A Data Science and Machine Learning Approach to Measure and Monitor Physical Activity in Children

Paul Fergus, Abir J. Hussain*, Dhiya Al-Jumeily, Ahmed J. Aljaaf, Jan Lunn

Liverpool John Moores University, Byrom Street, Liverpool, L3 3AF, United Kingdom.

ABSTRACT- Physical Activity is a fundamental component for the maintenance of a healthy lifestyle. Recommendations for physical activity levels are issued by most governments as part of public health measures. Therefore, it is vital for regulatory purposes, that there are reliable measurements of physical activity. However, the techniques and protocols used in existing physical activity research, including laboratory-based measurement, have received increasingly critical scrutiny in recent times. Consequently, physical activity researchers have begun to explore the use of wearable sensing technology to capture large amounts of data and the use of machine learning techniques, specifically artificial neural networks, to produce classifications for specific physical activity events. This paper explores this idea further and presents a supervised machine learning approach that utilises data obtained from accelerometer sensors worn by children in free-living environments. The paper posits a rigorous data science approach that presents a set of activities and features suitable for measuring physical activity in children in free-living environments. A Multilayer Perceptron neural network is used to classify physical activities by activity type, using ecologically valid data from body worn accelerometer sensors. A rigorous reproducible data science methodology is presented for subsequent use in physical activity research. Our results show that it was possible to obtain an overall accuracy of 92% using the initial data set, and 99.8% using interpolated cases.

Keywords: Physical Activity, Overweight, Obesity, Machine Learning, Classification, Neural Networks, Sensors.

*Corresponding author Tel. +44(0)231 2458
Email: a.hussain@ljmu.ac.uk

According to the recent McKinsey Global Institute report[1], the global cost of obesity is comparable with combined costs of smoking and armed conflict, which are both greater than the collective costs of alcoholism and climate change. The report states that obesity costs £1.3tn, or 2.8% of annual global economic activity. In the UK, the cost is £47bn. The prevalence of overweight and obesity is alarming. With 2.1bn people (30% of the world's population) being overweight or obese. According to the World Health Organisation (WHO), at least 2.8 million people worldwide die from being overweight or obese. It is also the cause of a further 35.8 million of global Disability-Adjusted Life Years (DALY)[2]. In the United Kingdom, data extracted from the Health Survey for England (HSE) in 2012 [1] reported that the percentage of obese children between the age of 2 and 15 has increased since 1995. Fourteen percent of boys and girls were classified as obese and 28% as either overweight or obese. In addition, 19% of children aged between 11 and 15 were more likely to be obese than children between 2-10 years.

The physical condition of adults is strongly conditioned by the early stages of life. In this sense, the results provided by the HSE revealed that in 2012 almost a quarter of men, specifically 24%, and a quarter of women (25%) were obese, while 42% of men and 32 % of women were overweight. Therefore, an increase in levels of obesity and overweight in the UK is observed which depends, among other factors, on the lack of physical activity. This growing trend of obesity within the UK and other countries and its associated health risks are cause for national and international concern.

Consequently, the accurate measurement of physical activity (PA) in children (particularly in free-living environments) is of great importance to health researchers and policy-makers [2]. One of the most reliable means of activity measurement in children can be captured using accelerometers to measure movement intensity and frequency [3]. This method offers the advantage of providing quantified values for activity intensity over time. However, methodological variations in data gathering and analysis methods used between studies have led to a growing saturation of

[1] Source: McKinsey Global Institute
[2] http://www.who.int/

conflicting cut-points (the quantitative boundaries between PA intensity classes) in the literature [4].

Consequently, several new research directions have been proposed that attempt to address this challenge. One such approach is the use of machine learning techniques for the prediction or detection of activity types and their associated intensity [5]-[7], [44], [45]. This paper builds on this growing research direction and previous works to present a rigorous data science methodology for predicting physical activity types and intensity using a dataset obtained via the field-based protocol described in the literature [8]. Exploratory data analysis is utilised to determine what activities and feature sets produce the best results when activity types are predicted using a Multilayer Perceptron (MLP).

The structure, of the remainder, of this paper is as follows. Section 2 provides and analysis of the problem while Section 3 describes the methods used. Section 4 presents the results before they are discussed in Section 5. The paper is concluded in Section 6.

2. ANALYSIS

2.1 Physical Activity

Physical activity (PA) is defined as any bodily movement produced by skeletal muscles that results in energy expenditure [9] and is measured in Kilojoules (Kj). The measurement of physical activity has become a fundamental component in healthy lifestyle management. Recommendations for physical activity levels are issued by most governments as part of public health measures [10]. However, they tend to be quite frequently updated or adjusted due to external circumstances, such as changes in diet and food pricing [11], sedentary lifestyle [12], technology [13], the built environment [14], family structure [15] and social influences [16].

With frequent changes in these factors, it has become increasingly important from a public health policy-makers perspective to develop a means of reliably measuring physical activity intensity to public health guidelines. Since the mid-90’s, advances have been made in physical activity measurement, particularly with the proliferation of accelerometer-based measurement protocols that provides accurate data capture.

This technological intervention has made it easier to estimate the frequency, duration and intensity of physical activity [17].

There are three intensity levels that describe physical activity, “Light Physical Activity” (LPA), “Moderate Physical Activity” (MPA) and “Vigorous Physical Activity (VPA). These three levels are distinguished using cut-points; threshold levels derived using a single-regression equation or ROC curve analysis [18]. These are mainly used in laboratory-based physical activity research however, there is growing interest in the use of field-based protocols for ecologically valid, free-living physical activity intensity measurement [19]. Underpinning many of these studies is the use of accelerometers, activity monitoring logs and diaries, as a relatively unobtrusive means of obtaining additional data [20]. However, assessing children’s activities, which are generally regarded as being more difficult to classify than the free-living physical activities of adults, is difficult. This is primarily because children tend to possess high variability in their physical activity behaviours over time, which may be easily miscategorised, i.e. classifying short burst of MPA and VPA as inactivity [21].

In free-living environments, it is considered ecologically invalid to use multiple pieces of measurement hardware, due to weight and encumbrance constraints. This is a significant challenge in physical activity research, where V02 masks and heart rate monitors are standard measurement tools. Consequently, interest in the use of accelerometers, as an objective alternative, in current physical activity research has become common practice.

2.2 Accelerometry and Cut Points

Accelerometry allows the measurement and quantification of movement and has become a useful tool for estimating activity intensity over time using cut-points [22]–[24]. There is a degree of controversy surrounding the specification of cut-points [25][26]. The variation between cut-point sets is due to the use of different equations, where variables, such as age, are used. More importantly, cut-point definitions vary even between studies using the same equation. This is an important factor when creating many, significantly different, sets of cut-points [4], where age range and gender need to be clearly defined. This is typically achieved using either linear regression techniques to derive cut-points, or via the use of Receiver Operating Characteristics (ROC) analysis. The most common approach is to use linear

regression techniques, however its validity has been questioned [27]. The use of ROC analysis has also been criticised as a noisy classification measure [28][29]. Consequently, this has led to the use of artificial neural networks (ANNs) and their potential to improve accuracy [8][30], which we will discuss later in the paper.

Concerns have also been raised about the validity of intensity estimation, using accelerometers, across different activity types. Studies have shown their usefulness, but only when specific activities are used [31]. Consequently, the type of activities used may also cause variance in the accelerometer readings obtained. This variability may have significant consequences when considering field-based measurement; although variability may be high in controlled experimental conditions, measuring PA in free-living situations is more problematic, due to the broad range of behaviours and movement types [32]. Moreover, children's activity is especially variable, due to the variable nature of children's play activities [8]. As such, developing a reliable measure of children's free-living physical activity is an ongoing research goal.

2.3 Direct Observation

The variable nature of accelerometer findings has led to a proliferation of varied and conflicting cut-points in the literature. Consequently, a need has been recognised for a more objective protocol for the determination of cut-points. Mackintosh et al. developed a field-calibration protocol, intended for use by researchers in generating objective and inexpensive, population-specific cut-points for sedentary time, MPA and VPA [8]. This field protocol comprises of two elements. The first is a broad set of structured and unstructured activities, representative of free-play situations. The second is the use of direct observation (DO) as a criterion measure for evaluating the subjects physical activity by a corresponding code [33]. DO is regarded as an important tool in PA research; being a direct measure of behaviour it requires little interpretation, having high internal validity [34]. Moreover, DO offers objective data on activity which may be intermittent or otherwise hard to monitor using accelerometers alone [35][36].

2.4 Physical Activity Classification

Artificial neural networks (ANN) have been used to classify physical activity in several studies, with good results [31]. In one such study, Staudenmayer et al. developed an ANN, which classified activity type in adults, using time windows, with 88% overall accuracy and a consistently low Root Mean Squared Error (rMSE) measure [37]. In a study carried out by De Vries et al. a series of ANNs were developed to predict PA in children across a range of activity types. However, the results reported were significantly lower than those reported in Staudenmayer et al [37] with classification accuracies between 57.2% and 76.8% [30], [31].

Trost et al. [5] conducted a rigorous study in which 90 ANN designs (different hidden layer and weight sizes) were developed and trained to predict PA type and intensity. Data from 100 children was used, comprising of data for 12 activities each of child performed. The best performing design was trained using features extracted from a range of time windows (10, 15, 20, 30 and 60 seconds). This rigorous methodology yielded particularly good results, with the most successful network able to predict PA type with 88.4% accuracy over a 60-second window. The classification of PA intensity was also successful, with the network able to classify moderate to vigorous intensity activities 93% of the time.

While Trost's study is one of the forerunners for artificial neural network usage in the classification of physical activity types and intensity, it does not fully investigate the full range of ANN techniques available. Crucially, the study does not explore the potential for using alternative algorithms or parameters to enhance the classification accuracy of the ANN. They discuss the shortcomings of their approach and point out their ANNs high error margins (as high as 44.6% in the case of sedentary activity), recommending that a combination of triaxial accelerometer use and different pattern recognition algorithms be used to generate more precise ANN outputs.

3. METHODOLOGY

Despite the advances made in laboratory-based physical activity classification and the use of artificial neural networks, more in-depth studies are required. This is

especially true when compared to research in free-living environments. The aim of most studies, in physical activity classification and intensity measurement, has been to carry out measurements in controlled environments. However, the Mackintosh et al. study is different, as it involves the assessment of physical activity in young children in free-living environments [8].

The Mackintosh et al. dataset contains records for twenty-eight children aged between 10 and 11 years old from a North-West England primary school. Children completed seven different physical activities performed in a randomised order, which took place in the school playground or classroom with 5 minutes seated rest between each activity. To capture both the sporadic nature of children's activity [38] and locomotive movement best suited to accelerometers [39], the activities incorporated both intermittent and continuous (i.e., walking and jogging) movements representative of culturally-relevant-free-play situations. Sedentary activities were watching a DVD and drawing, which were consistent with those used previously [26].

In summary, the dataset contains 28 records of children, age11.4±0.3 years, height 1.45±0.09 meters, body mass 42.4±9.9 kg, and BMI 20.0 ±4.7, where 46% of the population were male and 54% were female. The dataset also contains physical activity codes from the System for Observing Fitness Instruction Time (SOFIT) [33] to directly observe the children's physical activity behaviours during the activities. The physical activity coding element of SOFIT uses momentary time sampling to quantify health-related physical activity where codes 1 to 3 represent participants' body positions (lying down, sitting, standing), code 4 is walking, and code 5 (very active) is used for more intense activity than walking [33]. These DO physical activity codes have been validated with heart rate monitoring [40], oxygen consumption [40], [41], and accelerometry [42], [43] with preschool to 12th grade children, including those with development delays [34]. Throughout the protocol each child's activity was coded every 10-s by a trained observer.

Prior to observation of each child, ActiGraphs and a digital watch were synchronized to allow data alignment. The data was downloaded from the ActiGraph, and ActiLife 5.5.5 software was used to merge 5-s data to 10-s data in order to align mean activity counts with DO data. For each 10-2 accelerometer counts, DO codes

of 1 and 2 were categorised as sedentary time, code 3 as light intensity activity (LPA), 4 as MPA, and 5 as VPA.

An initial data capture process was performed, to obtain data for each of the subjects' weight, height, gender, age, sitting height, leg length and waist. An additional feature of BMI was calculated using height and weight data. During the performance of activities, accelerometers and a DO protocol were used to obtain activity data. Subjects were supervised throughout the performance of each activity; each student was also observed by an observer using the SOFIT DO protocol. This observer assessed the recorded student DO values of 10-second intervals. During the performance of activities, left hand accelerometer count, right hand accelerometer count, left waist accelerometer count, right waist accelerometer count and direct observation values were captured.

The accelerometer and DO values obtained during the recording and observation period were processed to generate mean values per epoch. Some values were subsequently used to approximate values for an additional set of features. This approximation was achieved via the use of established calculations for mean hand accelerometer count (HAC), mean waist accelerometer count (WAC), direct observation values (DO), body mass index (BMI), heart rate count (HR), moderate physical activity percentage (MPA%), vigorous physical activity percentage (VPA%), indirect calorimetry oxygen consumption (V02), and energy expenditure (EE). The dataset details are presented in Table 1, however, for a complete description of the dataset the reader is referred to Machkintosh et al [8].

	Attributes	Types	Descriptions
1	ID	Nominal	Participant code
2	Age	Numeric	Age in years
3	Weight	Numeric	Body mass weight in kg
4	Height	Numeric	Height in meters
5	Gender	Binary	Participant gender (1 for male and 2 for female)
6	BMI	Numeric	Body mass index calculated according to World health organisation criteria
7	Leg Length	Numeric	Leg length in meters
8	Sitting	Numeric	Sitting height in meters

	Height		
9	Waist	Numeric	Waist in centimetre
10	Accel RH	Numeric	Right hand accelerometer count
11	Accel LH	Numeric	Left hand accelerometer count
12	Accel RW	Numeric	Right waist accelerometer count
13	Accel LW	Numeric	Left waist accelerometer count
14	HR	Numeric	Heart rate count
15	MPA	Numeric	Moderate physical activity percentage
16	VPA	Numeric	Vigorous physical activity percentage
17	DO	Numeric	Direct observation
18	EE	Numeric	Energy expenditure
19	VO2	Numeric	Indirect calorimetric oxygen consumption
20	Class	Categorical	Physical activities (i.e., resting, drawing, free play, DVD watching, playground activities, Jogging and walking).

Table 1: Dataset attributes.

3.1 Data Pre-Processing

One notable concern with the dataset is that a significant number of values were missing. As previously described the study involved 28 participants. However, not all subjects performed every activity, and in some cases, values were missing for a number of features and/or activities. One subject who performed no activities and two subjects who performed only one activity were removed from the dataset. A further six subjects had a significant number of missing values for some or all of the features HR, MPA %, MPA time, VPA% and VPA time. Four of the six subjects were missing values for all activities and were therefore removed from the dataset. The remaining two subjects were missing values from three activities. Substitute values for these students were computed using cubic spline interpolation. For the features EE and V02, five subjects had missing values for some or all of the activities performed. As a result, these five records were removed, leaving a total dataset containing 16 cases per activity, with no missing or null values.

A correlation analysis was performed on the four-accelerometer features contained in the dataset for each of the subjects. Figure one, shows that there is a perfect correlation between the values for right and left hand accelerometers, and a very high level of correlation (0.97) between the values for right and left waist accelerometers. The degree of correlation suggests that using both right and left hand accelerometers or both right and left waist accelerometers is redundant.

	rh	lh	lw	rw
rh	1.0	1.0	0.9	0.9
lh	1.0	1.00	0.9	0.9
lw	0.9	0.9	1.00	0.97
rw	0.9	0.9	0.97	1.00

Table 2: Correlation Matrix Plot of Correlation Coefficients of Accelerometer Features: Left Hand, Right Hand, Left Waist and Right Waist.

Consequently, the mean value of both hand accelerometers and similarly, the mean value of both waist accelerometers were adopted to reduce the dimensionality of the data set.

3.2 Activity Selection

The dataset contains seven activities; resting (Rest), drawing (Draw), free play (Free), DVD watching (DVD), playground activities (Play), Jogging (Jog), and walking (Walk). These activities were selected to capture a range of actions consistent with children's activity, including both intermittent activities representative of free-play situations (e.g. Playground, Free Play), and continuous motions consistent with more typical activity situations (e.g. Walking, Jogging). Sedentary activities were also included, specifically drawing and DVD watching. The seventh activity, resting, was an additional sedentary activity, used to derive basal rates for features such as HR and V02.

A statistical evaluation is performed to determine the differentiability between features. This step allows us to understand the level of overlap between activities. A more easily separable set of activities that contain reduced overlap between values from different activities, or clear patterns that differentiate classes from one another, will yield better results during the classification phase. This was achieved using data

visualisation and trial classification tests using a Multi-Layer Perceptron (MLP) Neural Network. In each case, a subset of the feature set was used – specifically, the features HAC, WAC and DO. This subset contains the three gathered features, which vary by activity.

Figure 1 shows the distribution of different feature values by activity. It is clear that values for a number of activities occupy coincident regions of the feature space. This is most pronounced in Drawing, DVD Watching and Resting. Drawing and DVD watching, in particular, show high overlap, while resting is distinguishable from others by lower DO values. The results suggest that a classification algorithm will find it difficult to make distinctions between drawing and DVD watching. The activity resting however may be easier to distinguish due to slightly differentiated DO values.

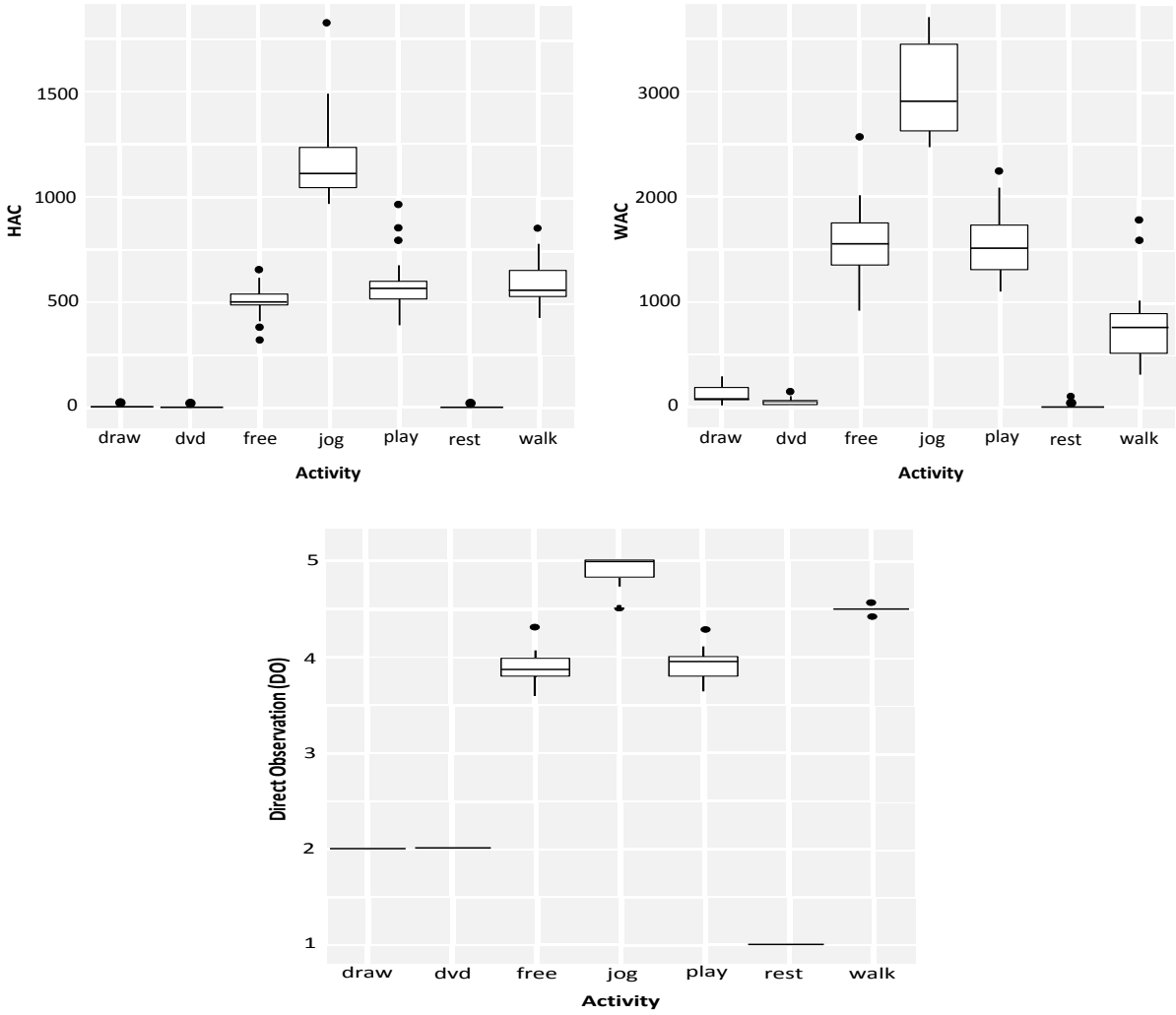

Figure 1: Mean Hand, Waist Accelerometer and Direct Observation Values, subset by Activity

Figure 1 also shows some separation between values for the activities Jogging and Free Play, which should support high classification accuracy. However, there are slight or near overlaps between Free Play and Walking values, which may be sufficient to cause a small reduction in classification accuracy. Having used statistical analysis to identify these trends, the following section determines whether the concerns identified above translate to reduced classifier performance in MLP trials.

3.2.1 MLP Classification Trial for Activity Selection

The classifier trial uses the mean hand and waist accelerometers, BMI and Direct Observation. A four-class classification problem was performed using combinations of four activities from the initial seven activities. Thirty iterations of 35 permutations of the 4-class classification problem were performed, and in each case, sensitivity and specificity data for each activity class was obtained. Figure 2 provides the sensitivity and specificity data, respectively, for each of the seven activities across these trials.

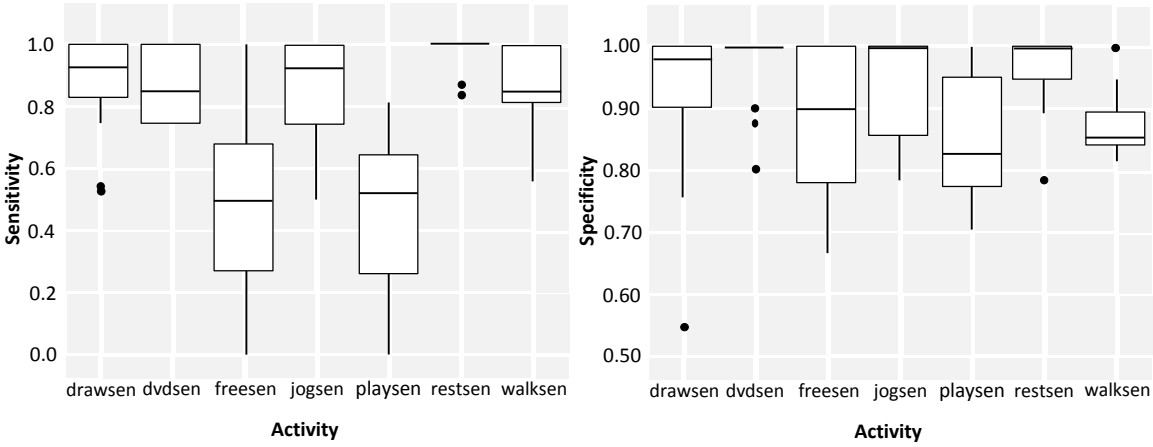

Figure 2: Mean Sensitivity and Specificity values for each activity across 30 MLP Classification Trials per 4-activity combination.

These results show that the activities Free Play and Playground demonstrate a greater spread of both specificity and sensitivity values across a range of classification problems, as well as having lower mean sensitivity and specificity values, than almost all other activities.

Activity values surrounding Free Play and Playground (Jogging and Walking) show substantial variance in specificity, while, Jogging also shows variance in sensitivity. This is related to the significant overlap between feature values for the activities Free

Play, Playground and other activities, as shown in Figure 2. Classification sensitivity (true positive rate) tends to be high for most activities, with mean values above 0.8 for all activities except Free Play and Playground. While mean classification specificity (false positive rate) is very high (>0.95) for all activities except Free Play, Playground and Walking, the variance in classification specificity values is significant for all activities except DVD watching. Observations for Drawing, Resting and DVD watching were confused for one another, which lowered the sensitivity of both features. DVD watching was often mistaken for Drawing or Resting and vice versa, which led to DVD watching having far higher specificities than either Drawing or Resting.

This analysis suggests that the activity set currently in use is not appropriate for ANN classification analysis. This is due to the presence of multiple classes whose observation cases occupy the same region of values. For this reason, only one of the three sedentary activities (Drawing, DVD Watching and Resting) is used. Despite the excellent sensitivity values for the activity Resting showed, it was decided that the activity Drawing would be used. The Resting activity was initially intended by Mackintosh et al. for use in calibrating basal rates for various features, and was not intended for classification analysis. Furthermore, a significant number of features (MPA, MPA time, VPA and VPA time) are missing from the Resting data, which would significantly complicate MLP analyses using those features. Conversely, both Drawing and DVD watching possess a full complement of feature data and were intended for use in classification of sedentary activities. The final dataset following this analysis contains four activities, Drawing, Free Play, Jogging, and Walking. This set covers a good breadth of activity intensities, while minimising the risk of value overlap or classification error.

3.3 Feature Selection

Using sets of faceted Kernel Density Estimation showed that the majority of features had an overlap in values between different activities, i.e. Jogging and Free Play, except for DO. This suggests that the input combinations that contain DO may classify with greater accuracy. One caveat to this is that a single outlying case of Jogging is likely to be misclassified as an instance of Free Play. The features HAC, WAC and DO show good distinction of classes, and generally possess normal

distributions with minimal outliers. This suggests that these features will perform well in classification analysis. The features MPA and MPA time, and VPA and VPA time were sufficiently correlated (0.97 and 1 for each pair respectively) to render the use of all four features redundant. Therefore, the features MPA and VPA are used in subsequent analysis. The BMI feature is normally distributed and is appropriate for use in classification analysis. The feature HR and V02 show significant overlap between Free Play, Jogging and Walking, while EE shows slightly reduced overlap. This suggests that these three features may not be conducive to high-accuracy classification.

Having carried out some exploratory data analysis of individual features, this study proceeded to develop a cross-feature, comparative analysis. This makes it possible to identify both features and feature combinations that potentially display relationships, which may be modelled by a machine-learning algorithm with less or greater difficultly.

3.3.1 Statistical Comparison of Features

Statistical comparison of features is performed on a per-activity basis. Figure 3 shows the plots for the statistical analysis of our features. The feature BMI was retained for all four activities, although naturally the range and distribution of values is identical across all activities. This was done to establish a common scaling for all four plots. The stationary, sedentary nature of the drawing activity was intended to provide a resting comparison to the more vigorous activities used. As such, all features have values at or around the minimum value of -1. For the most part, this suggests that Drawing may be easily distinguished from the other three activities.

Feature values show a broader spread for Free Play than for any other activity. The features HR, MPA, VPA, V02 and EE show significant coincidence between Free Play and Jogging, while some degree of coincidence between values for Free Play and Walking is present for almost every feature; suggesting that misclassification may occur at those class boundaries. The features HAC, WAC, and DO possess a small interquartile range, entailing that the majority of the data falls within a limited space of values. This is a positive finding for classification purposes, but one, which requires validation through MLP analysis.

Conversely, the interquartile range of the features MPA and VPA varies greatly. These features are measures of what proportion of the activity time was spent at vigorous or moderate levels of physical activity. In some activities this leads to an unusual distribution of values for both features; if an activity is vigorous, for instance, the VPA value may be at the maximum value for all subjects, if an activity is not vigorous, the values for all subjects performing the activity may be at the minimum of -1. However, activities which may or may not be vigorous, or which alternate between vigorous and non-vigorous activity states, tend to contain a range of MPA or VPA values.

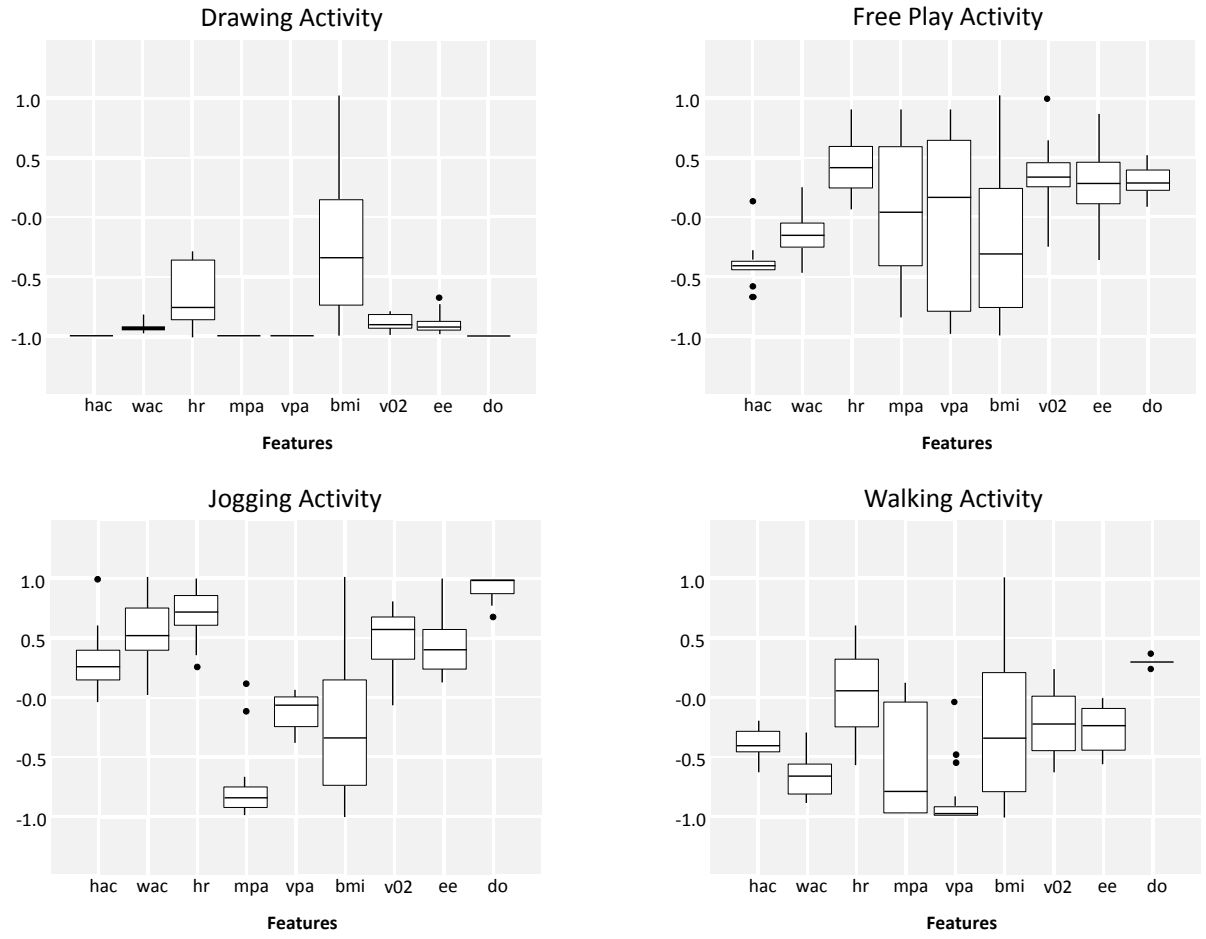

Figure 3: Boxplots per activity for feature statistical analysis

In the case of Free Play, the range of values stretches between (approx.) 0.9 and -1, which implies that the activity was classed as vigorous for some participants, and not vigorous for others. In the case of MPA, the spread of values is less pronounced;

with a mean value of -0.5, participants performing Free Play were classified as non-MPA more often than as MPA.

Nonetheless, the spread of values for both features is likely to cause significant classification problems when using VPA or MPA as features. This problem is particularly pronounced for VPA, where classification of Free Play using the feature is likely to be confused for any other feature with a similar, semi-vigorous profile.

From this analysis, the features HAC, WAC and DO are likely to yield the best results during classification. However, the following section will evaluate several feature combinations and provide empirically evident feature sets and associated classification accuracies to demonstrate their usefulness in classifying activity types.

4. RESULTS

This section describes the classification of activity types using MLP analysis and different feature combinations. Input layer sizes between 1 and 4 features were considered. Results are also presented using a larger input data set, computed via cubic spline interpolation techniques.

4.1 MLP Network Analysis Using 2-4-4 Architecture

This evaluation uses feature pairs. The performance for the classifier is evaluated, using the mean accuracy of 30 simulations with each simulation comprising randomly selected training and test sets.

4.1.1 Classifier Performance

The first evaluation uses all the features in the data set to construct feature pairs. Table 3, shows the top 10 highest mean accuracies obtained over 30 simulations (the remainder were excluded because of their low accuracy values).

	Feature One	Feature Two	Accuracy
1	hr	hac	74
2	hac	hr	74
3	hr	ee	67
4	ee	hr	67
5	hac	ee	61
6	ee	hac	61
7	hr	v02	60
8	v02	hr	60
9	hac	do	59

10	do	hac	59

Table 3: Mean Percentage Classification Correctness by Feature Pair

Table 3 shows that the mean classification accuracy rarely exceeded 70% and in many cases was between 40 and 60%. Variance between classification accuracy during trials was also high, with some feature combinations. This combination of high variance and low classifier accuracy indicate that feature pairs are insufficiently consistent and insufficiently accurate for use in subsequent MLP analysis.

4.2 MLP Network Analysis Using 3-4-4 Architecture

The feature space was increased to triple feature combinations. The performance for the classifier is determined, using the mean accuracy obtained from 30 simulations. The metric includes Sensitivity, Specificity and Kappa estimates. Again, randomly selected training and test sets are used for each simulation.

4.1.1 Classifier Performance

Using the triple feature combinations, Table 4, shows the top 10 highest mean accuracies, sensitivity, specificity and kappa values obtained from 30 simulations.

	Features	Accuracy	Sensitivity	Specificity	Kappa
1	wacbmido	96	0.95	0.99	0.94
2	waceedo	96	0.95	0.99	0.94
3	v02eedo	96	0.95	0.99	0.94
4	hreedo	94	0.93	0.97	0.92
5	bmieedo	94	0.94	0.97	0.91
6	wacv02do	93	0.93	0.97	0.89
7	bmiv02do	93	0.93	0.97	0.89
8	hacv02do	92	0.95	0.97	0.89
9	hacv02ee	91	0.88	0.92	0.87
10	haceedo	90	0.91	0.94	0.86

Table 4: Mean Percentage Classification Correctness by Three Features

Table 4 shows that the classification accuracy using triple feature combinations improves the results significantly. While mean classifier accuracies in the low 60th percentile were observed in a number of cases, several cases displayed mean classification accuracies >90%. These findings are highly positive, suggesting that modification or sophistication of the classification techniques used may further improve classification accuracy.

4.3 MLP Network Analysis Using 4-4-4 Architecture

This set of results extends the feature space to four to determine whether further improvements can be made. Table 5 presents the results.

4.3.1 Classifier Performance

Using a combination of four features, Table 5 again, shows the top 10 highest mean accuracies, sensitivity, specificity and kappa values.

	Features	Accuracy	Sensitivity	Specificity	Kappa
1	bmiv02doee	96	0.95	0.99	0.94
2	bmiv02wachac	96	0.95	0.99	0.94
3	bmidoeewac	96	0.95	0.99	0.94
4	v02doeewac	96	0.95	0.99	0.94
5	v02doeehac	96	0.95	0.99	0.94
6	doeewachac	96	0.95	0.99	0.94
7	hrdoeewac	96	0.95	0.98	0.93
8	bmiv02eehac	95	0.94	0.98	0.93
9	bmieewachac	95	0.94	0.97	0.93
10	v02dowachac	95	0.94	0.97	0.93

Table 5: Mean Percentage Classification Correctness by Four Features

The results show that 4-feature combinations improve the results further. Of all the combinations empirically tested no feature combination showed mean classification accuracies below 87%

4.4 Interpolated Data and MLP Network Analysis Using 4-4-4 Architecture

Having found larger input feature combinations yield improving classification accuracy and reduced variation, the results in this section considers interpolated data and 4-feature input combinations. The original dataset of 16 observations is extended to 64 using interpolation.

4.4.1 Classifier Performance

The design of MLP analysis trials using the interpolated data set and 4-feature input combinations is modelled on previous classification trials that used the un-interpolated data set. The results are shown in Table 6.

	Features	Accuracy	Sensitivity	Specificity	Kappa
1	doeewachac	98	0.98	0.99	0.96
2	hrdoeehac	97	0.98	0.99	0.95
3	hrdoeewac	97	0.98	0.99	0.95

4	bmidowachac	97	0.98	0.99	0.95
5	v02dowachac	97	0.98	0.99	0.95
6	bmiv02dohac	95	0.96	0.99	0.93
7	bmiv02doee	95	0.96	0.99	0.93
8	v02doeehac	95	0.95	0.98	0.93
9	bmiv02dowac	95	0.96	0.99	0.93
10	bmieewachac	95	0.96	0.99	0.93

Table 6: Mean Percentage Classification Correctness by Four Features

These results clearly show that the use of interpolation to generate an extended set of training cases has had a significant impact on classification accuracy across all feature combinations. A few feature combinations did show (which we found in our empirical evaluations) less than 90% accuracy. Strikingly, the 96% classification accuracy barrier seen in previous trials was exceeded, with the best performing feature combination (do, ee, hac, wac) showing 98% classification accuracy. A minority of feature combinations showed perfect classification of validation data in some trials.

4.5 Interpolated Data and MLP Network Analysis Using 3 ecologically valid-4-4 Architecture

This set of results re-runs the previous interpolated experiment using ecologically valid triple-feature input combinations and the interpolated data set. The results are shown in Table 7.

4.5.1 Classifier Performance

	Features	Accuracy	Sensitivity	Specificity	Kappa	Prev. Acc
1	wachacdo	99.8	0.99	0.99	0.99	88
2	wachacbmi	95.0	0.95	0.98	0.93	64
3	wacbmido	95.0	0.95	0.98	0.92	96
4	hacbmido	88.0	0.88	0.96	0.83	73

Table 7: Classification Performance for Ecologically Valid 3-Feature Input Combinations Using Interpolated Data

The result from this analysis is highly promising. Classification analysis for each ecologically valid triple-feature input combinations significantly improves the results compared with all previous evaluations. Moreover, MLP ANNs utilising the input feature combination HAC, WAC and DO achieved over 99% classification accuracy, with a Kappa, Sensitivity and Specificity value >0.99%. Classification error stemmed

from the misclassification of two test cases in 10% (3 out of 30) trials. In 90% of trials, this feature combination classified test cases with 100% accuracy.

4.6 Performance comparison

Although MLP ANNs has achieved over 99% of classification accuracy using ecologically valid triple-feature input combinations, this section explores other supervised machine learning methods using the same combination of features. Four well-known machine learning methods have been targeted in this section, namely: Support vector machine (SVM), Decision tree (DT), Naïve Bayes (NB) and Nearest Neighbour (NN) methods. The following Table 8 shows their overall classification performance.

	Methods	Accuracy	Sensitivity	Specificity	Kappa
1	SVM	70.4	0.70	0.52	0.65
2	DT	75.4	0.75	0.62	0.71
3	NB	79.5	0.79	0.68	0.76
4	NN	82.7	0.82	0.73	0.79

Table 8: Classification Performance of Four machine learning methods Using Ecologically Valid 3-Feature Input Combinations

Despite all the methods have achieved lower classification performance in comparison with MLP ANNs, they have obtained reasonably good results. Among them, NN method has achieved the highest overall performance with slightly less than 83% of classification accuracy and 0.82, 0.73 of sensitivity and specificity, respectively. NB has achieved the second highest overall performance with approximately 3% less than NN in classification accuracy, sensitivity and Kappa, followed by DT and SVM, which comes at the end of the list with 70% of accuracy and sensitivity.

5. DISCUSSION

The initial classifications on the dataset obtained a relatively low accuracy with HR and HAC providing the best pair of features. Heart rate, displayed higher classification accuracy across a number of feature combinations. While MPA and VPA failed to perform sufficiently well to justify their inclusion in further trials. These features followed different trends to other features. For example, MPA reached

maximum values during the performance of activities such as walking and free play, where other features tended to show mid-range values. This is considered an advantage due to the potential additional information content of features with this pattern in conjunction with more normally distributed features. The preceding feature pair analysis demonstrates that neither MPA nor VPA provide useful classifications of activities by type and should thus be excluded from the feature set.

Extending the feature space to three showed a marked improvement in classifier performance. In particular, it should be observed that classification accuracy peaked at 96%, with a maximum kappa value of 0.94. This value was seen consistently across all trials of a small number of feature combinations. This ceiling was due to the consistent misclassification of a single value; each of the network designs in question successfully classified all other values correctly across trials, but misclassified this single record on every occasion. Specifically, one record captured from participants performing the activity Jogging was consistently misclassified by the MLP networks as an instance of the activity Free play.

However, what these findings show is that larger input feature combinations produce higher classification accuracy and reduce variance between MLP trial iterations. A logical extension of the proceeding analysis, was to extend the input feature combination to a total of four features. The results showed that the top end classifiers as seen in Table 5 continue to fail to classify certain data values. Only one input feature combination (HR, DO, EE, V02) enabled perfect classification of the dataset, and perfect classification occurred in less than 7% of the cases (2/30). However, these instances of perfect classification do demonstrate that improved classification accuracy is attainable although it may not be achievable without the use of new techniques.

One option was to alter the dataset to allow the ANN architecture to classify the difficult-to-classify cases. This alteration was performed by adding interpolated training cases. This allowed the ANN to identify trends in data features, yielding improved overall classification accuracy as can be seen in Table 6. However, despite the excellent results yielded by this data set, all of the four-feature input combinations used in preceding analyses raised at least one of the following concerns. Some of the features used contain entirely generated values, for instance

EE, HR, and V02, which contain computed values derived from observed data features. This generation of features was necessitated by the low-encumbrance nature of the field-based protocol under examination in the Mackintosh et al. study. However, this practice led to the creation of a set of features calculated based on the values of other features, and furthermore introduces questions regarding statistical validity.

Furthermore, in order to capture data for each of the computed features, the use of specialised techniques and equipment (for instance oxygen masks) would be required. However, the use of such equipment and techniques is deemed to pose a burden on participants, which is incompatible with low-encumbrance activity. This is especially true in studies involving youth, who have lower encumbrance limits. As a result, it became apparent that the collection of ecologically valid data for the full range of features used in the preceding analysis is not feasible within the requirements of the field-based study. In short, there are significant theoretical concerns with the ecological validity of including certain combinations of the data features as inputs to preceding analyses. At the same time, there are feature combinations that may be gathered within the confines of a free-living study. These features are WAC, HAC, DO, and BMI.

Excluding 2-feature combinations (due to having previously demonstrated low classification accuracy), there are five potential combinations of these features, WAC, HAC, DO, BMI; WAC, HAC, DO; WAC, HAC, BMI; HAC, DO, BMI; and WAC, DO, BMI. The single 4-feature input combination, which utilised these features on an interpolation extended data set, showed strong classification accuracy of 97%. The 3-feature combinations, when used as inputs to analyse un-interpolated data, showed classification accuracies between 96 and 64%.

Therefore, in the final evaluation using the ecologically valid triple feature combinations WAC, HAC and DO, the results were highly positive. A low-encumbrance set of data features, which may be gathered in an ecologically valid way, demonstrate that good classification accuracy in classifying activity types is possible. However, there is still a concern about the use of interpolated training cases computed from the limited initial sample. Any deviation or outliers in the initial data sample will be expanded in the interpolated data set. Without the collection of a

larger set of real PA data using the same protocol, for comparison purposes, it is not possible to guarantee that the data used for training and classification is representative of children's PA. At the same time, it is unlikely that the data set obtained is wholly or even largely misrepresentative of children's PA; outliers do not invalidate the entire data set. This means, at worst, that the data used is largely (but not wholly) representative of the distribution and statistical properties of feature data obtained during children's performance.

6. CONCLUSIONS AND FUTURE WORK

This study used an existing dataset from recent research into physical activity in youth to classify activity data by the type of activity engaged upon. The dataset was analysed using rigorous data science techniques, which led to an improved understanding of activity types and features. Data items whose properties impeded classification were removed. This did affect the size of the dataset and consequently, interpolation techniques were applied to generate an extended set of training cases.

A series of machine learning analyses were performed. A range of classifier types, input feature combinations and architectural parameters were employed and refined to develop improved classification accuracy. Upon developing a set of parameters compatible with high-accuracy classification, MLP, analysis was performed using a specific subset of ecologically valid data features. Classification using an MLP architecture and imputed dataset, using the input feature combination WAC, HAC, and DO yielded a classification accuracy of 99.8%.

While the results show, specific activities and features tailored around a machine learning approach are promising, a great deal of research remains. Further development of Data Science techniques will help provide a varied and broad range of possibilities, particularly validating the preceding results using a substantial non-interpolated data set. While this study focused on youth, it would be interesting to look at other population groups, such as adults and the elderly. Finally, one important point would be to standardise the use of activities and cut points in PA research that is underpinned with strong Data Science evidence and advanced machine learning techniques.

REFERENCES

[1] A. Ryley, “HEALTH SURVEY FOR ENGLAND 2012: Chapter 11, Children’s BMI, overweight and obesity,” 2013.

[2] O. Oyebode and J. Mindell, “Use of data from the Health Survey for England in obesity policy making and monitoring,” *Obes. Rev.*, vol. 14, no. 6, pp. 463–476, 2013.

[3] K. Konstabel, T. Veidebaum, V. Verbestel, L. A. Moreno, K. Bammann, M. Tornaritis, and G. Eiben, “Objectively measured physical activity in European children: the IDEFICS study,” *Int. J. Obes.*, vol. 38, pp. S135–S143, 2014.

[4] S. G. Trost, P. D. Loprinzi, R. Moore, and K. A. Pfeiffer, “Comparison of accelerometer cut points for predicting activity intensity in youth,” *Med Sci Sport. Exerc.*, vol. 43, no. 7, pp. 1360–1368, 2010.

[5] S. G. Trost, “Artificial Neural Networks to Predict Activity Type and Energy Expenditure in Youth,” *Med Sci Sport. Exerc.*, vol. 44, no. 5, pp. 1801–09, 2012.

[6] A. Dalton and G. O’Laighin, “Comparing Supervised Learning Techniques on the Task of Physical Activity Recognition,” *Biomed. Heal. Informatics*, vol. 17, no. 1, pp. 46–52, 2013.

[7] B. Barshan and M. C. Yuksek, “Recognizing Daily and Sports Activities in Two Open Source Machine Learning Environments Using Body-Worn Sensor Units,” *Comput. J.*, 2014.

[8] K. A. Machkintosh, S. J. Fairclough, G. Stratton, and N. D. Ridgers, “A Calibration Protocol for Population-Specific Accelerometer Cut-Points in Children,” *PLoS One*, vol. 7, no. 5, p. e36919, 2012.

[9] C. J. Caspersen, K. E. Powell, and G. M. Christenson, “Physical Activity, Exercise and Physical Fitness: Definitions and Distinctions for Health-Realted Research,” *Public Heal. Rep.*, vol. 100, no. 2, pp. 126–131, 1985.

[10] R. R. Pate, M. Pratt, S. N. Blair, and E. Al., “Physical Activity and Public Health: A Recommendation from the Centers for Disease Control and

Prevention and the Amercial College of Sports Medicine," *JAMA*, vol. 273, no. 5, 1995.

[11] K. J. Duffey, P. Gordon-Larsen, J. M. Shikany, D. Guilkey, and E. Al., "Food Price and Diet and Health Outcomes: 20 Years of the CARDIA Study.," *Arch. Intern. Med.*, vol. 170, no. 5, pp. 420–426, 2010.

[12] M. A. Martinez-Gonzalez, "Physical Inactivity, Sedentary Lifestyle and Obesity in the European Union," *Int. J. Obes.*, vol. 23, no. 11, pp. 1192–201, 1999.

[13] S. Kautiainen, L. Koivusilta, T. Lintonen, S. M. Virtanen, and A. Rimpela, "Use of Information and Communication Technology and Prevalance of Overweight and Obesity Among Adolescents," *Int. J. Obes.*, vol. 29, no. 8, pp. 925–33, 2005.

[14] B. E. Saelens, J. F. Sallis, J. B. Black, and D. Chen, "Neighborhood-based Differences in Physical Activity: an Environment Scale Evaluation," *Am. J. Public Health*, vol. 93, no. 9, pp. 1552–58, 2003.

[15] I. Lissau and T. I. Sorensen, "Parental Neglect During Childhood and Increased Risk of Obesity in Young Children," *Lancet*, vol. 343, no. 8893, pp. 324–7, 1994.

[16] B. Mcferran, D. W. Dahl, G. J. Fitzsimons, and A. C. Morales, "I'll Have What She's Having: Effects of Social Influence and Body Type on the Food Choices of Others," *J. Consum. Res.*, vol. 36, no. 6, 2010.

[17] S. G. Trost, K. L. McIver, and R. R. Pate, "Conducting Accelerometer-based Activity Assessments in Field-based Research," *Med. Sci. Sport. Exerc.*, vol. 37, no. 11, p. S531, 2005.

[18] R. Jago and E. Al., "Decision Boundaries and Receiver Operating Characteristic Curves: New Methods for Determining Accelerometer Cutpoints," *J. Sports Sci.*, vol. 25, no. 8, pp. 937–944, 2007.

[19] C. S. Layne and E. Al., "Development of an Ecologically valid Approach to Assess Moderate Physical Activity Using Accelerometry in Community Dwelling Womenof Color: A Cross-sectional Study," *Int. J. Behav. Nutr. Phys. Act.*, vol. 8, no. 1, p. 21, 2011.

[20] G. Leticia and E. Al., “Assessment of intensity, prevalence and Duration of Everyday Activities in Swiss School Children: A Cross-Analysis of Accelerometer and Diary Data,” *Int. J. Behav. Nutr. Phys. Act.*, vol. 6, no. 50, 2009.

[21] C. J. Riddoch, C. Mattocks, K. Deere, J. Saunders, J. Kirkby, K. Tilling, S. D. Leary, S. N. Blair, and A. R. Ness, “Objective measurement of levels and patterns of physical activity.,” *Arch. Dis. Child.*, vol. 11, no. 2, pp. 963–969, 2007.

[22] S. G. Trost, “Objective measurement of physical activity in youth: current issues, future directions.,” *Exerc. Sport Sci. Rev.*, vol. 29, no. 1, pp. 32–36, 2001.

[23] G. J. Welk and C. B. Corbin, “The validity of the Tritrac-R3D activity monitor for the assessmetn of physical activity in children,” *Res. Q. Exerc. Sport*, vol. 66, no. 3, pp. 202–209, 1995.

[24] A. E. Ott, R. R. Pate, S. G. Trost, D. S. Ward, and R. Saunders, “The use of uniaxial and triaxial accelerometers to measure children’s free-play physical activity,” *Pediatr. Exerc. Sci.*, vol. 12, no. 4, pp. 360–370, 2000.

[25] D. R. Bassett, A. V. Rowlands, and S. G. Trost, “Calibration and validation of wearable monitors,” *Med Sci Sport. Exerc.*, vol. 44, no. 1, pp. S32–8, 2012.

[26] K. R. Evenson, D. Cattellier, K. Gill, K. Ondrak, and R. G. McMurray, “Calibration of two objective measures of physical activity for children,” *J. Sports Sci.*, vol. 26, no. 14, pp. 1557–1565, 2008.

[27] J. R. Zakeri, T. Baranowski, and K. Watson, “Decision boundaries and receiver operating characteristic curves: new methods for determining accelerometer cut points,” *J. Sports Sci.*, vol. 25, no. 8, pp. 937–944, 2007.

[28] D. M. W. Powers, “The problem with area under the curve,” in *International Conference on Information Science and Technology*, 2012, pp. 567–573.

[29] J. A. Hanley and B. J. McNeil, “A method of comparing the areas under receiver operating charactersitic curves derived from the same cases,” *Radiology*, vol. 148, no. 3, pp. 839–843, 1983.

[30] S. I. De Vries, F. G. Garre, L. H. Engbers, H. V. H., and S. Van Buuren, “Evaluation of neural networks to identify types of activity using accelerometers,” *Med Sci Sport. Exerc.*, vol. 43, no. 1, pp. 101–107, 2001.

[31] S. I. De Vries, M. Engels, and F. G. Garre, “Identification of children’s activity type with accelerometer-based neural networks,” *Med Sci Sport. Exerc.*, vol. 43, no. 10, pp. 1994–1999, 2011.

[32] J. R. Sirard, S. G. Trost, K. A. Pfeiffer, M. Dowda, and R. R. Pate, “Calibration and evaluation of an objective measure of physical activity in preschool children,” *J. Phys. Act. Heal.*, vol. 2, no. 3, pp. 345–357, 2005.

[33] T. L. Mckenzie, J. F. Sallis, and P. R. Nader, “SOFIT: System for observing fitness instruction time,” *J. Teach. Phys. Educ.*, vol. 11, no. 2, pp. 195–205, 1992.

[34] T. L. McKenzie, “2009 C. H. McCloy Lecture. Seeing is beliving: observing physical activity and its contexts,” *Res Q Exerc Sport*, vol. 8, no. 12, pp. 113–122, 2010.

[35] G. J. Welk, J. C. . Eisenmann, J. Schaben, S. G. Trost, and D. Dale, “Calibration of the Biotrainer Pro activity monitor in children,” *Pediatr. Exerc. Sci.*, vol. 19, no. 2, pp. 145–148, 2007.

[36] S. E. Crouter, G. C. Kurt, and D. R. Basset, “A novel method for using acceleromter data to predict energy expenditure,” *J. Appl. Physiol.*, vol. 100, no. 4, pp. 1324–1331, 2006.

[37] J. Staudenmayer, D. Prober, S. Crouter, D. Bassett, and P. S. Freedson, “An artificial neural network to estimate phuysical activity energy expenditure and identify physical activity tuype from an accelerometer,” *J. Appl. Physiol.*, vol. 107, no. 4, pp. 1300–1307, 2009.

[38] M. Orme, K. Wijndaele, S. J. Sharp, K. Westgate, U. Ekelund, and S. Brage, “Combined influence of epoch length, cut-point and bout duration on accelerometry-derived physical activity,” *Int. J. Behav. Nutr. Phys. Act.*, vol. 11, no. 34, pp. 11–34, 2014.

[39] G. J. Welk, “Principles of design and analyses for the calibration of accelerometry-based activity monitors,” *Med Sci Sport. Exerc.*, vol. 37, no. 11, pp. 501–511, 2005.

[40] H. Rowe, P. van der Mars, J. Schuldheisz, and S. Fox, “Measuring students’ physical activity levels: validating SOFIT for use with high-school students,” *J. Teach. Phys. Educ.*, vol. 23, no. 3, pp. 235–251, 2004.

[41] J. J. Honas, R. A. Washburn, B. K. Smith, J. L. Greene, G. Cook-Wiens, and E. Al., “The system for observing fitness instruction time (SOFIT) as a measure of energy expenditure during classroom-based physical activity,” *Pediatr. Exerc. Sci.*, vol. 20, no. 4, pp. 439–445, 2008.

[42] P. W. Scruggs, S. K. Beveridge, and B. D. Clocksin, “Tri-axial accelerometry and heart rate telemetry: relation and agreement with behavoural observation in elementarty physical education,” *Meas Phys Educ Exerc Sci*, vol. 9, no. 4, pp. 203–218, 2005.

[43] S. V. Sharma, R. J. Chuang, and K. Skala, “Measuring physical activity in preschoolers: reliability and validity of the sysetm for observing fitness instruction time for preschoolers (SOFIT-P),” *Sci, Meas Phys Educ Exerc*, vol. 15, no. 4, pp. 257–273, 2011.

[44] D.S.Huang, Ji-Xiang Du, “A constructive hybrid structure optimization methodology for radial basis probabilistic neural networks,” IEEE Transactions on Neural Networks, vol. 19, no.12, pp 2099-2115, 2008.

[45] D.S.Huang,“Radial basis probabilistic neural networks: Model and application,”International Journal of Pattern Recognition and Artificial Intelligence, 13(7), pp.1083-1101, 1999 .