Published in final edited form as:

Physiol Meas. 2016 March ; 37(3): 360–379. doi:10.1088/0967-3334/37/3/360.

Comparative evaluation of features and techniques for identifying activity type and estimating energy cost from accelerometer data

Rohit J. Kate¹, Ann M. Swartz^{2,3}, Whitney A. Welch², and Scott J. Strath^{2,3}

¹Department of Health Informatics and Administration, University of Wisconsin-Milwaukee, Milwaukee WI, USA

²Department of Kinesiology, University of Wisconsin-Milwaukee, Milwaukee, WI, USA

³Center for Aging and Translational Research, University of Wisconsin-Milwaukee, Milwaukee, WI, USA

Abstract

Wearable accelerometers can be used to objectively assess physical activity. However, the accuracy of this assessment depends on the underlying method used to process the time series data obtained from accelerometers. Several methods have been proposed that use this data to identify the type of physical activity and estimate its energy cost. Most of the newer methods employ some machine learning technique along with suitable features to represent the time series data. This paper experimentally compares several of these techniques and features on a large dataset of 146 subjects doing eight different physical activities wearing an accelerometer on the hip. Besides features based on statistics, distance based features and simple discrete features straight from the time series were also evaluated. On the physical activity type identification task, the results show that using more features significantly improve results. Choice of machine learning technique was also found to be important. However, on the energy cost estimation task, choice of features and machine learning technique were found to be less influential. On that task, separate energy cost estimation models trained specifically for each type of physical activity were found to be more accurate than a single model trained for all types of physical activities.

Keywords

activity identification; energy cost estimation; machine learning; feature representation

Address for correspondence: Rohit J. Kate, PhD, Department of Health Informatics and Administration, University of Wisconsin-Milwaukee, 2025 E. Newport Avenue, Milwaukee, WI 53211, USA, kate.rj@uwm.edu, phone: 414-229-4216, fax: 414-229-5100.

Conflict of Interest

The results of the present study do not constitute endorsement by the authors of the products described in this paper. The authors have no financial conflict of interest with any of the monitor manufacturers and have received no research funding from these companies.

1. Introduction

Accurate assessment of physical activity (PA) in free-living conditions has several important applications. For instance, it can be used in studies determining the link between PA levels and common health problems (Barengo *et al* 2004, Yancey *et al* 2004), in evaluating effectiveness of health interventions aimed at increasing levels of PA (van Sluijs *et al* 2007), in monitoring PA levels of hospitalized patients and older adults in nursing facilities (Hirsch *et al.* 1990, Hoenig and Rubenstein 1991), as well as in conducting PA surveillance estimates. For all these and other applications, widely varying physical activities need to be assessed under their natural free-living conditions, and incorporate the assessment of different activity domains (namely, household, occupation, transportation and leisure-time) with minimal intrusion and discomfort to the subjects. Wearable accelerometers offer such a way of assessing PA and hence its use has seen tremendous growth as a means to assess PA in the last 10–15 years.

Accelerometers typically generate data in the form of counts per unit of time. Early methods to assess PA using accelerometers used regression formulas that related accelerometer counts to various levels of PA or energy cost (Crouter *et al* 2006, Hendelman *et al* 2000, Matthews 2005, Swartz *et al* 2000). However, these methods were not found to be very accurate (Lyden *et al* 2011, Strath *et al* 2003) and would not work across different activity domains (Ward 2005). As a result, more recently, machine learning techniques have been employed which automatically learn the relation between input of accelerometer data and output of PA type or energy cost from the provided training data. There are two separate tasks in PA assessment – identifying the type of PA (e.g., walking or computer work or vacuuming etc.), and/or estimating the energy cost for performing the PA (typically measured in metabolic equivalents). PA identification is a classification task (i.e. predicting a class) and energy estimation task is a regression task (i.e. predicting a numerical value). A wide variety of machine learning techniques are available for both classification and regression tasks. Machine learning techniques differ in how they mathematically relate input to output and how they learn this relation from the training data. Some well-known machine learning techniques include support vector machines (Cristianini and Shawe-Taylor 2000), neural networks (Haykin 1999), logistic regression (Hosmer and Lemeshow 2004), decision trees (Quinlan 1993), naïve Bayes (Lewis 1998), k-nearest neighbors (Aha and Kibler 1991) and random forest (Breiman 2001).

Most machine learning techniques require input data in the form of a vector of features. However, for the task of assessing PA, the data from the accelerometer is not in this form but rather in the form of a time series which is typically a series of counts per second for the duration of a PA. Hence in order to employ machine learning techniques, accelerometer time series data is first converted into a vector of features that is representative of the time series. Summary statistics of the accelerometer time series have been the most commonly used features for PA assessment (Staudenmayer *et al* 2009, Trost *et al* 2012).

Performance of machine learning techniques can vary widely depending upon the task, the dataset, and the feature representation. Hence there is a need to do a comparative evaluation of different machine learning techniques and features for PA assessment from accelerometer

data. In this paper, several machine learning techniques are compared along with different feature representations for both the PA identification task and for the energy cost estimation task. Although a few papers have previously compared machine learning techniques on the PA identification task (Bao and Intille 2004, Parkka *et al* 2006, Pirttikangas *et al* 2006, Ermes *et al* 2008, Zhang *et al* 2012), they used fewer subjects (8 to 60) compared fewer machine learning techniques (2 to 5), and did not compare various types of features. In contrast, in this paper, the comparison is done with a large data of 146 subjects, using 8 different types of machine learning techniques, and in addition, various types of feature representations have also been compared. These feature representations include, besides commonly used summary statistics features, less commonly used distance based features as well as discrete features straight from the time series data. All these comparisons have also been done for the energy cost estimation task. Additionally, for that task, the paper also presents comparison between models trained for all types of activities and models specifically trained for each type of activity.

2. Methods

2.1. Data Collection

One hundred and forty six adults (67 men, 79 women) between ages of 18 and 79 participated in the study. They were recruited through local advertisement. Table 1 presents physical characteristics of the subjects. All participants provided written informed consent prior to the study and all procedures of the study were approved by the local institutional ethics review board. Participants completed walking activities and typical activities of daily living for the study. The energy cost of performing these activities was assessed as well as data from accelerometer worn by them were recorded as described next.

Before performing the activities, resting metabolic rate (RMR) for each participant was assessed under recommended measuring conditions (Compher *et al* 2006) using an open-circuit computerized indirect calorimeter method (ParvoMedics 2400 TrueOne, Sandy UT). During the measurement, each participant was awake in supine position for 30 minutes. The first 10 minutes of data was discarded and a minimum of 10 min of steady-state conditions (<5% coefficient of variation) were used for calculating RMR.

Each participant then performed six treadmill walking activities, 7 min each, 1.5 miles per hour (mph) to 4 mph in 0.5 mph increments. For the purpose of this study, these walking activities were categorized as slow walking (1.5 & 2 mph), medium walking (2.5 & 3 mph) and fast walking (3.5 & 4 mph); to total six walking activities of three categories. After a rest period of 15 min, each participant was then asked to complete five activities of daily living for 7 min each: computer work, vacuuming, mopping/sweeping, carrying/lifting boxes of 3 different weights (4.5, 6.8, and 9.1 kg), and walking/intermittent stair climbing. This resulted in a total of 1,553 activities performed by 146 subjects out of a possible 1,606 (some data were lost due to accelerometer malfunction and/or indirect calorimetry malfunction).

In order to assess energy cost, while doing these activities each participant wore a portable computerized indirect calorimeter metabolic measurement system (Cosmed K4b², Cosmed,

S.r.l., Italy) which has been shown to be valid across a wide range of physiological intensity (McLaughlin *et al* 2001). The breath-by-breath data from this system was averaged over 1-min-periods to derive VO_2 values in $\text{mL O}_2 \cdot \text{kg}^{-1} \cdot \text{min}^{-1}$. The values for min 3-6 for each activity were then averaged and divided by the participant's measured RMR value to derive measured metabolic equivalents (METs) per activity.

While doing the activities each participant also wore an ActiGraph GT3x triaxial accelerometer (ActiGraph LLC, Fort Walton Beach, FL) which is small ($3.8 \times 3.7 \times 1.8$ cm) and lightweight (28 grams). The same accelerometer was used for all participants, and was found to be in compliance with manufacturer's precision pre and post testing. It was secured via a belt on participant's superior iliac spine along the anterior axillary line on non-dominant hip and was initialized to record data in 1-second epochs. For the data analysis, only the combination of individual components along the three axes was used and not the individual components separately.

2.2. Machine Learning Techniques

Several machine learning techniques of different types have been developed, which include statistical, rule-based and probability-based techniques. The following techniques were used in this study which are well-known and are representative of different types of machine learning techniques. All of these have been used previously for PA assessment. The free and publicly available Weka machine learning and data mining software (Hall *et al* 2009) was used for all these techniques.

Support vector Machines (SVMs)—In this statistical technique (Cristianini and Shawe-Taylor 2000), training examples are first mapped to points in a multi-dimensional space of features. A separating plane is then numerically determined that maximizes the margin between the points belonging to different classes. A test example is then classified based on the side of the separating plane where it falls. Although designed for classification, SVMs have been also extended for doing regression. SVMs have sound theoretical properties and have been also shown to work well on tasks involving thousands of features. They can also model non-linearity in the data through kernel functions of higher degrees.

Neural Networks—This is a biologically motivated statistical technique (Haykin 1999). A typical neural network consists of input (features) and output nodes connected via a few hidden nodes akin to biological neurons. The amount of activation of a node depends on the activation of its incoming nodes and the weights of their connections. The weights of these connections are tuned by an iterative procedure using the training data to approximate the relation between inputs and outputs. Thus neural network is essentially a mathematical function approximation technique. It has been used in wide variety of applications for classification as well as for regression tasks. It, however, typically requires large amount of computational time for training.

Logistic Regression—This is a well-known statistical technique (Hosmer and Lemeshow 2004) in which the relationship between the output and the input features is determined in terms of their odds ratios from the training data and this information is then used to predict

probability of a test example's class. Despite its name, it is not a technique for doing regression but only for doing classification. It cannot directly model non-linearity if present in the data.

Decision Trees—This is a rule-based technique in which the relation between input features and output is automatically learned in the form of a decision tree structure (Quinlan 1993). Unlike the previously mentioned techniques, the learned model of this technique is not mathematical and is very human interpretable. Although a classification technique, it has been extended to do regression. Out of several types of decision trees available in Weka, its J48 decision tree implementation was used in this study.

Bagged Decision Trees—Bagging (Breiman 1996) is an ensemble learning method (Dietterich 2000) in which several models are built using the same machine learning technique but using different samples of the training data. The outputs of all the models are then combined through voting (for classification) or by averaging (for regression). This method has been shown to work generally better than an individual model. In this study, Weka's default bagged reduced error pruning (REPTree) decision trees were used.

Random Forest—This is also an ensemble learning method (Dietterich 2000) in which several decision tree models are built by sampling the training data (as in bagging) and by randomly selecting features (Breiman 2001). This technique as implemented in Weka can only do classification but not regression.

Naïve Bayes—This technique is based on probability theory in which probability of output given the input features is computed from the training data as a conditional probability distribution (Lewis 1998). To simplify this computation, it makes the naïve assumption that the features are conditionally independent of each other given the output. In spite of this naïve assumption, this technique has been found to perform competitively on many classification tasks and compared to other techniques requires very little computational time to train and test. It can only do classification not regression.

K-Nearest Neighbors—This is an example of instance-based machine learning technique (Aha and Kibler 1991) in which for a given test example its k nearest training examples (neighbors) are first determined, where k is a parameter. Then the output for the test example is computed by voting (for classification) or averaging (for regression) the outputs of its neighbors. A suitable distance metric is employed to compute the distances between examples.

Linear Regression—This is a well-known statistical technique in which a linear function is fitted using the training data that relates the input features to the output. It cannot be used for classification.

Although hidden Markov model (HMM) (Rabiner and Juang 1986) has been used in some previous work on PA identification (Poher *et al* 2006, Lukowicz *et al* 2004, Lester *et al* 2005), it was not an appropriate model for this study because HMM is a sequence labeling model meant to classify multiple connected entities together, for example, to simultaneously

identify multiple PAs done in a sequence. But in this work only single PAs were to be independently identified. It should be also noted that when used for a single entity classification task, HMM reduces to the naïve Bayes model (Lafferty *et al* 2001).

The results for all the experiments in this study were obtained using the standard 10-fold cross-validation (Japkowicz and Shah 2011). In this procedure, the data is first randomly divided into ten equal parts, a technique is then trained on nine parts and tested on the tenth part, and this process is repeated ten times using a different part for testing each time. The results of the ten folds are then combined for computing the evaluation scores. Some of the machine learning techniques described above have a few important parameters on which their performance depends. The optimal values for all these parameters were determined using 10-fold internal cross-validation within the training data of each fold (separately for classification and regression). In the following, we describe the parameters of different techniques and the range of values out of which their optimal values were searched.

Decision trees as well as bagged decision trees have an important parameter - the minimum number of instances per leaf, which specifies at how many training examples the method should stop further growing the tree at a node. If a very low value (say, one) is specified then the tree will overfit the training data and if a very high value is specified then the tree will underfit the training data (Provost and Fawcett 2013). The optimal value for this parameter was determined through internal cross-validation within each fold among the values from 1, 2, 4, 6, 8, and 10. SVMs have a noise parameter as well as the degree of the polynomial kernel. The optimal noise parameter was searched out of 0.1, 1, 10, 20 and 30, while the degree was searched from 1, 2 and 3. The optimal number of hidden nodes for neural network was searched out of 1, 3, 5 and 10 while its learning rate was searched from values 0.1, 0.3 and 0.5. The parameter k of k -nearest neighbors was searched out of 1, 3, 5 and 7. Logistic regression, naïve Bayes, random forest and linear regression methods do not have important parameters to set. For all the techniques, the less important parameters (e.g., various thresholds) were set to Weka's default values.

2.3. Feature Representations

Most machine learning techniques require training and test examples to be represented in terms of features (also referred to as covariates or independent variables). Features are the only information that a machine learning technique receives about the examples, and hence the performance of a machine learning technique is largely dependent on the feature representation employed. For the PA assessment task, each example is a time series obtained from the accelerometer, i.e. it is a long sequence of numerical values. As such the numbers themselves are not directly representative of the PA hence one needs to design appropriate features. Designing features is a task also known as feature engineering. Most previous work on PA assessment had used summary statistics of time series as the features. In this study, besides those summary statistics, features based on time series distances and discrete values straight from the time series were also used. All these types of features are described below.

Basic Statistical Features (Basic-Stats)—The following summary statistics computed over time series were used as features in a few well-known previous work on PA assessment

(Staudenmayer *et al* 2009, Trost *et al* 2012): 10th, 25th, 50th (median), 75th and 90th percentiles, and lag one autocorrelation. We call these the basic statistical features or Basic-Stats for short.

All Statistical Features (All-Stats)—Zheng *et al* (2013) extended the above feature set by adding to it the following summary statistics computed over time series: sum of values, mean, standard deviation, coefficient of variation, peak-to-peak amplitude, interquartile range, skewness, kurtosis, signal power, log energy, peak intensity and zero crossings. These are standard statistical measures and their definitions are also given in table 1 of (Zheng *et al* 2013). We call these as all statistical features or All-Stats for short.

The above features representations reduce an entire time series to a few summary statistics thus losing various details and intricacies present in the time series. It is possible for two very different time series to exhibit very similar statistical features, hence using these features can sometimes adversely affect performance on time series classification and regression. Hence we also considered the following types of features which are less reductionist.

Distance Based Features (DTW and DDTW)—One simple way of doing time series classification is by classifying a test time series to belong to the same class as its nearest time series in the training data, where nearness is determined using a distance measure. For example, an activity will be classified as “vacuuming” if its distance from a vacuuming activity is shortest out of all the activities present in the training data. Note that this is an instance of the one-nearest neighbor (1NN) technique (i.e. k -nearest neighbor with $k=1$) using a time series distance measure. Several types of distance measures can be defined between two time series. Perhaps the simplest distance measure is the well-known Euclidean distance which is square-root of the sum of the squares of point-to-point differences between the two time series. However, it requires the two series to be of equal lengths, and it is also overly sensitive to small mis-matches between two time series. Figure 1(a) shows the direct point-to-point alignment between two time series that will be considered to compute Euclidean distance between them. Notice that the two time series are only slightly different but the point-to-point alignment will over-estimate the differences between them. This drawback is elegantly overcome by the dynamic time warping (DTW) distance (Berndt and Clifford 1994). As shown in figure 1(b), instead of computing point-to-point differences, DTW allows for some elasticity (warping) and instead computes the differences between the points which are better aligned. This alignment is efficiently determined using a dynamic programming algorithm. DTW distance in combination with 1NN is known to be an excellent method for time series classification which has been found hard to outperform on several time series classification benchmark datasets (Xi *et al* 2006, Xing *et al* 2010, Wang *et al* 2013). However, it has not been widely used for PA identification task using accelerometer time series data.

Recently, Kate (2015) showed that the time series classification performance improves on many datasets if instead of using DTW distances with 1NN technique, they are used as features with a more powerful technique such as SVM. The DTW features for a time series are defined simply as its DTW distances from each of the training examples (Kate 2015);

thus the DTW distance from the first training example becomes the first feature, the DTW distance from the second training example becomes the second feature, and so on. There will be thus as many features as the number of training examples. After computing this feature representation, any of the machine learning techniques can be employed for classification (without such a feature representation only 1NN or kNN could be used with DTW distances).

It has been observed that sometimes the alignment between points obtained by DTW is not intuitive and may have “singularities”, for example, several points of one time series being aligned to the same point in the other time series (Keogh and Pazzani 2001). Figure 1(b) illustrates two such cases. Derivative dynamic time warping (DDTW) (Keogh and Pazzani 2001) overcomes some of these problems of DTW. In DDTW, instead of differences between points, differences between their derivatives are computed which often leads to a more natural alignment, see figure 1(c). DDTW distances in combination with 1NN has been also shown to work better than DTW distances in combination with 1NN on the PA identification task using accelerometer data (Muscillo *et al* 2007). In this study, we define DDTW features for a time series simply as its DDTW distances from each of the training examples, similar to the way DTW features are defined as described earlier. No previous work had used DTW or DDTW as features for PA assessment. On our data, we did not find the parameterized versions of DTW and DDTW (which put a restriction on the amount of warping allowed) perform any better than the unparameterized versions, hence we only included results of the unparameterized versions.

Ngram Based Features—Besides summary statistics and distance based features, it is possible to obtain features straight from a time series which are similar to the ngram features used in text document classification (Joachims 2002). To classify a text document, individual words (called unigrams), contiguous two words (called bigrams), contiguous three words (called trigrams) etc., have been widely used as features which are also collectively called ngrams. Same type of features can be analogously obtained for time series by treating a time series like a text document and its individual data points as words (Lin *et al* 2012). However, while there is a finite vocabulary of words, there are infinite numerical values possible for time series data points. Hence it is necessary to first discretize these values to a finite set before using them as features. In this work, we discretize the values by mapping them to bin numbers, where a chosen number of bins are first created by uniformly dividing the interval from minimum to maximum values (as found over the entire time series data). The number of bins, b , is a parameter. Figure 2 illustrates this with an example.

In figure 2, the chosen number of bins is 10, hence the interval 0.32 (minimum) to 0.92 (maximum) is uniformly divided into 10 bins shown on the top. The data points of the original time series (a) are then mapped to their corresponding bin numbers, thus discretizing the original time series into a series of bin numbers (b). Using these discretized bin numbers, one obtains unigrams (single bin numbers) and bigrams (contiguous two bin numbers) shown in (c) and (d) respectively as squares and ellipses. The values of these ngram features are their frequencies; for example, unigram “8” appears once and unigram “3” appears twice in (b); similarly bigram “8 0” appears once and bigram “3 9” appears twice in (b). A less demanding PA (e.g., computer work) is likely to exhibit high frequencies

for lower bin numbers (e.g., “0”, “1”), while a very active PA (e.g., running) is likely to exhibit high frequencies for higher bin numbers (e.g., “8”, “9”). If a PA requires frequent change in acceleration counts (say jumping from bin “4” to “6”) then the frequency of the corresponding bigram (“4 6”) will be high for that activity. Machine learning techniques will be able to thus use these features to learn to distinguish between different physical activities or to estimate their energy costs.

In this study, the optimal value for the number of bins b was determined through 10-fold internal cross-validation within the training data in each fold out of the values 20, 30, ..., 70. Only unigrams and bigrams were used as features and going beyond that was not found to be useful. These ngram features are similar to the “bag-of-patterns” features (named after “bag-of-words” term used in text classification) used by Lin *et al* (2012) for time series classification. There are two main differences, first, instead of mapping individual points to bins, in their method points within a small window size are first aggregated which are then mapped to bins; and second, instead of uniform distribution of bins their method uses a Gaussian distribution. Note that our accelerometer data has been already aggregated over one second windows hence this, in effect, removes the first difference.

Besides separately using these different types of features, one can also use them in various combinations by simply concatenating their feature vectors. All our code is available on our website http://www.uwm.edu/~katerj/pa_assessment for downloading.

3. Results

3.1. Identifying PA Type

Table 2 shows the 10-fold cross-validation classification accuracies (average of the 10-folds) of the different machine learning techniques on the PA identification task when using the Basic-Stats and All-Stats features as described in the previous section. Instead of using all the features, one can also use a feature selection method to first automatically select only the most informative features. The middle column of the table shows results when feature selection was done over All-Stats features using the Weka’s “AttributeSelectedClassifier” method (Hall 1999). The best accuracy in each column is shown in bold and italics, and the techniques that did not differ significantly from it using the paired two-tailed t-test at $p < 0.05$ are shown in bold. The same statistical significance test was used for all the comparisons in this paper. All statistical significance tests were based on the accuracy values over the individual 10 folds. SVM, neural networks and logistic regression gave the top accuracies in each column. A baseline random classifier that outputs random PA type will have around 14% accuracy on this data, hence all the techniques shown in table 2 worked well and obtained far better accuracy than a random classifier. Each technique also gave better accuracy with All-Stats features than with Basic-Stats features (most were statistically significant at $p < 0.05$). But feature selection did not improve results for most of the classifiers and helped by only small amounts for the others (none was statistically significant at $p < 0.05$).

While the comparison between different machine learning techniques using the same features is shown in table 2, the comparison between different types of features and their

combinations using the same machine learning technique is shown in table 3. SVM was used as the technique to compare the features because it did best in table 2. The best accuracy was obtained by the combination of All-Stats and DDTW features (86.67%) which was also statistically significant ($p < 0.05$) compared to the accuracy obtained using All-Stats features alone (83.98%). DDTW features alone, however, do worse. The combination of All-Stats and DTW also obtained better accuracy than All-Stats features alone ($p < 0.05$). Adding ngram features to All-Stats feature or to the combination of All-Stats and DDTW features did not improve the accuracy. The combination of DTW and ngram features obtained accuracy (85.12%) which was not significantly different from the best accuracy. But ngram features alone obtained a lower accuracy (81.84%). DTW distances directly used in combination with one-nearest neighbor (i.e. not as features) obtained 70.58% accuracy (not shown in table 3), while when used as features with SVM obtained a much higher accuracy of 85.27% (third row of table 3). This is in accordance to what was found to be true in (Kate 2015) for many time series benchmark datasets.

Figure 3 shows the confusion matrix of the model that used only All-Stats features. It can be seen that most of the confusion was between different walking speeds, and between moving boxes, vacuuming, and mop/sweeping. Figure 4 shows the confusion matrix that used the combination of All-Stats and DDTW features which obtained the best accuracy in table 3. It can be observed that the improvement in accuracy is mainly because of some reduction in confusion between the walking speeds and between moving boxes and vacuuming. Thus the added features were able to make finer distinctions between some of these activities.

3.2. Estimating Energy Cost

Table 4 shows biases (mean errors) and root mean square errors (rMSE) obtained by various machine learning techniques on the energy cost estimation task. The energy cost was estimated in metabolic equivalents. Some of the techniques used for PA identification could not be used for energy cost estimation task because they can only do classification and not regression, namely naïve Bayes, logistic regression, random forest (as in Weka) and decision trees (Weka's J48 implementation). However, Weka's implementation of bagged decision trees can do regression which we included for comparison. We also included the linear regression technique which can only do regression and not classification. As can be observed from table 4, bagged decision trees obtained the lowest errors but it was not statistically different ($p < 0.05$) from some other techniques. There is, however, not much difference between errors obtained using Basic-Stats and All-Stats features for most of the techniques (none was statistical significant except for linear regression). Feature selection also did not make much difference.

A very simple baseline for energy cost estimation is to always predict the average energy cost of all the activities in the training data (this is, in fact, equivalent to k -nearest neighbor with k equal to the number of training examples). In our data, this baseline gave 0.00197 bias and 1.7256 rMSE. Hence all the techniques shown in table 4 performed better than this simple baseline (all statistically significant with $p < 0.05$).

Table 5 compares different feature sets using bagged decision trees which performed best in table 4. Lowest error was obtained using All-Stats and ngram feature combination although

most other features types obtained errors not statistically different from this. Hence in this task the choices of machine learning technique and features were not found to be as consequential as they were found to be in the PA identification task.

We also determined energy cost estimation errors on individual activities obtained by bagged decision trees using All-Stats features. This model (as well as all the models shown in tables 4 & 5) was a *joint* model that used all the activities of the training data to learn to estimate energy cost of an activity of any type. We also built separate *activity-specific* models which were trained on each type of PA to predict energy cost of the same type of PA (i.e. one model trained on slow walking activities to predict its energy cost, another model trained on medium walking activities to predict its energy cost, etc.). All these models used bagged decision trees with All-Stats features. Table 6 shows the results of errors obtained on individual activities using the joint model, using the activity-specific models, as well as using the baseline model that simply predicts the average energy cost of that activity computed from the training data. Additionally, table 7 shows the corresponding results when all the walking activities and all the activities of daily living (ADLs) were merged in two separate groups.

Surprisingly, table 6 shows that the baseline of simply predicting average energy cost of an activity, in fact, obtained lower errors than the joint model on most of the activities. However, this is not true in table 7. But it is clear from both the tables that the activity-specific models did better than the joint model on each type of activity.

4. Discussion

In this study, we evaluated several machine learning techniques and features for PA identification and energy cost estimation tasks using accelerometer data.

Comparing machine learning techniques on the PA identification task (table 2), the statistical techniques (SVM, neural networks and logistic regression) performed better than other techniques, although other techniques were not too far behind. We want to point out that one cannot proclaim a particular machine learning technique to be universally better than others even on a specific task such as PA identification. This is because every machine learning technique makes some assumptions (called inductive bias; Mitchell 1980) about the data to enable predictions on unseen data, and it will do very well on a particular dataset if those assumptions indeed hold in the dataset. For example, SVM assumes that the data is linearly separable in the feature space, naïve Bayes assumes that the features are conditionally independent given the output, etc. Hence for a given dataset, it is always best to apply multiple machine learning techniques and empirically determine which works best through cross-validation.

Adding more features was found to significantly improve accuracy on the PA identification. This was found to be true when increasing the number of summary statistics based features (i.e. going from Basic-Stats to All-Stats; table 2), as well as when adding DTW and DDTW based features to All-Stats features (table 3). This makes sense because through more features a machine learning technique gets more information about the examples, and if

those features are indeed indicative of the PA type then the accuracy will improve. However, more features, if not relevant for the classification task, can sometimes be distracting to a machine learning technique leading to overfitting (Provost and Fawcett 2013) and thus a drop in accuracy. Hence feature selection is often employed to narrow down the features to only the most relevant ones for the task. But in this study feature selection did not help when performed on the All-Stats features (table 2). It should be also noted that SVM is a technique known to be particularly resistant to overfitting even in the presence of many features (Cristianini and Shawe-Taylor 2000). Hence it was particularly suitable for comparing different types of features and their combinations in this study because there were a few hundred DTW, DDTW and ngram features in our dataset.

Although DTW, DDTW and ngram feature combinations improved results over All-Stats features (table 3), it may be noted that All-Stats features were only 2 to 3 percentage points behind. This shows that in this dataset PA identification can be largely done with summary statistics alone (83.98%). Adding more features that capture finer details of time series are beneficial but only to a small extent. To look into this further, in figure 5 we plotted average of the Basic-Stats “mean” feature over all the activities of the same type. This feature is essentially the average accelerometer signal magnitude over an activity and we are plotting average of these averages for each PA type. As can be seen from the figure, this feature alone is sufficient to distinguish between many of the activities. Hence in this dataset, the types of physical activities were distinct enough that summary statistics were largely effective in identifying them. However, we expect that when the activities are such that they would generate similar summary statistics, then DTW, DDTW and ngram features will improve accuracy by larger amounts by helping to distinguish finer differences between the accelerometer time series generated by different activities. Of note, DDTW features alone give poor accuracy (65.77%) but give the best accuracy (86.67%) when combined with All-Stats features. This shows that only including the features that can make fine distinctions is not sufficient. For a given dataset, it is recommended to try multiple types of features and their combinations and empirically determine which gives the best accuracy.

On the energy cost estimation task, as noted before, the choice of the machine learning technique or the features did not affect the results by much (tables 4 & 5). This indicates that a few summary statistics gathered in the form of features from time series are sufficient for estimating energy cost. However, several interesting things can be noted from table 6 which shows activity-wise results as well as results of baseline and activity-specific models. Activity-wise models clearly outperformed the joint model. Similar results were obtained by Rumo *et al* (2011) in which regression formulas were used for energy cost estimation. Most of the biases obtained by the joint model are significantly different from zero in table 6 (in tables 4 & 5 when one bias was computed for all the activities, it was not significantly different from zero). The joint model is clearly over-estimating the energy costs of slow and medium walking while under-estimating the energy cost of fast walking. Similarly, it is over-estimating for computer, vacuuming and moving boxes while under-estimating for mop/sweeping and walk/stairs. On the other hand, the baseline and activity-specific models, which were trained and tested on each PA type separately, had no bias significantly different from zero.

This indicates, not surprisingly, that different types of physical activities have different associated energy costs and there is less variation in energy cost within the same PA. This is also depicted in figure 6 which shows average energy costs of different activities. Thus estimating energy cost for one type of PA is easier than estimating it for several types of PAs using the same model. Hence on activity-wise comparison, even a baseline model that estimates energy cost of any test example to be simply the average energy cost of all the training examples does better than the joint model. For future studies, we recommend using such a baseline model for comparison. However, activity-specific trained models did significantly better than the baseline model for most of the activities. It may be noted that root mean square errors are lower for activities that have low energy costs (e.g., computer) and higher for activities that have high energy costs (e.g., fast walking). This is not surprising because there is less scope of variation in energy cost when it is low and more scope of variation when it is high.

Table 7 shows that the baseline model does significantly worse than joint model when activities are grouped as walking and ADLs. In this case also activity-specific model obtained lower errors than the joint model, however, the differences were not significant. The results shown in tables 6 and 7 are important for designing frameworks for estimating energy costs of PA in practice. They indicate that if the energy costs are to be estimated for a particular type of PA then it is better to build a model specific for that PA rather than using a joint model trained for estimating costs of multiple types of PAs even if that includes the desired PA. This is true even when it is not for a specific type of PA but for a group of PAs.

Given the results of Table 6, one may wonder whether one can improve energy cost estimation by first identifying the PA type and then applying the appropriate activity-specific energy cost model. We tried this on our dataset, but unlike some previous work (Bonomi *et al* 2009, Albinali *et al* 2010), this two-step model for energy cost estimation did not do better than the joint (all activity) energy cost estimation model. This was primarily because the PA identification model made most of the errors exactly where distinguishing between them would have helped in estimating energy cost, for example, between medium walking and fast walking (see confusion matrix in figure 4). Note that our joint energy cost estimation model got the same training data in the form of input features as the PA identification model. Hence it is possible that the joint energy cost estimation model already implicitly learned what the PA identification model could have helped in energy cost estimation, thus making PA identification as a first step needless. In contrast, previous work that reported benefit of PA identification as a first step for energy cost estimation had either used different sets of features for the two tasks (Bonomi *et al* 2009) or had used machine learning technique for PA identification and regression formulas for energy cost estimation (Albinali *et al* 2010). Hence their PA identification models could learn something to help energy cost estimation that their energy cost estimation models could not have directly learned.

We point out a few limitations of the current study. Although it evaluated multiple types of features, it did not evaluate wavelet based features (Preece *et al* 2009). It used two ensemble methods - bagged decision trees and random forest, both being ensembles of decision trees, but it is also possible to create an ensemble of different machine learning techniques (Dietterich 2000) which was not done in the current study. Another limitation is that only a

limited number of activities were used under laboratory experimental conditions. It has been shown that classifiers trained on such activities perform poor under real life conditions (van Hees et al 2013). Hence in future, a wider range of activities performed in real-world scenarios should be considered. The data used in the current study was obtained from a unilateral monitoring site on the hip. Future studies would benefit from bi-lateral (e.g., left and right) monitor placements as well as by using multiple sites. In the current study, we used count based approaches for accelerometers which limit cross-device comparability. In future, use of non-proprietary acceleration units will allow wider comparability.

In conclusion, the current study compared multiple machine learning techniques and multiple types of features and their combinations for identifying PA and estimating their energy cost using accelerometer data. On the PA identification task, using more summary statistics features led to better accuracies than using fewer summary statistics features. Adding distance based features further improved the accuracy. All machine learning techniques were competitive on this task but statistical techniques of SVM, neural networks and logistic regression gave the best accuracies. On the energy cost estimation task, the performance was not found to be too dependent on different types of features or techniques. Models trained for estimating energy costs of specific activities performed better than a joint model trained for estimating energy cost of all the activities. Future research design would benefit from narrowing energy cost estimation models to specific activities whenever possible.

Acknowledgments

This work was partially supported by grant R01-HL091019, and UL1RR031973 from the Clinical and Translational Science Award (CTSA) program of the National Center for Research Resources and the National Center for Advancing Translational Sciences. The authors would like to thank Nora E. Miller, Teresa L. Hart, Elizabeth E. Lenz, Jason Jones, and Aubri Rote for assistance with data collection.

References

- Aha D, Kibler D. Instance-based learning algorithms. *Machine Learning*. 1991; 6:37–66.
- Albinali, F.; Intille, SS.; Haskell, W.; Rosenberger, M. Using wearable activity type detection to improve physical activity energy expenditure estimation. 12th ACM international conference on Ubiquitous computing (UbiComp'10); Copenhagen, Denmark. 2010. p. 311–320.
- Bao, L.; Intille, SS. *Pervasive Computing (Lecture Notes in Computer Science. Vol. 3001)*. Berlin: Springer; 2004. Activity recognition from user-annotated acceleration data; p. 1–17.
- Barengo NC, Hu G, Lakka TA, Pekkarinen H, Nissinen A, Tuomilehto J. Low physical activity as a predictor for total and cardiovascular disease mortality in middle-aged men and women in Finland. *Eur Heart J*. 2004; 25:2204–11. [PubMed: 15589637]
- Bonomi AG, Plasqui G, Goris AHC, Westerterp KR. Improving assessment of daily energy expenditure by identifying types of physical activity with a single accelerometer. *Journal of Applied Physiology*. 2009; 107(3):655–661. [PubMed: 19556460]
- Berndt DJ, Clifford J. Using dynamic time warping to find patterns in time series. *Knowledge Discovery and Data Mining Workshop*. 1994; 10(16):359–370.
- Breiman L. Bagging predictors. *Machine Learning*. 1996; 24(2):123–140.
- Breiman L. Random Forests. *Machine Learning*. 2001; 45(1):5–32.
- Compher C, Frankenfield D, Keim N, Roth-Yousey L. Best practice methods to apply to measurement of resting metabolic rate in adults: a systematic review. *J Am Diet Assoc*. 2006; 106:881–903. [PubMed: 16720129]

- Cristianini, N.; Shawe-Taylor, J. An introduction to support vector machines and other kernel-based learning methods. Cambridge: Cambridge University Press; 2000.
- Crouter SE, Clowers KG, Bassett DR Jr . A novel method for using accelerometer data to predict energy expenditur. *J Appl Physiol.* 2006; 100:1324–1331. [PubMed: 16322367]
- Dietterich, TG. Multiple classifier systems. Springer; Berlin Heidelberg: 2000. Ensemble methods in machine learning; p. 1-15.
- Ermes M, Parkka J, Mantjarvi J, Korhonen I. Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions. *IEEE Trans Inf Tech Biomed.* 2008; 12:20–6.
- Hall, MA. Doctoral Dissertation. The University of Waikato; Hamilton, New Zealand: 1999. Correlation-based Feature Subset Selection for Machine Learning.
- Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter.* 2009; 11(1):10–18.
- Haykin, S. *Neural Networks; a Comprehensive Foundation.* Englewood Cliffs, NJ: Prentice-Hall; 1999.
- Hendelman D, Miller K, Baggett C, Debold E, Freedson P. Validity of accelerometry for the assessment of moderate intensity physical activity in the field. *Med Sci Sports Exerc.* 2000; 32:S442–449. [PubMed: 10993413]
- Hirsch CH, Sommers L, Olsen A, Mullen L, Winograd CH. The natural-history of functional morbidity in hospitalized older patients. *J Am Geriatr Soc.* 1990; 38:1296–303. [PubMed: 2123911]
- Hoenig HM, Rubenstein LZ. Hospital-associated deconditioning and dysfunction. *J Am Geriatr Soc.* 1991; 39:220–2. [PubMed: 1991956]
- Hosmer, DW., Jr; Lemeshow, S. *Applied logistic regression.* Hoboken, NJ: John Wiley & Sons; 2004.
- Japkowicz, N.; Shah, M. *Evaluating learning algorithms: a classification perspective.* Cambridge University Press; 2011. p. 172
- Joachims, T. *Learning to classify text using support vector machines: Methods, theory and algorithms.* Kluwer Academic Publishers; 2002.
- Kate RJ. Using dynamic time warping distances as features for improved time series classification. *Data Mining and Knowledge Discovery.* 2015; online first. doi: 10.1007/s10618-015-0418-x
- Keogh EJ, Pazzani MJ. Derivative Dynamic Time. *Warping First SIAM International Conference on Data Mining (Chicago, IL, USA).* 2001; 1:5–7.
- Lafferty, J.; McCallum, A.; Pereira, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *18th International Conference on Machine Learning; Williamstown, MA, USA.* 2001. p. 282-289.
- Lester, J.; Choudhury, T.; Kern, N.; Borriello, G.; Hannaford, B. A hybrid discriminative/generative approach for modeling human activities. *19th International Joint Conference on Artificial Intelligence; Edinburgh, UK.* 2005. p. 766-72.
- Lewis, DD. Naive (Bayes) at forty: The independence assumption in information retrieval. *10th European Conference on Machine Learning; Chemnitz, Germany.* 1998. p. 4-15.
- Lin J, Khade R, Li Y. Rotation-invariant similarity in time series using bag-of-patterns representation. *Journal of Intelligent Information Systems.* 2012; 39(2):287–315.
- Lukowicz, P.; Ward, J.; Junker, H.; Stager, M.; Troster, G.; Atrash, A.; Starner, T. *Pervasive Computing (Lecture Notes in Computer Science. Vol. 3001. Berlin: Springer; 2004. Reconizing workshop activity using body worn microphones and accelerometers; p. 18-32.*
- Lyden K, Kozey SL, Staudenmeyer JW, Freedson PS. A comprehensive evaluation of commonly used accelerometer energy expenditure and MET prediction equations. *Eur J Appl Physiol.* 2011; 111:187–201. [PubMed: 20842375]
- Matthews CE. Calibration of accelerometer output for adults. *Med Sci Sports Exerc.* 2005; 37:S512–522. [PubMed: 16294114]
- McLaughlin JE, King GA, Howley ET, Bassett DR, Ainsworth BE. Validation of the Cosmed K4b2 portable metabolic system. *Int J Sports Med.* 2001; 22:280–283. [PubMed: 11414671]
- Mitchell, TM. *The need for biases in learning generalizations.* Department of Computer Science, Laboratory for Computer Science Research, Rutgers University; 1980. p. 184-191.

- Muscillo, R.; Conforto, S.; Schmid, M.; Caselli, P.; D'Alessio, T. Classification of motor activities through derivative dynamic time warping applied on accelerometer data. 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society; Lyon, France. 2007. p. 4930-4933.
- Quinlan, R. C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann Publishers; 1993.
- Parkka J, Ermes M, Korpipaa P, Mantyjarvi J, Peltola J, Korhonen I. Activity classification using realistic data from wearable sensors. *IEEE Trans Inf Technol Biomed*. 2006; 10:119–28. [PubMed: 16445257]
- Preece SJ, Goulermas JY, Kenney LP, Howard D, Meijer K, Crompton R. Activity identification using body-mounted sensors—a review of classification techniques. *Physiological measurement*. 2009; 30(4):R1–33. [PubMed: 19342767]
- Pirttikangas, P.; Fujinami, K.; Nakajima, T. Ubiquitous Computing Systems (Lecture Notes in Computer Science. Vol. 4239. Berlin: Springer; 2006. Feature selection and activity recognition from wearable sensors; p. 516-27.
- Pober DM, Staudenmayer J, Raphael C, Freedson PS. Development of novel techniques to classify physical activity mode using accelerometers. *Med Sci Sports Exerc*. 2006; 38:1626–34. [PubMed: 16960524]
- Provost, F.; Fawcett, T. Data Science for Business: What you need to know about data mining and data-analytic thinking. Vol. Chapter 5. Sebastopol, CA: O'Reilly Media, Inc; 2013. Overfitting and its Avoidance.
- Rabiner LR, Juang BH. An introduction to hidden Markov models. *ASSP Magazine, IEEE*. 1986; 3(1): 4–16.
- Rumo M, Amft O, Troster G, Mader U. A stepwise validation of a wearable system for estimating energy expenditure in field-based research. *Physiological Measurement*. 2011; 32(12):1983–2001. [PubMed: 22056999]
- Staudenmayer J, Pober D, Crouter S, Bassett D, Freedson P. An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer. *Journal of Applied Physiology*. 2009; 107:1300–1307. [PubMed: 19644028]
- Strath SJ, Bassett DR Jr, Swartz AM. Comparison of MTI accelerometer cut-points for predicting time spent in physical activity. *Int J Sports Med*. 2003; 24:298–303. [PubMed: 12784173]
- Swartz AM, Strath SJ, Bassett DR Jr, O'Brien WL, King GA, Ainsworth BE. Estimation of energy expenditure using CSA accelerometers at hip and wrist sites. *Med Sci Sports Exerc*. 2000; 32:S450–456. [PubMed: 10993414]
- Trost SG, Wong WK, Pfeiffer KA, Zheng Y. Artificial neural networks to predict activity type and energy expenditure in youth. *Med Sci Sports Exerc*. 2012; 44(9):1801–1808. [PubMed: 22525766]
- van Hees VT, Golubic R, Ekelund U, Brage S. Impact of study design on development and evaluation of an activity type classifier. *Journal of Applied Physiology*. 2013; 114(8):1042–1051. [PubMed: 23429872]
- van Sluijs EM, McMin AM, Griffin SJ. Effectiveness of interventions to promote physical activity in children and adolescents: systematic review of controlled trials. *Br Med J*. 2007; 335:703. [PubMed: 17884863]
- Wang X, Mueen A, Ding H, Trajcevski G, Scheuermann P, Keogh E. Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery*. 2013; 26(2):275–309.
- Ward D. Objective monitoring of physical activity: Closing the gaps in the science of accelerometry. *Med Sci Sports Exerc*. 2005; 37:S487–S588. [PubMed: 16294111]
- Yancey AK, Wold CM, McCarthy WJ, Weber MD, Lee B, Simon PA, Fielding JE. Physical inactivity and overweight among Los Angeles County adults. *Am J Prev Med*. 2004; 27:146–52.
- Xi, X.; Keogh, E.; Shelton, C.; Wei, L.; Ratanamahatana, CA. Fast time series classification using numerosity reduction. 23rd international conference on Machine learning; Pittsburgh, PA, USA. 2006. p. 1033-1040.
- Xing Z, Pei J, Keogh E. A brief survey on sequence classification. *ACM SIGKDD Explorations Newsletter*. 2010; 12(1):40–48.

- Zhang S, Rowlands AV, Murray P, Hurst T. Physical activity classification using GENEa wrist-worn accelerometer. *Med Sci Sports Exerc.* 2012; 44(4):742–748. [PubMed: 21988935]
- Zheng, Y.; Wong, WK.; Guan, X.; Trost, S. Physical activity recognition from accelerometer data using a multi-scale ensemble method. 25th Innovative Applications of Artificial Intelligence Conference; Bellevue, WA, USA. 2013. p. 1575-1581.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1.

Illustrative sketches showing alignments between points of two time series that will be considered for computing (a) Euclidean distance, (b) Dynamic Time Warping (DTW) distance, and (c) Derivative Dynamic Time Warping (DDTW) distance.

Figure 2. An illustration ngram features for a time series. (a) An original time series. The interval from its minimum (0.32) to maximum (0.92) values is uniformly divided into 10 bins as shown on the top. (b) The bin numbers corresponding to the time series points. (c) Unigram features (squares) and their feature values. (d) Bigram features (ellipses) and their feature values.

a	b	c	d	e	f	g	h	<-- classified as
262	27	1	0	0	0	1	0	a = Slow Walking
30	215	34	0	1	0	1	4	b = Medium Walking
2	44	206	0	1	0	0	2	c = Fast Walking
0	0	0	121	0	5	13	4	d = Moving Boxes
0	0	0	1	144	0	1	0	e = Computer
0	0	0	7	0	117	20	0	f = Mop/Sweeping
0	0	0	23	1	19	102	0	g = Vacuuming
0	0	0	3	1	2	1	137	h = Walk/Stairs

Figure 3.
Confusion matrix obtained by SVM using only All-Stats features on the physical activity identification task.

	a	b	c	d	e	f	g	h	<-- classified as
266	22	1	1	0	0	1	0	0	a = Slow Walking
24	227	30	0	1	0	1	2	2	b = Medium Walking
2	38	214	0	1	0	0	0	0	c = Fast Walking
0	0	0	125	0	7	8	3	3	d = Moving Boxes
0	0	0	0	144	0	2	0	0	e = Computer
0	0	0	4	0	118	21	1	1	f = Mop/Sweeping
0	0	0	8	2	19	114	2	2	g = Vacuuming
1	0	0	4	1	0	0	138	1	h = Walk/Stairs

Figure 4.

Confusion matrix obtained by SVM using All-Stats and DDTW features which obtained the best accuracy on the physical activity identification task.

Figure 5. Accelerometer signal magnitude averages for different types of physical activities (mean and one standard deviation).

Figure 6.
Energy costs of different physical activities (mean and one standard deviation).

Table 1

Participant demographics (means and standard deviations).

Variable	Men (<i>n</i> =67)	Women (<i>n</i> =79)	All (<i>n</i> =146)
Age (yrs)	47.52±18.56	50.47±16.55	49.12±17.37
Mass (kg)	84.99±15.44	67.20±12.69	75.36±16.59
Height (m)	1.78±0.07	1.65±0.06	1.70±0.09
BMI (kg/m ²)	26.8±4.1	24.8±4.4	25.7±4.4
Average Gait (m/s ²)	1.38±0.17	1.38±0.2	1.38±0.19

Table 2

Physical activity identification accuracy obtained by various machine learning techniques when using basic and all summary statistics features.

Machine Learning Technique	Basic-Stats Features	All-Stats with Feature Selection	All-Stats Features
SVM	80.96	83.99⁺	83.98⁺
Neural networks	82.56	83.45	83.25
Logistic Regression	81.38	83.14⁺	82.63
Decision Trees	77.12	78.89	79.57 ⁺
Bagged Decision Trees	79.42	81.35	82.16 ⁺
Random Forest	78.55	79.56	80.24 ⁺
Naïve Bayes	72.46	79.70 ⁺	78.52 ⁺
K-nearest neighbors	78.22	83.11⁺	82.08 ⁺

The best accuracy in each column is shown in bold and italics, and the techniques in the same column that did not differ significantly ($p < 0.05$) from it are shown in bold. Accuracies significantly ($p < 0.05$) different from Basic-Stats accuracy in the same row are indicated with a ⁺ symbol. There was no significant ($p < 0.05$) difference between All-Stats accuracies with and without feature selection in each row.

Table 3

Physical activity identification accuracy using various types of features and their combinations obtained by SVM.

Feature Set	Accuracy
Basic-Stats	80.96
All-Stats	83.98
DTW	85.27
DDTW	65.77
DTW + DDTW	85.50
All-Stats + DTW	85.96
All-Stats + DDTW	<i>86.67</i>
All-Stats + DDTW + DTW	86.28
Ngram	81.84
All-Stats + Ngram	83.18
DTW + Ngram	85.12
All-Stats + DDTW + Ngram	85.38

The best accuracy is shown in bold and italics; those that did not differ significantly ($p < 0.05$) from it are shown in bold.

Table 4

Biases and root mean square errors obtained by different machine learning techniques on the energy cost estimation task using basic and all statistical features.

Machine Learning Technique	Bias Basic-Stats	Bias All-Stats with Feature Selection	Bias All-Stats	rMSE Basic-Stats	rMSE All-Stats-with Feature Selection	rMSE All-Stats
SVM	-0.0568	-0.0440	-0.0336	1.02321	0.9933	1.0234
Neural networks	0.1782 *	0.2011 *	0.2200 *	1.00215	1.0168	1.0090
Bagged decision trees	0.0007	0.0075	0.0039	0.97785	0.9662	0.9652
K-nearest neighbors	0.0083	0.0006	0.0133	1.0038	0.9971	0.9899
Linear regression	0.0038	0.0055	0.0029	1.01446	0.9702 †	0.9726 †

The lowest rMSE in each column is shown in bold and italics, and the techniques in the same column that did not differ significantly ($p < 0.05$) from it are shown in bold. Other symbols:

* bias is significantly different from zero ($p < 0.05$),

† rMSE is significantly different from Basic-Stats rMSE ($p < 0.05$) in the same row.

There was no significant ($p < 0.05$) difference between All-Stats rMSE with and without feature selection in each row.

Table 5

Biases and root mean square errors obtained by bagged decision trees on the energy cost estimation task using various types of features and their combinations.

Feature Set	Bias	rMSE
Basic-Stats	0.0007	0.9779
All-Stats	0.0039	0.9652
DTW	−0.0042	0.9895
DDTW	−0.0083	1.3461
DTW + DDTW	0.0084	0.9810
All-Stats + DTW	0.0047	0.9731
All-Stats + DDTW	−0.0019	0.9728
All-Stats + DDTW + DTW	0.0094	0.9696
Ngram	−0.0086	0.9862
All-Stats + Ngram	0.0050	0.9633

No bias was significantly different from zero ($p < 0.05$). The lowest rMSE is shown in bold and italics, and those that did not differ significantly ($p < 0.05$) from it are shown in bold.

Table 6

Activity-wise biases and root mean square errors obtained by the baseline model of average energy, by the joint model, and by the activity-specific models.

Activity	Bias Baseline Model	Bias Joint Model	Bias Activity-Specific Model	rMSE Baseline Model	rMSE Joint Model	rMSE Activity-Specific Model
Slow Walking	0.0013	0.1316 [*]	-0.0124	0.7344	0.7443	0.7019 ⁺
Medium Walking	0.0054	0.1520	-0.0071	0.9335	1.0455	0.9213
Fast Walking	0.0079	-0.3706 [*]	0.0174	1.3873	1.4370	1.2904 ⁺ [≠]
Computer	-0.0008	0.1225 [*]	0.0025	0.2042	0.2961	0.2061
Vacuuming	0.0028	0.2146 [*]	0.0147	0.7560	0.7475	0.6751 ⁺
Mop/Sweep	0.0012	-0.1110	0.0337	0.8205	0.8211	0.7990
Moving Boxes	-0.0001	0.2842 [*]	0.0272	0.7360	0.7665	0.6750 ⁺ [≠]
Walk/Stairs	-0.0001	-0.3691 [*]	0.0397	0.9126	0.9634	0.8502 ⁺ [≠]

All the models used bagged decision trees with All-Stats features. Symbols:

^{*} bias is significantly different from zero ($p < 0.05$),

⁺ rMSE is significantly different from rMSE of baseline model ($p < 0.05$) in the same row,

[≠] rMSE is significantly different from rMSE of joint model in the same row.

Table 7

Biases and root mean square errors obtained by the baseline model of average energy, by the joint model, and by the activity-specific models on two groups of activities. All the models used bagged decision trees with All-Stats features.

Activity	Bias Baseline Model	Bias Joint Model	Bias Activity-Specific Model	rMSE Baseline Model	rMSE Joint Model	rMSE Activity-Specific Model
Walking	0.0038	-0.0170	0.0131	1.6356	1.1024 ⁺	1.0925 ⁺
ADLs	0.0004	0.0284	-0.0165	1.6108	0.7686 ⁺	0.7456

No bias was significantly different from zero ($p<0.05$). Symbols ⁺ indicates that rMSE is significantly ($p<0.05$) different from the rMSE of Basic-Stats in the same row. There was no significant ($p<0.05$) difference between rMSE of the joint model and the activity-specific model in each row.