Ensemble Methods for Classification of Physical Activities from Wrist Accelerometry

ALOK KUMAR CHOWDHURY[1], DIAN TJONDRONEGORO[1], VINOD CHANDRAN[1], and STEWART G. TROST[2]

[1]Science and Engineering Faculty, Queensland University of Technology, Brisbane, AUSTRALIA; and [2]Institute of Health and Biomedical Innovation at QLD Centre for Children's Health Research, School of Exercise and Nutrition Sciences, Queensland University of Technology, Brisbane, AUSTRALIA

ABSTRACT

CHOWDHURY, A. K., D. TJONDRONEGORO, V. CHANDRAN, and S. G. TROST. Ensemble Methods for Classification of Physical Activities from Wrist Accelerometry. *Med. Sci. Sports Exerc.*, Vol. 49, No. 9, pp. 1965–1973, 2017. **Purpose**: To investigate whether the use of ensemble learning algorithms improve physical activity recognition accuracy compared to the single classifier algorithms, and to compare the classification accuracy achieved by three conventional ensemble machine learning methods (bagging, boosting, random forest) and a custom ensemble model comprising four algorithms commonly used for activity recognition (binary decision tree, k nearest neighbor, support vector machine, and neural network). **Methods**: The study used three independent data sets that included wrist-worn accelerometer data. For each data set, a four-step classification framework consisting of data preprocessing, feature extraction, normalization and feature selection, and classifier training and testing was implemented. For the custom ensemble, decisions from the single classifiers were aggregated using three decision fusion methods: weighted majority vote, naïve Bayes combination, and behavior knowledge space combination. Classifiers were cross-validated using leave-one subject out cross-validation and compared on the basis of average F1 scores. **Results**: In all three data sets, ensemble learning methods consistently outperformed the individual classifiers. Among the conventional ensemble methods, random forest models provided consistently high activity recognition; however, the custom ensemble model using weighted majority voting demonstrated the highest classification accuracy in two of the three data sets. **Conclusions**: Combining multiple individual classifiers using conventional or custom ensemble learning methods can improve activity recognition accuracy from wrist-worn accelerometer data. **Key Words:** MOTION SENSORS, MACHINE LEARNING, PATTERN RECOGNITION, RANDOM FOREST, BAGGING, BOOSTED DECISION TREES

Physical inactivity is recognized as a critical population health risk factor, and a significant contributor to the direct and indirect health care costs associated with management of a wide range of chronic health conditions (14). In addition, there is a growing body of evidence to suggest that, sedentary behavior, characterised by prolonged bouts of sitting, is associated with serious health conditions, independent of the effects of physical activity (PA) (20,35). Hence, valid and reliable measures of PA and sedentary behavior are a necessity in studies designed to: 1) document the frequency and distribution of PA and sedentary behavior in defined population groups, 2) identify the psychosocial and environmental factors that influence PA and sedentary behavior, and 3) evaluate the efficacy or effectiveness of programs and policies to increase habitual PA and reduce sedentary behavior (36).

Accelerometer-based motion sensors have become the method of choice for measuring PA and sedentary time in free-living contexts, because they are small, robust, and low-cost (18,37). However, differences in accelerometer data processing methods have hindered research efforts to quantify, understand, and intervene on PA and sedentary behavior. Existing approaches can be categorized into two groups: 1) threshold-based and 2) machine learning approaches. Threshold or "cutpoint" methods use regression methods to map accelerometer outputs to energy expenditure (3,18,37). Machine learning approaches extract features or patterns from the acceleration data and use supervised or unsupervised

Address for correspondence: Stewart G. Trost, Ph.D., Institute of Health and Biomedical Innovation at QLD Centre for Children's Health Research, Level 6, 62 Graham Street, South Brisbane, QLD 4101, Australia; E-mail: s.trost@qut.edu.au.

Submitted for publication October 2016.
Accepted for publication March 2017.
Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.acsm-msse.org).

0195-9131/17/4909-1965/0
MEDICINE & SCIENCE IN SPORTS & EXERCISE®
Copyright © 2017 by the American College of Sports Medicine

DOI: 10.1249/MSS.0000000000001291

learning algorithms to predict PA type and/or energy expenditure (19,32–34). Relative to cutpoint methods, machine learning approaches can provide a greater variety of PA metrics (e.g., activity type, walking speed) and more accurate predictions of energy cost (9,33,38). Nevertheless, the adoption of machine learning methods in PA studies has been low because they are not as easily implemented as threshold-based methods.

To date, a range of machine learning algorithms have been used in the PA classification and measurement domain. They include k nearest neighbor (kNN), artificial neural networks (ANN), support vector machines (SVM), Markov models, decision trees, and so on. Preece et al. (24) summarized the relative strengths, weaknesses, and performance characteristics of 11 different machine learning approaches. The authors concluded that it was impossible to declare one particular machine learning technique as universally better than others for any given PA recognition problem. Most recently, Kate and colleagues (15) compared the accuracy of eight different machine learning techniques for activity recognition and energy cost estimation from accelerometer data. Their results indicated that no single machine learning technique works best in all testing situations.

Because there is no single, optimal machine learning algorithm for any given classification or estimation problem, ensemble learning approaches are gaining popularity (6). An ensemble of classifiers is a set of base level classifiers (known as weak learners) whose individual decisions are combined to improve overall decision accuracy. In this respect, an ensemble of machine learning models can be conceptualized as a committee of experts brought together to make a final decision. If the weak learners are combined appropriately, the fusion of outputs is constructive, leading to better overall decisions and generalization.

Three commonly used ensemble learning schemes are *bagging*, *boosting*, and *random forests*. Bagging stands for *bootstrap aggregation*. It involves taking multiple random samples of training instances (with replacement) and applying a weak learning algorithm (typically a decision tree) to the data. The decisions of each classifier are combined to make a final class prediction using the majority vote rule (4). Boosting also applies a voting procedure to combine the decisions of multiple weak learners. However, boosting adopts an iterative approach in which each new model is influenced by the performance of previously built models. The boosting algorithm begins by assigning equal weights to all instances in the training data. It then builds a classifier (typically a decision tree), and instances are reweighted based on the classifiers performance on the training data. The weights of correctly classified instances are decreased, whereas the weights of misclassified instances are increased. This weighting scheme allows subsequent classifiers to be more proficient at classifying instances misclassified by earlier models. The final class prediction is based on the weighted majority vote of each model, where the weights are determined by the accuracy of the model (31). Random Forests are another widely used ensemble learning method. The random forests algorithm is similar to bagging in that multiple weak learners (decision trees) are trained on randomly sampled instances from the training data. However, unlike bagging, where all features in the training data are considered for splitting a node, the random forest algorithm selects the best among a random sample of features. The decisions generated by each tree are recorded, and the final class prediction is based on majority vote (5).

Although bagging and boosting combine base learners of the same type, it is possible to construct custom ensembles featuring learning algorithms of different types. The decisions of each classifier are subsequently combined using an established decision fusion method such as weighted majority voting (WMV), naive Bayes (NB), behavior knowledge space (BKS), and so on. (17). Unlike conventional ensembles, custom ensemble methods achieve diversity by using heterogeneous classification algorithms as base classifiers, which may lead to better generalized performance (39).

Although ensemble learning methods are starting to emerge in PA research, no previous studies in the exercise and movement sciences have compared the performance of different ensemble methods and decision fusion rules. Therefore, the purpose of this study was to systematically compare the classification accuracy achieved by conventional ensemble methods (bagged decision tree, boosted decision tree, and random forest) and a custom multiclassifier ensemble combining four machine learning algorithms (binary decision tree [BDT], kNN, SVM, and neural network) using three decision fusion rules (WMV, NB, and BKS). Performance was evaluated in three independent PA recognition data sets.

METHODS

Data Sets

This study used three independent accelerometer data sets collected from different participant groups (adults and children), performing different activity in different contexts (laboratory and outdoor). A brief description of each data set is provided in Table 1.

Data set 1. The PAMAP2 data set is a fully annotated, publicly available PA monitoring data set. The data was downloaded from the UCI machine learning repository (https://archive.ics.uci.edu/ml/datasets/PAMAP2+Physical+Activity+Monitoring). Detailed information about the study can be found elsewhere (1,26). Nine participants (1 girl, 8 boys; age, 27.2 ± 3.3 yr; body mass index [BMI], 25.1 ± 2.6 kg·m^{-2}) performed 12 different types of physical activities. Participants wore three Colibri wireless inertial measurement units (IMU) on their dominant-arm wrist, dominant-side ankle, and chest. Each IMU contained two 3D acceleration sensor (scale: ±6g and ±16g) with a resolution of 13 bits, a 3D gyroscope sensor, a 3D magnetometer sensor, temperature, orientation and HR monitor sensors. The sampling rate of accelerometer and HR sensor was approximately 100 Hz and 9 Hz, respectively. Only 3D accelerometer (±16g) data of wrist

TABLE 1. Comparison across three data sets.

	Data Set 1	Data Set 2	Data Set 3
Data collection environment	PAMAP2 is a public data set collected in a laboratory	Private data set collected outdoor	Private data set collected in a laboratory
Participants	9 Adult participants (1 female, 8 male), Age: 27.2 ± 3.3 yr, BMI: 25.1 ± 2.6 kg·m^{-2}	8 Adult participants (4 women and 4 men), Age: 29.9 ± 4.2 yr, BMI: 22.8 ± 1.9 kg·m^{-2}	17 Children (9 boys, 8 girls), Age: 14.6 ± 2.4 yr BMI percentile: 66.8 ± 25.9
Sensors and placements	Colibri wireless IMU contains 3D accelerometer, 3D gyroscope sensor, 3D magnetometer sensor, temperature, orientation and HR monitor sensors Placements: dominant-arm wrist, dominant-side ankle and chest	Empatica E4 contains 3D accelerometer, electrodermal activity, HR and temperature sensors Placements: nondominant wrist	ActiGraph GT3X+ triaxial accelerometer Placements: right hip and non-dominant wrist
Accelerometer sensor specifics	Scale: ±6*g* and ±16*g* Sampling rate: 100 Hz	Scale: ±2*g* Sampling rate: 32 Hz	Scale: ±6*g* Sampling rate: 30 Hz
Physical activities performed	Lying down, sitting, standing, walking, running, cycling, ascending stairs, descending stairs, Nordic walking, vacuum cleaning, ironing clothes and jumping rope	sit or stand still, self-paced comfortable walk, self-paced brisk walk, jogging, and fast-run	Lying down, sitting (handwriting, computer game), standing with upper body movements (throw and catch, laundry task, floor sweeping), walking (comfortable overground walk, brisk overground walk, brisk treadmill walk), running, basketball, and dance

accelerometer was used in this study. The physical activities included in the data sets were: lying down, sitting, standing, walking, running, cycling, ascending stairs, descending stairs, Nordic walking, vacuum cleaning, ironing clothes, and jumping rope. For the purposes of this study, the first eight basic activity classes were selected for evaluation because these activity classes were widely used in past studies.

Data set 2. The second data set comprised wrist accelerometer data collected on eight individuals (mean age, 29.9 ± 4.2 yr, 50% male, mean BMI, 22.8 ± 1.9 kg·m^{-2}) during an outdoor PA session in a park. The data collection protocol included the activities in the following order: stationary activity (sit or stand still) for 5 min, self-paced comfortable walk for 5 min, self-paced brisk walk for 5 min, jogging for 5 min, and fast-run for 2 min. In between each activity, participants rested for 5 to 15 min. During each trial, motion and HR were recorded using Empatica E4 monitor. The Empatica E4 (Empatica Inc., Boston, MA), a light-weight (25 g) wristwatch, was placed on participant's nondominant wrist to record 3D acceleration (±2*g*), HR, electrodermal activity, and temperature. This study used only the 3D acceleration data. The sampling rate of the acceleration data was 32 Hz.

Data set 3. The third accelerometer data set was collected from 17 children (9 boys, 8 girls; age, 14.6 ± 2.4 yr; BMI percentile, 66.8 ± 25.9) (37,38). A total of 12 activity trials were performed over two laboratory visits. On visit 1, participants completed the following six trials: lying down, handwriting, laundry task, throw and catch, comfortable overground walk, and aerobic dance. On the second visit, the following six trials were completed: seated computer game, floor sweeping, brisk over-ground walk, basketball, over-ground run/jog, and brisk treadmill walk. The duration of each trial was 5 min. Based on the movement pattern, activities were categorized into seven categories: lying down, sitting (handwriting, computer game) standing with upper body movements (throw and catch, laundry task, floor sweeping), walking (comfortable over-ground walk, brisk over-ground walk, brisk treadmill walk), running, basketball, and dance. Further details of the activities trials can be found in Trost et al (37). During the trials, participants wore an ActiGraph GT3X+ tri-axial accelerometer (ActiGraph Corporation, Pensacola, FL) on the right hip and nondominant wrist. The sampling rate was set to 30 Hz. In this study, only wrist-worn acceleration data were used.

Classification Framework

The four steps of the classification framework, shown in Figure 1, were data preprocessing, feature extraction, normalization and feature selection, and activity classification. In the classification step, both conventional and custom ensemble methods were implemented. In the custom ensemble method, decisions from four "state-of-the-art" single classifiers, including BDT, kNN, SVM, and ANN, were aggregated together using three decision fusion techniques, namely, WMV, NB combiner, and BKS combiner. Among the conventional ensemble methods, boosted decision trees, bagged decision trees, and random forest were investigated. All steps of the framework were implemented using Matlab (The MathWorks Inc., USA). Example features and code for using the proposed framework can be found in the following link: https://github.com/alokchy04/Decision-Fused-Ensembles-for-PA-Classification-from-Wrist-Worn-Accelerometer.

Preprocessing. In the preprocessing step, the accelerometer data were annotated with activity labels and converted to time series data structure. If the data set contained missing accelerometer data, linear interpolation was used to find the intermediate missing values in the data. Missing values at the end of each labeled activity were replaced by previous value. In addition, 10 s of data at the beginning and end of each labeled activity was discarded from analysis to remove non–steady-state data.

Feature extraction. A range of time and frequency domain features were extracted from 10-s sliding window with 50% overlapping. A 10-s window was chosen as this period is sufficient to capture multiple periodic movements for all activities (38). In total, 45 features were extracted from each accelerometer. Consistent with previous studies, mean,

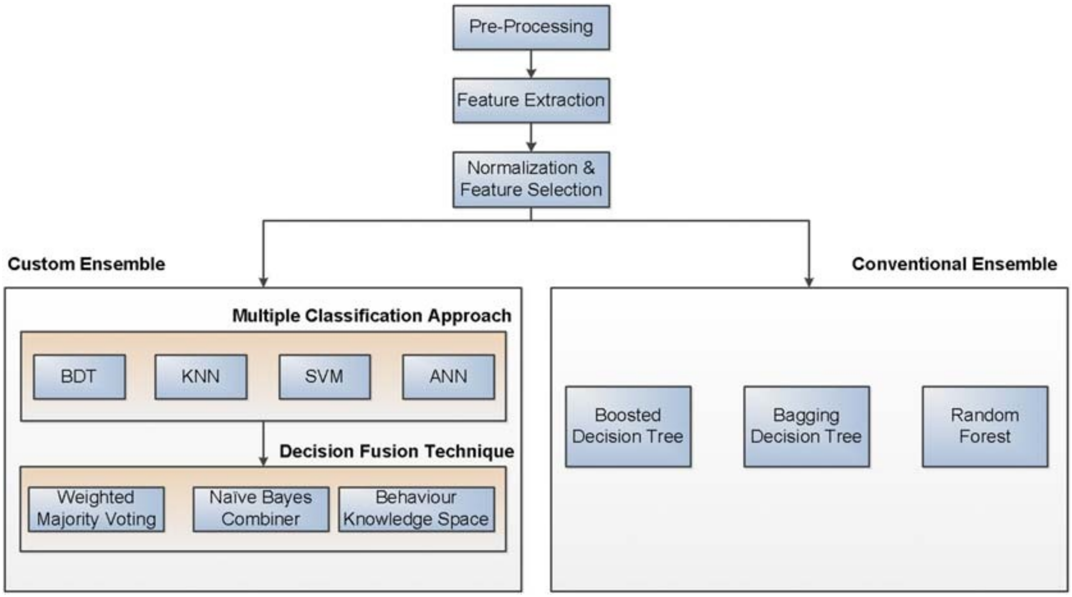

FIGURE 1—Flow diagram of the proposed framework.

standard deviation, minimum, maximum, variance, median, skewness, 25th and 75th percentile, and kurtosis were extracted from each axis of a three-axis accelerometer (10,22). In addition to these simple time domain features, frequency domain features including spectral energy, dominant frequency, dominant frequency magnitude, zero crossings, and cross-axis correlations were calculated. Spectral energy was calculated by summing the squared discrete Fast Fourier Transform component magnitudes of the signal (2). Spectral energy was normalized by dividing it by window length. The frequency with highest FFT magnitude was considered as principle frequency (23). Zero-crossing for each accelerometer axis represented the number of times the signal changed sign, and accelerometer axis cross-correlations (*corrxy, corrxz, corryz*) (2,34) were calculated and also included in the feature list. A detailed description of the features is provided in Document, Supplemental Digital Content 1, feature descriptions, http://links.lww.com/MSS/A922.

Normalization and feature selection. Normalization of the features before classification is useful when the feature values vary in different dynamic ranges. In this study, the training features were normalized to a zero mean and unit variance by subtracting the corresponding mean and dividing by the standard deviation. Features in the testing data were normalized using the same approach using the training data means and standard deviations. Feature selection is another important step necessary to improve time and space complexity of the classification algorithms. A correlation-based feature selection method (12) was applied on the training data to select features for classification. Features with a correlation of ≥0.25 with the activity classes were selected as inputs to the classifiers. A list of features selected for inclusion in each training data set is provided in Table, Supplemental Digital Content 2, feature list, http://links.lww.com/MSS/A923.

Conventional Ensemble Methods

The performance of three standard ensemble methods were evaluated—bagged decision trees, random forests and boosted decision trees. "Treebagger" classification class of Matlab was used as the bagging decision tree and random forest implementation. The number of decision trees in the ensemble was empirically set to 20 because it provided optimum performance. Although the bagging decision tree considered all features for splitting a node, random forest implementation used the number of features to sample equal to the square root of the total number of features available. For the boosting decision tree implementation, Adaboost.M2 multiclass classification method with 100 learning cycle and "Discriminant" weak learners was chosen.

Custom Ensemble Methods

Heterogeneity in the decisions of multiple single classification algorithms (base classifier) on the same data set can be used to improve classification performance in PA recognition problems (6,17). When each base classifier has good individual performance and also sufficient diversity (due to having different algorithms), fusion will significantly improve performance. This study used four well-known, widely used supervised learning algorithms of different complexity (BDT, kNN, SVM, and ANN) as base classifiers, which were fused together using three decision-fusion techniques (WMV, NB, and BKS). Detailed information related to the implementation of the single classifier models can be found in Document, Supplemental Digital Content 3, multiple classification algorithms, http://links.lww.com/MSS/A924.

Decision Fusion Techniques

In a N-classifier ensemble, let the classifier set and set of classes are $E = \{E_1, E_2, \ldots, E_N)$ and $C = \{c_1, c_2, \ldots, c_m)$ respectively. Each classifier E_i produces a class label $l_i \in C$, $i = 1, \ldots, m$ without any further information. When classifying an object x, the N classifier outputs a vector $L = [l_1, l_2, \ldots, l_N]$. Then, decision fusion techniques combine the classifier's output and provide a single class label. The current study evaluated the performance of three different decision fusion techniques including WMV, NB combination, and BKS combination (16,17,29).

WMV. The weighted majority vote is one of the most widely used decision fusion combiners, often useful when all classifiers in the ensemble do not have equal performance. This approach measures the individual accuracy of each classifier on the training data and uses these as weights, $W = \{w_1, w_2, \ldots, w_N)$, to give the more competent classifiers more authority in making the final decision. Then, when predicting for an object x, for all predicted class labels, it calculates the score using following equation 1,

$$\text{score}(k) = \sum_{l_i = C_k} w_i \qquad k = 1, 2, \ldots, m \qquad [1]$$

Finally, it selects the class label which has maximum score.

$$\text{final label} = \arg\max{}_{k=1}^{m} \text{score}(k) \qquad [2]$$

NB combination. This fusion method assumes the classifiers are mutually independent. For each classifier E_i, a $m \times m$ confusion matrix CM_i is calculated by applying it to the training data set. Let, T is the total number of objects in training data, where the number of objects in each class is denoted by $T_1, T_2, \ldots, T_m$. During testing for an object x, this method calculates posterior probability for all predicted class labels using following formula.

$$\text{score}(k) = \frac{T_k}{T} \prod_{i=1}^{N} cm^i(k, l_i) \qquad k = 1, 2, \ldots, m \qquad [3]$$

Where $cm^i(k, l)$ is the number of elements of the training data set whose true class label was c_k and were assigned by E_i to class c_l. Finally, NB combiner assigns the class label with maximum score to object x using equation 2.

BKS combination. Unlike most fusion methods, BKS (13) does not require an assumption of independence of the decisions of individual classifiers. The accuracy of the BKS combiner is very high when data set is large, but on small data sets BKS often overtrains. It creates a knowledge space using a lookup table based on the classification of training data. The look up table provides information on how often each labelling combination is produced by the classifiers. When testing for an object x (window of test data), it looks for a combination of predicted class labels in the look up table and selects the most frequent true label corresponding to that combination as a final result. The challenge with this fusion technique comes when the testing data evokes label combinations that do not appear in the look up table. To address this problem, WMV was used for combinations of labels not in the look up table.

Performance Evaluation

The performance was evaluated using leave-one-subject-out (LOSO) cross-validation (27). In LOSO, data from one user are used for testing, the other users' samples are used for training. In this way, samples of each subject are used exactly once for testing. This study used F1 score (11) to measure the performance of the ensemble learning methods. The study favored F1 score over classification accuracy because unlike accuracy or percentage of agreement, it is not influenced by class distribution. The F1 score was computed from precision and recall by keeping a balance between them.

$$\text{F1 Score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \times 100\% \qquad [4]$$

Where precision describes the exactness of a classifier. A lower value of precision indicates a high false-positive rate. Recall or sensitivity is useful to measure the completeness of classifiers. Low recall indicates a high false-negative rate.

RESULTS

Data set 1 results. Table 2 reports F1 scores of the four single classifiers, the custom ensemble, and the three conventional ensemble methods for all activities in data set 1. In this data set, the custom ensembles using WMV and NB fusion were effective and outperformed all of the single classifiers, but the conventional ensemble methods failed to exceed all of the single classifiers. Among the four single classifiers, the performance of SVM was best.

TABLE 2. Classification results (F1 score) using wrist acceleration sensor of data set 1.

	Conventional Ensembles			Individual Classifiers				Custom Ensembles		
	Random Forest	Bagging Decision Tree	Boosted Decision Tree	BDT	kNN	SVM	ANN	WMV Fusion	NB Fusion	BKS Fusion
Lying	80.18	72.76	89.31	73.28	87.36	92.78	91.22	92.69	91.55	86.4
Sitting	76.92	74.34	79.57	70.94	78.97	85.71	82.39	85.5	85.8	79.4
Standing	87.65	82.08	81.6	76.02	84.91	86.04	85.16	88.89	88.7	87.65
Walking	84.5	86.55	88.96	70.07	76.99	87.45	84.34	87.4	84.38	82.02
Running	100	99.12	99.71	85.25	99.71	96.02	94.15	99.12	99.12	99.12
Cycling	95.68	92.93	93.67	92.89	95.62	95.96	86.4	96.61	96.8	96.12
Ascending Stairs	59.89	58.15	53.61	44.44	39.26	48.87	58.51	58.46	57.39	50.87
Descending Stairs	66.67	75.7	56.2	64.26	68.46	72.88	69.6	76.42	79.84	74.49
Average	**81.44**	80.2	80.33	72.14	78.91	**83.22**	81.47	**85.64**	85.45	82.01

Bold values indicate the best results in conventional ensembles, individual classifiers, and custom ensembles.

TABLE 3. Classification results (F1 score) using wrist acceleration sensor of data set 2.

	Conventional Ensembles			Individual Classifiers				Custom Ensembles		
	Random Forest	Bagging Decision Tree	Boosted Decision Tree	BDT	kNN	SVM	ANN	WMV Fusion	NB Fusion	BKS Fusion
Stationary (sit and stand)	94.52	95.83	93.41	91.35	90.5	92.24	89.55	94.07	92.58	92.37
Comfortable walking	70.73	67.17	62.78	65.17	57.08	64.52	58.91	68.63	66.15	63.02
Fast walking	75.64	67.94	70.75	67.75	64.7	74.36	54.2	74.56	72.96	65.77
Jogging	83.52	82.09	74.03	79.21	72.98	69.94	66.92	75.94	75.1	76.4
Running	73.82	72.9	66.07	73.56	67.47	66.67	55.83	71.22	71.14	70.69
Average	**79.65**	77.18	73.41	**75.41**	70.54	73.54	65.08	**76.88**	75.58	73.65

Bold values indicate the best results in conventional ensembles, individual classifiers, and custom ensembles.

Data set 2 results. Table 3 reports F1 scores of the four single classifiers, the custom ensemble, and the three conventional ensemble methods in data set 2. In this data set, the random forest model was the best ensemble classifier overall, with the custom ensemble with WMV fusion also providing better recognition accuracy than the four single classifiers. The random forest model provided better recognition accuracy for all physical activities, with the exception of stationary activities, for which the bagged decision tree was best. The custom ensemble with WMV fusion performed well for sitting and standing, comfortable walking, and fast walking, but failed to outperform BDT for jogging and running. Of the four single classifiers, BDT was the best performer.

Data set 3 results. Table 4 reports F1 scores of the four single classifiers, the custom ensemble, and the three conventional ensemble methods in data set 3. In this data set, the custom ensemble with WMV fusion provided the highest performance, whereas the random forest and custom ensemble with NB custom also performed better than the single classifiers. Compared with the other classifiers, the random forest model exhibited the best recognition accuracy for lying down, sitting, and walking activities. Of the four single classifiers, SVM exhibited the highest classification accuracy.

Classification accuracies and confusion matrices for three data sets can be found in Document, Supplemental Digital Content 4, confusion matrices, http://links.lww.com/MSS/A925.

Statistical comparison. The comparative performance of the different ensemble models and the single classifiers across different folds/subjects were tested for statistical significance using one-way repeated-measures ANOVA. To increase statistical power and enhance the generalizability of the findings, F1 scores for each hold out subject/fold from all three data sets were pooled.

Overall, mean F1 scores differed significantly between the ensemble and single classifier models (Wilks lambda = 0.270, F (9,24) = 7.204, P < 0.0001). LSD *post hoc* comparisons revealed that the custom ensemble with WMV or NB provided statistically significant improvements in performance relative to the single classifiers. The custom ensemble with WMV significantly outperformed the conventional ensemble models with the exception of the random forest classifier. NB fusion significantly outperformed Adaboost, but not random forest or bagged decision trees. The custom ensemble with BKS fusion offered no significant improvements in performance relative to the conventional ensemble models and, with the exception of BDT, failed to outperform the single classifiers. Among the conventional ensembles, the random forest ensemble significantly outperformed the custom ensemble with BKS and the single classifiers, with the exception of SVM. Bagged decision tree and Adaboost significantly outperformed BDT, but not SVM, kNN, or ANN.

DISCUSSION

This study systematically examined the performance accuracy achieved by several conventional ensembles and a custom ensemble method in three data sets featuring wrist-worn accelerometer data. Across the three data sets, random forest ensembles and the custom ensemble with weighted majority vote provided consistently higher classification accuracy than bagged and boosted decision trees, and with the exception of SVM in data set 1, significantly outperformed the four single classifiers. Of the three decision fusion techniques examined, weighted majority vote provided marginally better performance than NB fusion. However, both weighted majority vote and NB fusion significantly outperformed BKS fusion.

TABLE 4. Classification results (F1 score) using wrist acceleration sensor of data set 3.

	Conventional Ensembles			Individual Classifiers				Custom Ensembles		
	Random Forest	Bagging Decision Tree	Boosted Decision Tree	BDT	kNN	SVM	ANN	WMV Fusion	NB Fusion	BKS Fusion
Lying down	79.46	78.18	74.45	68.17	66.94	76.66	70.97	78.26	75.98	72.13
Sitting+	92.13	90.04	89.97	85.52	87.09	91.19	89.87	90.63	90.39	86.74
Standing+	86.69	85.12	82.68	79.34	84.3	86.19	85	87.67	87.83	85.32
Walking	95.39	93.14	93.41	91.95	94.9	94.88	93.14	95.34	95.31	93.5
Running	71.35	64.12	71.48	57.45	66.54	67.16	64.41	73.18	69.93	64.41
Basketball	85.63	84.31	89.38	76.52	87.22	89	89.54	91.16	91.14	86.72
Dance	84.47	79.59	84.58	74.45	78.88	80.9	82.05	85.71	81.74	81.2
Average	**85.02**	82.07	83.71	76.2	80.84	**83.71**	82.14	**85.99**	84.62	81.43

Bold values indicate the best results in conventional ensembles, individual classifiers, and custom ensembles.

Our results are consistent with previous studies demonstrating that combining multiple classifiers with different induction bias provides better PA recognition than conventional ensemble methods and single model classifiers. Ruch et al. (28) used majority voting (MV) to combine the decisions of kNN, normal density discriminant function, and custom decision tree classifiers. In free living conditions, MV provided a maximum 67% classification accuracy when employing both hip and wrist accelerometers. Most recently, Catal et al. (6) reported that the combination of three PA classifiers (logistic regression, decision tree, and multilayer perceptron) using voting provided better performance than a single model approach.

The poorest-performing custom ensemble in our experiment was BKS. The limitations of BKS are well documented (25). It frequently suffers from generalization error if the training data set is not sufficiently large and/or representative. For example, when the number of classes (m) and classifiers (N) are large, the combinations of classifier's outputs for all classes in the look-up table become very large ($m^N \times m$). In this case, if the training data set is not representative and sufficiently large to estimate all or most of the combinations of classifiers outputs, BKS fusion can provide poor performance. BKS fusion also does not perform well when the combinations of classifier's outputs are ambiguous, that is, multiple occurrences of the same combination of classifier outputs correspond to different true labels in the look-up table, and has low confidence/probability on the most representative true class. Considering the number of classes (eight in data set 1, five in data set 2, and seven in data set 3) and four classifiers used in this study, the data sets were small for a reliable BKS fusion. Also, in our experiment, BKS fusion occasionally misclassified test instances due to the ambiguous cells in the look-up table. To avoid the problems related to ambiguous cells, some existing papers propose to use a local classifier in the original feature space associated with ambiguous cells (25); however, this was not investigated in the current study.

Among the conventional ensembles, the random forest algorithm provided strong classification performance, with F1 scores ranging from 79.6% to 85% across three data sets. This finding is consistent with results of recent studies developing and testing random forest classifiers for use in the exercise and movement sciences. Ellis and colleagues (8) developed a random forest classifier for recognition of four broad classes of physical activities (household duties, stair climbing, walking, running) in healthy adults. Separate classifiers were trained using frequency and time domain features in accelerometer data collected on the hip and wrist. Using LOSO cross-validation, the average overall accuracy for the hip and wrist classifier was 92.7% and 87.5%, respectively. In a follow-up investigation (9), a two-step activity recognition model comprising a random forest classifier and a hidden Markov model provided a balanced accuracy of 88.1% and 83.6% for the hip and wrist, respectively. Most recently, Pavey et al. (21) developed a random forest activity classifier for recognition four activity classes from accelerometer data collected on the wrist. Recognition accuracy for sedentary, stationary plus, walking, and running was 80.1%, 95.7%, 91.7%, and 93.7%, respectively. When evaluated on 24-h free-living data, recognition of stepping events (walking and running) exceeded 90%.

However, it is important to note that random forest classifiers do not perform well in all testing scenarios. Sasaki et al. (30) used time and frequency domain features in accelerometer signal collected on the dominant hip, wrist, and ankle to train random forest PA classifiers for older adults (65–85 yr). In the LOSO cross-validation of the laboratory-based activity trials, recognition accuracy for five activity classes (sedentary, standing, household chores, locomotion, recreational activities) was 87%, 84%, and 89% for the hip, wrist, and ankle models, respectively. However, when the models were deployed in free-living conditions, the overall classification accuracy declined significantly to just over 50%.

Although the focus of this study was ensemble learning methods, the strong performance (F1-score) of the base classifiers is worth noting. Of the four single classifiers examined, SVM provided the highest averaged F1-score for data set 1 (83%) and data set 3 (84%), which is consistent with the results of previous investigations comparing the performance of different supervised learning algorithms (7). BDT on other hand, performed best (75%) for data set 2, but exhibited the worst performance of all the base classifiers in data sets 1 and 3. The superior performance of BDT in data set 2, may be explained, at least in part, by the relatively homogeneous nature of the activities represented in the training data (rest versus walking and running at different speeds). It may be that ensemble methods are more suitable for more complex activity recognition problems requiring the detection of more fine-grained activities. Future research should explore this hypothesis.

Although the ensemble methods consistently achieved better performance accuracy, the magnitude of improvement over the single model classifiers was relatively small. This is because the single classifiers were trained with sufficient data and exhibited relatively high recognition accuracy in their own right. Nevertheless, when investigating performance on a classwise basis, notable performance differences were observed for several activity classes. In data set 1, the custom ensemble with WMV fusion improved the recognition of stair climbing to 61% compared with the best single classifier (ANN, 57%). Similarly, for data set 3, recognition accuracy for running increased to 73%, where best single classifier (SVM) provided only 68% accuracy. Although the increment in performance afforded by ensemble methods varied by data set and activity class, the results confirm the general principle that ensemble methods work best when the decisions from individual classifiers are complimentary.

A strength of the current study was the use of three diverse PA data sets, collected on different participant groups (adults and children) performing different physical activities in different contexts (laboratory-based vs outdoors). The examination of three different decision fusion methods to combine four widely used "state-of-the-art" classification algorithms was an additional strength. There were, however, some limitations that warrant consideration. First, although the study was conducted using training data collected in under different conditions, all three data sets comprised activities that were completed in predetermined sequences. Thus, additional work is required to evaluate the relative performance of ensemble methods in true free living contexts. Second, the activities in the selected data sets were primarily ambulatory in nature. Only data set 3 included nonambulatory lifestyle activities, such as basketball, dance, and so on. Future studies should include a more diverse set of physical activities to recognise using ensemble methods. Third, a simple correlation-based feature selection method was used. The use of a more sophisticated feature selection algorithm would likely have improved performance. Fourth and finally, our experiments focused on activity recognition or classification. It should be noted that ensemble methods can also be used for numerical prediction problems, such as estimating energy expenditure or PA intensity.

In summary, the results demonstrate that activity recognition accuracy can be improved through the implementation of ensemble learning methods. Conventional ensemble methods, such as bagging, boosting, and random forests improve activity recognition, in most, but not all situations. However, a custom ensemble using weight MV to fuse the decisions of four widely used "state-of-the-art" classification algorithms consistently outperformed the constituent base classifiers and most conventional ensemble models. Decision-fused ensemble methods thus have strong potential to improve PA recognition from wearable sensors.

No funding was received for completion of this project. Trost is a member of the ActiGraph Scientific Advisory Board. Chowdhury, Tjondronegoro, and Chandran declare no conflict of interest. The results from the present study do not constitute endorsement by the American College of Sports Medicine.

REFERENCES

1. Arif M, Kattan A. Physical activities monitoring using wearable acceleration sensors attached to the body. *PLoS One*. 2015; 10(7):e0130851.
2. Bao L, Intille SS. Activity recognition from user-annotated acceleration data. In. *Pervasive computing: Springer*; 2004, pp. 1–17.
3. Bassett DR Jr, Rowlands A, Trost SG. Calibration and validation of wearable monitors. *Med Sci Sports Exerc*. 2012;44(1 Suppl 1): S32–8.
4. Breiman L. Arcing classifier (with discussion and a rejoinder by the author). *The Ann Stat*. 1998;26(3):801–49.
5. Breiman L. Random forests. *Machine learning*. 2001;45(1):5–32.
6. Catal C, Tufekci S, Pirmit E, Kocabag G. On the use of ensemble of classifiers for accelerometer-based activity recognition. *Appl Soft Comput*. 2015;37:1018–22.
7. Cleland I, Kikhia B, Nugent C, et al. Optimal placement of accelerometers for the detection of everyday activities. *Sensors (Basel)*. 2013;13(7):9183–200.
8. Ellis K, Kerr J, Godbole S, Lanckriet G, Wing D, Marshall S. A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers. *Physiol Meas*. 2014;35(11):2191.
9. Ellis K, Kerr J, Godbole S, Staudenmayer J, Lanckriet G. Hip and wrist accelerometer algorithms for free-living behavior classification. *Med Sci Sports Exerc*. 2016;48(5):933–40.
10. Ermes M, Pärkka J, Mantyjarvi J, Korhonen I. Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions. *IEEE Trans Inf Technol Biomed*. 2008; 12(1):20–6.
11. Forman G, Scholz M. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *ACM SIGKDD Explorations Newsletter*. 2010;12(1):49–57.
12. Hall MA, Smith LA. Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper. In: Kumar A, Russell I, eds. *Proceedings of the FLAIRS Conference*. Palo Alto, CA: American Association for the Advancement of Artificial Intelligence; 1999. pp. 235–9.
13. Huang YS, Suen CY. A method of combining multiple experts for the recognition of unconstrained handwritten numerals. *IEEE Trans Pattern Anal Mach Intell*. 1995;17(1):90–4.
14. Jefferis BJ, Whincup PH, Lennon L, Wannamethee SG. Longitudinal associations between changes in physical activity and onset of type 2 diabetes in older british men: the influence of adiposity. *Diabetes Care*. 2012;35(9):1876–83.
15. Kate RJ, Swartz AM, Welch WA, Strath SJ. Comparative evaluation of features and techniques for identifying activity type and estimating energy cost from accelerometer data. *Physiol Meas*. 2016;37(3):360–79.
16. Kuncheva LI, Bezdek JC, Duin RP. Decision templates for multiple classifier fusion: an experimental comparison. *Pattern Recog*. 2001;34(2):299–314.
17. Mangai UG, Samanta S, Das S, Chowdhury PR. A survey of decision fusion and feature fusion strategies for pattern classification. *IETE Technical review*. 2010;4:293–307.
18. Montoye AH, Moore RW, Bowles HR, Korycinski R, Pfeiffer KA. Reporting accelerometer methods in physical activity intervention studies: a systematic review and recommendations for authors. *Br J Sports Med*. 2016;doi:10.1136/bjsports-2015-095947.
19. Montoye AH, Pivarnik JM, Mudd LM, Biswas S, Pfeiffer KA. Comparison of activity type classification accuracy from accelerometers worn on the hip, wrists, and thigh in young, apparently healthy adults. *Meas Phys Educ Exerc Sci*. 2016;20(3):173–83.
20. Owen N, Healy GN, Matthews CE, Dunstan DW. Too much sitting: the population health science of sedentary behavior. *Exerc Sport Sci Rev*. 2010;38(3):105–3.
21. Pavey TG, Gilson ND, Gomersall SR, Clark B, Trost SG. Field evaluation of a random forest activity classifier for wrist-worn accelerometer data. *J Sci Med Sport*. 2017;20(1):75–80.
22. Pirttikangas S, Fujinami K, Nakajima T. Feature selection and activity recognition from wearable sensors. In: Youn HY, Kim M, Morikawa H, eds. *Ubiquitous Computing Systems*. New York, NY: Springer; 2006, pp. 516–27.
23. Preece SJ, Goulermas JY, Kenney LP, Howard D. A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data. *IEEE Trans Biomed Eng*. 2009;56(3):871–9.
24. Preece SJ, Goulermas JY, Kenney LP, Howard D, Meijer K, Crompton R. Activity identification using body-mounted sensors—a review of classification techniques. *Physiol Meas*. 2009;30(4):R1–33.
25. Raudys Š, Roli F. The behavior knowledge space fusion method: analysis of generalization error and strategies for performance improvement. In: Windeatt T, Roli F, eds. *Proceedings of the International Workshop on Multiple Classifier Systems*. New York, NY: Springer; 2003. pp. 55–64.

26. Reiss A, Stricker D. Creating and benchmarking a new dataset for physical activity monitoring. In: *Proceedings of the 5th International Conference on Pervasive Technologies Related to Assistive Environments*. 2012. pp. 1–8.
27. Reiss A, Weber M, Stricker D. Exploring and extending the boundaries of physical activity recognition. In: *Proceedings of the Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on*. 2011. pp. 46–50.
28. Ruch N, Rumo M, Mäder U. Recognition of activities in children by two uniaxial accelerometers in free-living conditions. *Eur J Appl Physiol*. 2011;111(8):1917–27.
29. Ruta D, Gabrys B. An overview of classifier fusion methods. *Comput Inform Systems*. 2000;7(1):1–10.
30. Sasaki JE, Hickey A, Staudenmayer J, John D, Kent JA, Freedson PS. Performance of activity classification algorithms in free-living older adults. *Med Sci Sports Exerc*. 2016;48(5):941–50.
31. Schapire RE, Freund Y, Bartlett P, Lee WS. Boosting the margin: a new explanation for the effectiveness of voting methods. *Ann Stat*. 1998;26(5):651–86.
32. Skotte J, Korshøj M, Kristiansen J, Hanisch C, Holtermann A. Detection of physical activity types using triaxial accelerometers. *J Phys Act Health*. 2014;11(1):76–84.
33. Staudenmayer J, He S, Hickey A, Sasaki J, Freedson P. Methods to estimate aspects of physical activity and sedentary behavior from high-frequency wrist accelerometer measurements. *J Appl Physiol (1985)*. 2015;119(4):396–403.
34. 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. *J Appl Physiol (1985)*. 2009;107(4):1300–7.
35. Tremblay MS, LeBlanc AG, Kho ME, et al. Systematic review of sedentary behaviour and health indicators in school-aged children and youth. *Int J Behav Nutr Phys Act*. 2011;8:98.
36. Trost SG. State of the art reviews: measurement of physical activity in children and adolescents. *Am J Lifestyle Med*. 2007;1(4):299–314.
37. Trost SG, Loprinzi PD, Moore R, Pfeiffer KA. Comparison of accelerometer cut points for predicting activity intensity in youth. *Med Sci Sports Exerc*. 2011;43(7):1360–8.
38. Trost SG, Zheng Y, Wong W-K. Machine learning for activity recognition: hip versus wrist data. *Physiol Meas*. 2014;35(11):2183–9.
39. Yang P, Hwa Yang Y, Zhou B, Zomaya YA. A review of ensemble methods in bioinformatics. *Curr Bioinform*. 2010;5(4):296–308.