Physical Activity Monitoring

CS 4780 Final Project Report

Akshay Dhawan
Samuel Jones
Nicholas Tombari

Introduction

Physical activity monitoring is a cutting-edge research topic due to the advance of low cost and easy-to-use sensors that can record various features of a subject such as movement and heart rate [1]. Though it is generally recognized that exercise leads to a healthier lifestyle, recent reports have elaborated the recommended amount of exercise. One report recommends the following: 30 minutes of moderate activity for 5 days a week and 20 minutes of vigorous activity for 3 days a week [2]. Maintaining a record of physical activity is a difficult task which can be aided by the use of physical monitoring. A device can tell you how much exercise you have done daily as well as the intensity of those exercises. Such knowledge can help a subject make decisions toward a healthier lifestyle.

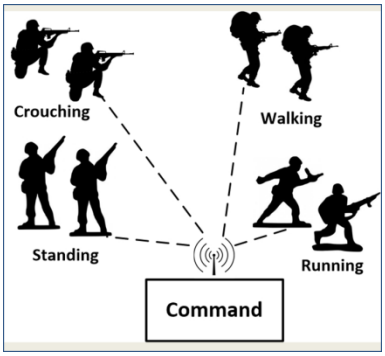

Figure 1-Depiction of Soldier Activities

Monitoring activities has utility in other applications as well including soldier monitoring, firefighter monitoring, and nursing homes. The stance or movement of a soldier can provide vital information for an officer in charge. Crouching might indicate that the soldier is in the ready position, while sprinting might indicate that the soldier is in trouble [3]. With a monitoring device, officers can keep an eye on each soldier and plan moves accordingly. Another application is firefighter monitoring, in which a home base can keep track of the movements of a firefighter and know when he or she gets stuck [4]. Finally, combining activity and physiological monitoring in nursing homes can provide useful information on the state of elderly patients [5].

This project looks at the prediction of daily activities such as walking, running, cycling, ascending stairs, etc. We attempt to classify activities as well as their relative intensity using data recorded from Inertial Measurement Units (IMU). IMUs basically sample movement information from subjects as they perform various activities. Our overall task is to collect the sampled data, perform feature extraction, and classify the activity and intensity.

Methodology

We divided our project into 2 main parts. The first part uses a dataset called PAMAP2 Physical Activity Monitoring Data Set found on the UCI Machine Learning Repository [1]. The second part uses a dataset that we created using sensors on our iPhones.

Classification Tasks

We have 2 main classification tasks

1 – Classify the specific activity being performed
2 – Classify the intensity of the activity being performed (i.e. light, moderate, and vigorous). Intensity categories are taken from [1] and are based on Metabolic Equivalents (MET). [6]

PAMAP2 Data

Our main resource is the open source PAMAP 2 dataset on the UCI Machine Learning Repository Website. It contains features from 8 subjects performing 18 activities including lying, sitting, standing, walking, running, cycling, Nordic walking, watching TV, computer work, car driving, ascending stairs,

descending stairs, vacuum cleaning, ironing, folding laundry, house cleaning, playing soccer, and rope jumping. The intensities for tasks are organized as shown in Table 1. The protocol was for each subject to perform each activity for 3 minutes. However, the data set is realistic, so many of the activities were performed for less time or even left out. Thus, our algorithms and methods had to account for left out data.

Table 1- Intensity Classification for Activities

Activity	Intensity
Lying Sitting Standing Ironing	Light
Walking Nordic Walk Descend Stairs Vacuuming Cycling	Moderate
Running Ascend Stairs Rope Jump	Vigorous

Subjects were all male and the average age was 27.22 ± 3.31 years. Each subject performed the activities with 3 IMU attached to them on their hand, chest, and ankle. Each IMU recorded acceleration, gyroscope, and magnetometer data all sampled at 100 Hz. The IMU also distinguished the x, y, and z axes. The acceleration scale was ±16 g and the units were ms^{-2}. The gyroscope units were rad/s and the magnetometer units were μT. Finally, all subjects also wore a heart rate monitor sampling at 9 Hz.

The questions we wanted to ask with this data set were as follows:

1 – Is this data separable, and, if so, are any of the algorithms out of Decision Tree, Naïve Bayes, and Support Vector Machine superior than the others? In particular, how well does an SVM perform?
2 – Which features are most important for classification?

Some of the previous questions had been addressed in [1], but we wanted to go through the data preprocessing, learning, and analysis portions ourselves to reinforce the material we learned in class. Previous work did not address SVM, so we wanted to ask if we could see an improvement through this method. A burning question we wanted to ask but could not address with this data set was whether subject dependent training would outperform subject independent training. Specifically, we wanted to test 2 different scenarios. The first scenario assembles all subject data and then divides into train and testing. The second scenario focuses on a single subject and divides into train and test only for that subject. However, each subject only performed an activity for 3 minutes, and it is likely that the subjects performance changed significantly across those 3 minutes. Therefore, dividing data for subjects would be difficult. This led to the 2nd part of our project.

iPhone Data

We used an iPhone 4S to collect accelerometer, gyroscope and magnetometer data with the phone placed in our pocket. We used an app by Chris Wonzy called Data Collection in order to collect the data. We recorded ourselves sitting, walking, and running. Walking and running were both performed on a treadmill in order to generate the least noisy data. The iPhone was placed in our right pocket with the face inward and the top of the phone downward. The sampling rate was set at 100 Hz (with an error range defined by the iPhone electronics).Each activity was recorded for 5 minutes twice in order to have sufficient examples to train and test. This data set allowed us to ask the following questions simultaneously.

1 – Does subject dependent training outperform subject independent training?
2 – Can we exhibit high accuracy with just 1 device as opposed to 3 (the case in PAMAP2)?

Feature Extraction

For each activity and subject, we took 5s windows shifted by 1s and extracted time domain features from those windows including mean, standard deviation, absolute integral, integral, and max. Along with heart rate, the PAMAP2 data set contained 3 Inertial Measurement Units (IMU) placed on the hand, chest, and leg. Each IMU contained accelerometer, gyroscope, and magnetometer data for x, y, and z axes. Each IMU results in 3 types * 3 axes = 9 measurements that we can extract features from. 3 IMU + HR result in 28 total measurements. Taking 5 time domain features for each plot, we generate 140 features. For the iPhone data set, we only had 9 measurements, which led to 45 features. By windowing the subject's activity, we can create single examples and label it with the activity being performed. This is shown in more detail in Figure 2. Here, the orange box on the data shows an example window. This window produces time domain features and a single example. The window is shifted to produce another example.

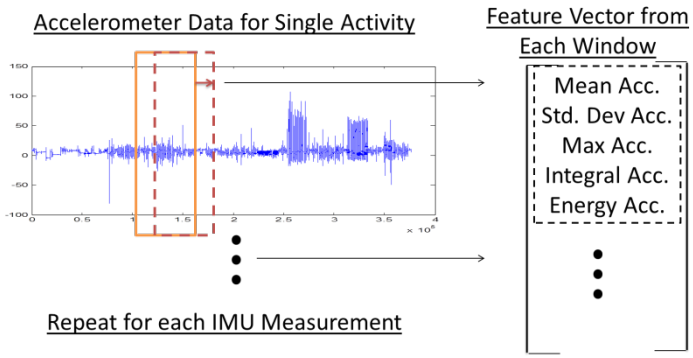

Figure 2-Illustration of Feature Extraction

Feature extraction code was written in MATLAB. Overall, we generated 20105 examples across all 8 subjects for the PAMAP2 data set. We generated 17617 examples across 3 subjects for the iPhone data set. We also attempted to classify non-overlapping windows for the iPhone data; we had 3434 examples in that case. Because the ranges varied widely across features, we normalized each feature by subtracting the mean and dividing the standard deviation. These examples were then written to a text file in the format used in class and understood by SVMlight. There were several dropped data points (indicated by NaN) in the PAMAP2 data set, which we got around by using the nanmean class of functions in MATLAB, which calculate time domain features by ignoring the NaN numbers.

Train and Test Division

For the PAMAP2 data set, we implemented subject independent training, which means we aggregated all examples and divided the set 50% train and 50% test. The fear with this approach, of course, is the violation of the i.i.d. assumption. We will address this with a test described in a later section.

For the iPhone data set, we recorded 2 sets of activities for each subject. One set for training and another for testing. This does not violate the i.i.d assumption and, thus, allows us to make stronger statistical conclusions. The division was 66% train and 33% test. We ran 2 separate types of tests. The first trained and tested on only a single subject. The second aggregated all subjects and then divided into train and test.

Learning Algorithms

We coded all of our learning algorithms in Python 3.2. We coded an ID3 Decision Tree (DT), a Linear Discriminant Analysis (LDA), and shell code to run SVMlight [7]. Both of our classification tasks were multi-class and used continuous attributes, which was not a problem for the DT or LDA. For the SVM, we implemented a 1 vs. all approach as done in Homework 3 to address the multi-class issue.

Results/Discussion

This section will be divided into the results for the PAMAP2 data set and results for the iPhone data set.

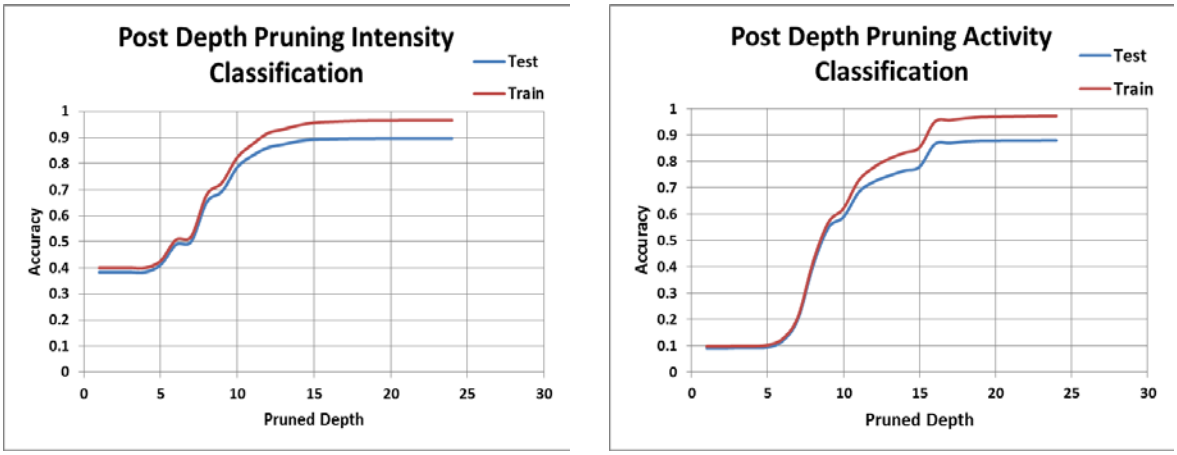

Figure 3-Pruning Accuracies

PAMAP2 Results

The DT algorithm took around 10 hours to train on 10092 examples. The tree was built to a full depth in order to achieve perfect or near perfect training accuracy**.** We implemented post-pruning validation to find the depth to use on the test set. We simply cut the tree at a certain depth and tested on a portion of the training set. The results of this validation method on the 2 classification tasks are shown in Figure 3. Results indicate that a tree of full depth was optimal. For intensity and activity, the optimal depth was 24.

The SVM algorithm took around 10 minutes to train 12 different classifiers on 10092 examples. Validation to find the appropriate C was used and the results are shown in Figure 4**.** A C of 500 was used for the activity task and a C of 10 was used for the intensity task.

The LDA took around 20 seconds to train on 10092 examples. A Leave One Out Subject (LOSO) test was performed in order to address the i.i.d. violation on the PAMAP2 data set. We tested on 1 subject, trained on the remaining 7 subjects, and repeated for each subject. A schematic of this test is shown in Figure 5**.** The averaged results on the testing set for the activity and intensity task respectively were 82.39% and 85.31%.

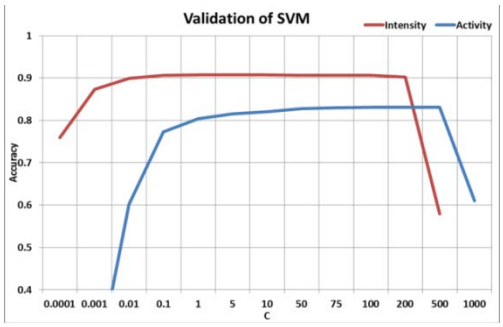

Figure 4 - SVM Validation Accuracies vs. C

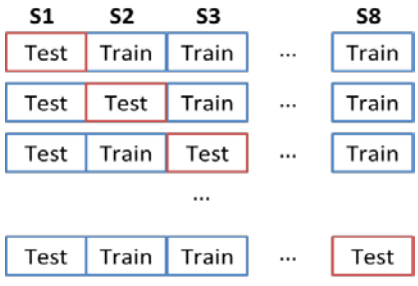

Figure 5 - Leave One Subject Out Test

The summarized results on the training and testing sets for each algorithm and classification task are shown in Table 2**.** Additionally, the bounds on the testing error for each classification task are shown in Table 3**.** Taking a glance at these tables tells us that DT and SVM perform the best. The LDA does not perform as well as the others, but this is expected given the nature of the LDA. The upper error bound of the DT and

Table 2 – Accuracy results for train/test sets on activity and intensity tasks

	Activity Train	Activity Test	Intensity Train	Intensity Test
DT	0.9927	0.8834	0.9923	0.8982
SVM	0.9564	0.8742	0.9458	0.9002
LDA	0.8551	0.8467	0.8580	0.8556

Table 3 – Error and bounds for test sets on activity and intensity tasks

	Activity Test	±	Intensity Test	±
DT	0.1166	0.0063	0.1018	0.0059
SVM	0.1258	0.0065	0.0998	0.0058
LDA	0.1533	0.0070	0.1444	0.0069

the lower error bound of the SVM seem to overlap on both classification tasks and this merits a closer look at the 2 algorithms through McNemar's test [7].

To compare SVM and DT performance on the activity and intensity tasks, we utilized McNemar's test and formed the contingency tables. The Null hypotheses in these cases are that the DT and the SVM have the same error rate. For the activity classification task, the chi square statistic is 7.78 (larger than $X^2_{1,0.95}$=3.8415), which means that we can reject the Null hypothesis and DT has a lower error rate than the SVM. However, for the Intensity classification task, the statistic is 0.376, which means we have to accept the Null hypothesis that the DT and SVM have equal error rates. Utilizing the same methodology, we found that the DT and SVM both have lower error rates than the LDA for both classification tasks.

Finally, in order to get an idea of how our algorithms performed on each specific intensity or activity level, we formed the confusion matrices as shown in Tables 4 and 5. These tables are only for Decision Tree performance. A quick glance shows that performance is very high for most tasks except for ascending and descending stairs, which the algorithms switch frequently. Also, performance degrades with increasing intensity.

Table 4 – Confusion matrix intensity

	Estimated Intensity			
Annotated Intensity	Light	Moderate	Vigorous	Performance [%]
Light	3909	40	6	98.84
Moderate	63	4090	438	89.09
Vigourous	11	469	1066	68.95

PAMAP2 Discussion

DT was found to have the highest accuracy on the activity task, while SVM and DT had comparable accuracies on the intensity task. In practice, however, it might make more sense to use SVM due to its reasonable training time. The LDA actually exhibits high accuracy as well, and, due to its very low training time, it could be used in an online setting as an adaptation to a subject. The LOSO results show us that the violation of the i.i.d. assumption actually did not have too much of an effect because we tested the LDA on completely new subjects and still exhibited high accuracies. The confusion matrices show us that most light tasks are quite easy to classify, which makes sense since most of the IMU are probably stationary. The more intense a task becomes, however, the more difficult it is to classify. This could be because a subject's behavior changes more across an intense task than a light task. Alternately, we could have trouble because different subjects have very different behaviors during intense tasks.

Table 5 – Confusion matrix activity

	Estimated Activity												
Annotated Activity	Lying	Sitting	Standing	Walking	Running	Cycling	Nordic Walking	Ascending Stairs	Descending Stairs	Vacuum Cleaning	Ironing	Rope Jumping	Performance [%]
Lying	919	3	6	1	0	0	0	0	0	2	3	0	98.39
Sitting	0	901	1	5	0	2	1	0	0	9	5	0	97.51
Standing	6	6	872	3	2	0	1	8	3	7	7	0	95.30
Walking	0	0	1	1179	3	4	7	4	8	1	0	0	97.68
Running	0	0	2	1	463	3	2	6	4	0	0	1	96.06
Cycling	0	1	0	0	3	746	4	4	5	10	0	1	96.38
Nordic Walking	0	1	3	6	0	2	915	3	0	1	1	1	98.07
Ascending Stairs	12	2	5	3	4	5	2	386	385	6	4	0	47.42
Descending Stairs	9	2	6	5	5	3	4	394	409	7	1	1	48.35
Vacuum Cleaning	1	11	5	4	3	10	5	15	7	752	17	1	90.49
Ironing	1	0	26	0	0	0	1	5	1	9	1139	0	96.36
Rope Jumping	0	0	0	1	3	1	2	1	2	1	4	235	94.00

To answer the 2nd question we posed, the most important features as retrieved from the top of the DT were maximum features. These presented the highest information gain. Specifically, max acceleration in all 3 IMU and max heart rate were found to be important towards classification.

iPhone Results

The same DT, SVM and LDA algorithms were used to classify the iPhone data. The code was slightly modified because the iPhone data only had 45 features, and we were only classifying three activities. Each learning algorithm performed general classification and then subject specific classification. This was done for both the data created using overlapping windows for feature extraction and non-overlapping windows.

The DT took around three hours to train on 11964 examples and 86 seconds to train on 2398 examples. The same post-pruning techniques were used to validate the tree. The results again showed that full depth was optimal.

The SVM trained in a few minutes on 11964 examples and less than a minute with 2398. Validation was done for each case and the optimal C values ranged from 1 to 50.

The LDA reported the fastest training time on both data sets; it was around 1 second for both. It is worth noting that the LDA had a significant drop in accuracy for the task of general classification.

The results are summarized in the Figure 6 for both the overlapping data set and the non-overlapping data set. The general classification is reported as “Everyone”. In this case, all subjects train examples were aggregated and all subjects test examples were aggregated. The subject specific classifications are shown by our names. The DT again reported the best accuracies for the overlapping data set while the SVM reported the best on the non-overlapping. An example of an important feature, the maximum acceleration in y-axis is shown in Figure 7. It can be seen that the activities are quite distinguishable just from this feature alone.

	Sam	Akshay	Nick	Everyone
DT	.9967	.9964	.9987	.9928
SVM	.9882	.9957	.9983	.9862
LDA	.951	.9491	.9963	.687

	Sam	Akshay	Nick	Everyone
DT	.9765	.9964	.9959	.9791
SVM	.9695	.9929	.9979	.9849
LDA	.9367	.9646	1	.6755

Figure 6 - (left) Accuracy Table for overlapped data and (right) non-overlapped data

To further investigate the difference between subject dependent training and subject independent training we ran a t-test on our iPhone results accounting for different test samples [8]. The t-test showed that the LDA performs better with subject dependent training with 95% confidence, but no significant difference exists with SVM or DT. For our t-test, the distributions for the two types of tests are different since one test works on a single subject and the other works on all subjects. This means the conclusions from this test need to be taken with a grain of salt.

iPhone Discussion

Even though we did not do as much testing and analysis with the iPhone, we answered the two questions that motivated this extension of our project. First, we found that subject dependent training part outperformed subject independent training only for the LDA. The t-test only allowed us to make statements with 95% confidence about the LDA, but we have to accept the null hypothesis that DT and SVM perform equally well on subject dependent and subject independent tasks. Some subjects proved harder to classify than others, so this question can be better answered with a data set that includes a collection from many more subjects.

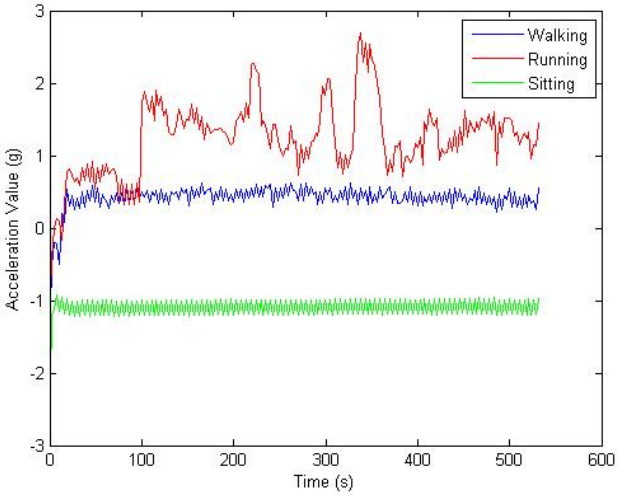

Figure 7 - Plot of Max Acceleration in Y axis for different activities for single subject

Second, we wanted to see if one device recording our motion could classify with similar high accuracy. Our results surprised us; the accuracy was high for both the overlapping and non-overlapping windows. There are a few explanations for why the accuracy was so high. With our iPhone, we only tried to classify three different activities: sitting, walking and running. This is more or less a classification of intensity limited to 3 activities. As noted before, out data was only collected by three subjects, this certainly made the general classification task easier. We also kept the data recording experiments pretty controlled with us only walking and running on treadmills and the orientation of the phone being the same every time.

We again investigated the most important attributes for the DT. They include max acceleration, max rotation, absolute sum of acceleration, standard deviation of acceleration and maximum magnetometer reading. The magnetometer reading often showed up in our decision tree because a lot of work was done on treadmills. We realized this after and the effects that the electric motor would have on our readings, so this is a flaw in the recording of our data.

Future Work

There were 2 major shortcomings to our project that future work could focus on. The first includes the lack of data available in the PAMAP2 data set. Because each subject only performed each activity once, we were unable to split up the activities because behaviors could have changed from the beginning to end of a performed activity. Instead, we had to create overlapping windows and then split the data into train and test files, which violated the i.i.d. assumption as some parts of the training data could have ended up in the testing data. We used a LOSO test to show that the algorithms still classified well on new subjects; however, for future comparison, we would like to have much more data with the PAMAP2 protocol in order to get a realistic idea of how good our algorithms are. Additionally, more data would allow us to delve deeper into the comparison between subject dependent and subject independent training.

Our second major shortcoming was the lack of performed activities with the iPhone data. Due to time constraints we were only able to perform sitting, walking, and running. This gave us a good range of intensities; however, in the future we would like to see the performance of a single measurement device on a range of activities. This will allow us to make more rigorous statistical assumptions about the performance of a single device compared to multiple devices as in the PAMAP2 dataset. We predict that the performance will degrade a little, but since most activities use lower body motion, a single measurement device could in theory be trained to differentiate all activities. In the end, the most useful applications of physical monitoring involve the monitoring of intensity rather than specific activities, and in this realm, a single iPhone can perform quite well.

Finally, future work should focus on more expanded machine learning methods such as hidden Markov models or SVM with kernels. Since all activities are performed as a sequence from one to another, a hidden Markov model could efficiently classify activities with the added knowledge of prior actions/state transitions. Our SVM performed well on all tasks for both parts of the project, but it would be interesting to see if a RBF or polynomial kernel could improve performance even further. Another interesting application would be to combine approaches to make an online classifier: one that could be taken of the shelf and adapted towards an individual. One could take advantage of the high accuracy of the large training time methods to build a baseline model and use an easily updateable method such as LDA to adapt to a subject.

Conclusions

Our project involved the classification of activities and intensities of activities using measurement data from sensors placed on subject's bodies. We developed an SVM, DT, and LDA implementation to test data from 2 separate datasets: the PAMAP2 dataset obtained online and an iPhone dataset that we developed. The PAMAP2 dataset contained recordings of subjects wearing 3 measurement devices and performing various activities. We exhibited very high accuracy on this dataset. While DT and SVM performed the best, the LDA exhibited much lower training time than the other algorithms. Statistically, there wasn't much difference between the DT and SVM performance on the data. We also found that the important features included maximum values of accelerometers and heart rate. The iPhone dataset contained recordings of us performing sitting, walking, and running with the phone in our pocket. We found that we could exhibit very high accuracy using just a single measurement device. DT and SVM again exhibited high accuracy on both the subject dependent and independent tasks, which bodes well for the general use of a learned classifier. Finally, we found that there was not a significant difference

between accuracies for subject dependent and subject independent classification tasks for DT and SVM. Future work should focus on expanding both the PAMAP2 and the iPhone dataset with more subjects and activities in order to draw more rigorous statistical assumptions about the performance of machine learning algorithms.

Acknowledgements

We would like to thank Professor Thorsten Joachims for his help and advice for our project. We would also like to thank Declan Boyd for providing consulting and advice throughout the course of the project.

References

[1] A. Reiss and D. Stricker, "Creating and Benchmarking a New Dataset for Physical Activity Monitoring," in *The 5th Workshop on Affect and Behaviour Related Assistance*, 2012.

[2] W. L. Haskell, I. M. Lee, R. R. Pate, K. E. Powell, S. N. Blair, B. A. Franklin, C. A. Macera, G. W. Heath, P. D. Thompson and A. Bauman, "Physical activity and public health: updated recommendation for adults from the American College of Sports Medicient and the American Heart Association," *Medicine and Science in Sports and Exercise,* vol. 39, no. 8, pp. 1423-1434, 2007.

[3] S. Biswas and M. Quwaider, "Remote Monitoring of Soldier Safety through Body Posture Identification using Wearable Sensor Networks," in *Wireless Sensing and Processing*, 2008.

[4] J. Duckworth and P. Nahass, "Integrated Firefighter Location and Physiological Monitor," Worcester Polytechnic Institute, 2012.

[5] M. Stachura and E. Khasanshina, "Telehomecare and Remote Monitoring: An Outcomes Overview," The Advanced Medical Technology Association, Augusta, GA.

[6] M. Jette, K. Sidney and G. Blumchen, "Metabolic Equivalents in Exercise Testing, Exercise Prescription, and Evaluation of Functional Capacity," *Clinical Cardiology,* vol. 13, pp. 555-565, 1990.

[7] T. Joachims, "Making large-Scale SVM Learning Practical," in *Advances in Kernel Methods - Support Vector Learning*, B. Scholkopf, C. Burges and A. Smola, Eds., MIT-Press, 1999.

[8] T. G. Dietterich, "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms," *Neural Computation,* vol. 10, pp. 1895-1923, 1998.