, T., & Bull, F. (2006). Development of the World Health Organization Global Physical Activity Questionnaire (GPAQ). *Journal of Public Health 2006 14:2*, *14*(2), 66–70. https://doi.org/10.1007/S10389-006-0024-X

Artificial neural network. (2021, November 11). In Wikipedia. https://en.wikipedia.org/wiki/Artificial_neural_network

Bai, J., Di, C., Xiao, L., Evenson, K. R., LaCroix, A. Z., Crainiceanu, C. M., & Buchner, D. M. (2016). An Activity Index for Raw Accelerometry Data and Its Comparison with Other Activity Metrics. *PLoS ONE*, *11*(8). https://doi.org/10.1371/JOURNAL.PONE.0160644

Bakrania, K., Yates, T., Rowlands, A. V., Esliger, D. W., Bunnewell, S., Sanders, J., Davies, M., Khunti, K., & Edwardson, C. L. (2016). Intensity thresholds on raw acceleration data: Euclidean norm minus one (ENMO) and mean amplitude deviation (MAD) approaches. *PLoS ONE*, *11*(10). https://doi.org/10.1371/JOURNAL.PONE.0164045

Bastian, T., Maire, A., & Dugas, J. (2015). Automatic identification of physical activity types and sedentary behaviors from triaxial accelerometer: laboratory-based calibrations are not enough. *J Appl Physiol*, *118*(6), 716–722. https://doi.org/10.1152/japplphysiol.01189.2013

Bianchim, M. S., McNarry, M. A., Larun, L., Mackintosh, K. A., group, on behalf of A. S., Technology, A. S. S., & Centre, M. R. (2019). Calibration and validation of accelerometry to measure physical activity in adult clinical groups: A systematic review. *Preventive Medicine Reports*, *16*, 101001. https://doi.org/10.1016/J.PMEDR.2019.101001

Billiet, L., Swinnen, T. W., Westhovens, R., Vlam, K. D., & Huffel, S. van. (2016). Accelerometry-based activity recognition and assessment in rheumatic and musculoskeletal diseases. *Sensors*, *16*(12), 2151. https://doi.org/10.3390/s16122151

Bonomi, A. G., Goris, A. H., Yin, B., & Westerterp, K. R. (n.d.). Detection of type, duration, and intensity of physical activity using an accelerometer. *Med Sci Sports Exerc*, *41*(9), 1770–

1777. https://doi.org/10.1249/MSS.0b013e3181a24536

Brond, J. C., & Arvidsson, D. (2016). Sampling frequency affects the processing of Actigraph raw acceleration data to activity counts. *Journal of Applied Physiology*, *120*(3), 362–369. https://doi.org/10.1152/JAPPLPHYSIOL.00628.2015

Brondeel, R., Kestens, Y., Anaraki, J. R., Stanley, K., Thierry, B., & Fuller, D. (2021). Converting Raw Accelerometer Data to Activity Counts Using Open-Source Code: Implementing a MATLAB Code in Python and R, and Comparing the Results to ActiLife. *Journal for the Measurement of Physical Behaviour*, *4*(3), 205–211. https://doi.org/10.1123/JMPB.2019-0063

Bull, F. C., Maslin, T. S., & Armstrong, T. (2009). Global Physical Activity Questionnaire (GPAQ): Nine Country Reliability and Validity Study. *Journal of Physical Activity and Health*, *6*(6), 790–804. https://doi.org/10.1123/JPAH.6.6.790

Choi, L., Liu, Z., Matthews, C. E., & Buchowski, M. S. (2011). Validation of accelerometer wear and nonwear time classification algorithm. *Medicine and Science in Sports and Exercise*, *43*(2), 357–364. https://doi.org/10.1249/MSS.0B013E3181ED61A3

Chowdhury, A. K., Tjondronegoro, D., Chandran, V., & Trost, S. G. (2017). Ensemble methods for classification of physical activities from wrist accelerometry. *Med Sci Sports Exerc*, *49*(9), 1965–1973. https://doi.org/10.1249/MSS.0000000000001291

Chowdhury, A. K., Tjondronegoro, D., Chandran, V., & Trost, S. G. (2018). Physical activity recognition using posterior-adapted class-based fusion of multiaccelerometer data. *IEEE J Biomed Health Inform*, *22*(3), 678–685. https://doi.org/10.1109/JBHI.2017.2705036

Cleland, I., Kikhia, B., & Nugent, C. (2013). Optimal placement of accelerometers for the detection of everyday activities. *Sensors*, *13*(7), 9183–9200. https://doi.org/10.3390/s130709183

Clevenger, K. A., Pfeiffer, K. A., & Montoye, A. H. K. (2020). Cross-generational comparability of hip- and wrist-worn ActiGraph GT3X+, wGT3X-BT, and GT9X accelerometers during free-living in adults. *Journal of Sports Sciences*, *38*(24), 2794–2802. https://doi.org/10.1080/02640414.2020.1801320

Colley, R. C., Butler, G., Garriguet, D., Prince, S. A., & Roberts, K. C. (2018). Comparison of self-reported and accelerometer-measured physical activity in Canadian adults. *Health Reports*, *29*(12), 3–15.

Colley, R. C., Garriguet, D., Janssen, I., Craig, C. L., Clarke, J., & Tremblay, M. S. (2011). Physical activity of canadian adults: Accelerometer results from the 2007 to 2009 canadian health measures survey. *Health Reports*, *22*(1).

Corder, K., Ekelund, U., Steele, R. M., Wareham, N. J., & Brage, S. (2008). Assessment of physical activity in youth. *Journal of Applied Physiology (Bethesda, Md. : 1985)*, *105*(3), 977–987. https://doi.org/10.1152/JAPPLPHYSIOL.00094.2008

Craig, C. L., Marshall, A. L., Sjöström, M., Bauman, A. E., Booth, M. L., Ainsworth, B. E., Pratt, M., Ekelund, U., Yngve, A., Sallis, J. F., & Oja, P. (2003). International physical activity questionnaire: 12-country reliability and validity. *Medicine and Science in Sports and Exercise*, *35*(8), 1381–1395. https://doi.org/10.1249/01.MSS.0000078924.61453.FB

Dalton, A., & Ólaighin, G. (2013). Comparing supervised learning techniques on the task of physical activity recognition. *IEEE J Biomed Health Inform*, *17*(1), 46–52. https://doi.org/10.1109/TITB.2012.2223823

Dyrstad, S. M., Hansen, B. H., Holme, I. M., & Anderssen, S. A. (2014). Comparison of self-reported versus accelerometer-measured physical activity. *Medicine and Science in Sports and Exercise*, *46*(1), 99–106. https://doi.org/10.1249/MSS.0B013E3182A0595F

Ekelund, U., Steene-Johannessen, J., Brown, W. J., Fagerland, M. W., Owen, N., Powell, K. E., Bauman, A., Lee, I. M., Ding, D., Heath, G., Hallal, P. C., Kohl, H. W., Pratt, M., Reis, R., Sallis, J., Aadahl, M., Blot, W. J., Chey, T., Deka, A., … Yi-Park, S. (2016). Does physical activity attenuate, or even eliminate, the detrimental association of sitting time with mortality? A harmonised meta-analysis of data from more than 1 million men and women. *Lancet (London, England)*, *388*(10051), 1302–1310. https://doi.org/10.1016/S0140-6736(16)30370-1

Ekelund, U., Tarp, J., Steene-Johannessen, J., Hansen, B. H., Jefferis, B., Fagerland, M. W., Whincup, P., Diaz, K. M., Hooker, S. P., Chernofsky, A., Larson, M. G., Spartano, N., Vasan, R. S., Dohrn, I. M., Hagströmer, M., Edwardson, C., Yates, T., Shiroma, E., Anderssen, S. A., & Lee, I. M. (2019). Dose-response associations between accelerometry measured physical activity and sedentary time and all cause mortality: systematic review and harmonised meta-analysis. *BMJ (Clinical Research Ed.)*, *366*. https://doi.org/10.1136/BMJ.L4570

Ellis, K., Kerr, J., Godbole, S., Staudenmayer, J., & Lanckriet, G. (2016). Hip and wrist accelerometer algorithms for free-living behavior classification. *Med Sci Sports Exerc*, *48*(5), 933–940. https://doi.org/10.1249/MSS.0000000000000840

Fida, B., Bernabucci, I., Bibbo, D., Conforto, S., & Schmid, M. (2015). Varying behavior of different window sizes on the classification of static and dynamic physical activities from a single accelerometer. *Med Eng Phys*, *37*(7), 705–711. https://doi.org/10.1016/j.medengphy.2015.04.005

Freedson, P., Bowles, H. R., Troiano, R., & Haskell, W. (2012). Assessment of physical activity using wearable monitors: Recommendations for monitor calibration and use in the field. *Medicine and Science in Sports and Exercise*, *44*(SUPPL. 1). https://doi.org/10.1249/MSS.0b013e3182399b7e

Freedson, P. S., Melanson, E., & Sirard, J. (1998). Calibration of the Computer Science and Applications, Inc. accelerometer. *Medicine and Science in Sports and Exercise*, *30*(5), 777–781. https://doi.org/10.1097/00005768-199805000-00021

Fuller, D., Anaraki, J. R., Simango, B., Rayner, M., Dorani, F., Bozorgi, A., Luan, H., & A

Basset, F. (2021). Predicting lying, sitting, walking and running using Apple Watch and Fitbit data. *BMJ Open Sport & Exercise Medicine*, *7*(1), e001004. https://doi.org/10.1136/BMJSEM-2020-001004

Fullerton, E., Heller, B., & Munoz-Organero, M. (2017). Recognizing human activity in free-living using multiple body-worn accelerometers. *IEEE Sens J*, *17*(16), 5290–5297. https://doi.org/10.1109/JSEN.2017.2722105

Gao, L., Bourke, A. K., & Nelson, J. (2014). Evaluation of accelerometer based multi-sensor versus single-sensor activity recognition systems. *Med Eng Phys*, *36*(6), 779–785. https://doi.org/10.1016/j.medengphy.2014.02.012

Garcia-Ceja, E., & Brena, R. F. (2018). An improved three-stage classifier for activity recognition. *Int J Pattern Recogn*, *32*(1), 1860003. https://doi.org/10.1142/S0218001418600030

Grydeland, M., Hansen, B. H., Ried-Larsen, M., Kolle, E., & Anderssen, S. A. (2014). Comparison of three generations of ActiGraph activity monitors under free-living conditions: do they provide comparable assessments of overall physical activity in 9-year old children? *BMC Sports Science, Medicine and Rehabilitation*, *6*(1), 26. https://doi.org/10.1186/2052-1847-6-26

Gupta, P., & Dallas, T. (2014). Feature selection and activity recognition system using a single triaxial accelerometer. *IEEE Trans Biomed Eng*, *61*(6), 1780–1786. https://doi.org/10.1109/TBME.2014.2307069

Guthold, R., Stevens, G. A., Riley, L. M., & Bull, F. C. (2018). Worldwide trends in insufficient physical activity from 2001 to 2016: a pooled analysis of 358 population-based surveys with 1·9 million participants. *The Lancet. Global Health*, *6*(10), e1077–e1086. https://doi.org/10.1016/S2214-109X(18)30357-7

Guthold, R., Stevens, G. A., Riley, L. M., & Bull, F. C. (2020). Global trends in insufficient physical activity among adolescents: a pooled analysis of 298 population-based surveys with 1·6 million participants. *The Lancet Child & Adolescent Health*, *4*(1), 23–35. https://doi.org/10.1016/S2352-4642(19)30323-2

Gyllensten, I. C., & Bonomi, A. G. (2011). Identifying types of physical activity with a single accelerometer: evaluating laboratory-trained algorithms in daily life. *IEEE Trans Biomed Eng*, *58*(9), 2656–2663. https://doi.org/10.1109/TBME.2011.2160723

Hagströmer, M., Troiano, R. P., Sjöström, M., & Berrigan, D. (2010). Levels and patterns of objectively assessed physical activity--a comparison between Sweden and the United States. *American Journal of Epidemiology*, *171*(10), 1055–1064. https://doi.org/10.1093/AJE/KWQ069

Health Organization World. (2010). *Global recommendations on physical activity for health.* Geneva: World Health Organization. https://www.who.int/publications/i/item/9789241599979

Hernando, C., Hernando, C., Collado, E. J., Panizo, N., Martinez-Navarro, I., & Hernando, B. (2018). Establishing cut-points for physical activity classification using triaxial accelerometer in middle-aged recreational marathoners. *PLoS ONE*, *13*(8). https://doi.org/10.1371/JOURNAL.PONE.0202815

Hu, F., Wang, L., Wang, S., Liu, X., & He, G. (2016). A human body posture recognition algorithm based on BP neural network for wireless body area networks. *China Commun*, *13*(8), 198–208. https://doi.org/10.1109/CC.2016.7563723

Ismail, M., Sana'a, A., Loucks-Atlinson, A., Atkinson, M., Kelly, L., Alkanani, T., Basset, F. A., Alsubheen salsubhe, A., Loucks-Atkinson aloucksa, A., Atkinson matkinso, M., Kelly lpkelly, L. P., Alkanani talkanan, T., Basset, F. A., & Author, C. (2019). *Multiple propane gas flow rates procedure to determine accuracy and linearity of indirect calorimetry systems : An experimental assessment of a method.* https://doi.org/10.7287/PEERJ.PREPRINTS.27550

Jalloul, N., Poree, F., Viardot, G., L'Hostis, P., & Carrault, G. (2018). Activity recognition using complex network analysis. *IEEE J Biomed Health Inform*, *22*(4), 989–1000. https://doi.org/10.1109/JBHI.2017.2762404

John, D., & Freedson, P. (2012). ActiGraph and Actical physical activity monitors: a peek under the hood. *Medicine and Science in Sports and Exercise*, *44*(1 Suppl 1). https://doi.org/10.1249/MSS.0B013E3182399F5E

John, D., Sasaki, J., Staudenmayer, J., Mavilia, M., & Freedson, P. S. (2013). Comparison of Raw Acceleration from the GENEA and ActiGraph™ GT3X+ Activity Monitors. *Sensors*, *13*(11), 14754–14763. https://doi.org/10.3390/s131114754

John, D., Tang, Q., Albinali, F., & Intille, S. (2019). An Open-Source Monitor-Independent Movement Summary for Accelerometer Data Processing. *Journal for the Measurement of Physical Behaviour*, *2*(4), 268–281. https://doi.org/10.1123/JMPB.2018-0068

k-nearest neighbors algorithm. (2021, October 24). In Wikipedia. https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

Karas, M., Muschelli, J., Leroux, A., Urbanek, J. K., Wanigatunga, A. A., Bai, J., Crainiceanu, C. M., & Schrack, J. A. (2022). Comparison of Accelerometry-based Measures of Physical Activity. *MedArxiv*, 2022.03.16.22272518. https://doi.org/10.1101/2022.03.16.22272518

Keating, X. D., Zhou, K., Liu, X., Hodges, M., Liu, J., Guan, J., Phelps, A., & Castro-Piñero, J. (2019). Reliability and Concurrent Validity of Global Physical Activity Questionnaire (GPAQ): A Systematic Review. *International Journal of Environmental Research and Public Health*, *16*(21). https://doi.org/10.3390/IJERPH16214128

Kelly, P., Fitzsimons, C., & Baker, G. (2016). Should we reframe how we think about physical activity and sedentary behaviour measurement? Validity and reliability reconsidered. *International Journal of Behavioral Nutrition and Physical Activity*, *13*(1). https://doi.org/10.1186/S12966-016-0351-4

Kerr, J., Carlson, J., Godbole, S., Cadmus-Bertram, L., Bellettiere, J., & Hartman, S. (2018). Improving hip-worn accelerometer estimates of sitting using machine learning methods. *Med Sci Sports Exerc*, *50*(7), 1518–1524. https://doi.org/10.1249/MSS.0000000000001578

Kerr, J., Patterson, R. E., & Ellis, K. (2016). Objective assessment of physical activity: classifiers for public health. *Med Sci Sports Exerc*, *48*(5), 951–957. https://doi.org/10.1249/MSS.0000000000000841

Khan, A. M., Lee, Y. K., Lee, S. Y., & Kim, T. S. (2010). A triaxial accelerometer-based physical-activity recognition via augmented-signal features and a hierarchical recognizer. *IEEE Trans Inf Technol Biomed*, *14*(5), 1166–1172. https://doi.org/10.1109/TITB.2010.2051955

Khataeipour, S. J., Anaraki, J. R., Bozorgi, A., Rayner, M., Basset, F. A., & Fuller, D. (2022). Predicting lying, sitting and walking at different intensities using smartphone accelerometers at three different wear locations: hands, pant pockets, backpack. *BMJ Open Sport & Exercise Medicine*, *8*(2), e001242. https://doi.org/10.1136/BMJSEM-2021-001242

Kim, Y., Beets, M. W., & Welk, G. J. (2012). Everything you wanted to know about selecting the "right" Actigraph accelerometer cut-points for youth, but…: A systematic review. *Journal of Science and Medicine in Sport*, *15*(4), 311–321. https://doi.org/10.1016/J.JSAMS.2011.12.001

Lavesson, N., & Davidsson, P. (2006). Quantifying the impact of learning algorithm parameter tuning. *Proceedings of the National Conference on Artificial Intelligence*, *1*(May 2014), 395–400.

Lee, I.-M., Shiroma, E. J., Lobelo, F., Puska, P., Blair, S. N., & Katzmarzyk, P. T. (2012). Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. *The Lancet*, *380*(9838), 219–229. https://doi.org/10.1016/S0140-6736(12)61031-9

Lee, I. M., & Shiroma, E. J. (2014). Using accelerometers to measure physical activity in large-scale epidemiological studies: issues and challenges. *British Journal of Sports Medicine*, *48*(3), 197–201. https://doi.org/10.1136/BJSPORTS-2013-093154

Lee, M. W., Khan, A. M., & Kim, T. S. (2011). A single tri-axial accelerometer-based real-time personal life log system capable of human activity recognition and exercise information generation. *Pers Ubiquit Comput*, *15*(8), 887–898. https://doi.org/10.1007/s00779-011-0403-3

Liu, S., Gao, R. X., & Freedson, P. S. (2012). Computational methods for estimating energy expenditure in human physical activities. *Medicine and Science in Sports and Exercise*, *44*(11), 2138–2146. https://doi.org/10.1249/MSS.0B013E31825E825A

Liu, S. H., & Chang, Y. J. (2009). Using accelerometers for physical actions recognition by a neural fuzzy network. *Telemed J E Health*, *15*(9), 867–876. https://doi.org/10.1089/tmj.2009.0032

Mannini, A., Intille, S. S., Rosenberger, M., Sabatini, A. M., & Haskell, W. (2013). Activity recognition using a single accelerometer placed at the wrist or ankle. *Med Sci Sports Exerc*, *45*(11), 2193–2203. https://doi.org/10.1249/MSS.0b013e31829736d6

Mannini, A., & Sabatini, A. M. (2010). Machine learning methods for classifying human physical activity from on-body accelerometers. *Sensors*, *10*(2), 1154–1175. https://doi.org/10.3390/s100201154

Marschollek, M. (2013). A Semi-Quantitative Method to Denote Generic Physical Activity Phenotypes from Long-Term Accelerometer Data - The ATLAS Index. *PLoS ONE*, *8*(5). https://doi.org/10.1371/journal.pone.0063522

Mesanza, A. B., Lucas, S., Zubizarreta, A., Cabanes, I., Portillo, E., & Rodriguez-Larrad, A. (2020). A Machine Learning Approach to Perform Physical Activity Classification Using a Sensorized Crutch Tip. *IEEE Access*, *8*, 210023–210034. https://doi.org/10.1109/ACCESS.2020.3039885

Migueles, J. H., Cadenas-Sanchez, C., Ekelund, U., Delisle Nyström, C., Mora-Gonzalez, J., Löf, M., Labayen, I., Ruiz, J. R., & Ortega, F. B. (2017). Accelerometer Data Collection and Processing Criteria to Assess Physical Activity and Other Outcomes: A Systematic Review and Practical Considerations. *Sports Medicine (Auckland, N.Z.)*, *47*(9), 1821–1845. https://doi.org/10.1007/S40279-017-0716-0

Migueles, J. H., Cadenas-Sanchez, C., Tudor-Locke, C., Löf, M., Esteban-Cornejo, I., Molina-Garcia, P., Mora-Gonzalez, J., Rodriguez-Ayllon, M., Garcia-Marmol, E., Ekelund, U., & Ortega, F. B. (2019). Comparability of published cut-points for the assessment of physical activity: Implications for data harmonization. *Scandinavian Journal of Medicine and Science in Sports*, *29*(4), 566–574. https://doi.org/10.1111/SMS.13356

Montoye, A., Conger, S. A., & Connolly, C. P. (2017). Validation of accelerometer-based energy expenditure prediction models in structured and simulated free-living settings. *Meas Phys Educ Exerc Sci*, *21*(4), 223–234. https://doi.org/10.1080/1091367X.2017.1337638

Montoye, A., Dong, B., Biswas, S., & Pfeiffer, K. A. (2014). Use of a wireless network of accelerometers for improved measurement of human energy expenditure. *Electronics*, *3*(2), 205–220. https://doi.org/10.3390/electronics3020205

Montoye, A., Dong, B., Biswas, S., & Pfeiffer, K. A. (2016). Validation of a wireless accelerometer network for energy expenditure measurement. *J Sports Sci*, *34*(21), 2130–2139. https://doi.org/10.1080/02640414.2016.1151924

Montoye, A., Mudd, L. M., Biswas, S., & Pfeiffer, K. A. (2015). Energy expenditure prediction using raw accelerometer data in simulated free living. *Med Sci Sports Exerc*, *47*(8), 1735–1746. https://doi.org/10.1249/MSS.0000000000000597

Montoye, A., Pivarnik, J. M., Mudd, L. M., Biswas, S., & Pfeiffer, K. A. (2016a). Comparison of activity type classification accuracy from accelerometers worn on the hip, wrists, and thigh in young, apparently healthy adults. *Meas Phys Educ Exerc Sci*, *20*(3), 173–183.

https://doi.org/10.1080/1091367X.2016.1192038

Montoye, A., Pivarnik, J. M., Mudd, L. M., Biswas, S., & Pfeiffer, K. A. (2016b). Validation and comparison of accelerometers worn on the hip, thigh, and wrists for measuring physical activity and sedentary behavior. *AIMS Public Health*, *3*(2), 298–312. https://doi.org/10.3934/publichealth.2016.2.298

Montoye, A., Pivarnik, J. M., Mudd, L. M., Biswas, S., & Pfeiffer, K. A. (2017). Evaluation of the activPAL accelerometer for physical activity and energy expenditure estimation in a semi-structured setting. *J Sci Med Sport*, *20*(11), 1003–1007. https://doi.org/10.1016/j.jsams.2017.04.011

Montoye, A., Westgate, B. S., Fonley, M. R., & Pfeiffer, K. A. (2018). Cross-validation and out-of-sample testing of physical activity intensity predictions with a wrist-worn accelerometer. *J Appl Physiol*, *124*(5), 1284–1293. https://doi.org/10.1152/japplphysiol.00760.2017

Muscillo, R., Schmid, M., Conforto, S., & D'Alessio, T. (2010). An adaptive Kalman-based Bayes estimation technique to classify locomotor activities in young and elderly adults through accelerometers. *Med Eng Phys*, *32*(8), 849–859. https://doi.org/10.1016/j.medengphy.2010.05.009

Narayanan, A., Desai, F., Stewart, T., Duncan, S., & MacKay, L. (2020). Application of raw accelerometer data and machine-learning techniques to characterize human movement behavior: A systematic scoping review. *Journal of Physical Activity and Health*, *17*(3), 360–383. https://doi.org/10.1123/jpah.2019-0088

Pavey, T. G., Gilson, N. D., Gomersall, S. R., Clark, B., & Trost, S. G. (2017). Field evaluation of a random forest activity classifier for wrist-worn accelerometer data. *Journal of Science and Medicine in Sport*, *20*(1), 75–80. https://doi.org/10.1016/J.JSAMS.2016.06.003

Preece, S. J., Goulermas, J. Y., Kenney, L. P. J., & Howard, D. (2009). A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data. *IEEE Trans Biomed Eng*, *56*(3), 871–879. https://doi.org/10.1109/TBME.2008.2006190

Probst, P., Boulesteix, A. L., & Bischl, B. (2018). Tunability: Importance of Hyperparameters of Machine Learning Algorithms. *Journal of Machine Learning Research*, *20*. https://doi.org/10.48550/arxiv.1802.09596

Random forest. (2021, November 2). In Wikipedia. https://en.wikipedia.org/wiki/Random_forest

Rosenberg, D., Godbole, S., & Ellis, K. (2017). Classifiers for accelerometer-measured behaviors in Older women. *Med Sci Sports Exerc*, *49*(3), 610–616. https://doi.org/10.1249/MSS.0000000000001121

Rothney, M. P., Neumann, M., Béziat, A., & Chen, K. Y. (2007). An artificial neural network model of energy expenditure using nonintegrated acceleration signals. *J Appl Physiol*, *103*(4), 1419–1427. https://doi.org/10.1152/japplphysiol.00429.2007

Sasaki, Jeffer E., John, D., & Freedson, P. S. (2011). Validation and comparison of ActiGraph activity monitors. *Journal of Science and Medicine in Sport*, *14*(5), 411–416. https://doi.org/10.1016/J.JSAMS.2011.04.003

Sasaki, Jeffer Eidi, Hickey, A. M., Staudenmayer, J. W., John, D., Kent, J. A., & Freedson, P. S. (2016). Performance of activity classification algorithms in free-living older adults. *Medicine and Science in Sports and Exercise*, *48*(5), 941–949. https://doi.org/10.1249/MSS.0000000000000844

Siscovick, D. S., Laporte, R. E., Newman, J., Health ; Iverson, D. C., & Fielding, J. E. (1985). Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research. *Public Health Reports*, *100*(2), 126. /pmc/articles/PMC1424733/?report=abstract

Staudenmayer, J., He, S., Hickey, A., Sasaki, J., & Freedson, P. (2015). Methods to estimate aspects of physical activity and sedentary behavior from high-frequency wrist accelerometer measurements. *J Appl Physiol*, *119*(4), 396–403. https://doi.org/10.1152/japplphysiol.00026.2015

Stewart, T., Narayanan, A., Hedayatrad, L., Neville, J., Mackay, L., & Duncan, S. (2018). A dual-accelerometer system for classifying physical activity in children and adults. *Med Sci Sports Exerc*, *50*(12), 2595–2602. https://doi.org/10.1249/MSS.0000000000001717

Strath, S. J., Kaminsky, L. A., Ainsworth, B. E., Ekelund, U., Freedson, P. S., Gary, R. A., Richardson, C. R., Smith, D. T., & Swartz, A. M. (2013). Guide to the assessment of physical activity: Clinical and research applications: A scientific statement from the American Heart association. *Circulation*, *128*(20), 2259–2279. https://doi.org/10.1161/01.cir.0000435708.67487.da

Support-vector machine. (2021, November 16). In Wikipedia. https://en.wikipedia.org/wiki/Support-vector_machine

Troiano, R. P., Berrigan, D., Dodd, K. W., Mâsse, L. C., Tilert, T., & Mcdowell, M. (2008). Physical activity in the United States measured by accelerometer. *Medicine and Science in Sports and Exercise*, *40*(1), 181–188. https://doi.org/10.1249/MSS.0B013E31815A51B3

Trost, S. G., Cliff, D. P., Ahmadi, M. N., Tuc, N. V, & Hagenbuchner, M. (2018). Sensor-enabled activity class recognition in preschoolers: hip versus wrist data. *Med Sci Sports Exerc*, *50*(3), 634–641. https://doi.org/10.1249/MSS.0000000000001460

Trost, S. G., Zheng, Y., & Wong, W.-K. (2014). Machine learning for activity recognition: Hip versus wrist data. *Physiol Meas*, *35*(11), 2183–2189. https://doi.org/10.1088/0967-3334/35/11/2183

Vähä-Ypyä, H., Vasankari, T., Husu, P., Suni, J., & Sievänen, H. (2015). A universal, accurate intensity-based classification of different physical activities using raw data of accelerometer. *Clinical Physiology and Functional Imaging*, *35*(1), 64–70. https://doi.org/10.1111/cpf.12127 van Hees, V. T., Gorzelniak, L., Dean León, E. C., Eder, M., Pias, M., Taherian, S., Ekelund, U., Renström, F., Franks, P. W., Horsch, A., & Brage, S. (2013). Separating Movement and Gravity Components in an Acceleration Signal and Implications for the Assessment of Human Daily Physical Activity. *PLoS ONE*, *8*(4). https://doi.org/10.1371/journal.pone.0061691

Wang, Z. L., Wu, D. H., Chen, J. M., Ghoneim, A., & Hossain, M. A. (2016). A triaxial accelerometer-based human activity recognition via EEMD-based features and game-theory-based feature selection. *IEEE Sens J*, *16*(9), 3198–3207. https://doi.org/10.1109/JSEN.2016.2519679

Wijndaele, K., Westgate, K., Stephens, S. K., Blair, S. N., Bull, F. C., Chastin, S. F. M., Dunstan, D. W., Ekelund, U., Esliger, D. W., Freedson, P. S., Granat, M. H., Matthews, C. E., Owen, N., Rowlands, A. V., Sherar, L. B., Tremblay, M. S., Troiano, R. P., Brage, Sø., & Healy, G. N. (2015). Utilization and Harmonization of Adult Accelerometry Data: Review and Expert Consensus. *Medicine and Science in Sports and Exercise*, *47*(10), 2129–2139. https://doi.org/10.1249/MSS.0000000000000661

Willetts, M., Hollowell, S., Aslett, L., Holmes, C., & Doherty, A. (2018). Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96, 220 UK Biobank participants. *Sci Rep*, *8*, 10. https://doi.org/10.1038/s41598-018-26174-1

Wong, M. Y., Day, N. E., Luan, J. A., Chan, K. P., & Wareham, N. J. (2003). The detection of gene-environment interaction for continuous traits: should we deal with measurement error by bigger studies or better measurement? *International Journal of Epidemiology*, *32*(1), 51–57. https://doi.org/10.1093/IJE/DYG002

World Health Organization. (2020, November 25). *Every move counts towards better health – says WHO*. [Press Release]. https://www.who.int/news/item/25-11-2020-every-move-counts-towards-better-health-says-who

Wu, Y., Chen, R., Wang, J., Sun, X., & She, M. F. H. (2012). Intelligent clothing for automated recognition of human physical activities in free-living environment. *J Text Inst*, *103*(8), 806–816. https://doi.org/10.1080/00405000.2011.611641

Wullems, J. A., Verschueren, S. M. P., Degens, H., Morse, C. I., & Onambélé, G. L. (2017). Performance of thigh-mounted triaxial accelerometer algorithms in objective quantification of sedentary behaviour and physical activity in older adults. *PLoS One, 12*(11), e0188215. https://doi.org/10.1371/journal.pone.0188215

Xiao, W., & Lu, Y. (2015). Daily human physical activity recognition based on kernel discriminant analysis and extreme learning machine. *Math Probl Eng*, *2015*, 790412. https://doi.org/10.1155/2015/790412

Yang, J. Y., Wang, J. S., & Chen, Y. P. (2008). Using acceleration measurements for activity recognition: an effective learning algorithm for constructing neural classifiers. *Pattern Recognit Lett*, *29*(16), 2213–2220. https://doi.org/10.1016/j.patrec.2008.08.002

Zhang, S., Murray, P., Zillmer, R., Eston, R. G., Catt, M., & Rowlands, A. V. (2012). Activity classification using the GENEA: optimum sampling frequency and number of axes. *Med Sci Sports Exerc*, *44*(11), 2228–2234. https://doi.org/10.1249/MSS.0b013e31825e19fd

Zhang, S., Rowlands, A. V, Murray, P., & Hurst, T. L. (2012). Physical activity classification using the GENEA wrist-worn accelerometer. *Med Sci Sports Exerc*, *44*(4), 742–748. https://doi.org/10.1249/MSS.0b013e31823bf95c

Zhao, W., Adolph, A. L., Puyau, M. R., Vohra, F. A., Butte, N. F., & Zakeri, I. F. (2013). Support vector machines classifiers of physical activities in preschoolers. *Physiological Reports*, *1*(1), 6. https://doi.org/10.1002/PHY2.6

APPENDICES

Appendix A: Ethical Approval

MEMORIAL UNIVERSITY

Interdisciplinary Committee on Ethics in Human Research (ICEHR)

St. John's, NL Canada A1C 5S7
Tel: 709 864-2561 icehr@mun.ca
www.mun.ca/research/ethics/humans/icehr

ICEHR Number:	**20222477-HK**
Approval Period:	February 10, 2022 – February 28, 2023
Funding Source:	CFI & CIHR [RGCS# 20160392 & 20181241]
Responsible Faculty:	Dr. Daniel Fuller School of Human Kinetics and Recreation
Title of Project:	*Comparing Accelerometer Processing Metrics for Physical Activity Classification Accuracy using Machine Learning Methods*

Title of Parent Project:	*Are Smartphones capable of providing valid and reliable objective physical activity data?–a cross-over validation study*
ICEHR Number:	*20180188-EX*

February 10, 2022

Ms. Sumayyah Musa
School of Human Kinetics and Recreation
Memorial University

Dear Ms. Musa:

Thank you for your submission to the Interdisciplinary Committee on Ethics in Human Research (ICEHR) seeking ethical clearance for the above-named research project. The Committee has reviewed the proposal and agrees that the project is consistent with the guidelines of the *Tri-Council Policy Statement on Ethical Conduct for Research Involving Humans* (TCPS2). *Full ethics clearance* is *granted to* February 28, 2023. ICEHR approval applies to the ethical acceptability of the research, as per Article 6.3 of the *TCPS2*. Researchers are responsible for adherence to any other relevant University policies and/or funded or non-funded agreements that may be associated with the project. If funding is obtained subsequent to ethics approval, you must submit a Funding and/or Partner Change Request to ICEHR so that this ethics clearance can be linked to your award.

The *TCPS2* **requires** that you **strictly adhere to the protocol and documents as last reviewed** by ICEHR. If you need to make additions and/or modifications, you must submit an Amendment Request with a description of these changes, for the Committee's review of potential ethical issues, before they may be implemented. Submit a Personnel Change Form to add or remove project team members and/or research staff. Also, to inform ICEHR of any unanticipated occurrences, an Adverse Event Report must be submitted with an indication of how the unexpected event may affect the continuation of the project.

The *TCPS2* **requires** that you submit an Annual Update to ICEHR before February 28, 2023. If you plan to continue the project, you need to request renewal of your ethics clearance and include a brief summary on the progress of your research. When the project no longer involves contact with human participants, is completed and/or terminated, you are **required** to provide an annual update with a brief final summary and your file will be closed. All post-approval ICEHR event forms noted above must be submitted by selecting the ***Applications: Post-Review*** link on your Researcher Portal homepage. We wish you success with your research.

Yours sincerely,

Kelly Blidook, Ph.D.
Chair, ICEHR

KB/bc

cc: Supervisor – Dr. Daniel Fuller, School of Human Kinetics and Recreation

Appendix B: Description of studies. Adapted from (Narayanan et al., 2020)

Reference	**Device name; Number of devices; Axis, sampling frequency; Placement position**	**Features generated from raw accelerometer data**	**Machine-learning classifier used; Validation method**
(Bonomi et al., n.d.)	Tracmor (Philips Research, Eindhoven, The Netherlands); 1; 3-Axis accelerometer, 20 Hz; Lower back	27 time and frequency domain features	Decision Tree; Leave-one-subject-out cross-validation
(Andreu-Perez et al., 2017)	Axivity AX3 (Axivity Ltd, York, United Kingdom); 1; 3-axis accelerometer, 100 Hz; Lower back	Several time and frequency domain features	DMF-DL, dichotomy mapped forest-metric Learning, Convolutional Deep Belief Networks, RF, SVM, and cHMM; Leave-one-subject-out cross-validation
(Arif & Kattan, 2015)	Colibri (Trivisio Prototyping, Trier, Germany) wireless inertial measurement units (IMUs); 3; 3-axis accelerometer, 100 Hz; Wrist, chest, and ankle	Several time domain features are segregated into 3 different feature sets respective to each sensor location	KNN classifier, rotation forest, and NN; Validation method: 70% training data set 30% test data set
(Bastian et al., 2015)	MotionLogs (Movea, Grenoble, France); 1; 3-axis accelerometer, 100 Hz; Hip	9 time and frequency domain features	Bayesian classifier; Leave-one-subject-out cross-validation
(Billiet et al., 2016)	SenseWear Pro 3 Armband (Bodymedia Inc, Pittsburgh, PA); 1; 2-axis accelerometer, 32 Hz; Biceps of the dominant arm	39 (18 signal patternbased features and 21 time domain features)	3-stage classifiers (stage 1—RF with rejection, stage 2—linear discriminant model, and stage 3—binary classifiers); Leave-one-subject-out cross-validation

(Chowdhury et al., 2017)	Data set 1: Colibri wireless IMU sensor; 1; 3-axis accelerometer, 100 Hz; Wrist Data set 2: Empatica E4 sensor (Empatica, Milano, MI); 1; 3-axis accelerometer, 32 Hz; Wrist	45 features extracted from both time domain and frequency domain	Boosted DT, bagging DT, RF, BDT, KNN, SVT, ANN, custom ensemble classifiers (weighted majority voting, NB combiner, and behaviour knowledge space) Leave-one-subject-out cross-validation
(Chowdhury et al., 2018)	Data set 1: Colibri wireless IMU sensor; 3; 3-axis accelerometer, 100 Hz; Dominant-side wrist, ankle, and chest Data set 2: Shimmer 2R (Realtime Technologies, Dublin, Ireland); 3; 3-axis accelerometer, 50 Hz; Chest, right wrist, and left ankle	45 features from both time domain and frequency domain	BDT, SVM, deep NN, RF, and AdaBoost Leave-one-subject-out cross-validation
(Cleland et al., 2013)	Shimmer 2R (Realtime Technologies, Dublin, Ireland); 6; 3-axis accelerometer, 51.2 Hz; Chest, wrist, lower back, hip, thigh, and foot	26 features from both time domain and frequency domain	DT (J48), NB, NN (multilayer perceptron), and SVM; 10-fold cross-validation
(Dalton & Ólaighin, 2013)	Witilt (version 2.5; SparkFun Electronics); 5; 3-axis accelerometer, 135 Hz; Just below the suprasternal notch, left side of the chest over the lower ribs, directly above the right hip, wrist of the dominant hand, and ankle of the dominant leg	160 features from both time domain and frequency domain	C4.5 Graft, NB, BayesNET, IBI, IBK, KStar, JRip, SVM, Multi Perception, AdaBoost, Ada-BoostM1, Bagging, MultiBoost, Vote; Leave-one-subject-out cross-validation
(Ellis et al., 2016)	ActiGraph (Pensacola, FL); 2; 3-axis accelerometer, 30 Hz; Right hip and nondominant wrist	40 total features from both time domain and frequency domain	RF coupled with HMM; Leave-one-subject-out cross-validation

(Fida et al., 2015)	Internally developed Inertial Measurement Unit; 1; 3-axis accelerometer, 100 Hz; Waist	22 features from time domain	DT classifier, NB classifier, KNN classifier, SVM, and NNs; Leave-one-subject-out cross-validation, 70% training data set 30% test data set
(Fullerton et al., 2017)	RunScribe™ inertial sensors (Scribe Labs, CA) 9; 3-axis accelerometer, 10 Hz; Left and right lateral ankle, left and right hip, left and right wrist, left and right upper arm, and spine (T10)	Several time and frequency domain features	Complex DT, SVM, fine KNN classifier, and ensemble-bagged trees Validation method: 80% of the data was used for training and 20% was tested
(Gao et al., 2014)	Shimmer™ (wireless sensor platform); 4; 3-axis accelerometer, 200 Hz; Chest, left under arm, waist, and thigh	Several time and frequency domain, and heuristic features	DT classifier, NB classifier, KNN classifier, SVM, ANN; 10-fold crossvalidation
(Garcia-Ceja & Brena, 2018)	Uses publicly available datasets: Data set 1: N/M; 1; 3-axis accelerometer, 52 Hz; Chest Data set 2: N/M; 1; 3-axis accelerometer, 32 Hz; Wrist	16 total features extracted from time domain	3-stage classifiers of each of the below single classifier model Recursive partitioning, DT, bagging with DTs, SVM, NB, LDA, and RF. 10-fold cross-validation; Leave-one-subject-out cross-validation
(Gupta & Dallas, 2014)	MEMS—Freescale MMA7260 accelerometer 1 (Freescale Semiconductor, Austin, TX); 3-axis accelerometer, 126 Hz; Waist	Several time domain features	NB classifier and KNN classifier Leave-one-subject-out cross-validation

(Gyllensten & Bonomi, 2011)	Tracmor (Philips Research, Eindhoven, The Netherlands) 1; 3-axis accelerometer, 20 Hz; Waist; IDEEA monitor (MiniSun, Fresno, CA); 5; N/M, 32 Hz; Soles of the feet, thighs, and upper sternum	113 features from both time domain and frequency domain.	SVM, Feed-forward NN, DT, Majority Voting (Combining all of the above 3 classifiers) Leave-one-subject-out cross-validation, Validation testing on data sets 2 and 3.
(Hu et al., 2016)	VG350 (MEMSIC, San Jose, CA); acceleration sensor; 1; 3-axis accelerometer, 100 Hz; Waist	120 total input features per axes	Backpropagation NN; 10- fold cross-validation
(Jalloul et al., 2018)	Shimmer3 IMUs; 6; 3-axis accelerometer, 512 Hz; Dominant ankle, nondominant thigh, dominant wrist, nondominant arm, hip, and neck	108 features from the time domain	RF; Validation method 1- Training data: data set1 Testing data: Data set 2 Validation method 2- Training data: Data set 2 Testing data: Data set 1
(John et al., 2013)	GENEA (Unilever Discover, Colworth, United Kingdom); 1; 3-axis accelerometer, 80 Hz; Wrist; ActiGraph™ GT3X+ (ActiGraph™ Inc, Pensacola, FL); 1; 3-axis accelerometer, 80 Hz; Wrist	8 time domain and 6 frequency domain features	RF classifier; Leave-one-subject-out cross-validation,
(Kerr et al., 2018)	ActiGraph GT3X+ (Acti-Graph™ Inc, Pensacola, FL); 1; 3-axis accelerometer, 30 Hz; Hip activPAL (PAL Technologies, Glasgow, Scotland); 1; 1-axis-accelerometer; Thigh	41 time and frequency domain features	RF classifier; Leave-one-subject-out cross-validation

(Kerr et al., 2016)	ActiGraph GT3X+ (Acti-Graph™ Inc, Pensacola, FL); 1; 3-axis accelerometer, 30 Hz; Hip	43 time and frequency domain features	KNN, SVM, NB, DTs, RF, HMM; Leave-one-subject-out cross-validation
(Khan et al., 2010)	Witilt (version 2.5; Sparkfun Electronics, Boulder, CO) 1; 3-axis accelerometer, 20 Hz; Chest	Time domain and augmented features	ANNs; 6-fold cross-validation
(Lee et al., 2011)	SerAccel (version 5; Sparkfun Electronics, Boulder, CO); 1; 3-axis accelerometer, 20 Hz; Chest	Time domain and augmented features	ANNs; Subject-independent validation: training data from 10 subjects (group 1), tested on the remaining 10 subjects (group 2) Subject-dependent validation: training data from 20 subjects, tested on 10 subjects (group 1)
(Liu & Chang, 2009)	Data set 1: KXM52-L20 (Kionix, Ithaca, NY); 3; 3-axis accelerometer, 250 Hz; Chest, waist, and thigh Data set 2 KXM52-L20 (Ithaca,NY); 1; 3-axis accelerometer, 250 Hz;Waist	Features comprise parameters of the autoregression model of the raw signal	SONFIN; Validation method: Training data: 50% of the data set for each activity from each subject Testing data: The remaining 50% of the data set for each activity from each subject
(Mannini et al., 2013)	Wockets; 2; 3-axis accelerometer, 90 Hz; Wrist and ankle	Several time and frequency domain feature sets	SVM; Leave-one-subject-out cross-validation

(Mannini & Sabatini, 2010)	ADXL210E accelerometers (Analog Devices, Norwood, MA); 5; 2-axis accelerometer, 76.25 Hz; Right hip, dominant wrist, nondominant arm, dominant ankle, and nondominant thigh	Several time and frequency domain feature	Naive Bayesian, Gaussian mixture model, logistic classifier, Parzen classifier, SVM, BDT (C4.5), Nearest mean, KNN, ANN (multilayer perceptron), and cHMM-based sequential classifier; Validation method: Training data: 7 windows/activity class/subject Testing data: Remaining windows/activity class/subject
(Montoye et al., 2014)	MICA2 motes (Crossbow Inc, Milpitas, CA); 3; 2-axis accelerometer, 10 Hz; Right wrist, right thigh, and right ankle; ActiGraph (ActiGraph LLC, Fort Walton Beach, FL); 1; 3-axis accelerometer, 30 Hz; Waist	14 total input features (12 time domain from raw accelerometer data, and height and weight of participants); 8 total input features (6 time domain from raw accelerometer data, and height and weight of participants)	ANN; Leave-one-subject-out cross-validation
(Montoye et al., 2017)	ActiGraph GT9X Link (ActiGraph LLC, Pensacola, FL); 4; 3-axis accelerometer, 60 Hz; Right ankle, hip, right wrist, and left wrist	18 features from the time domain	ANN; Leave-one-subject-out cross-validation

(Montoye, Dong, et al., 2016)	MICA2 motes (Crossbow Inc, Milpitas, CA); 3; 2-axis accelerometer, 10 Hz; Right wrist, right thigh, and right ankle; ActiGraph (ActiGraph LLC, Fort Walton Beach, FL); 1; 3-axis accelerometer, 30 Hz; Hip	14 input features (12 time domain from raw accelerometer data, and height and weight of participants); 8 total input features (6 time domain from raw accelerometer data, and height and weight of participants)	ANN; Leave-one-subject-out cross-validation
(Montoye et al., 2015)	GENEActiv; (Activinsights Ltd, Cambridgeshire, United Kingdom); 2; 3-axis accelerometer, 20 Hz; Right wrist and left wrist ActiGraph (ActiGraph LLC, Fort Walton Beach, FL; 2; 3-axis accelerometer, 40 Hz; Right thigh, Right hip	39 input features (36 features from raw accelerometer data, and height, weight, and sex of participants)	ANN; Leave-one-subject-out cross-validation
(Montoye, Pivarnik, et al., 2016b)	GENEActiv (Activinsights Ltd, Kimbolton, Cambridgeshire, United Kingdom); 2; 3-axis accelerometer, 20 Hz; Right wrist and left wrist ActiGraph (ActiGraph LLC, Fort Walton Beach, FL); 2; 3-axis accelerometer, 40 Hz; Right thigh and right hip	15 input features	ANN; Leave-one-subject-out cross-validation

(Montoye, Pivarnik, et al., 2016a)	GENEActiv (Activinsights Ltd, Kimbolton, Cambridgeshire, United Kingdom); 2; 3-axis accelerometer, 20 Hz; Right wrist and left wrist ActiGraph (ActiGraph LLC, Fort Walton Beach, FL); 2; 3-axis accelerometer, 40 Hz; Right thigh and right hip	Multiple features sets were extracted from each accelerometer: Set 1 = 36 features, set 2 = 6 features, set 3 = 12 features, and set 4 = 15 features	ANN; Leave-one-subject-out cross-validation
(Montoye et al., 2017)	activPAL3 accelerometer (PAL Technologies Ltd, Glasgow, United Kingdom); 1; 3-axis accelerometer, 20 Hz; Right thigh	6 total features	ANN; Leave-one-subject-out cross-validation
(Montoye et al., 2018)	GENEActiv (Activinsights Ltd, Kimbolton, Cambridgeshire, United Kingdom); 2; 3-axis accelerometer, 20 Hz; Left wrist GENEActiv (Activinsights Ltd, Kimbolton, Cambridgeshire, United Kingdom); 2; 3-axis accelerometer, 60 Hz; Left wrist	Feature set 1: 27 time domain features Feature set 1: 21 time domain features Feature set 3: 12 time domain features Feature set 4: 6 time domain features Feature set 5: 39 both time domain and frequency domain features Feature set 6: 33 both time domain and frequency domain features	DTs with boosting, RF, ANNs, and SVCs; In-sample leave-one-subject-out cross-validation performed separately on data sets 1 and 2 Out-of-sample validation: training data— data set 1; testing data —data set 2 (vice versa)
(Muscillo et al., 2010)	ADXL202 (Analog Devices, Norwood, MA); 1; 2-axis accelerometer, 100 Hz; Dominant leg at the shin level	32 features (16 time domain and frequency domain features were extracted for each axis)	Naïve 2D Bayes classifier; Model tested on the whole data set

(Pavey et al., 2017)	GENEActiv (Activinsights Ltd, Cambridgeshire, United Kingdom); 1; 3-axis accelerometer, 30 Hz; Nondominant wrist; activPAL (version 3; Pal Technologies Ltd, Glasgow, United Kingdom); 1; 3-axis accelerometer, N/M; Right thigh	Several time and frequency domain features	RF classifier; Leave-one-subject-out cross-validation
(Preece et al., 2009)	Pegasus activity monitors (ETB Technologies, Dalbeattie, United Kingdom); 3; 3-axis accelerometer, 64 Hz; Waist, thigh, and ankle	Several feature sets were generated comprising of features from the time and frequency domains and wavelet transformation	KNN classifier; Leave-one-subject-out cross-validation
(Rosenberg et al., 2017)	ActiGraph GT3X+ (ActiGraph, Pensacola, FL); 1; 3-axis accelerometer, N/M; Right hip	41 features from both time domain and frequency domain	RF combined with HMM; Leave-one-subject-out cross-validation
(Rothney et al., 2007)	IDEEA monitor (MiniSun, Fresno, CA); 3; 2-axis accelerometer, 32 Hz; Hip (at anterior and posterior and at medial/lateral locations)	30 features from both time domain and frequency domain	ANN; Leave-one-subject-out cross-validation
(Sasaki et al., 2016)	ActiGraph GT3X+ (ActiGraph, Pensacola, FL); 3; 3-axis accelerometer, 80 Hz; Hip, wrist, and ankle	Several time and frequency domain features	RF classifier and SVM Leave-one-subject-out cross-validation

(Staudenmayer et al., 2015)	ActiGraph GT3X+ (ActiGraph, Pensacola, FL); 1; 3-axis accelerometer, 80 Hz; Dominant wrist	Several time and frequency domain features	RF classifier; Leave-one-subject-out cross-validation
(Stewart et al., 2018)	Axivity AX3 (Axivity Ltd, York, United Kingdom); 2; 3-axis accelerometer, 100 Hz; Lower back and dominant thigh	142 total time and frequency domain features	RF; Leave-one-subject-out cross-validation
(Trost et al., 2014)	ActiGraph GT3X+ (ActiGraph, Pensacola, FL); 2; 3-axis accelerometer, 30 Hz; Right hip and nondominant wrist	Several time domain features (number N/M)	Regularized logistic regression classifier; Modified 10-fold cross-validation
(Trost et al., 2018)	ActiGraph GT3X+ (ActiGraph, Pensacola, FL); 2; 3-axis accelerometer, 100 Hz; Right hip and nondominant wrist	36 time domain and frequency domain features (18/sensor)	RF classifier and SVM Leave-one-subject-out cross-validation
(Wang et al., 2016)	IMU, N/M; 2; 3-axis accelerometer, 100 Hz; Waist and left ankle	Several features were extracted based on Ensemble Empirical Mode Decomposition method and Game-Theory-Based Feature Selection method	SVM and KNN; Leave-one-subject-out cross-validation
(Willetts et al., 2018)	Axivity AX3 (Axivity Ltd, York, United Kingdom) 1; 3-axis accelerometer, 100 Hz; Wrist	126 features extracted from both time domain and frequency domain	RF and HMM; Leave-one-subject-out cross-validation

(Wu et al., 2012)	MMA7260 (Sparkfun Electronics, Boulder, CO); 1; 3-axis accelerometer, 50 Hz; Waist	48 features were extracted based on discrete wavelet transform and principle component analysis	SVM; Trial 1: 4-fold cross validation Trial 2: out-of-sample validation—classifier trained in trial 1 tested on data set collected in trial 2
(Wullems et al., 2017)	GENEActiv (Activinsights Ltd, Kimbolton, United Kingdom); 2; 3-axis accelerometer, 60 Hz; Right thigh and left thigh	55 features extracted from both time domain and frequency domain	RF classifier; Leave-one-subject-out cross-validation
(Xiao & Lu, 2015)	MPU-6000 sensor (Inven-Sense Inc, Sunnyvale, CA); 1; 3-axis accelerometer, 50 Hz; Thigh	Several time domain and frequency domain features. Some features were also extracted using Kernel discriminant analysis	ELM, NN, and SVM; Data randomly split into training and test data sets (split% N/M)
(Yang et al., 2008)	MMA7260Q (Freescale Semiconductor, Austin, TX); 1; 3-axis accelerometer, 100 Hz; Dominant wrist	24 features extracted from both time domain and frequency domain	Neural classifier and k-NN; Leave-one-subject-out cross-validation
(Zhang, Rowlands, et al., 2012)	GENEA (Colworth, United Kingdom); 1; 3-axis accelerometer, 80 Hz (multiple data sets were created from original data by modifying number of axes [1–3] and sampling frequencies [5, 10, 20, and 40 Hz]); Wrist;	Several time domain and frequency domain features (number N/M)	Logistic regression, DT, SVM, Bayesian belief network, and NN; 10-Fold cross-validation, Split validation (training data—randomly selected two-third of the samples from each activity; test data—remaining one-third samples)

(Zhang, Murray, et al., 2012)	GENEA (Colworth, United Kingdom); 3; 3-axis accelerometer, 80 Hz; Wrists and waist;	Several time domain and frequency domain features (number N/M)	Logistic regression, DT, SVM, Bayesian belief network, and NN; 10-Fold cross-validation, Split validation (training data—randomly selected two-third of the samples from each activity; test data—remaining one-third samples)

Notes. Abbreviations: ANN, artificial neural networks; BDT, binary decision tree; cHMM, KNN, k-nearest neighbour; DMF-DL, dichotomy mapped forest-deep Learning; DMF-Metric, dichotomy mapped forest-metric Learning; DT, decision tree; ELM, extreme learning machine; HMM, Hidden Markov model; LDA, linear discriminant analysis; LPA, light-intensity physical activity; METs, metabolic equivalents; MVPA, moderate-vigorous physical activity; N/A, not applicable; N/M, not mentioned; NB, Naïve Bayes; NMF, nonnegative matrix factorization; NN, neural network; RF, random forest; SB, sedentary behaviour; SONFIN, self-constructing neural fuzzy inference network; SVM, support vector machine; WN, wireless network.

Appendix C: Features included in each model for ENMO, Actigraph counts, and MIMS units

Extracted Features from a 1-second window	**ENMO**	**ActiGraph counts**	**MIMS units**
Sum of X axis data - sum_x	✓	✓	✓
Sum of Y axis data - sum_y	✓	✓	✓
Sum of Z axis data - sum_z	✓	✓	✓
Signal power of X axis data - snp_x		✓	✓
Signal power of Y axis data - snp_y	✓	✓	✓
Signal power of Z axis data - snp_z	✓	✓	✓
Mean of X axis data - mean_x	✓		
Mean of Y axis data - mean_y			
Mean of Z axis data - mean_z			✓
Coefficient of variation of X axis data - cv_x		✓	✓
Coefficient of variation of Y axis data - cv_y		✓	✓
Coefficient of variation of Z axis data - cv_z	✓	✓	✓
Standard Deviation of X axis data - sd_x	✓	✓	✓
Standard Deviation of Y axis data - sd_y	✓	✓	✓
Standard Deviation of Z axis data - sd_z	✓	✓	✓
Skewness of X axis data - skw_x			✓

Skewness of Y axis data - skw_y			
Skewness of Z axis data - skw_z			
Kurtosis of X axis data - krt_x			✓
Kurtosis of Y axis data - krt_y			✓
Kurtosis of Z axis data - krt_z			✓
Interquartile Range of X axis data - iqr_x	✓		
Interquartile Range of Y axis data - iqr_y			
Interquartile Range of Z axis data - iqr_z			✓
Peak-to-peak amplitude of X axis - amp_x	✓		
Peak-to-peak amplitude of Y axis - amp_y	✓	✓	✓
Peak-to-peak amplitude of Z axis - amp_z		✓	✓
Autocorrelation of X axis data - acf_x		✓	
Autocorrelation of Y axis data - acf_y	✓	✓	✓
Autocorrelation of Z axis data - acf_z			✓
Correlation of X and Y axis data - cor_xy			
Correlation of X and Z axis data - cor_xz			
Correlation of Y and Z axis data - cor_yz			

Vector magnitudes of all axes - vec_mag			
Zero-crossing in X axis data - zcr_x	✓	✓	
Zero-crossing in Y axis data - zcr_y		✓	
Zero-crossing in Z axis data - zcr_z	✓	✓	
Peak Intensity of X axis data - pin_x			
Peak Intensity of Y axis data - pin_y			
Peak Intensity of Z axis data - pin_z			
Sum Log Energy of X axis data - sle_x	✓	✓	✓
Sum Log Energy of Y axis data - sle_y		✓	✓
Sum Log Energy of Z axis data - sle_z			
Dominant Frequency of X axis data - dfr_x	✓	✓	
Dominant Frequency of Y axis data - dfr_y		✓	
Dominant Frequency of Z axis data - dfr_z		✓	
Amplitude of Dominant Frequency in X axis - adf_x			✓
Amplitude of Dominant Frequency in Y axis - adf_y		✓	✓
Amplitude of Dominant Frequency in Z axis - adf_z		✓	✓
Entropy of X axis data - ent_x			

Entropy of Y axis data - ent_y				
Entropy of Z axis data - ent_z				