A Study in Human Activity Recognition:
Hierarchical Classification and Statistical Analysis

by

Anzah Hayat Khan Niazi

(Under the direction of Frederick W. Maier and Khaled Rasheed)

Abstract

A three-stage hierarchical classifier, using Random Forests, was constructed to classify 23 different physical activities of various types. This classifier was built using triaxial accelerometer data from 77 subjects collected during trials in Phoenix, Arizona. The activities were hierarchically divided and five Random Forest classifiers were trained for each level. The classifier performed well compared to similar classification studies in this domain, achieving 94% for activity groups and 87% at the individual activity level.

Furthermore, the effect of sampling rate and window size on activity recognition was also analyzed. Window size and sampling rate were varied, and a two-way weighted least squares analysis of variance was carried out. This analysis was carried out across a variety of activity types and demographic features. It was found that data collected at 50Hz, using 10 second windows performed statistically better than other data. There is, however, some statistical margin to allow for lower sampling rates and window sizes to be used without a significant reduction in classifier performance.

Index words: Data mining; Data classification; body-worn accelerometer; human activity recognition; random forests;

A Study in Human Activity Recognition: Hierarchical Classification and Statistical Analysis

by

Anzah Hayat Khan Niazi

B.Eng, National University of Science and Technology,
Islamabad, Pakistan, 2013

A Thesis Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

Master of Science

Athens, Georgia

2016

© 2016

Anzah Hayat Khan Niazi

All Rights Reserved

A Study in Human Activity Recognition: Hierarchical Classification and Statistical Analysis

by

Anzah Hayat Khan Niazi

Approved:

Major Professors:	Frederick W. Maier
	Khaled Rasheed
Committee:	Walter D. Potter
	Jennifer L. Gay

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
August 2016

Acknowledgments

I would like to thank Dr. Maier and Dr. Rasheed for their immense help and patience in assisting me to complete this thesis, as well as Dr. Potter and Dr. Gay for their valuable feedback and support. Further thanks is required for Dr. Matt Buman, Dr. Lakhsmish Ramaswamy, Dr. Jaxk Reeves and his team at the Statistics Consulting Center, and Delaram Yazdensepas for their crucial contributions to this work. Finally, I can never repay the love and support that my family and friends, both in Pakistan and the States, have provided me in these past few years.

This work is dedicated to my pillars of strength, Mama, Baba and Mona.

Table of Contents

Page

Acknowledgments iv

List of Figures vii

List of Tables viii

Chapter

1 Introduction and Literature Review 1

1.1 Human Activity Recognition 1

1.2 Random Forests 4

1.3 Two-Way ANOVA 5

1.4 References 6

2 A Hierarchical Meta-Classifier for Human Activity Recognition 10

2.1 Introduction 12

2.2 Related Work 13

2.3 Data Collection, Preprocessing and Experiments 16

2.4 Experimental Results 21

2.5 Conclusion and Future Work 26

2.6 References 27

3 Statistical Analysis of Window Sizes And Sampling Rates in Human Activity Recognition 30

3.1 Introduction 32

3.2 Related Work 33

3.3 Data Collection, Preprocessing and Methodology 34
3.4 Statistical Analysis of Results 37
3.5 Conclusion 45
3.6 References 46
4 Conclusions and Future Research 48
Bibliography 49

List of Figures

2.1 Accuracy performance for feature subsets across four classifiers 20

2.2 Performance of classification techniques on the 16 subject training set with 10-fold cross validation. 22

3.1 Distribution of BMI groups over age groups 43

List of Tables

2.1 Description of Activities Performed 17
2.2 Subject Demographics 18
2.3 Division of activities in the clusters 21
2.4 Performance analysis for Level 1 23
2.5 Performance analysis for Level 2 24
2.6 Performance on non-ambulatory activities 24
2.7 Performance on running activities 24
2.8 Performance on walking activities 25
3.1 Number of Records in the Datasets 35
3.2 Standard Deviations 37
3.3 All Activities/Demographics 38
3.4 Ambulatory vs. Non-Ambulatory Activites 39
3.5 Ambulatory Activity Groups 39
3.6 Stairs: Ascent vs. Descent 40
3.7 Non-Ambulatory Activites 40
3.8 Walking Activites 40
3.9 Running Activites 41
3.10 Gender: Female Subjects 41
3.11 Gender: Male Subjects 41
3.12 Age: 18-26 Years 42
3.13 Age: 27-33 Years 42
3.14 Age: 34-44 Years 42
3.15 Age: 49-63 Years 43

3.16 BMI: Normal 44
3.17 BMI: Overweight 44
3.18 BMI: Obese 44

Chapter 1

Introduction and Literature Review

This thesis describes a hierarchical classifier using hip-worn accelerometer data to identify physical activities (e.g. jogging, walking upstairs) and analyzes the effects of sampling rate and window size on classification accuracy. Chapter 2 discusses the hierarchical classifier, constructed using Random Forest algorithms, and its performance on the data. Chapter 3 relates a statistical analysis of the effects of window size and sampling rate on the classification of accelerometer data. The remainder of this chapter presents basic concepts related to human activity recognition, machine learning classification and its analysis. These concepts are needed to understand the later sections of the thesis.

1.1 Human Activity Recognition

Human Activity Recognition (HAR) is the field of study pertaining to the classification and prediction of physical activity [11]. With the advent of compact, commercially available sensor devices, this area has recently received a lot of interest. HAR has various practical applications — personal activity tracking [5, 14, 16, 24], public health monitoring [12, 20, 22], security analyses [2, 13, 18] and patient risk detection [9, 15, 26] to name a few. HAR can be used to perform immediate classification while data is being collected (*online learning*) or processed post-collection (*offline learning*) depending on requirements and resources. It must also be decided whether the model being obtained is to be universally applied to all users (*between-subjects*) or personalizable for individuals (*within-subjects*). The methods of collecting and analyzing data in this domain also vary according to objectives of research. We shall discuss some important aspects of this process.

1.1.1 Data Collection

In HAR studies, data is first collected for experiments, most commonly in controlled conditions. This data can be obtained through *external* or *wearable* sensors [11]. The decision of which type of sensor to use depends upon the objectives of the experiment, domain requirements as well as possible limitations. For example, if the purpose is security monitoring at a restricted facility, external sensors can be used for access monitoring (at doors and entryways), motion/presence detection (tactile door pads, camera image processing, GPS tracking) [2], and usage detection (machine access logging, time-based locks). However, if the objective is, for instance, health monitoring of elderly patients, worn sensors to monitor heart rate and detect falls might be needed [26].

Our research concerns physical activity recognition for public health purposes. We utilize wearable accelerometers for our work, in particular, a single hip-worn device, the Actigraph GT3X+ [1]. Other commonly used sensors in this domain are wrist-worn accelerometers, smartphones and heart-rate monitors. Wrist-worn accelerometers, however, are not efficient in detecting lower body movement which limits their practicality for studies like ours, which feature many locomotive activities, such as walking and running. [11]. Smartphones, while easily available, are not dedicated instruments of motion detection, require fixed placement for proper monitoring, have to share processing power with other applications and lack the resolution and reliability of accelerometers used for scientific purposes. Heart rate monitors have been shown to be of limited use for physically demanding activities as heart rates can remain high long after an activity has been performed [23].

A crucial step in accelerometer selection is the sampling rate at which data is collected. The range for this can vary from 100Hz to lower [16]. While high sampling rates are better for certain high energy activities, they can be noisy in less active scenarios, take up more data storage and are a drain on power. Conversely, low sampling rates consume less power and storage but do not provide sufficient information to recognize some activities. Ultimately, a trade-off between power and accuracy would be made according to the needs of the study.

Most HAR studies use *supervised learning* [11], which means that the training data used in the machine learning algorithms consists of observations whose classes are known. Labeling the data in a format suitable for use in an ML algorithm is itself a difficult problem. One method is to record activity durations and start/stop times using an independent observer. Another would be to fix the order and duration of activities performed. Both methods are susceptible to human error, though it can be argued that by having minor amounts "noisy" data, the classifier is in less danger of overfitting the training data and more robust for use in test environments.

1.1.2 Preprocessing Data

"Raw" data from accelerometers consist only of axial values at every timestamp. However, the raw data is not necessarily usable in a classification algorithm. A typical pre-processing step is to divide the raw data into temporal *windows*. Windows can be long or short and can be *sliding* (overlap exists between windows) or *disjoint* (no overlap exists). Windows that are too short might be insufficient to contain enough information for an activity while windows that are too long might contain more than one activity. Overlapping may possibly be helpful in handling transitional data but, more often than not, are shown to be redundant. Past research has advocated all of these forms of windowing for different purposes [11], so an informed decision can be made using the literature and domain-specific requirements.

The raw data in each window is used to extract features. These are typically time-based or frequency-based features. Common examples of time-based features are the means, standard deviations, percentiles of accelerometer readings, as well as median crossings. Common frequency-based features are the dominant frequency and its magnitude and wavelet coefficients. There is, however, a law of diminishing returns on the efficiency of the learning method using features, and, by Occam's Razor, a feature set having a balance of high performance and low number of features should be selected to save power consumption and processing time.

1.1.3 Learning and Evaluation

A wide variety of machine learning algorithms have been applied in HAR research [19, 11]. Section 2.2 of the next chapter provides a review of some methods that have been utilized while section 1.2 below discusses the learning method most pertinent to this study, *Random Forests*.

Regardless of the specific algorithm used, one of the most commonly used methods in dealing with classification is to divide the data into *training*, *validation* and *testing* sets. The learner trains over the training set, while the validation set is used to keep the performance of the learner in check. The test set is used as unseen data to provide an estimate of how the classifier would perform on independent data. A better way to get this estimate is through *k-fold cross-validation*, [10]. The dataset is divided into k number of folds. Then the classifier is trained on *k-1* folds and tested on the remaining fold. This is repeated k times and the performance results for each fold are averaged.

Some typical performance measures used in machine learning for HAR are accuracy, precision, recall and F-Measures. See equations 2.1, 2.2, 2.3 and 2.4 in Chapter 2.

1.2 Random Forests

Random forests are ensemble learners which utilize multiple decision trees and classify using the mode of all the trees [8, 4]. These are used in the hierarchical classifier described in Chapter 2. Decision trees classify data by splitting "leaves" (nodes) of a tree according to a function such as the *information gain ratio* [17]. As one traverses from the "root" of the tree to the "leaf", the classes to which a given instance can belong get smaller. While decision trees are used widely in classification, they are susceptible to *overfitting*, that is to say they tend to favor the training set to the point where the trained learner performs poorly on test sets and independent data. Random Forests seek to rectify this by using multiple trees with a random selection of features [25].

A brief summary of the random forest learning method is as follows :

1. Randomly sample, with replacement, n observations from the training set.
2. Use these samples to construct a decision tree using a random set of features. For each node in the tree:
   (a) Randomly select m features from the feature set.
   (b) Select the feature that maximizes the information gain ratio (or other function).
3. Repeat steps 1 and 2 for k number of trees. Typical random forests have a hundred trees or more.
4. Average the results or calculate the mode target value from the forest.

Sampling with replacement from the training and feature sets (also known as bagging or bootstrap aggregating [3]) decreases sensitivity of the learner to noise. This means that the variance of the model has been decreased. However, by choosing the feature from the random feature set using a function, such as *information gain ratio*, the correlation between the trees can be further reduced, thus resulting in a strong ensemble learning technique.

In initial test, Random Forests performed well compared to other classifiers and was subsequently used as the primary classifier throughout this study. For more information about the selection of classifiers, see Chapter 2.

1.3 Two-Way ANOVA

Analysis of variance (ANOVA) is the name of a collection of statistical models, most useful in analyzing the whether the means of multiple groups are equal [7]. While the t-test is a common method for testing statistical significance, it is limited to application on two groups. ANOVA extends this analysis across multiple groups and results in less type-I errors (false positives) than t-tests. Two-way ANOVA is an extension of the standard one-way ANOVA in which two independent variables influence one dependent variable. It looks at the effect

each of the independent variables have on the dependent variable as well the effects of the interactions of both variables. ANOVA has been recommended for use in classifier analysis previously [21, 28] but has rarely been used in ML research [6]. The specific variety of ANOVA used in the present study is elaborated on in Chapter 3.

1.4 References

[1] Actisoft analysis software 3.2 user's manual. Fort Walton Beach, FL: MTI Health Services.

[2] Robert Bodor, Bennett Jackson, and Nikolaos Papanikolopoulos. Vision-based human tracking and activity recognition. *Proceedings of the 11th Mediterranean Conference on Control and Automation*, 1, 2003.

[3] Leo Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[4] Leo Breiman. Random forests. *Machine Learning*, 14:5–32, 2001.

[5] Keith M. David, David J. Krupa, Melinda J. Chang, James Peacock, Yao Ma, Jeff Goldsmith, Joseph E. Schwartz, and Karina W. Davidson. Fitbit: An accurate and reliable device for wireless physical activity tracking. *International journal of cardiology*, 185:138–140, 2015.

[6] Janez Demsar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[7] Ronald A. Fisher. *Statistical methods for research workers*. Genesis Publishing Pvt Ltd, 1925.

[8] Tin Kam Ho. Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 1, pages 278–282, Aug 1995.

[9] Ning Jia. Detecting human falls with a 3-axis digital accelerometer. *A forum for the exchange of circuits, systems, and software for real-world signal processing*, 43:719–722, 2009.

[10] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 2:1137–1143, 1995.

[11] Oscar D. Lara and Miguel A. Labrador. A survey on human activity recognition using wearable sensors. *IEEE Communication Surveys and Tutorials*, 15:1192–1209, 2013.

[12] Jonathan Lester, Tanzeem Choudhury, and Gaetano" Borriello. Pervasive computing: 4th international conference, pervasive 2006, dublin, ireland, may 7-10, 2006. proceedings. In *A Practical Approach to Recognizing Physical Activities*, pages 1–16. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[13] Weiyao Lin, Ming-Ting Sun, Radha Poovandran, and Zhengyou Zhang. Human activity recognition for video surveillance. *IEEE International Symposium on Circuits and Systems*, pages 2737–2740, 2008.

[14] Jeffrey W. Lockhart and Gary M. Weiss. The benefits of personalized smartphone-based activity recognition models. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 614–622. 2014.

[15] Mitja Lutrek, Hristijan Gjoreski, Simon Kozina, Boidara Cvetkovi, Violeta Mirchevska, and Matja Gams. Detecting falls with location sensors and accelerometers. *Twenty-Third IAAI Conference*, pages 1662–1667, 2011.

[16] Uwe Maurer, Asim Smailagic, Daniel P. Siewiorek, and Michael Deisher. Activity recognition and monitoring using multiple sensors on different body position. *International Workshop on Wearable and Implantable Body Sensor Networks*, 2006.

[17] Tom M. Mitchell. *Machine Learning*. Tim Mc-Graw-Hill Companies, Inc., 1997.

[18] Wei Niu, Jiao Long, Dan Han, and Yuan-Fang Wang. Human activity detection and recognition for video surveillance. *Proceedings of IEEE International Conference on Multimedia and Expo*, 1:719–722, 2004.

[19] Stephen J Preece, John Y Goulermas, Laurence P J Kenney, Dave Howard, Kenneth Meijer, and Robin Crompton. Activity identification using body-mounted sensorsa review of classification techniques. *Physiological Measurement*, 30(4):R1, 2009.

[20] Nishkam Ravi, Nikhil Dandekar, Preetham Mysore, and Michael L. Littman. Activity recognition from accelerometer data. In *Proceedings of the 17th Conference on Innovative Applications of Artificial Intelligence - Volume 3*, IAAI'05, pages 1541–1546. AAAI Press, 2005.

[21] Steven L. Salzberg. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data mining and knowledge discovery*, 1:317–328, 2007.

[22] John Staudenmayer, David Pober, Scott Crouter, David Bassett, and Patty Freedson. An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer. *Journal of Applied Physiology*, 107(4):1300–1307, 2009.

[23] Emmanuel Munguia Tapia, Stephen S. Intille, William Haskell, Kent Larson, Julie Wright, Abby King, and Robert Friedman. Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart rate monitor. *11th IEEE international symposium on wearable computers*, pages 37–40, 2007.

[24] Gary M. Weiss, Jeffrey W. Lockhart, Tony T. Pulickal, Paul T. Mchugh, Isaac H. Ronan, and Jessica L. Timko. Actitracker: A smartphone-based activity recognition system for improving health and well-being, 2014.

[25] Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kauffman, 2005.

[26] Falin Wu, Hengyang Zhao, Yan Zhao, and Haibo Zhong. Development of a wearable-sensor-based fall detection system. *11th IEEE international symposium on wearable computers*, 2, 2015.

[27] Jhun-Ying Yang, Jeen-Shing Wang, and Chen Yen-Ping. Using acceleration measurements for activity recognition: An effective learning algorithm for constructing neural classifiers. *Pattern Recognition Letters*, 29(16):2213–2220, 2008.

[28] Jerrold H. Zar. *Biostatistical analysis*. Pearson Education, India, 1997.

Chapter 2

A Hierarchical Meta-Classifier for Human Activity Recognition[1]

[1]A.H. Niazi, D. Yazdensepas, J. L. Gay, F.W. Maier, L. Ramaswamy, K. Rasheed and M. P. Buman. Submitted to *15th IEEE International Conference on Machine Learning and Applications* 7/6/2016.

Abstract

This paper proposes a multi-level meta-classifier for identifying human activities based on accelerometer data. The training data consists of 77 subjects performing a combination of 23 different activities and monitored using a single hip-worn triaxial accelerometer. Time and frequency based features were extracted from two-second windows of raw accelerometer data and a subset of the features, together with demographic information, was selected for classification. The activities were divided into five activity groups: non-ambulatory activities, walking, running, climbing upstairs, and climbing downstairs. Multiple classification techniques were tested for each classifier level and groups. Random forests were found to perform comparatively better at each level.

Based upon those tests, a 3-level hierarchical classifier, consisting of 5 random forest classifiers, was built. At the first level, the non-ambulatory activities are separated from the rest. At the second, the ambulatory activities are divided into four activity groups. At the final level, the activities are classified individually.

Accuracy on test sets was found to be approximately 87% overall for individual activities and 94% at the activity group level. These results compare favorably to contemporary results in classifying human activity.

2.1 Introduction

Interest in human activity recognition has seen growth in recent years. Both the commercial market and the research sector have seen an increase in demand for efficient techniques in the classification of activities. The public health sector, in particular, is in need of such models. The target of this work is the development of a classification model for triaxial accelerometer data gathered from a demographically diverse group of subjects which could be used for test subjects under free-living conditions. Such a classification could be used to better estimate energy expenditure for individuals and determine the necessary steps for the individual to take towards a healthy lifestyle.

Activity recognition remains a difficult classification problem, mainly due to the sheer variance in how subjects perform the same activities and the relative difficulty in effectively monitoring subjects. Additional hardware (e.g., multiple body-worn accelerometers, heart-rate monitors, etc.) can improve classification, but this is typically not practical under free-living conditions. Using fewer measuring instruments is better from a subject's point of view, though it tends to be insufficient from a research perspective. A highly accurate classifier based on data from a single unobtrusive set of sensors would be ideal.

There are many ways to approach activity recognition. An important distinction to be made is between *within-subject* and *between-subject* recognition. For *within-subject* classification, individual models are produced for each subject. As is shown in Section 2, this can lead to very accurate results. However, by its very nature, the model development involves extensive training and calibration on each subject, which is difficult to achieve outside of controlled conditions.

In the case of *between-subject* classification, a universal model for all subjects is created. This model would (ideally) be trained only once and be applicable to all future subjects. Arguably, this is a more practical approach, as little calibration would be needed post-training. However, the generalized nature of the single model typically results in lower overall accuracy compared to individualized models.

For our research, the data used consists of triaxial data obtained from 77 subjects, quite fairly spread demographically, for 23 activities in studies performed at Arizona State University in 2013 and 2014. The activities performed are various types of non-ambulatory (sitting, standing, etc.) and ambulatory (walking, running, etc.) activities. This is a fairly large data set compared to similar research in the field.

The ultimate objective of this research is the development of models to be used in free-living conditions by subjects for the purpose of health monitoring and energy expenditure calculation. This would eliminate the need for self-reporting as is customary in free-living studies which tends to be inaccurate due to the nature of human error. It should be noted that the data was collected in a controlled environment, though the subjects had significant choice in the selection and order of the activities they performed.

The classifier we have built uses data extracted from a single triaxial accelerometer, an ActiGraph GT3X+, along with demographic information from the subjects. Our approach was for between-subject classification and grouping similar activities into distinct groups. The final 3-level classifier was built using random forests to cater to each group of activities as well as the individual activities that they contain.

Section 2.2 provide an analysis of prior research done in this domain. Section 2.3 explains how the data was collected and prepared for training. Section 2.4 goes on to describe how the classifier was trained and the results obtained in testing. Finally, Section 2.5 considers potential areas for future research and the methods by which results could be further improved.

2.2 Related Work

Data mining and machine learning techniques have been applied extensively to physical activity classification. Commonly, classification is based on body-worn accelerometer data.

One of the initial problems in data mining is determining, for a given domain, the most suitable classifier type as well as the optimal features to use for classification. Preece et al. [12] provide a good overview of the tools that prove useful in activity classification as well the

features that are frequently used. They present a comparison of different studies consisting of multiple routines, features, accelerometer placement locations and classification techniques. Some of the techniques analyzed include artificial neural networks, decision trees, k-nearest neighbor and support vector machines.

Bao & Intille [2] use five biaxial accelerometers for experiments, trained on 20 subjects and 20 activities in semi-natural settings. Decision trees gave them their best result of 84%. Additional experiments were carried out using only two accelerometers; it was found that a thigh- and hip-worn device combination provided the least decrease in accuracy.

In Ravi et al. [13], feature selection was used to identify that the mean, standard deviation, energy, and correlation are the best features for activity recognition. This study used data collected from two subjects wearing waist-mounted triaxial accelerometers. 8 activities were performed over the course of several days. Accuracy levels of 99% were obtained using Plurality Voting from a number of base classifiers trained by K-Nearest-Neighbors, Decision Trees, Naive Bayes, and Support Vector Machines. Though obtaining very high results, the study was only carried out on 2 subjects, which would not translate well to a general population.

Lester et al. [9] describe a personal activity recognition system using a custom-made module worn on the shoulder, consisting of an accelerometer, microphones and barometers (a combination of 7 devices). They combined multiple static classifiers using a Hidden Markov Model (HMM) and claim an accuracy of 90% (though their test accuracy reaches a high point of 84%). They tested their model on 12 subjects with 8 activities and indicate that accuracy drops to around 65% if only the accelerometer is used.

Yang et al. [17] use neural classifiers to classify eight domestic activities with data gathered from 7 subjects by wrist-worn accelerometers. They achieved a *within-subject* average accuracy of 95.24% by initially separating dynamic activities (running and walking) from static activities (standing and sitting) before using separate feature subsets for both types of activities.

Staudenmayer et al. [14] used artificial neural networks to classify groups of activities. They gather multiple activities into five groups (low level/non-ambulatory, locomotion, household activities, and vigorous sports). They had 48 subjects equally divided across genders and used triaxial accelerometers mounted at the waist. The accuracy of this system was 88.8% at the activity group level. The activities were not classified at the individual level.

Khan et al. [6] classifies a group of 15 activities consisting of 3 static (non-ambulatory) activities and 3 dynamic (ambulatory) activities with the rest being transitional activities that do not relate to this study. The data was collected from an accelerometer set at 20Hz worn on the chest by 6 subjects performing a specific sequence of activities each day for a month. They used a somewhat hierarchical approach to initially distinguish between static and dynamic activities followed by utilizing artificial neural networks and an augmented-feature vector to achieve accuracies averaging 97.9%. The data was collected on relatively few subjects for few activities with much attention given to transitional activities which are not the focus of our study. Additionally, the data was collected in specific activity sequences rather than the freer method applied by the subjects in our dataset. However, the paper displays the improvement in accuracy achieved by a hierarchical approach by comparing results obtained at a single-level classification (71.6%).

Kwapisz, Weiss & Moore [8] use the data mining software WEKA on a data set of 29 subjects performing 6 activities. The data collected came from a pocketed Android phone application. The activities performed would be regarded as high-level (grouped) activities in our work. They achieved an average accuracy of 91.7% using a multilayer perceptron.

Weiss et al. [15] developed a smartphone-based system using random forests on 5 grouped activities walking, jogging, climbing stairs, Standing, and sitting/lying down. They use both personalized (*within-subject*) and universal (*between-subject*) models. They extracted 43 features from triaxial accelerometer data. The universal model shows an accuracy of 76% with similar results on new subjects [10]. The personalized data is said to have accuracies generally higher than 95% for new subjects who train on themselves.

Phan [11] uses a pruned decision tree to classify 5 (grouped) activities performed by 20 subjects. The data was gathered using a Samsung mobile phone at 32 Hz. They achieved an accuracy of 96.8% by pruning off and discarding data after training with a C4.5 decision tree.

Deng et al [3] uses various KELMs (kernel extreme learning machines) on a dataset of 30 subjects performing 6 activities collected at 50Hz on a waist-worn smartphone accelerometer. They achieved accuracies of 99% using this approach.

Zheng [18] features a hierarchical approach similar to ours in classifying HAR. The dataset consists of hip-worn accelerometer data collected at 100Hz from 14 subjects on 10 self-paced activities. The study divides the activities into 4 "states". Using multi-layered Least Squares Support Vector Machines (LS-SVM) and Naive Bayes (NB) classifiers, they obtained an average accuracy of 95.6% across individual activities.

Our study, compared to those just discussed, consists of a significantly larger dataset with 23 activities categorized at a more sophisticated level than most datasets implemented by other studies. The primary reason for this is that this study is aimed at recognizing activities for public health purposes rather than general classification. While this certainly provides a bigger challenge at achieving high accuracies on this dataset, we intend to show that a hierarchical approach significantly improves our results over a single-layer classification and compares favorably to other studies despite the higher level of discrepancy involved in activity separation and the use of a single accelerometer.

2.3 Data Collection, Preprocessing and Experiments

Our data was trained on 77 subjects performing 4 grouped and 23 individual activities. This section details the collection and preprocessing of this data.

Table 2.1: Description of Activities Performed

#	Activity	Duration or Distance	# of subjects
1	Treadmill at 27 mmin-1 (1mph) @ 0% grade	3 min	29
2	Treadmill at 54 mmin-1 (2mph) @ 0% grade	3 min	21
3	Treadmill at 80 mmin-1 (3mph) @ 0% grade	3 min	28
4	Treadmill at 80 mmin-1 (3mph) @ 5% grade (as tolerated)	3 min	29
5	Treadmill at 134 mmin-1 (5mph) @ 0% grade (as tolerated)	3 min	21
6	Treadmill at 170 mmin-1 (6mph) @ 0% grade (as tolerated)	3 min	34
7	Treadmill at 170 mmin-1 (6mph) @ 5% grade (as tolerated)	3 min	26
8	Seated, folding/stacking laundry	3 min	74
9	Standing/Fidgeting with hands while talking.	3 min	77
10	1 minute brushing teeth + 1 minute brushing hair	2 min	77
11	Driving a car	-	21
12	Hard surface walking w/sneakers	400m	76
13	Hard surface walking w/sneakers hand in front pocket	100m	33
14	Hard surface walking w/sneakers while carry 8 lb. object	100m	30
15	Hard surface walking w/sneakers holding cell phone	100m	24
16	Hard surface walking w/sneakers holding filled coffee cup	100m	26
17	Carpet w High heels or dress shoes	100m	70
18	Grass barefoot	134m	20
19	Uneven dirt w/sneakers	107m	23
20	Up hill 5% grade w high heels or dress shoes	58.5m x 2 times	27
21	Down hill 5% grade w high heels or dress shoes	58.5m x 2 times	26
22	Walking up stairs (5 floors)	5 floors x 2 times	77
23	Walking down stairs (5 floors)	5 floors$\times$2 times	77

2.3.1 Participants and procedures

Participants were recruited from the Phoenix, AZ and surrounding areas through community sources, email distribution lists, and social media outlets. Participants were 18-64 years of age and free of any contraindications for exercise. Participants were fitted with the accelerometer and completed a series of activities for 3 min in duration (see Table 2.1). Virtually all participants completed the following activities: standing, fidgeting with hands while talking; 1 min of brushing teeth and 1 min brushing hair; some form of hard surface or carpet walking;

and walking up and down stairs. An additional three treadmill activities and three other activities were randomly assigned. Timestamps for the beginning and end of activities were captured using a custom-built Android application which was synced to the same computer as the activity monitor.

2.3.2 Activity monitoring

Participants were fitted with the ActiGraph GT3X+ (ActiGraph LLC, Pensacola, FL) activity monitor positioned along the anterior axillary line of the non-dominant hip. The monitor was fixed using an elastic belt. The ActiGraph GT3X+ is a lightweight monitor (4.6cm x 3.3cm x 1.5 cm, 19g) that measures triaxial acceleration ranging from -6g to +6g. Devices were initialized to sample at a rate of 100hz. Accelerometer data were download to and extracted using Actilife 5.0 software (ActiGraph, LLC, Pensacola, FL) [1].

310 subjects participated in the study. From them, data from 77 subjects, 53 females and 24 males, were used to train our classifiers. Table 2.2 provides demographic information on the subjects.

Table 2.2: Subject Demographics

	Mean	Standard Deviation	Range
Age (Years)	33.2	9.7	18.2 - 63.2
Height (cm)	167.9	7.9	152.6 -188.9
Weight (kg)	72.1	12.1	48.3 - 105.5
BMI	25.6	3.9	17.7 - 35.4

2.3.3 Feature Extraction

A total of 246 features were initially extracted from the raw data. A summary of the features is as follows:

- **Features in the time domain**: features extracted from the axes and their first differentials and the vector magnitude. These include the mean, maximum, minimum values, standard deviations, median crossings and the 10th, 25th, 50th, 75th, 90th

percentiles. Also included are the correlations between the each axes as well as the correlations between their first differentials.

- **Features in the frequency domain**: these include the dominant frequency and its magnitude for the axes, their first differentials and the vector magnitude.
- **Features of wavelet analysis**: features extracted from the 1st level to the 5th level of wavelet decomposition coefficients for each accelerometer signal (x,y and z). These include the mean, maximum, minimum, standard deviations, median crossings and the 10th, 25th, 50th, 75th, 90th percentiles.
- **Demographic features**: these include the age, height, weight, gender and BMI of the subject.

2.3.4 Feature Selection

The initial 246 features, while helpful for initially gauging the data, increase complexity. Generating the features requires extra work and significantly increases the time needed to train and run a classifier. Feature selection was carried out to whittle down the features to a smaller but more significant subset.

Initially, correlation-based [4] and relief-based [7] feature selection methods were used. Additionally, a 42-feature subset was selected using domain knowledge and an eye to standardize the preprocessing step. This feature set performed comparably well to the others in preliminary testing and was subsequently the feature set used in creating the final classifier. Figure 2.1 shows the classifiers performance on the feature subsets compared to the entire feature set.

Random Forests (100 trees) trained on the self-selected subset outperformed the other subsets across the classifiers. While other subsets also performed substantially well with Random Forests, the self-selected was also a robust performer with other classifiers, signifying its strength as a subset.

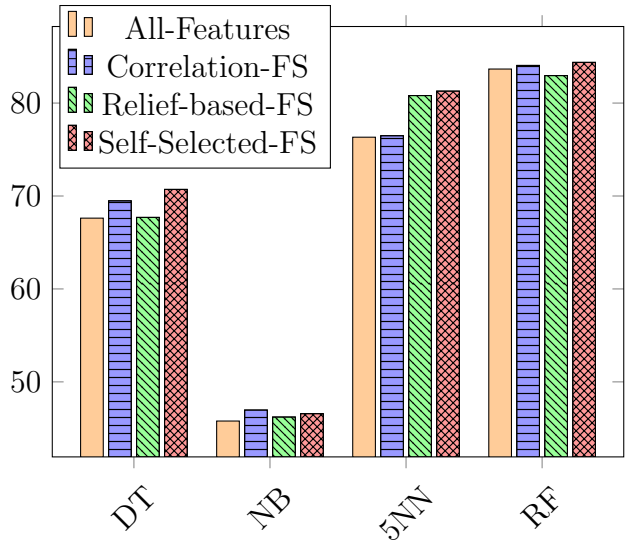

Figure 2.1: Accuracy performance for feature subsets across four classifiers. DT = Decision Trees; NB = Naive Bayes; 5NN = 5 Nearest Neighbor; RF = Random Forests,100 trees

2.3.5 Activity Groups

The 23 activities were separated into 5 activity groups, each group containing similar types of activities. Grouping permits layered classification—a higher level classifier learns to efficiently distinguish between groups of activities, while lower level classifiers specialize in classifying in-group activities. A possible disadvantage of this arrangement is that misclassification of instances at higher levels can propagate errors to lower levels. Because of this, it is imperative that the higher-level classifiers be much more accurate.

The activities were grouped as shown in Table 2.3. Activities which involved minimal physical activity were classified as non-ambulatory. Ambulatory activities were further divided into 4 subgroups; walking (all locomotive & treadmill activities less than 4 mph), running (treadmill activities more than 4 mph), climbing upstairs, and climbing downstairs.

Table 2.3: Division of activities in the clusters

Non-Ambulatory Activities
8,9,10,11

Ambulatory Activities	
Walking	1,2,3,4,12,13,14,15, 16,17,18,19,20,21
Running	5,6,7
Upstairs	22
Downstairs	23

2.4 Experimental Results

WEKA [16], a data mining software application, was used for classifier development. WEKA provides extensive libraries of machine learning techniques and an easily manageable GUI to carry out training experiments.

Though random forests was indicated as a preferable base classifier during the feature selection, other preliminary tests were carried out. A 16 subject subset was used to train many different classifiers with 10-fold cross-validation (as shown in Figure 2.2). In these initial tests, a 100 tree Random Forest outperformed all other classifiers, including meta-classifiers such as bagged decision trees and stacked classifiers.

Random Forests [5] work by generating many decision trees and classifying according to the modal class of the "forest". Subsequent experiments at the activity group level repeatedly indicated that random forests were a reasonable choice for at all levels of the meta-classifier.

Our data set contained 87,943 records. 90% of the data was used as training data and the remaining 10% was reserved as the test set, the data separated by stratification. Due to high number of activities in the walking group, the training set was weight-balanced to avoid a bias.

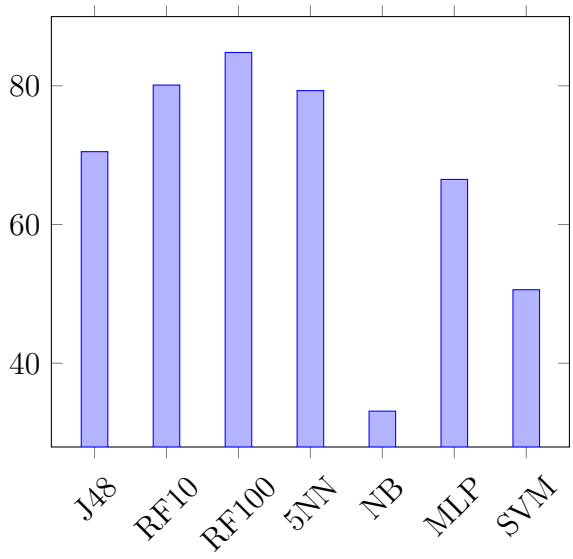

Figure 2.2: Performance of classification techniques on the 16 subject training set with 10-fold cross validation. RF10 = Random Forest, 10 trees; RF100 = Random Forest 100 trees; 5NN = 5 Nearest Neighbor; NB = Naive Bayes; MLP = Multilayer perceptron; SVM = Support Vector Machine

Five random forest classifiers were obtained for level 1, 2 and 3 for the walking, running and non-ambulatory groups. The trained classifiers were then set up in a Java program. The test set was run through this Java program which resulted in confusion matrices for each level. The accuracy, precision and recall for each level and activity were extracted from the confusion matrices using standard equations.

$$Precision = \frac{tp}{tp + fp} \tag{2.1}$$

$$Recall = \frac{tp}{tp + fn} \tag{2.2}$$

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{2.3}$$

$$F - measure = 2\frac{precision \times recall}{precision + recall} \tag{2.4}$$

tp=true positive, fp=false positive, fp=true negative, fn=true negative.

2.4.1 LEVEL ONE

At level one, the activities are divided into ambulatory and non-ambulatory activities. The accuracy at this level is 97.899%. The misclassified instances will contribute to some trickle-down errors in the levels below. However, the accuracy is low enough for overfitting not to be an issue. As the data set does not account for transitional activities (e.g., transitioning from standing to running), these discrepancies are welcome. The performance of level one is shown in table 2.4.

Table 2.4: Performance analysis for Level 1

LEVEL 1	Precision	Recall	F-Measure
Non-Ambulatory	0.987	0.944	0.965
Ambulatory	0.978	0.995	0.986
Accuracy: 98.03%			

2.4.2 LEVEL TWO

At level two, the activities are divided into groups shown in Table 2.3. Note that the non-ambulatory activities were separated at level one. Therefore, the errors shown for that group are a result of the trickle-down effect of level one. The accuracy at level 2 is almost 94%, which we believe compares favorably to other research, e.g., the 88% of Staudenmayer, et al. [14] and the 91.7% of Kwapisz, Weiss & Moore [8]. Phan [11], Ravi et al. [13], Khan et al [6] and [18] all achieve higher accuracies on their datasets, but these studies involve fewer subjects and overall less data. Given these discrepancies, we believe our results to be competitive.

The results in Table 2.5 can also be compared to results based on models using multiple sensors. E.g., Lester et al. [9] achieved 84% with 7 sensors. Bao & Intille [2] achieved 84% as well with 5 sensors.

Table 2.5: Performance analysis for Level 2

LEVEL 2	Precision	Recall	F-Measure
Non-Ambulatory	0.944	0.987	0.965
Walking	0.915	0.978	0.945
Running	0.993	0.847	0.914
Upstairs	0.958	0.813	0.880
Downstairs	0.969	0.755	0.849
Accuracy: 93.673%			

2.4.3 LEVEL THREE

At this level, activities are classified individually inside their respective groups.By doing this, we can train classifiers specifically for a group of activities which would otherwise be difficult to separate. The results for the non-ambulatory, running and walking groups are shown in Tables 2.6, 2.7 and 2.8 respectively.

Table 2.6: Performance on non-ambulatory activities

Non-Ambulatory Activities	Precision	Recall	F-Measure
Seated, folding/stacking laundry	0.924	0.951	0.937
Standing/Fidgeting while talking	0.933	0.958	0.945
brushing, 1min teeth + 1min hair	0.934	0.851	0.891
Driving a car	0.990	0.997	0.993
Overall Accuracy: 93.739%			

Table 2.7: Performance on running activities

Running Activities	Precision	Recall	F-Measure
Treadmill 5mph @ 0% grade	0.963	0.991	0.977
Treadmill 6mph @ 0% grade	0.976	0.965	0.970
Treadmill 6mph @ 5% grade	0.979	0.971	0.975
Overall Accuracy: 97.368%			

Non-Ambulatory Activities: Non-ambulatory activities are those that require little to no locomotion. The hierarchical approach allows the individual classifier to discern the minute differences between activities with a relatively low number of distinguishing features,

Table 2.8: Performance on walking activities

Walking Activities	Precision	Recall	F-Measure
Treadmill 1mph @ 0%	0.900	0.997	0.946
Treadmill 2mph @ 0%	0.904	0.943	0.923
Treadmill 3mph @ 0%	0.917	0.871	0.893
Treadmill 3mph @ 5%	0.932	0.927	0.929
Hard surface	0.842	0.967	0.900
Hard surface, hand in pocket	0.956	0.752	0.842
Hard surface, carrying 8 lbs.	0.913	0.652	0.761
Hard surface, cell phone	0.921	0.648	0.761
Hard surface, coffee	0.852	0.730	0.786
Carpet, heels/dress shoes	0.869	0.816	0.842
Grass barefoot	0.944	0.878	0.910
Uneven dirt w/sneakers	0.957	0.611	0.746
Uphill 5%, heels/dress shoes	0.948	0.895	0.921
Downhill 5%, heels/dress shoes	0.946	0.859	0.900
Overall Accuracy: 88.722%			

achieving an accuracy of 94%. Table 2.6 shows the performance of non-ambulatory activity classification.

Running Activities: Treadmill activities at 5-6 mph were regarded as running activities. The classifier achieved an very high accuracy of 97%, shown in table 2.7.

Walking Activities: The overall accuracy for classification of walking activities is 88.7%. Table 2.8 shows the comparatively low recall rates of the "Hard surface" activities, indicating that these activities are much more difficult to separate. The difference between these activities is arm position (holding cell phone, coffee cup, etc.) a feature that would probably be better detected by a wrist-worn accelerometer than a hip-worn one.

Note that as the "upstairs" and "downstairs" groups are single-activity groups, table results for them are not shown. The "upstairs" activity group had an accuracy of 81.33% and the "downstairs" activity group had an accuracy 75.5%. The downstairs activity has significantly lower recall. One possible explanation could be the varied approaches subjects

would have to climbing downstairs, demographically. For example, older and/or weightier subjects would descend stairs much more slowly than younger and/or lighter subjects.

Overall, the accuracy obtained was 86.63% at level 3. We believe our results here compare favorably with the few studies that have done activity classification at this minute level. Bao & Intille [2], for example, achieved 84% accuracy with 20 activities using data from 4 accelerometers.

2.5 Conclusion and Future Work

The hierarchical meta-classifier achieved an accuracy of 93.7% at the activity group level, which compares favorably to other group-level studies. Furthermore, the meta-classifier was able to distinguish between intra-group activities with an accuracy of 86.6% as compared to 84% at a single-level as seen in Figure 2.2.

This work was done in view of domain requirements for public health research in physical activities, specifically tracking the activities of free-living subjects without relying on self-reporting. By classifying activities in Table 2.1 at an individual level rather than a group level, physical activity researchers can obtain specific energy expenditure information for their subjects. Future work includes testing the model on free-living data which can lead to calibrations to improve the model. Eventually, creating a streamlined process of collecting, preprocessing and classifying data would be extremely useful for public health research.

Other work that will be done on the data include investigations can be carried out to analyze the effect of sampling rates and window sizes on classification accuracy, a comparative study which dives into a demographic and classifier based analysis, and classification of the data set using deep neural networks.

2.6 References

[1] Actigraph. Actisoft analysis software 3.2 user's manual. Fort Walton Beach, FL: MTI Health Services.

[2] Ling Bao and Stephen S. Intille. Activity recognition from user-annotated acceleration data. In *International Conference On Pervasive Computing*, pages 1–17. Springer Berlin Heidelberg, 2004.

[3] Wan-Yu Deng, Qing-Hua Zheng, and Zhong-Min Wang. Cross-person activity recognition using reduced kernel extreme learning machine. *Neural Networks*, 53:1–7, 2014.

[4] Mark A Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

[5] Tin Kam Ho. Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 1, pages 278–282, Aug 1995.

[6] Adil Mehmood Khan, Young-Koo Lee, Sungyoung Y. Lee, and Tae-Seong Kim. A triaxial accelerometer-based physical-activity recognition via augmented-signal features and a hierarchical recognizer. *IEEE Transactions on Information Technology in Biomedicine*, 14(5):1166–1172, 2010.

[7] Kenji Kira and Larry A. Rendell. A practical approach to feature selection. In *Proceedings of the Ninth International Workshop on Machine Learning*, ML92, pages 249–256, San Francisco, CA, USA, 1992. Morgan Kaufmann Publishers Inc.

[8] Jennifer R. Kwapisz, Gary M. Weiss, and Samuel A. Moore. Activity recognition using cell phone accelerometers. *SIGKDD Explor. Newsl.*, 12(2):74–82, March 2011.

[9] Jonathan Lester, Tanzeem Choudhury, and Gaetano" Borriello. Pervasive computing: 4th international conference, pervasive 2006, dublin, ireland, may 7-10, 2006. proceedings. In *A Practical Approach to Recognizing Physical Activities*, pages 1–16. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[10] Jeffrey W. Lockhart and Gary M. Weiss. The benefits of personalized smartphone-based activity recognition models. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 614–622. 2014.

[11] Thomas Phan. Improving activity recognition via automatic decision tree pruning. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, UbiComp '14 Adjunct, pages 827–832, New York, NY, USA, 2014. ACM.

[12] Stephen J Preece, John Y Goulermas, Laurence P J Kenney, Dave Howard, Kenneth Meijer, and Robin Crompton. Activity identification using body-mounted sensorsa review of classification techniques. *Physiological Measurement*, 30(4):R1, 2009.

[13] Nishkam Ravi, Nikhil Dandekar, Preetham Mysore, and Michael L. Littman. Activity recognition from accelerometer data. In *Proceedings of the 17th Conference on Innovative Applications of Artificial Intelligence - Volume 3*, IAAI'05, pages 1541–1546. AAAI Press, 2005.

[14] John Staudenmayer, David Pober, Scott Crouter, David Bassett, and Patty Freedson. An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer. *Journal of Applied Physiology*, 107(4):1300–1307, 2009.

[15] Gary M. Weiss, Jeffrey W. Lockhart, Tony T. Pulickal, Paul T. Mchugh, Isaac H. Ronan, and Jessica L. Timko. Actitracker: A smartphone-based activity recognition system for improving health and well-being, 2014.

[16] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009.

[17] Jhun-Ying Yang, Jeen-Shing Wang, and Chen Yen-Ping. Using acceleration measurements for activity recognition: An effective learning algorithm for constructing neural classifiers. *Pattern Recognition Letters*, 29(16):2213–2220, 2008.

[18] Yuhuang Zheng. Human activity recognition based on hierarchical feature selection and classification framework. *Journal of Electrical and Computer Engineering 2015*, pages 34–43, 2015.

Chapter 3

Statistical Analysis of Window Sizes And Sampling Rates in Human Activity Recognition[1]

[1] A.H. Niazi, D. Yazdensepas, J. L. Gay, F.W. Maier, L. Ramaswamy, K. Rasheed and M. P. Buman. To be submitted to *HEALTHINF 2017: 10th International Conference on Health Informatics.*

Abstract

Accelerometers are the most common device for data collection in the field of Human Activity Recognition (HAR). This data is recorded at the sampling rate of the accelerometer and then usually separated into time windows before classification takes place. Though the sampling rate and window size can have a significant impact on the accuracy of the trained classifier, there has been relatively little research on their role in activity recognition. This paper presents a statistical analysis on the effect the sampling rate and window sizes on HAR data classification.

The raw data used in the analysis was collected at 100Hz from 77 subjects performing 23 different activities. It was then re-sampled and divided into windows of varying sizes and trained using a single data classifier. A weighted least squares linear regression model was developed and two-way factorial ANOVA was used to analyze the effects of sampling rate and window size for different activity types and demographic categories. Based upon this analysis, we find that 10-second windows recorded at 50Hz perform statistically better than other combinations of window size and sampling rate.

3.1 Introduction

The field of Human Activity Recognition (HAR) is dependent on a variety of instruments for data collection — heart rate monitors, GPS, light sensors, etc. — of which wearable triaxial accelerometers are the most commonly utilized [5, 9]. Accelerometers are commercially available in many formats, from modern smartphones and consumer-grade activity-monitoring products to high-grade research-oriented devices, the consequences of which are wide degrees of quality in data collection for HAR. When preparing for data collection in a HAR study, two aspects of the accelerometer to use should be strongly considered: the placement of the device and the sampling rate at which it gathers data.

The placement of the device depends greatly on the context of the study. Many studies focusing on ambulation activities (walking, running etc.) prefer hip-worn or wrist-worn devices [5], both of which have advantages and disadvantages. Wrist-worn devices have trouble distinguishing lower-body activities (for instance, walking and stair climbing), while hip-worn devices can be problematic when recognizing upper-body activities (for instance, eating and brushing teeth). The impact of sampling rate is discussed in later sections.

Once data has been collected — at a certain sampling rate — it is prepared for classification by extracting relevant features such as means and standard deviations and diving the accelerometer readings into windows. Often, windows of fixed length are used.

Both the sampling rate and window size of data are crucial decisions in HAR which directly affect the accuracy of developed classifiers. Though a literature review revealed some relevant analyses (Section 3.2), there appears to be a relative dearth of work directly addressing sampling rate and window size in HAR. This study is an attempt to remedy what we perceive as a gap in the research. We have attempted to statistically identify the window size and sampling rate combination which best suits activity recognition across demographical and activity divisions.

The data used in this study was obtained from 77 demographically diverse subjects for 23 activities in studies performed at Arizona State University in 2013 and 2014. Data was

collected from a single hip-worn triaxial accelerometer, an ActiGraph GT3X+, at a sampling rate of 100Hz. By artificially downsampling the data and creating differently sized windows, we have obtained datasets at a cross section of 6 window sizes and 5 sampling rates. We used a single classifier to train these datasets with 10-fold cross-validation and statistically observed the trends using repeated measures two-way ANOVA. We then further divided these datasets to observe how these effects change due to activity type or demographic features of the subject.

It should be noted that this study, by necessity, takes into account only certain aspects of HAR classification process. For example, we are utilizing data from a single hip-worn accelerometer, as opposed to other or multiple placements. Similarly, we use only time- and frequency-based features with a single classifier (Random Forests) to further standardize our tests. While feature sets and classifier selection certainly play a role in the outcomes of HAR classification research [9], to account for all of them would lead to an unworkable level of complexity.

Section 3.2 details the literature available in this domain. Section 3.3 describes the data collection and preprocessing done to the data to obtain our data sets. Section 3.4 gives the results of our classification and statistical analysis of these results. Finally, Section 3.5 states what we conclude from this work and how these conclusions can be implemented in HAR data classification.

3.2 Related Work

While a considerable amount of research has been done in HAR using accelerometers, there has been a lack of consensus on the methodology of collecting and preprocessing data and thus have largely remained unanalyzed [9]. Lara & Labrador [5] note that sampling rates in HAR studies vary from 10Hz to 100Hz while window sizes range from less than 1 second to 30 seconds. While there are some domain-related justifications for such decisions, there is a lack of standardization which likely impacts replicability.

Lau & David [6] attempted a study similar to ours, in the sense that multiple data sets of differing window sizes (0.5, 1, 2 and 4 seconds) and sampling rates (5,10,20 and 40 Hz) were generated from raw accelerometer data (gathered from a pocketed smart phone) and the effects studied. While they claim that these lower values are sufficient for good performance, their setup consisted of a single test subject performing 5 activities. Maurer et al. [8], using 6 subjects, state that recognition accuracy does not significantly increase at sampling rates above 15-20Hz when their biaxial accelerometer is used in conjunction with 3 other sensors (light, temperature and microphone). Bieber et al. [3] calculate that 32Hz should be the minimum sampling rate given human reaction time. Tapia et al. [11] varied window length from 0.5 to 17 seconds and tested the data sets with C4.5 decision tree classifier, concluding that 4.2 seconds was the optimum window size for their needs. Banos et al. [2] created data sets with window sizes ranging from 0.25 to 7 seconds at interval jumps of 0.25. They found that 1-2 seconds is the best trade-off speed and accuracy for online training. Larger windows were only needed if the feature set was small.

Statistical analysis of classifier performance appears rarely performed. Most studies, such as the ones cited above, simply state a performance measure (often accuracies and f-measures) but do not present any statistical evaluation. Demsar [4] comments on the lack of statistical analysis of classifier performance and suggests non-parametric tests for comparing classifiers and data sets.

3.3 Data Collection, Preprocessing and Methodology

3.3.1 Collecting Data

The data used in the present study was collected in Phoenix, AZ from volunteers recruited through Arizona State University. Participants were fitted with an ActiGraph GT3X+ activity monitor positioned along the anterior axillary line of the non-dominant hip. The monitor was fixed using an elastic belt. The ActiGraph GT3X+ [1] is a lightweight monitor (4.6cm x 3.3cm x 1.5 cm, 19g) that measures triaxial acceleration ranging from -6g to +6g.

Devices were initialized to sample at a rate of 100hz. Accelerometer data was downloaded and extracted using Actilife 5.0 software (ActiGraph, LLC, Pensacola, FL). The subjects performed a number of activities which can be observed in Table 2.1.

Table 3.1: Number of Records in the Datasets

Window Size (s)	No. of Records
1	175284
2	88557
3	59666
5	36533
10	19186

310 subjects participated in the study. From them, data from 77 subjects, 53 females and 24 males, was used to train the classifiers. Table 2.1 describes the activities performed while Table 2.2 provides demographic information on the subjects.

3.3.2 Generating Datasets

The data obtained was collected at 100Hz. From this, 30 data sets with varying window sizes (of 1, 2, 3, 5 and 10 seconds) with sampling rates (5, 10, 20, 25, 50 and 100Hz) were created. To create data sets for sampling rates < 100Hz, we downsampled from the original data sets, e.g., 50Hz is generated by using every 2nd accelerometer record (100/50), 25Hz using every 4th record (100/25), etc. The number of records in a window then depends on the sampling rate as well as the window size. E.g., A 1-second window at 100 Hz contains 100 records (100x1), a 3-second window at 25Hz contains 75 records (3x25), and so on. As summarized in Table 3.1, the window size effects the number of records in the data set, a fact that will become significant during analysis.

It should also be noted that, in some situations, partial windows are formed, where there is not enough data for a complete window. Such partial windows were discarded for the sake of consistency.

3.3.3 Feature Extraction and Selection

246 features were extracted using the raw accelerometer data which were then reduced to a 32 feature data set with time- and frequency-based features, which are listed below. More information about the feature extraction and selection can be found in Chapter 2.

- **Features in the time domain**: These features include the mean, standard deviation and 50th percentile of each axis (x, y and z) and their vector magnitude as well as the correlation values between the axes.
- **Features in the frequency domain**: These features include the dominant frequency and its magnitude for each axis (x, y and z) as well as their vector magnitude.

3.3.4 Methodology

Random forest classifiers perform very well with this data set as seen in Chapter 2 and so this was chosen as our standard classifier. Each data set was divided and evaluated in 10 folds. Further divisions were carried out for certain activity groups (see Table 2.3) or demographic groups. The accuracy on the test fold was recorded. WEKA software packages [12] were used in conjunction with Java for training and testing the data sets.

RStudio [10] was used to evaluate results. A two-way factorial ANOVA was carried out with weighted least squares to calculate the expected average value (EV) for every combination. It was found that window size and sampling rate as well as their interaction were statistically significant. By determining the maximum expected accuracy (the maximum EV), we discovered the accuracy remained significant at the 95% confidence level. The next section details the analysis and results of our experiments.

3.4 Statistical Analysis of Results

3.4.1 Weighting

From Table 3.1, it is clear that window size directly affects the number of records in the data set. Table 3.2 shows that the variance increases as window size increases, and so the weighting function should be inversely proportional to the variance. We use 1/WindowSize as a good approximation.[2] Although sampling rate can also be seen to have a small effect on the variance, it appears negligible. All experiments use this weighting function to normalize the distributions.

Table 3.2: Standard Deviations

		Sampling Rate (Hz)					
		5	10	20	25	50	100
Window Size (s)	1	0.0035	0.0034	0.0032	0.0029	0.0027	0.0021
	2	0.0051	0.0031	0.0048	0.0032	0.0057	0.0032
	3	0.0049	0.0071	0.0076	0.0066	0.0040	0.0054
	5	0.0045	0.0057	0.0092	0.0108	0.0107	0.0071
	10	0.0091	0.0129	0.0074	0.0082	0.0098	0.0096

Subsection 3.4.2 describes in detail the statistical process followed by all the experiments.

3.4.2 All Activities and Demographics

Our first test evaluated all the data available, i.e., for 23 activities as performed by 77 subjects. The objective was to find the maximum average expected value (EV) and use this to determine if other values can be considered statistically significant. A two-way analysis of variance (ANOVA) on a Weighted Least Squares (WLS) linear regression model shows that both window size and sampling rate have a significant effect on accuracy with 99% confidence ($p<0.001$), which has been found true for all our experiments. The linear model is then used to obtain EVs for all window size/sampling rate combinations. These values are show in Table 3.3.

[2]The weighting scheme was chosen after a consultation with the University of Georgia Statistics Consulting Center.

Table 3.3: All Activities/Demographics

		Sampling Rate (Hz)					
		5	10	20	25	50	100
Window Size (s)	1	0.5858	0.6868	0.7893	0.8050	0.8251	0.8292
	2	0.6324	0.7355	0.8219	0.8334	0.8456	0.8435
	3	0.6544	0.7551	0.8269	0.8385	0.8488	0.8411
	5	0.6848	0.7752	0.8322	0.8379	0.8473	0.8282
	10	0.7316	0.8050	0.8474	**0.8529**	**<u>0.8583</u>**	0.8126

The $10s/50Hz$ data set has the highest expected value (EV_{max}) for accuracy (in **<u>bold underline</u>** in Table 3.3) in this experiment. Next we determine if other accuracy EVs are significantly different than the maximum EV_{max}. As the alternate hypothesis is that other combinations will have lower EVs, we use a 1-sided interval with a 95% confidence interval.

$$\bar{X_{max}} - \bar{X_k} = t_{290,0.95} * \sqrt{MSE} * \sqrt{\frac{WS_{max}}{n_{max}} + \frac{WS_k}{n_k}} \tag{3.1}$$

Equation 3.1 is used to find the critical distance when the sample sizes are unequal but the variance is assumed equal. As each EV represents 10 folds, we have 290 degrees of freedom. The value of $t_{290,0.95}$ is found as 1.651 using a t-table. The MSE value is obtained from ANOVA. WS represents window size of EV_{max} while WS_k and n is the number of observations which in our case is always 10. Having found the critical distance, we can observe which EV values fall inside the margin.

In this experiment, the $10s/25Hz$ value (in **bold** in Table 3.3) is less than the critical distance away from EV_{max}. Hence, it can be concluded that it is statistically as good as EV_{max} with 95% confidence.

The procedure elaborated in this section is replicated for all of the following experiments.

Table 3.4: Ambulatory vs. Non-Ambulatory Activites

		Sampling Rate (Hz)					
		5	10	20	25	50	100
Window Size (s)	1	0.6408	0.7295	0.8228	0.8369	0.8559	0.8590
	2	0.6812	0.7735	0.8521	0.8634	0.8754	0.8730
	3	0.7016	0.7957	0.8605	0.8688	0.8791	0.8725
	5	0.7319	0.8127	0.8656	0.8727	0.8796	0.8634
	10	0.7792	0.8419	0.8805	**0.8876**	**0.8913**	0.8537

Table 3.5: Ambulatory Activity Groups

		Sampling Rate (Hz)					
		5	10	20	25	50	100
Window Size (s)	1	0.8345	0.8720	0.9065	0.9106	0.9165	0.9170
	2	0.8345	0.8872	0.9155	0.9177	0.9219	0.9195
	3	0.8609	0.8951	0.9181	0.9211	0.9254	0.9200
	5	0.8754	0.9045	0.9237	0.9267	0.9293	0.9180
	10	0.9022	0.9264	0.9412	**0.9411**	**0.9440**	0.9169

3.4.3 Activity Groups

In Table 3.4, ambulatory activities were separated from non-ambulatory activities while in Table 3.5 they were classified as walking, running or stairclimbing activities. Both experiments represent a macro-classification and as such exhibit similar patterns to Table 3.3 — the $10s/50Hz$ has EV_{max}.

Tables 3.7-3.12 show the results of experiments on different activity group classifications. These groups were divided as shown in Table 3.5.

However, classifications at a micro-level, within these activity groups, exhibit different results. Classifying between ascending and descending stairs (Table 3.6) achieves EV_{max} of 97% at $2s/50Hz$ but has a wide spread of equally significant values. Interestingly data at lower sampling rates are also deemed significant for larger window sizes. Statistical values

Table 3.6: Stairs: Ascent vs. Descent

Window Size (s) \ Sampling Rate (Hz)	5	10	20	25	50	100
1	0.9555	0.9640	0.9675	0.9682	**0.9690**	**0.9694**
2	0.9599	0.9652	0.9681	**0.9686**	**0.9697**	**0.9690**
3	0.9611	0.9651	0.9670	0.9675	**0.9690**	0.9673
5	0.9618	0.9655	0.9668	0.9672	0.9670	0.9647
10	0.9650	**0.9676**	**0.9676**	**0.9690**	**0.9687**	0.9624

for non-ambulatory activities (Table 3.7) show similar patterns. For walking and running activities, the spread is smaller and concentrated towards higher sampling rates, though there is a lot of variance in window size. Running in particular prefers smaller windows. This is in agreement with the claim by Bieber, et al. [3] that the sampling rate should be more than 32Hz for ambulatory activities.

Table 3.7: Non-Ambulatory Activites

Window Size (s) \ Sampling Rate (Hz)	5	10	20	25	50	100
1	0.7854	0.8298	0.8609	0.8647	0.8711	0.8723
2	0.8086	0.8471	0.8726	**0.8783**	**0.8795**	**0.8775**
3	0.8161	0.8476	0.8734	0.8732	0.8780	0.8746
5	0.8246	0.8525	0.8682	**0.8730**	**0.8726**	0.8594
10	0.8406	0.8571	0.8713	**0.8716**	**0.8716**	0.8514

Table 3.8: Walking Activites

Window Size (s) \ Sampling Rate (Hz)	5	10	20	25	50	100
1	0.5556	0.6656	0.7916	0.8105	0.8329	0.8385
2	0.5976	0.7162	0.8274	0.8407	**0.8581**	**0.8574**
3	0.6189	0.7415	0.8344	**0.8460**	**0.8598**	0.8543
5	0.6474	0.7594	0.8374	0.8408	0.8527	0.8353
10	0.6875	0.7746	0.8387	0.8491	**0.8557**	0.8159

Table 3.9: Running Activites

		Sampling Rate (Hz)					
		5	10	20	25	50	100
Window Size (s)	1	0.7081	0.7795	0.8522	0.8688	0.9070	**0.9140**
	2	0.7349	0.8191	0.8793	0.8961	**0.9185**	**0.9210**
	3	0.7418	0.8321	0.8891	0.8968	**0.9176**	**0.9177**
	5	0.7584	0.8266	0.8703	0.8863	0.8953	0.8972
	10	0.7728	0.8333	0.8639	0.8714	0.8759	0.8553

3.4.4 Demographics

Table 3.10: Gender: Female Subjects

		Sampling Rate (Hz)					
		5	10	20	25	50	100
Window Size (s)	1	0.6037	0.7132	0.8128	0.8227	0.8405	0.8430
	2	0.6509	0.7606	0.8388	0.8490	**0.8599**	0.8554
	3	0.6762	0.7762	0.8433	0.8529	**0.8598**	0.8498
	5	0.7052	0.7937	0.8441	0.8490	0.8539	0.8351
	10	0.7521	0.8164	**0.8586**	**0.8595**	**0.8667**	0.8169

Table 3.11: Gender: Male Subjects

		Sampling Rate (Hz)					
		5	10	20	25	50	100
Window Size (s)	1	0.6439	0.7248	0.8139	0.8265	0.8474	0.8508
	2	0.6857	0.7633	0.8412	0.8506	**0.8653**	0.8624
	3	0.7017	0.7815	0.8478	0.8569	**0.8675**	0.8597
	5	0.7226	0.7984	0.8484	0.8547	**0.8641**	0.8408
	10	0.7759	0.8183	**0.8636**	**0.8678**	**0.8736**	0.8253

For the next round of experiments, data was separated into demographic groups to observe any significant effects. The data sets were then used to classify all 23 activities. Division by gender, female (53 subjects) and male (24 subjects) (Tables 3.10 and 3.11 respectively) display similar results. EV_{max} is at $10s/50Hz$ for both experiments and there are very similar spreads in significant results. This indicates that there is an insignificant difference in HAR for genders and activity classification should be generalized for both cases.

-

Table 3.12: Age: 18-26 Years

		Sampling Rate (Hz)					
		5	10	20	25	50	100
Window Size (s)	1	0.6207	0.7174	0.8094	0.8236	0.8432	0.8457
	2	0.6662	0.7620	0.8362	0.8488	0.8588	0.8553
	3	0.6857	0.7824	0.8443	0.8559	0.8629	0.8551
	5	0.7196	0.8024	0.8484	0.8542	0.8623	0.8424
	10	0.7633	0.8292	**0.8627**	**0.8717**	**0.8753**	0.8250

Table 3.13: Age: 27-33 Years

		Sampling Rate (Hz)					
		5	10	20	25	50	100
Window Size (s)	1	0.6614	0.7513	0.8343	0.8428	0.8590	0.8618
	2	0.7043	0.7891	0.8564	0.8676	**0.8746**	**0.8731**
	3	0.7198	0.8051	0.8623	0.8678	**0.8779**	0.8677
	5	0.7390	0.8117	0.8573	0.8643	0.8679	0.8488
	10	0.7784	0.8292	0.8658	**0.8695**	**0.8720**	0.8250

Table 3.14: Age: 34-44 Years

		Sampling Rate (Hz)					
		5	10	20	25	50	100
Window Size (s)	1	0.6651	0.7660	0.8442	0.8547	0.8689	0.8722
	2	0.7085	0.8038	0.8654	0.8730	**0.8849**	**0.8805**
	3	0.7271	0.8193	0.8651	0.8730	**0.8807**	0.8696
	5	0.7482	0.8226	0.8596	0.8624	0.8721	0.8533
	10	0.7833	0.8424	0.8733	**0.8792**	**0.8822**	0.8375

Data was then divided into 4 age groups; $18-25$ (24 subjects), $26-32$ (24 subjects), $33-44$ (21 subjects) and $49-63$ (8 subjects). The results of these experiments are recorded in Tables 3.15-3.18, respectively. There is a visible trend of decreasing window size with increasing age. The spread of significant values gets larger as well.

Similar patterns are noted when the data is divided according to Body Mass Index (BMI) categories; Normal (40 subjects), Overweight (28 subjects) and Obese (9 subjects) (Tables 3.19-3.21). As BMI increases, the significance of the EV_{max} decreases along with the window size. Subjects with lower BMIs fare better with larger windows than those with higher BMIs.

Table 3.15: Age: 49-63 Years

		Sampling Rate (Hz)					
		5	10	20	25	50	100
Window Size (s)	1	0.7593	0.8382	0.8892	0.8981	0.9065	0.9063
	2	0.7856	0.8581	0.9043	**0.9084**	**0.9135**	**<u>0.9146</u>**
	3	0.8046	0.8689	0.9040	**0.9067**	**0.9101**	0.9030
	5	0.8201	0.8730	0.9031	0.9017	**0.9084**	0.8855
	10	0.8503	0.8986	**0.9114**	**0.9114**	**0.9119**	0.8725

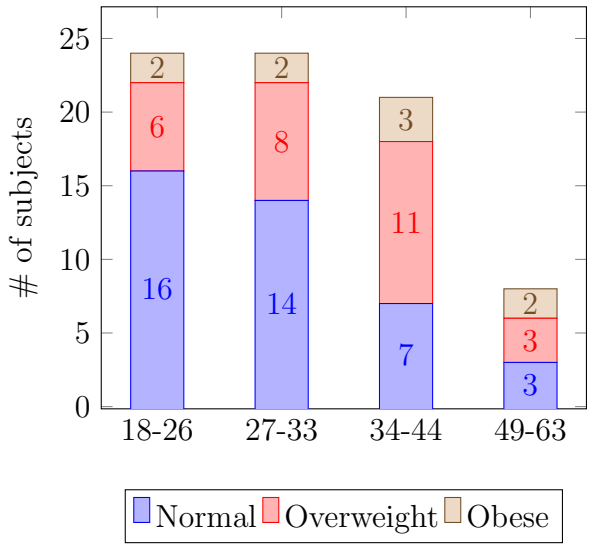

Figure 3.1: Distribution of BMI groups over age groups

This can suggest a correlation between age and BMI - elderly people are less likely to be active than young people and are thus more likely to have high BMIs. This hypothesis is supported in Figure 3.1 which shows that the proportion of normal weighted people decreases with age in the dataset.

Table 3.16: BMI: Normal

Window Size (s) \ Sampling Rate (Hz)	5	10	20	25	50	100
1	0.6031	0.7074	0.8056	0.8188	0.8363	0.8393
2	0.6531	0.7525	0.8320	0.8437	0.8531	0.8503
3	0.6776	0.7753	0.8395	0.8493	0.8553	0.8478
5	0.7138	0.7946	0.8446	0.8482	0.8549	0.8376
10	0.7617	0.8204	**0.8614**	**0.8615**	**0.8678**	0.8149

Table 3.17: BMI: Overweight

Window Size (s) \ Sampling Rate (Hz)	5	10	20	25	50	100
1	0.6419	0.7381	0.8256	0.8391	0.8564	0.8597
2	0.6831	0.7762	0.8520	0.8609	**0.8714**	**0.8689**
3	0.7002	0.7940	0.8523	0.8612	**0.8701**	0.8637
5	0.7225	0.8064	0.8549	0.8619	**0.8696**	0.8494
10	0.7612	0.8287	0.8607	**0.8674**	**0.8732**	0.8252

Table 3.18: BMI: Obese

Window Size (s) \ Sampling Rate (Hz)	5	10	20	25	50	100
1	0.7423	0.8279	0.8803	0.8900	0.8998	0.9015
2	0.7817	0.8532	0.9008	0.9039	**0.9164**	**0.9115**
3	0.7968	0.8648	0.9010	**0.9098**	**0.9167**	**0.9098**
5	0.8164	0.8663	0.8943	0.9001	0.9070	0.8878
10	0.8368	0.8774	0.8994	**0.9125**	**0.9091**	0.8648

3.4.5 Summary of Analysis

Viewing all experiments together suggests that $10s/50Hz$ is the optimal combination of window size and sampling rate, especially if the subjects of the study are young, able-bodied and physically active. Most high significant EV are spread around high sampling rates and window sizes, although there is enough evidence to suggest there is not a very significant loss in accuracy if the sampling rate is decreased to $25Hz$ or window size is decreased to $2s$.

3.5 Conclusion

This study provides some basis for the selection of sampling rates and window sizes for human activity recognition. The analysis indicates that $10s/50Hz$ is statistically the best combination for data collected with a hip-worn Actigraph GT3X+. Most of the experiments carried out preferred larger windows and high sampling rates though some low intensity activities and demographics can perform better with smaller windows. Our analysis further suggests that window size can vary between 2-10 seconds and sampling rate 25-100Hz for different situations without a significant loss in performance. While our study has shown that larger windows are preferable, smaller windows can still provide significant results if power consumption is an issue. Additionally, lower values are preferable for studies involving the less dynamic activities or subjects who are more liable to be less active.

Future work in this field should be done to understand aspects of Human Activity Recognition better. A different set of sensors could be used in a similar study. Feature set and classifier optimality can also be tested.

3.6 References

[1] Actisoft analysis software 3.2 user's manual. Fort Walton Beach, FL: MTI Health Service

[2] Oresti Banos, Juan-Manuel Galvez, Miguel Damas, Hector Pomares, and Ignacio Rojas. Window size impact in human activity recognition. *Sensors*, pages 6474–6499, 2014.

[3] Gerald Beiber, Jrg Voskamp, and Bodo Urban. Activity recognition for everyday life on mobile phones. *International Conference on Universal Access in Human-Computer Interaction*, pages 289–296, 2009.

[4] Janez Demsar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[5] Oscar D. Lara and Miguel A. Labrador. A survey on human activity recognition using wearable sensors. *IEEE Communication Surveys and Tutorials*, 15:1192–1209, 2013.

[6] S.L. Lau and K. David. Movement recognition using the accelerometer in smartphones. *IEEE Future Network ands Mobile Summit 2010*, pages 1–9, 2010.

[7] Jeffrey W. Lockhart and Gary M. Weiss. The benefits of personalized smartphone-based activity recognition models. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 614–622. 2014.

[8] Uwe Maurer, Asim Smailagic, Daniel P. Siewiorek, and Michael Deisher. Activity recognition and monitoring using multiple sensors on different body position. *International Workshop on Wearable and Implantable Body Sensor Networks*, 2006.

[9] Stephen J Preece, John Y Goulermas, Laurence P J Kenney, Dave Howard, Kenneth Meijer, and Robin Crompton. Activity identification using body-mounted sensorsa review of classification techniques. *Physiological Measurement*, 30(4):R1, 2009.

[10] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA, 2015.

[11] Emmanuel Munguia Tapia, Stephen S. Intille, William Haskell, Kent Larson, Julie Wright, Abby King, and Robert Friedman. Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart rate monitor. *11th IEEE international symposium on wearable computers*, pages 37–40, 2007.

[12] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009.

Chapter 4

Conclusions and Future Research

This study presented a novel method of data classification in the field of Human Activity Recognition. By splitting our classification problem into multiple levels, we can obtain classifiers that train to particular activity types. This allows us to cater to more specificity in activity recognition, e.g., effectively differentiating similar modes of walking.

This hierarchical model can certainly be improved upon. Heavy machine learning techniques such as neural networks and genetic algorithms can be tested for potentially better results. Testing can be carried out on actual free-living data to observe how applicable the model is in the real world. Additionally, in line of the purpose of HAR for free-living data, a streamlined process application of data processing and classification would prove to be very useful for public health research.

This study also described a statistical analysis of the window size and sampling rate of the data and how it has significant effects on the results of HAR, both for the types of activities and for multiple demographics. It could provide a basis of informed literature for selecting window sizes and sampling rates in future studies according to domain requirements and limitations.

We note again that this study used only a single type of classifier (Random Forests), and other restrictions or assumptions were made. As such, while this analysis might be a step in the right direction, there is considerable room left for additional study. E.g., the effects of using other machine learning algorithms and feature sets could be examined, and studies similar to this one could be carried out using wristworn accelerometers, pocket smartphones and other devices.

Bibliography

ActiGraph. Actisoft analysis software 3.2 user's manual. Fort Walton Beach, FL: MTI Health Services.

Oresti Banos, Juan-Manuel Galvez, Miguel Damas, Hector Pomares, and Ignacio Rojas. Window size impact in human activity recognition. *Sensors*, pages 6474–6499, 2014.

Ling Bao and Stephen S. Intille. Activity recognition from user-annotated acceleration data. In *International Conference On Pervasive Computing*, pages 1–17. Springer Berlin Heidelberg, 2004.

Gerald Beiber, Jrg Voskamp, and Bodo Urban. Activity recognition for everyday life on mobile phones. *International Conference on Universal Access in Human-Computer Interaction*, pages 289–296, 2009.

Robert Bodor, Bennett Jackson, and Nikolaos Papanikolopoulos. Vision-based human tracking and activity recognition. *Proceedings of the 11th Mediterranean Conference on Control and Automation*, 1, 2003.

Leo Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

Leo Breiman. Random forests. *Machine Learning*, 14:5–32, 2001.

Keith M. David, David J. Krupa, Melinda J. Chang, James Peacock, Yao Ma, Jeff Goldsmith, Joseph E. Schwartz, and Karina W. Davidson. Fitbit: An accurate and reliable device for wireless physical activity tracking. *International journal of cardiology*, 185:138–140, 2015.

Janez Demsar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

Wan-Yu Deng, Qing-Hua Zheng, and Zhong-Min Wang. Cross-person activity recognition using reduced kernel extreme learning machine. *Neural Networks*, 53:1–7, 2014.

Ronald A. Fisher. *Statistical methods for research workers*. Genesis Publishing Pvt Ltd, 1925.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009.

Mark A Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

Tin Kam Ho. Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 1, pages 278–282, Aug 1995.

Ning Jia. Detecting human falls with a 3-axis digital accelerometer. *A forum for the exchange of circuits, systems, and software for real-world signal processing*, 43:719–722, 2009.

Adil Mehmood Khan, Young-Koo Lee, Sungyoung Y. Lee, and Tae-Seong Kim. A triaxial accelerometer-based physical-activity recognition via augmented-signal features and a hierarchical recognizer. *IEEE Transactions on Information Technology in Biomedicine*, 14(5):1166–1172, 2010.

Kenji Kira and Larry A. Rendell. A practical approach to feature selection. In *Proceedings of the Ninth International Workshop on Machine Learning*, ML92, pages 249–256, San Francisco, CA, USA, 1992. Morgan Kaufmann Publishers Inc.

Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 2:1137–1143, 1995.

Jennifer R. Kwapisz, Gary M. Weiss, and Samuel A. Moore. Activity recognition using cell phone accelerometers. *SIGKDD Explor. Newsl.*, 12(2):74–82, March 2011.

Oscar D. Lara and Miguel A. Labrador. A survey on human activity recognition using wearable sensors. *IEEE Communication Surveys and Tutorials*, 15:1192–1209, 2013.

S.L. Lau and K. David. Movement recognition using the accelerometer in smartphones. *IEEE Future Network ands Mobile Summit 2010*, pages 1–9, 2010.

Jonathan Lester, Tanzeem Choudhury, and Gaetano" Borriello. Pervasive computing: 4th international conference, pervasive 2006, dublin, ireland, may 7-10, 2006. proceedings. In *A Practical Approach to Recognizing Physical Activities*, pages 1–16. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

Weiyao Lin, Ming-Ting Sun, Radha Poovandran, and Zhengyou Zhang. Human activity recognition for video surveillance. *IEEE International Symposium on Circuits and Systems*, pages 2737–2740, 2008.

Jeffrey W. Lockhart and Gary M. Weiss. The benefits of personalized smartphone-based activity recognition models. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 614–622. 2014.

Mitja Lutrek, Hristijan Gjoreski, Simon Kozina, Boidara Cvetkovi, Violeta Mirchevska, and Matja Gams. Detecting falls with location sensors and accelerometers. *Twenty-Third IAAI Conference*, pages 1662–1667, 2011.

Uwe Maurer, Asim Smailagic, Daniel P. Siewiorek, and Michael Deisher. Activity recognition and monitoring using multiple sensors on different body position. *International Workshop on Wearable and Implantable Body Sensor Networks*, 2006.

Tom M. Mitchell. *Machine Learning*. Tim Mc-Graw-Hill Companies, Inc., 1997.

Wei Niu, Jiao Long, Dan Han, and Yuan-Fang Wang. Human activity detection and recognition for video surveillance. *Proceedings of IEEE International Conference on Multimedia and Expo*, 1:719–722, 2004.

Thomas Phan. Improving activity recognition via automatic decision tree pruning. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, UbiComp '14 Adjunct, pages 827–832, New York, NY, USA, 2014. ACM.

Stephen J Preece, John Y Goulermas, Laurence P J Kenney, Dave Howard, Kenneth Meijer, and Robin Crompton. Activity identification using body-mounted sensorsa review of classification techniques. *Physiological Measurement*, 30(4):R1, 2009.

Nishkam Ravi, Nikhil Dandekar, Preetham Mysore, and Michael L. Littman. Activity recognition from accelerometer data. In *Proceedings of the 17th Conference on Innovative Applications of Artificial Intelligence - Volume 3*, IAAI'05, pages 1541–1546. AAAI Press, 2005.

RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA, 2015.

Steven L. Salzberg. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data mining and knowledge discovery*, 1:317–328, 2007.

John Staudenmayer, David Pober, Scott Crouter, David Bassett, and Patty Freedson. An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer. *Journal of Applied Physiology*, 107(4):1300–1307, 2009.

Emmanuel Munguia Tapia, Stephen S. Intille, William Haskell, Kent Larson, Julie Wright, Abby King, and Robert Friedman. Real-time recognition of physical activities and their

intensities using wireless accelerometers and a heart rate monitor. *11th IEEE international symposium on wearable computers*, pages 37–40, 2007.

Gary M. Weiss, Jeffrey W. Lockhart, Tony T. Pulickal, Paul T. Mchugh, Isaac H. Ronan, and Jessica L. Timko. Actitracker: A smartphone-based activity recognition system for improving health and well-being, 2014.

Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kauffman, 2005.

Falin Wu, Hengyang Zhao, Yan Zhao, and Haibo Zhong. Development of a wearable-sensor-based fall detection system. *11th IEEE international symposium on wearable computers*, 2, 2015.

Jhun-Ying Yang, Jeen-Shing Wang, and Chen Yen-Ping. Using acceleration measurements for activity recognition: An effective learning algorithm for constructing neural classifiers. *Pattern Recognition Letters*, 29(16):2213–2220, 2008.

Jerrold H. Zar. *Biostatistical analysis*. Pearson Education, India, 1997.

Yuhuang Zheng. Human activity recognition based on hierarchical feature selection and classification framework. *Journal of Electrical and Computer Engineering 2015*, pages 34–43, 2015.