HUMAN ACTIVITY RECOGNITION USING PSEUDO FREE-LIVING DATA

By

JOSHUA SHANNON

(Under the Direction of Frederick Maier)

Abstract

Human activity recognition involves the learning and classification of various activities performed in daily life. However, most research has focused on using carefully obtained data collected in a supervised, laboratory setting that is far from representative of data collected in real-world conditions. Therefore, this project has investigated several machine learning models and studied the hyperparameters involved in human activity recognition on a pseudo free-living dataset collected at the University of Georgia. On this data set, we found that standard, flat models outperformed hierarchical metaclassifiers with MLPs and SVMs achieving 64.62% and 63.95 percent accuracy on classifying 7 different activities. Ensemble models achieved only marginally better results. A window size of 10 seconds was found to be ideal for this dataset, and participant pre-training was revealed to be a promising method for improving classification accuracy. It was observed that activities on an incline, such as ascending and descending stairs, proved the most difficult to classify. Excluding this class improved accuracy to 80%, while folding it into the “walking comfortably” class further increased accuracy to 85.9%.

INDEX WORDS: Supervised Classification, human activity recognition, support vector machines, multi-layer perceptron, machine learning

HUMAN ACTIVITY RECOGNITION USING PSEUDO FREE-LIVING DATA

by

JOSHUA SHANNON

B. Eng, Vanderbilt University,

Nashville, TN, 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, GA

2019

HUMAN ACTIVITY RECOGNITION USING PSEUDO FREE-LIVING DATA

by

JOSHUA SHANNON

Major Professor:	Frederick Maier
Committee:	Khaled Rasheed Jennifer Gay

Electronic Version Approved:

Ron Walcott
Interim Dean of the Graduate School
The University of Georgia
December 2019

ACKNOWLEDGEMENTS

I would like to thank Dr. Maier for being my major advisor on this project, and for his invaluable insights and patience in helping me during my work on this thesis. He has always been encouraging at every point during my time here, and he is a great instructor as well. I would also like to express my gratitude to Dr. Rasheed for his support and advice for this thesis. He is an excellent teacher and I've enjoyed all three courses that I've taken under him. I would also like to thank Dr. Gay for agreeing to be on my committee and for her guidance and advice as well. Finally, thanks to my family and friends for their support, especially my parents who instilled in me the importance of knowledge and education, and who supported me throughout my entire life.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

This thesis involves the study of human activity recognition based on a body-worn triaxial accelerometer, and specifically encompasses the methods and models used on a dataset collected in a setting that occurred under pseudo free-living conditions rather than in a controlled, laboratory environment. Much research has focused on these clean datasets, which may be useful as benchmarks or for prototyping new algorithms but have limited practical applications because they do not accurately reflect the challenges of human activity recognition in free-living or pseudo free-living environments that are more representative of the real world. Thus, this study investigates several different machine learning models and analyzes their performance on a dataset that has been collected under more realistic conditions with numerous participants. Many hyperparameters, such as window size, are studied along with methods that aim to improve classification accuracy and reduce noise. Figure 1.1 provides a list of experiments performed and groups them into two categories: experiments focused on machine learning approaches, and those focuses on data and hyperparameter analysis.

1.1 Machine Learning and Data Analysis

The dataset for this thesis was obtained from 20 participants wearing a hip-worn accelerometer, which provided acceleration information in all three axial directions. A total of seven distinct activities were performed by each patient and recorded by a researcher, and each

participant performed them in a natural manner without input from the researcher in order to generate a more realistic dataset.

The differences in how participants performed activities allowed us to realistically test inter-subject accuracy, which is the strategy of training machine learning models on every participant except one, and then using that individual to test the performance of the classifier. This is a practical approach for human activity recognition because only one general model is trained before it can be used on new participants. Intra-subject validation occurs when the participants' data is randomly partitioned into training and testing sets so no variance between participants is considered. This approach would require people to individually train models and is less convenient.

1.1.1 Outlier detection

A brief, initial experiment was performed to determine if there were any outlier participants present in the data. Analysis showed that participant 16 only had 28.9% accuracy when their data was used in the test set, so this individual was excluded from the experiments presented in this paper and only 19 participants were used.

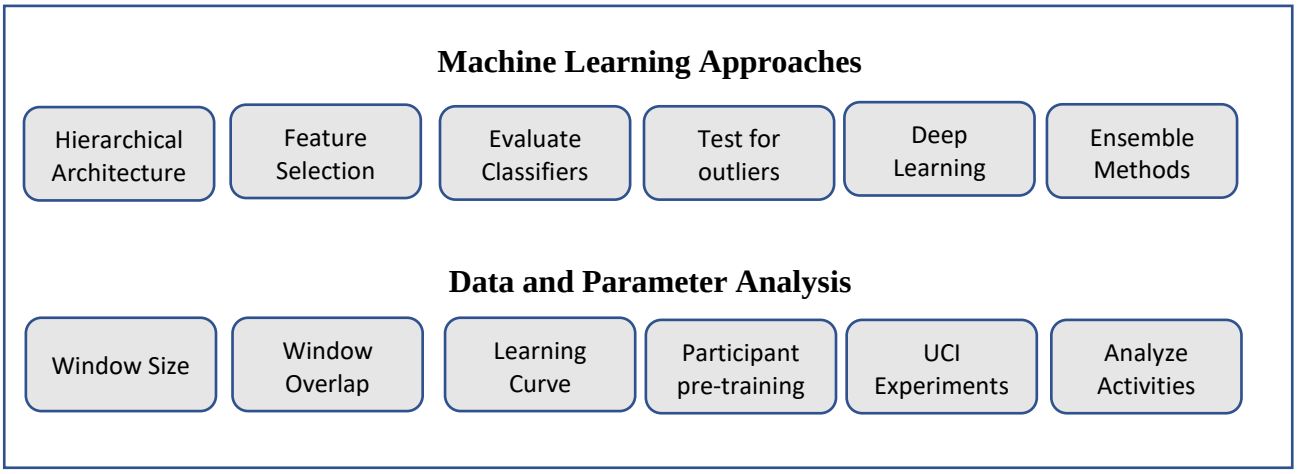

Figure 1.1: Experiments performed in this thesis. All 12 fall under two categories, machine learning approaches and data analysis

1.1.2 Hierarchical Classifier

This work is based earlier work of Niazi et al. [25], described in section 2.2.5, which used a hierarchical meta-classifier. A similar hierarchical classifier was initially used to classify our own dataset. Performance was evaluated based on inter-subject accuracy, where a single participant's data is left out of the training set and used only in the testing dataset; data for the remaining participants is used to train the classifier. The initial, 3-level hierarchy presented in Niazi [25] (figure 2.2) was used on this dataset with random forest classifiers used for all 5 base classifiers, achieving accuracies of 98.11%, 70.84%, and 54% for the first, second, and third levels respectively. After studying the confusion matrix, an alternative structure to the hierarchy was developed and all three levels saw an improvement in performance, with the top level reaching 98.25% accuracy, the second level had 84.24% and the third achieved 57.97% accuracy. This indicated that Niazi's [25] initial hierarchy did not generalize well to inter-subject testing, and that an alternative structure was able to attain higher accuracies across every level.

1.1.3 Feature Engineering

A total of 52 features in the time and frequency domain were extracted from the raw data and is described in more detail in section 3.3. Four different feature selection methods were then applied: expert-based, K-Best, correlation-based, and recursive feature selection. The expert-based and K-Best methods produced 16 features, correlation-based feature selection resulted in 21 features, and recursive feature elimination (RFE) resulted in 42 being used from the initial 52. Various machine learning models have different sensitivities to redundant or correlated features, so to properly test the feature subsets three different base classifiers were used within the hierarchy: support vector machines (SVMs), random forests, and multi-layer perceptrons (MLPs). It was found that the expert-selected features performed the worst, with the highest

level-3 classification accuracy being 57.97% when random forests were used, while the full feature set with no feature selection performed quite well, achieving the second highest classification accuracy at 63.94% when SVMs were used in the hierarchy. In this set of experiments on pseudo free-living data, RFE performed the best at 65% accuracy.

1.1.4 Evaluation of classifiers

Niazi et al. [25] found that a hierarchical classifier using random forests as base learners achieved the highest classification accuracy on a clean data set, which is described later in this study. Six traditional machine learning models, random forests, MLPs, SVMs, decision trees, XGBoost, and quadratic discriminant analysis (QDA), were tested on our dataset in a flat configuration and as base learners in the hierarchical approach described in Niazi [25]. Results showed that SVMs performed the best within the meta-classifier, achieving 63.94% accuracy at the third level in the hierarchy, while QDAs performed the worst, achieving only 41.59% accuracy. However, MLPs performed the best overall as a single, "flat" classifier which reached 64.62% accuracy on the data. The results indicated that the hierarchical approach didn't convey any benefits over simpler methods for HAR classification on our realistic dataset. A statistical test determined that the flat SVM and MLP classifiers were significantly better than the hierarchical approach used in Niazi et al. [25]. Furthermore, when intra-subject testing was performed via 10-fold cross-validation (discussed in section 3.6), the accuracy increased to 85%, which is comparable to the intra-subject accuracy found in Niazi [25] of 86%. The large difference between inter-subject and intra-subject accuracy highlights the challenges facing machine learning approaches for datasets collected in pseudo free-living environments.

1.1.5 Deep Learning

In a separate experiment, a convolutional neural network (CNN) was constructed with two convolutional layers and two max-pooling layers, followed by a softmax classifier. CNNs are powerful classifiers that can extract feature information from the raw triaxial data, so no manual feature extraction was performed for this classifier. Three participants were randomly chosen for the test group, while the rest were used in training and validation. After 250 epochs, the classification accuracy at the activity level was revealed to be 57.81%, which is below the best performing classifiers, MLPs and SVMs, but is on par with many others.

1.1.6 Ensemble methods

In addition to the classifiers described in previous sections, more ensemble methods were tested on our pseudo free-living dataset, including XGBoost, AdaBoost, Extra-trees, Voting classifier, and a stacking classifier. The voting classifier, based on an SVM, MLP and RF, performed best as part of the meta-classifier hierarchy, where it reached 65.7% accuracy at the activity level, but the stacking classifier, based on the MLP, SVM and voting classifiers, had the best results overall as a flat classifier, reaching 65.9% accuracy while the other ensemble methods achieved poor results. The voting and stacking classifier obtained slightly better results on the data, but at the cost of computational complexity and clarity. It seems unlikely that the configuration of these 2 ensemble approaches would generalize well to other, pseudo free-living datasets which makes any further uses restricted.

1.1.7 Window size experiments

Window size is a hyperparameter of the HAR problem and is defined as the length of time over which feature information is extracted from the body-worn accelerometers. Classifiers

performed the best on nine second windows, reaching a peak of 63% accuracy, while the 15 second window had the worst result at 58% accuracy.

1.1.8 Window Overlap and Learning Curve

Window overlap specifies how much data each successive window shares with the preceding window during feature extraction and can be useful because higher overlap allows more samples to be generated for the machine learning models. A range of values from no overlap to 50% overlap were tested, with results showing little change in classifier performance. A learning curve was also generated by starting with a training set of only one participant and incrementally adding participants until all were included in the training set except one. This test showed that at around 14-16 participants in the training set, the classification accuracy plateaued, indicating that there was sufficient training data present and reinforcing the results seen in the window overlap experiment.

1.1.9 Participant pre-training

Participant pre-training is the strategy where a small percentage of a test participant's movement data is included in training and excluded from the testing set, simulating what happens when a participant briefly trains a machine learning model. This approach aims to reduce inter-subject variability and boost accuracy of classifiers, and it yielded results showing that around 69% accuracy was produced when 30% of a test participant's data was used for training. This is over a 6% increase in classification accuracy over no pre-training at all, showing that this has potential to improve inter-subject accuracy in pseudo free-living datasets.

1.1.10 Experiments with a clean dataset

Anguita et al. [4] published a HAR dataset to UCI, and it has become a well-known, publicly available dataset that is often used as a benchmark for machine learning algorithms. Our

feature extraction methods and machine learning models recorded a 95% classification accuracy when tested on this dataset, indicating that our approaches work well on clean data. We then included this data in our pseudo free-living training set to determine whether this could boost accuracy, but we found that accuracy actually decreased when this was done, indicating that too many differences existed between the datasets for this to improve our results.

1.2.11 Analyzing activity classes

Most of the confusion in our models occurred when trying to classify the "walking up" and "walking down an incline" classes. When these classes were rolled into the "comfortably walking" class, activity classification greatly improved to 85.9% accuracy when using SVMs. However, those classes may be considered too different depending on the application, so we also ran tests to study the effects when the "walking up" and "walking down" instances were completely removed from the data. Results showed that classification accuracy still reached 80%, showing that in pseudo-free living data sets, this may be preferable if accuracy is paramount.

This paper addresses many of the different approaches and challenges when using pseudo free-living data for human activity recognition as compared to a clean data set. It has been found that a hierarchical approach for motion classification offered no benefits over flat, base classifiers, where SVMs and MLPs performed the best. Ensemble methods only slightly improved accuracy, but act as black boxes that are unlikely to generalize well in practical settings. Window size has a significant impact on classification accuracy, with a span of 10 seconds producing the best results. When using a noisier, pseudo free-living dataset, most feature selection methods only degraded performance and only RFE improved accuracy, while participant pre-training was found to be a useful tactic in improving accuracy. Walking on

inclines proved to be the hardest to classify, so folding those classes into the “walking comfortably” class or withholding those instances entirely both boosted accuracy over 80%.

Chapter 2

Motivation, Background, and Related Work

It is common that data collection for HAR happens in a laboratory setting under controlled conditions. Several large, publicly available datasets have been collected this way and are commonly used in HAR studies [4, 18, 24]. Although useful as a benchmark, these results are likely optimistic and are not representative of data acquired in realistic, real-world scenarios or pseudo free-living conditions. It is reasonable to suppose that differences in collection methods are one of the reasons why performance often varies widely in the literature. Many studies also focus solely on intra-subject classification accuracy [5,6,8,9,12,20,24,25,31] which has limited practical use because a new model needs to be trained for each participant. Thus, there is a need to have a thorough study investigating what models and parameters are needed to accurately classify human activities in a pseudo free-living environment for practical applications.

2.1 Background

2.1.1 Human Activity Recognition

Human activity recognition (HAR) involves studying, classifying, and predicting human activities [20] such as running, walking, and standing. HAR has seen a surge of interest in recent years due to ubiquitous access to smartphones and other wearable technologies, along with reduced costs in embedded systems. These devices have greatly increased the access and amount of biometric data, which presents promising opportunities for data science and machine learning. There are wide-ranging applications for this field, including public health monitoring [30], personal activity tracking [5,7,27], and patient risk detection [13], creating a need for robust

models to accurately classify supervised activities performed in a realistic environment. However, there are many different approaches used to collect and analyze HAR data, so a more thorough review is presented below.

2.1.2 Data collection

The manner of data collection can impact machine learning performance, such as reducing the variance in motion between different individuals performing the same activity [23, 30], or helping delineate different activities when a participant changes between them [23]. It is common in the research literature to have data generated from multiple participants, from as few as two [13] to over a hundred [25], performing predefined activities in a laboratory setting while being monitored by a researcher. Typically, each participant performs the activities in a fixed timeframe while the researcher uses a stopwatch to record when each activity begins and ends [8, 14, 22]. Classification using these types of datasets is referred to as supervised learning because the collected data is associated with a specific class (the activity), while only a minority of papers [5,22] have studied the ability to classify unsupervised data. Unsupervised data is significantly easier to collect without the need of a person meticulously labeling data, but the resulting models based on it normally have much lower performance than supervised learning [5, 22].

The number and types of sensors used in data collection also significantly affect the performance of activity classification. Multiple studies have required participants to wear several sensors at different positions on the body [3, 5, 7, 12, 23, 30] such as on the wrists (i.e. smart watches) [12], on the hip [25], or in pockets [20], while others have required only one sensor [4, 8, 13, 20, 25]. Having multiple sensors may lead to better models, but it is unlikely participants would agree to wear them for very long in most scenarios, limiting their practical application.

The most common wearable sensors for HAR are triaxial accelerometers and gyroscopes, which provide acceleration and orientation information along the three spatial axes. One reason for their widespread use is that much of the literature is focused on smartphone applications [8, 20, 22, 27, 31], which each have a native accelerometer and gyroscope. However, smartphones may achieve unreliable results if they're not fixed to the body and may not be realistic if a participant wants to use their phones for other purposes while they're being monitored, and they're not as accurate as research-grade sensors. They also tend to be less reliable than dedicated hardware since background processes may be running.

Although wearable sensors are common, external sensors can also be utilized for HAR. Vision based sensors can be used to study how a participant is moving in a free, unobstructed space and can be used to classify the activity [10,21]. The rise in computer vision accuracy has made this approach feasible, but several hurdles remain. For example, placing cameras in a room raises privacy concerns, and the need for a clear line-of-sight may make it impractical. An approach using WIFI signals has been attempted to overcome these issues [32], but faces problems related to noise and signal attenuation in buildings. This study focuses on the most practical and convenient scenario, which is a single, body-worn accelerometer.

2.1.2 Data Analysis and Hyperparameters

There are several parameters involved in HAR, each of which impacts the performance of any machine learning classifiers. Window size is one such parameter that has been shown to influence the performance in HAR [6] and is defined as the length of time over which feature information is extracted from the body-worn accelerometer. Windows are necessary because a complete human activity can occur over several seconds, so all the information provided by sensors during that timeframe can be used by the machine learning model to more accurately

classify human activities. A small window size commonly seen in the literature is one second [9, 16], while the largest values tend to be around 10 seconds [20]. Due to the lack of a thorough analysis of a free-living HAR dataset, the aim is to properly measure how window size can influence the performance of classifiers in such conditions. Some research [6] indicates a smaller window size increases the number of samples present while also allowing for faster activity recognition, while larger windows increase the ability to recognize more complex activities [6]. As this research is based around the work of Niazi et al. [25] we initially use window sizes of 10 seconds, which they found to produce the best results for a hip-worn accelerometer [25].

Window overlap specifies how much data each successive window shares with the preceding window during feature extraction. Overlap between windows increases the number of samples available to any classifiers, and which may be impactful in some scenarios. In theory, a higher percentage of overlap between windows can mitigate the randomness present when windows are applied to data that contain transitions between activities. However, its influence on accuracy hasn't been studied quite as extensively in the literature, so it is worth investigating, especially in datasets that may have higher noise and inter-subject variability present. Most studies typically use 50% overlap [4, 8, 13] for windows, but can range from no overlap to 80% overlap [5] in cases where more data is necessary.

The sampling rate, varying in the literature from 16Hz [12] to 126Hz [13], is how fast data is collected by a sensor and is measured in samples per second. It can "directly affect power consumption, data storage, and power or bandwidth requirements" [18] and has been shown to influence the performance of activity classifiers [18]. More information is provided with a higher sampling rate, but it increases computation time and may also include noise that can negatively impact performance in machine learning models. The Shannon-Nyquist theorem provides the

lower bound for sampling rate without losing any information, which is determined to be twice the highest frequency present in the signal. Humans tend to move slowly compared to the ability of sensors to collect information, so it is often possible to down-sample the data without losing any information. However, more complex activities generally require a higher sampling rate [29], so it is necessary to consider the objectives in a particular HAR project. A sampling rate that is too low can miss important information in the signal, while a high rate can waste resources and negatively impact accuracy.

The wide range of values of HAR parameters present in the literature often make it difficult to compare results and determine the best approaches for machine learning. Therefore, the purpose of this research is to analyze a dataset that is more realistically representative of uncontrolled conditions for applications related to public health and caloric expenditure. It is necessary to study the parameters involved in HAR in order to maximize the performance of machine learning models, especially with inter-subject accuracy.

2.2 Project Background

2.2.1 Data Collection Methodology

The data for this project was obtained from a single triaxial accelerometer, the ActiGraph GT3X+ [2], which was fastened using an elastic belt to the non-dominant hip so that the three spatial axes were pointing forward, sideways, and straight up. The hip was chosen because it allowed us to more accurately capture lower body movements.

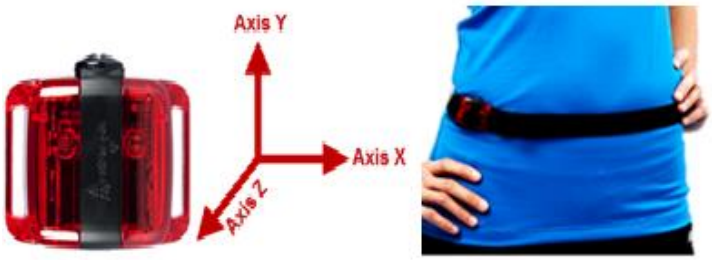

Figure 2.1: The Actigraph GT3X+ accelerometer [1]. On right, it's being worn around the hip, similar to this study

The information provided consisted of the acceleration in all three axial directions sampled at 100 Hz [2] and was obtained from 20 patients performing 9 different activities (Table 2.1) in a non-laboratory, office setting. The participant was given the option of what order to perform the activities and executed them in a manner that seemed natural and appropriate. The aim was to create a dataset that would reflect how people move in real-world settings. The types and number of activities are comparable to other studies, which often range from 3 [12] to over 30 different activities [25]. However, several of the activities are quite similar and produce similar results during feature extraction, so both walking uphill and walking upstairs were combined to create a "walking up an incline" activity. Similarly, walking downhill and walking downstairs were combined to create the new activity of walking down an incline, resulting in 7 total activities provided in Table 2.2.

Table 2.1 Nine Initial Activities Performed by Each Participant

#	**Activity Description**
0	Sitting
1	Standing
2	Walking at a Comfortable Pace
3	Brisk Walking
4	Walking Uphill
5	Walking Downhill
6	Jogging
7	Climbing Upstairs
8	Climbing Downstairs

Table 2.2 Updated List of Seven Activities Performed

#	**Activity Description**
0	Sitting
1	Standing
2	Walking at a Comfortable Pace
3	Brisk Walking
4	Up Incline
5	Down Incline
6	Jogging

This was also motivated by the fact that preliminary experiments indicated that the two different walking up and walking down activities were very similar so that it was difficult to classify them separately.

2.2.2 Classifiers

The classification of data can be performed by several different machine learning schemes. There is no clear indication that any specific model is best-suited for HAR, as SVMs [3, 17, 27], MLPs [8, 20], KNN [5, 6], and random forests (RF) [9, 25] have all performed as the best classifier in different studies. Therefore, it is usually necessary to study many different possible classifiers to obtain accurate results, especially when analyzing a unique dataset. Parameters for the models, such as number of trees in a RF or neurons in a MLP, are tuned to mitigate the possibility of overfitting the training data.

One obvious hurdle in HAR is optimizing classification time with accuracy, as overly complex learning schemes may take too long to be useful in online scenarios. Ideally, we would want a robust, online classifier that analyzed data from a single wearable sensor. Online learning occurs as data is being collected, while offline learning involves processing and classifying data afterwards. Online learning is constrained by resources, especially computation time, but is generally desirable due to its convenience.

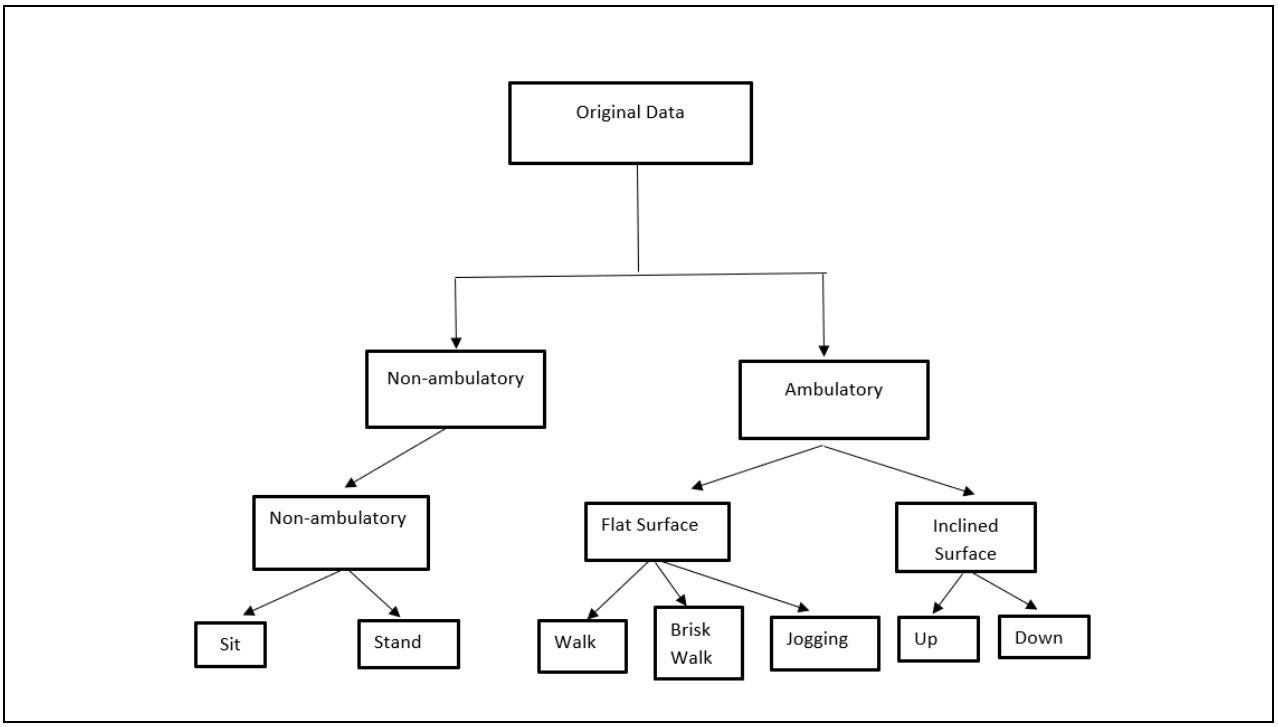

Figure 2.2: Hierarchical meta-classifier. Has 3 levels of classification requiring 5 base learners (1 for each split).

2.2.3 Hierarchical Meta-Classifier

Several different machine learning models have been studied for HAR classification, but some generally perform better than others. SVMs and MLPs tend to achieve good results and are often considered the standard [2, 3, 5]. However, hierarchical classifiers are an interesting possibility and have been explored in the past as a viable model [9, 24] since activities can be logically grouped together based on caloric expenditure. Based on previous work done by Niazi et al. [25], a hierarchical, three-level classifier was used in this study. The first level separates non-ambulatory activities (i.e. sitting) from the ambulatory ones. At the second level the ambulatory activities are further divided into 2 different groups, and at the third and final level, activities are classified individually. Figure 2.1 displays the hierarchy in a tree structure.

This meta-classifier requires a total of five different standard classifiers, one for each split in the tree. The benefit is that splitting the data gradually into more specific groups allows

models to specialize in classifying a smaller range of activities. However, error propagation occurs, so a significant amount of noise and errors may be present at lower levels if classifiers at the top of the hierarchy aren't sufficiently accurate.

2.2.4 Cross Validation

To evaluate the performance of classifiers, cross-validation is typically used in machine learning applications. Machine learning models work by trying to fit a function to specific instances of input data, which is generally called a training set. However, to get an accurate representation of the performance of the model, it needs to be evaluated on a set of data it did not encounter during training, and this is called the testing set. Cross-validation allows every instance of the data to be evaluated in the testing set, which is desirable since it reduces the odds of random chance affecting the performance of the model. For HAR, it makes sense to use leave-one-out cross validation, where a single participant is used in the testing set while all others are used to train the model. This strategy is known as inter-subject recognition because it characterizes how well a model can generalize to an individual that it has never seen before. This is particularly useful because it is difficult to obtain labelled data in HAR, so it removes the requirement that people must train the model themselves before using it. However, inter-subject accuracy tends to be much lower than intra-subject accuracy because only one general model is used for all potential individuals. Intra-subject recognition describes the ability of a model to classify activities performed by a single participant, so any possible variance between individuals is removed. Five-fold or ten-fold cross validation is typically used when intra-subject accuracy is desired. An advantage is that performance is almost always better than leave-one-out because a significant source of noise is removed (variability between individuals), and so a specific model is created

for each individual. However, this approach is often unrealistic in many applications and a substantial amount of time is required for the participant to train the model.

This study focuses on inter-subject recognition, so leave-one-subject-out cross-validation is performed. In this approach, one participant hasn't been seen by the model before and so we believe it presents an accurate way to see how well the model can generalize over the data to new participants.

2.2.5 Previous Work by Niazi et al.

Previous work on HAR that has influenced the development of this study has been done by Niazi et al. [25]. Niazi et al. performed a thorough analysis of HAR data collected at the University of Arizona where 310 participants were in the study. However, out of the 310, only 16 participants were used by Niazi, and they each performed 23 different activities while wearing the same brand of hip-worn accelerometer used in this study under controlled conditions [25]. Their data collection methods occurred in a controlled, laboratory environment that produced a clean dataset. This contrasts with our pseudo free-living data, which aims to be representative of conditions in a realistic, free-living environment.

They initially used a two second window to extract features from the raw accelerometer data in the time and frequency domain [25], and applied wavelet analysis to produce a total of 246 features. To make the feature set more manageable, Niazi et al. used expert-based, correlation-based, and relief-based feature selection methods to generate subsets of the attributes [25]. A random forest classifier performed the best on the expert-based subset of 42 features, while it performed the worst when the entire set of features were used in the classifier [25], so they proceeded to use the reduced subset for all of their remaining experiments.

In their study of classifiers, Niazi et al. utilized a three-level hierarchical meta-classifier, similar to the one that was referenced above in Section 2.2.3. After studying multiple classification techniques, they found that random forests performed better at every level in the hierarchy, even when compared to ensemble methods like stacked classifiers [25]. They employed 10-fold cross-validation to evaluate their model. The hierarchical meta-classifier achieved a level-1 accuracy of 97.899%, a level-2 accuracy of 94%, and a level-3 accuracy of 86.63% [25], which outperformed all of the other classifiers in the study.

Using the same dataset, Niazi et al. also performed an analysis on window sizes used for feature extraction and sampling rates [25]. For this part of the study, they only used the subset of time and frequency-based features from the original 246 features that were extracted from the raw accelerometer data, which resulted in a total of only 32 features [25]. Furthermore, only a random forest base classifier was used to analyze the results, and 10-fold cross validation was performed as before [25]. They determined that "window size and sampling rate have a significant effect on accuracy" [25]. Furthermore, after studying six different window sizes and five different sampling rates, Niazi determined that the optimum configuration was a 9 second window with 50Hz sampling rate [25].

2.3 Related Work

The field of HAR is widely studied and is primarily concerned with body-worn sensors, and Table 2.3 displays the categories many of the references fall into.

A huge contribution to this field was made by Anguita et al. [4] when they released a public domain dataset for such a purpose and have subsequently collected over 600 citations for their work. The collection was performed under controlled, laboratory conditions with a single triaxial accelerometer and a three-dimensional gyroscope as well.

Table 2.3 References displayed in different categories

Cross-Validation Strategy	Inter-subject	4,7,13,28,30	Best Model Tested	MLP	8,20
	Intra-subject	5,6,8,9,12,19,20,24,25,30,31		SVM	3,4,17,27
Sensor Type	Accelerometer	8,9,13,19,20,25		KNN	5,6,13,18
	Both	4,24,28,31		RF/Dec Tree	7,9,23,25,30
	Multiple (>2)	3,5,6,7,12,24		Deep	16,24,28
				Other	3,12,19,31

The participants performed six different activities. They also provided several hundred extracted features with a window size of 2.56 seconds and a 50% overlap between windows, with a sampling rate of 50Hz. Their dataset is often used as a benchmark for new machine learning models or methods, and an inter-subject accuracy over 95% is often achieved on this data.

Another significant, early contribution to this field occurred with Altun et al. [3]. They performed a very thorough analysis of HAR, using several different machine learning approaches and different preprocessing techniques. Five different sensors were used, including a magnetometer, and around 1170 features were extracted from a 5 second window. They used PCA for feature reduction, and tested Bayesian decision making (BDM), LSM, kNN, dynamic time warping, SVMs, and an ANN as models. They found that BDM performed better overall, while SVMs performed better in certain scenarios.

Much like this current study, Gupta et al. [13] used only a single accelerometer worn on the waist. They only had seven participants but were still able to analyze inter-subject accuracy. With six possible activities, a 6 second window with 50% overlap, and a sampling rate of 126Hz, they achieved 98% accuracy with a KNN classifier.

Bao et al. [7] utilized user-annotated acceleration data, which is important because it provides a more realistic data set than those made in carefully controlled settings. This group used 20 participants with 5 accelerometers placed on different locations on the body. They found

that decision tree classifiers achieved 85% accuracy, and that the accelerometer placed on the thigh was the most accurate.

Bayat et al. [8] used an accelerometer in a smartphone with only 4 participants to produce a realistic data set. Six activities were used, along with a 10 second window and a sampling rate of 100Hz. However, they didn't study inter-subject accuracy, but only used 10-fold cross validation. They found that multi-layer perceptrons performed best at 89% accuracy.

In another similar study, Kwapisz et al. [20] also used a smartphone accelerometer placed in the pocket to classify 6 different activities. However, this group used 29 participants and a 10 second window, one of the largest observed. They also found MLP to be the best performing classifier. Interestingly, their classifier had difficulty recognizing climbing stairs and descending stairs.

Khan et al. [11] used a hierarchical classifier to classify 15 activities with 97.9% accuracy. This is a unique classifier that was also used by Niazi et al. [25] and is also utilized in this study. Khan's classifier uses different features at different levels in the hierarchy. However, they only used 6 individuals, so they couldn't analyze inter-subject accuracy.

Banos et al. [6] studied the effect of window size in human activity recognition. They found that large windows were useful for classifying more complex activity, but ultimately recommended using a window size between 1 and 2 seconds for best accuracy. However, they used a more controlled data set that may not reflect real-world conditions and noise. They also found that kNN was the best classifier at over 95% accuracy.

In an influential, comprehensive study, Wang et al. [31] used an accelerometer and gyroscope in a smartphone to classify six different possible activities in a public data set. They found naïve Bayes to be the most successful at 90% accuracy.

In an early HAR study, Tapia et al. [30] used accelerometers and a heart rate monitor for their study. They found that their classifier could achieve a 94.6% classification accuracy on subject-dependent tests, but only 56.3% accuracy when it was subject-independent. This highlights the struggles in making a single, general classifier that can be efficacious on new individuals.

Gonzalez et al. [12] studied HAR for elderly patients at risk of stroke or falls. They used two wrist-worn sensors and classified based on three different activities with a sampling rate of 16Hz. They achieved high accuracy using a genetic fuzzy finite state machine.

Attal et al. [5] performed a review of HAR using 3 accelerometers worn on the body with 6 participants performing 12 possible activities. They used a one second window with 80% overlap, one of the highest seen in the literature. They found that kNN performed best at 99% accuracy, while hidden markov models performed the best among the unsupervised classifiers. However, they used 10-fold cross-validation, so inter-subject accuracy wasn't examined. Lu et al. [22] also studied unsupervised physical activity.

Liu et al. [21] and Chen et al. [10] both explored HAR using visual sensors instead of only body worn sensors. Chen explored the fusion of both types of sensors to improve accuracy. Wang et al. [32] used WiFi signals for HAR because they aren't limited by line-of-sight, which is a downside to visual cameras. They achieved 96% accuracy using 10-fold cross-validation.

Khan et al. [18] performed an in-depth study of sampling rates on 5 public, benchmark datasets. They used SVMs for their classification, and developed a framework to optimize this parameter independent of the dataset being used.

Ronao et al. [28] used a convolutional neural network to classify samples into six different activities. They used 30 volunteers who performed 6 different activities with a

smartphone in their pocket, providing accelerometer and gyroscopic data. They didn't perform any manual feature extraction because CNNs automatically perform this task in their hidden layer, and they fed 2.56 seconds of the data into the network at a time. Using inter-subject experiments, they achieved 94.8% accuracy with this deep learning approach.

Hammerla et al. [16] explored several different deep learning approaches on 3 public datasets. They used a one second window with 50% overlap. They found that CNNs perform better classifying differentiating between running and walking, while LSTMs perform better at classifying other tasks.

Reyes-Ortiz et al. [27] studied the classification of activities and the transitions between activities. They used three datasets and used an accelerometer and gyroscope as their sensors. SVMs were used as the classifiers and were able to achieve over 90% accuracy.

Casale et al. [9] used a single accelerometer and used 20 features after performing feature selection on their extracted information. They also used a one second window with 50% overlap. They found that random forests achieved 94% classification accuracy using 5-fold cross-validation.

Most studies that have been observed do not consider inter-subject accuracy or they use public datasets that were collected in well-controlled environments. Neither situation translates well to real-world scenarios where the data is noisy, and users will not be able to properly train any machine learning models. Even the Niazi et al. study, which this work expands upon, does not measure inter-subject accuracy and uses a clean dataset collected in a laboratory environment. For the studies that do achieve high performance when doing inter-subject testing [4,7,13,28], they use clean datasets collected in controlled conditions where participants tend to make more deliberate movements. Thus, this study focuses on a pseudo free-living data set with

20 participants, which allows for adequate testing of the performance of our machine learning models. We also used only a single accelerometer, which complicates classification but presents a more practical, realistic scenario. Furthermore, our goal is to characterize the effects of different parameters, such as window overlap and window size, on such a dataset.

Chapter 3

Classifying a HAR Dataset

3.1 Preprocessing

The data for this research was collected from 20 participants, ranging in age from 18 to 54, performing 9 activities which were later grouped into 7 activities for classification purposes (Table 2.2). The data was collected at the University of Georgia and participants were selected from the surrounding area. Each participant wore an accelerometer on the waist secured by an elastic strap, and they were able to perform the activities in any order they chose and had flexibility in choosing the duration of each activity. Each device was initially sampled at 100Hz.

Because this work is based off research done by Niazi et al. [25], a hierarchical structure was initially used on the data. At the highest level, activities are grouped into either non-ambulatory or ambulatory activities. At the second level, ambulatory is further divided into activities on a flat surface or inclined surface. The lowest level in the hierarchy consists of classifying the individual activities, and the initial hierarchy is displayed in Figure 2.1.

3.2 Determining Hierarchy Structure

The hierarchy used on the Niazi et al. [25] dataset may not be ideal for our current dataset, so the performance of it was evaluated using random forests at every level, consistent with Niazi's approach [25] (see Ch. 2 for more information). Five random forest base classifiers were used in the hierarchy, and leave-one-out cross validation was used to evaluate the model, resulting in the confusion matrix displayed in Figure 3.1. Ten second windows with no

i. Level - 1

0	1	
433	27	0
14	1700	1

0 – Non- ambulatory
1 - Ambulatory

ii. Level – 2

0	1	2	
433	10	17	0
9	622	295	1
5	298	485	2

0 – Non-ambulatory, 1- Group 1, 2 – Group 2,

iii. Level- 3

0	1	2	3	4	5	6	
144	72	2	0	3	2	0	0
57	160	6	1	5	7	1	1
2	4	303	44	153	87	4	2
0	1	30	76	14	41	17	3
0	2	166	20	149	63	0	4
0	3	76	31	61	212	5	5
0	2	0	7	0	0	141	6

0 – Sit, 1 – Stand, 2 – Walk at comfortable pace, 3 – Brisk walking, 4 – Up, 5 – Down, 6 – Jogging

OVERALL ANALYSIS –
In terms of Level-wise accuracy

LEVEL-1	LEVEL-2	LEVEL-3
Accuracy: 98.48% Total Correctly Classified: 2133.0 Total Number of Instances: 2174.0	Accuracy: 70.84% Total Correctly Classified: 1540 Total Number of Instances: 2174	Accuracy: 54% Total Correctly Classified: 1185 Total instances: 2174

Figure 3.1: Confusion matrix and performance metrics for 3-level meta-classifier

overlapping were used to extract 16 expert-selected features for the classifier (features explained in next section).

These results show that the hierarchical classification strategy produces weak results at the individual activity level (level 3), and even performs poorly at level 2 with only 70% accuracy for the groupings. This seems to indicate that the data is quite noisy, or that the machine learning model is not suitable for this dataset. This contrasts with Niazi's results, which achieved 94% accuracy at the grouping level (level-2) and had a level-3 accuracy of 86.63% [25]. In theory, a new hierarchy better suited for this data should produce better results, so a new tree structure was created based on the confusion matrix of the previous results, which is displayed in Figure 3.2. The confusion matrix suggests that the "walking up an incline" class is often confused with the "walk at a comfortable pace" class. These 2 were subsequently grouped together at level 2 in order to apply a random forest model to better classify the data. A more accurate level 2 model will reduce error propagation down to the third level, which should

improve the performance at that level too. The same models and cross-validation method were used on the update hierarchy, and the results are presented in Figure 3.3.

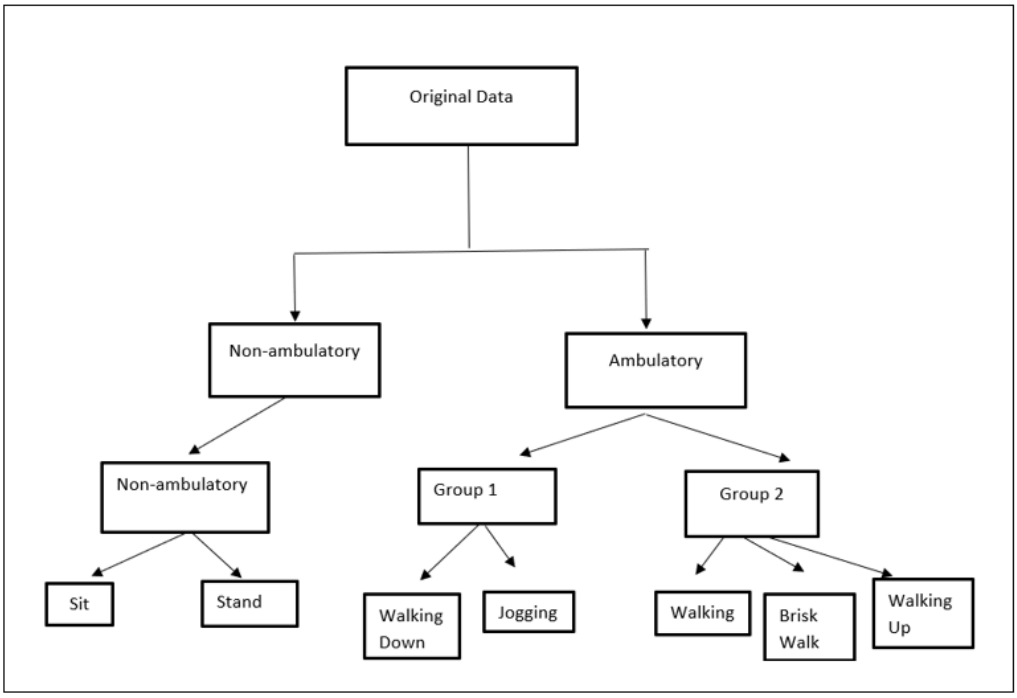

Figure 3.2: Updated hierarchy structure. Walking down and jogging now share a Level 2 group. The other 3 ambulatory activities share a different Level 2 group

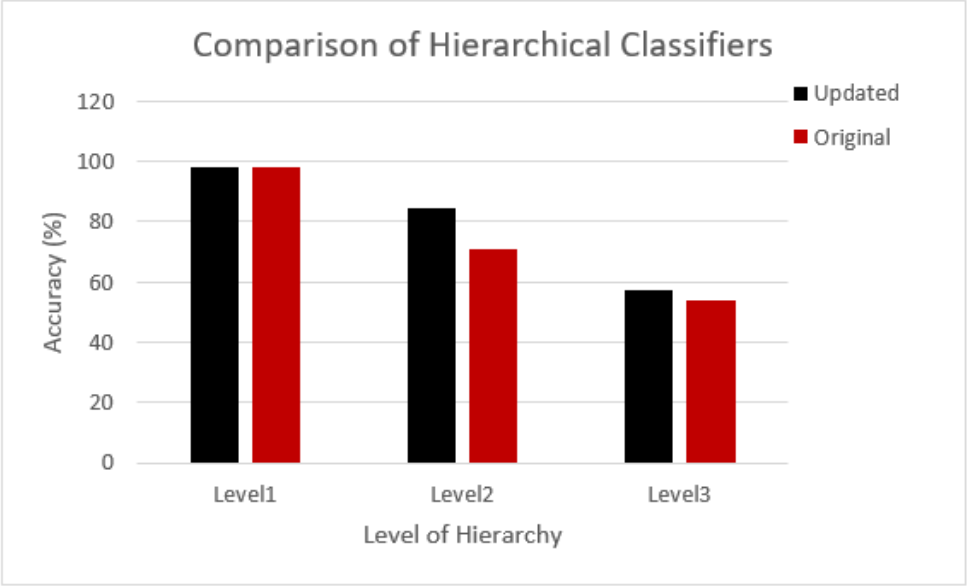

Figure 3.3: Comparison of Hierarchical Classifiers. The updated hierarchy (Fig 3.2) outperforms the original structure (Fig 2.1)

The results show improvement with the new tree structure, especially at level-2 which saw a 13% increase in classification accuracy. Even classification at the individual activity level saw a 2.13% improvement over the previous tree structure that had been used. This indicates that the groupings of activities can heavily influence accuracy results, although level-3 accuracy still remains relatively weak with the updated hierarchy. This does demonstrate how classifier performance in inter-subject testing can differ significantly from that in intra-subject testing.

3.3 Feature Extraction

Several features were extracted from the time domain among all three spatial axes and the vector magnitude. These features included the mean, maximum, and minimum values, as well as the standard deviations, median crossings and the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentile. A Fourier transform was used to extract features in the frequency domain, such as the dominant frequency and the magnitudes among all three axes. This resulted in a total of 36 features for the time domain, and 16 for the frequency domain for a total of 52 manually generated features. First differentials and wavelet analysis were used in feature extraction in the meta-classifier section of Niazi's study [25]. However, in the statistical analysis section, Niazi et al. only used features extracted from the time and frequency domain [25], which is the approach taken in this study due to its simplicity and faster computation times.

3.4 Feature Selection

Processing many features increases the complexity and training time for a model, and it may negatively impact performance because there is a risk for overfitting to occur on unimportant features. Correlations between features were measured, and we can see that some are strongly

correlated with each other (Figure 3.4). Consequently, several different feature selection methods were used to create a subset of features that may contribute more to activity classification.

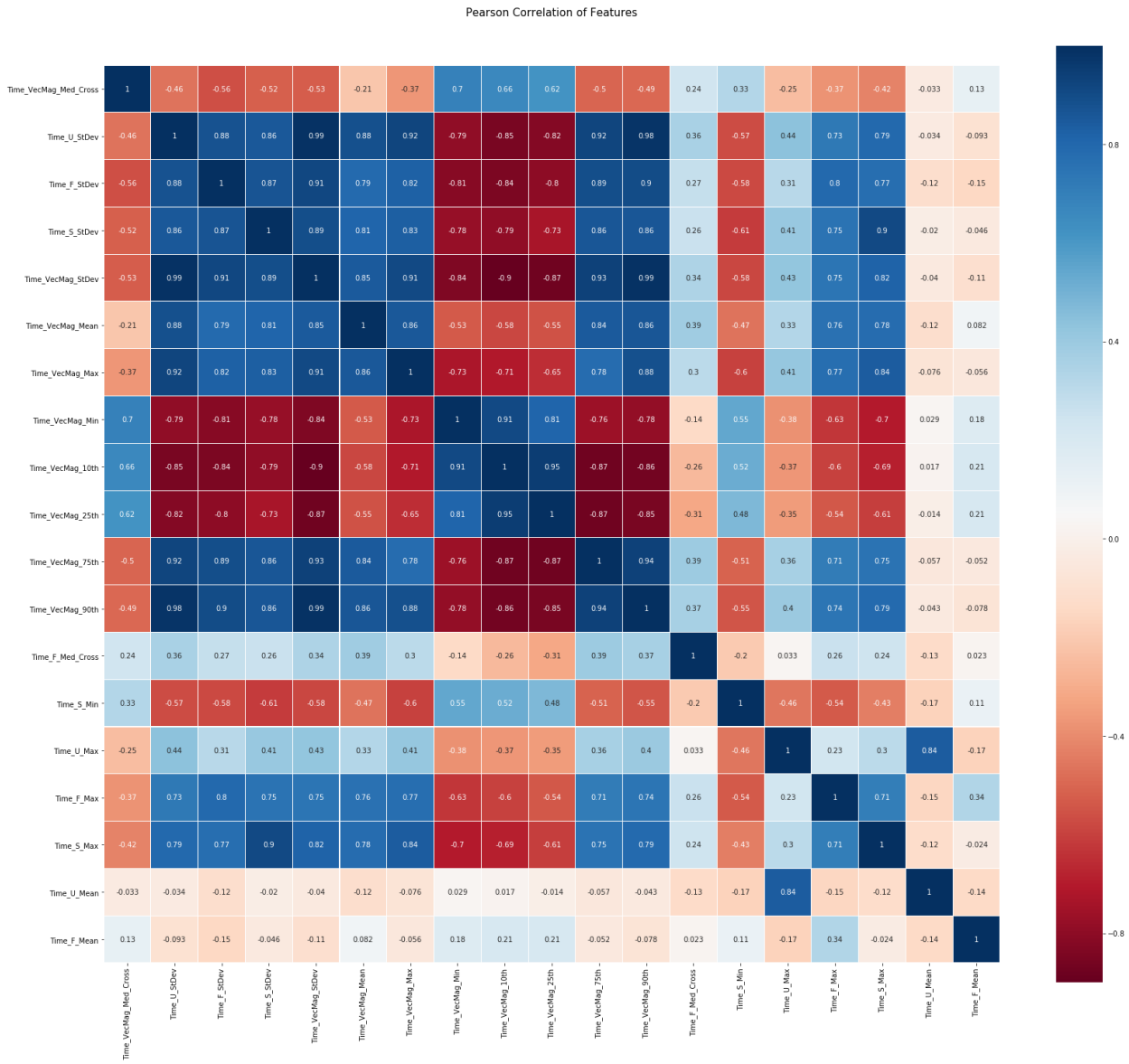

Figure 3.4: Correlation among subset of time-based features. This figure displays how the subset of extracted features in the time-domain are correlated with each other.

Expert-based feature selection used a researcher (a member of Niazi et al. [25]) with domain knowledge to choose the most relevant features, and 16 were selected. K-Best feature selection was also chosen to select 16 features in order to compare this method to the previous expert-based one, and it was found that this selection method did choose a different feature set than the expert. Furthermore, WEKA, an open-source machine learning tool that has many robust feature selection methods, was used to explore other ways to optimize the features [15]. From WEKA, correlation-based feature subset selection was used [14] and 21 features were produced. Two other methods, classifier and correlation attribute evaluation, didn't remove any of the features from the original 52, while recursive feature elimination (RFE) resulted in 42 features. Three different classifiers were used to test the efficacy of the feature sets in order to eliminate any possible biases. This is because random forests tend to be resistant to unnecessary features in the data, while other methods may be more sensitive to different features. The results are shown in Table 3.1.

The results consistently show that the expert-selected features perform the worst, even against K-Best feature selection. Surprisingly, using the entire feature set produced very accurate results at the 2nd and 3rd levels of the hierarchy, while RFE performed the best when using 42 features out of the original 52. This runs counter to many of the studies encountered which showed a large increase in accuracy when using feature-selection methods [31]. This is also at odds with the results obtained by Niazi et al. [25], which showed that expert-based feature selection produced the highest results. The reason for this discrepancy is likely due to the feature selecting abilities inherent to the classifiers, since machine learning models can be robust to some superfluous features. Niazi et al. [25] eliminate a large percentage of features, while this study shows that for a pseudo-free living dataset, retaining most of the features is a more promising approach.

Table 3.1: Evaluation of Feature-Selection Methods on 3 Different Classifiers. Results display the accuracies from each level of the meta-classifier with 3 different base classifiers tested.

Classifier	**Random Forest (Accuracy %)**	**MLP (Accuracy %)**	**SVM (Accuracy %)**
Expert FS	Level 1: 98.25% Level 2: 84.24% Level 3: **57.97%**	Level 1: 98.4% Level 2: 83.8% Level 3: 55.6%	Level 1: 98.5% Level 2: 84.2% Level 3: 57.0%
K-Best FS	Level 1: 98.48% Level 2: 82.77% Level 3: **58.98%**	Level 1: 98.45% Level 2: 84.1% Level 3: 56.8%	Level 1: 98.5% Level 2: 84.4% Level 3: 58.9%
Corr.-Based FS	Level 1: 98.56% Level 2: 82.55% Level 3: 58.89%	Level 1: 98.22% Level 2: 83.27% Level 3: 60.58%	Level 1: 98.37% Level 2: 83.80% Level 3: **60.87%**
No FS (All 52 Features)	Level 1: 98.46% Level 2: 82.36% Level 3: 58.98%	Level 1: 98.36% Level 2: 84.57% Level 3: 63.17%	Level 1: 98.37% Level 2: 85.38% Level 3: **63.94%**
RFE (42 Features)	Level 3: 58.37%	Level 3: 64.85	Level 3: **65.05**

Furthermore, RFE was able to perform the best only by retaining many of the features compared to the other selection methods (42) while only dropping the very worst features, slightly outperforming the "No FS" strategy. This offers a slight performance boost for a slight reduction in complexity. However, if there are serious constraints on time for an online learner, then it may be beneficial to use correlation-based feature selection, which reduced the number of features to only 21 while dropping accuracy between 2 and 3 percent for MLPs and SVMs. Random forests appear to be unaffected by the different feature selection strategies, which is likely because they already randomly assign feature subsets to the different decision trees in the forest.

3.5 Experimental Results of Classifiers

Scikit-learn is an open-source machine learning library that was used in this study for experimenting with different classification algorithms and their parameters [26].

The hierarchical meta-classifier requires 5 base classifiers, and to reduce computation time each of the 5 nodes used the same classifier while the parameters of each base learner could

be unique from each other. Previous studies have shown the benefits of using the hierarchical approach [19, 25], but we tested this approach against multiple "flat" classifiers. In this instance, a flat classifier is just the individual base learner by itself, which is useful for evaluating the effectiveness of the hierarchical meta-classifier. Furthermore, participant 16 was excluded from each of the below tests, as this always improved accuracy and will be explained further in the study. As a preprocessing step, the data was standardized before being fed into the model.

Six different base classifiers were used, random forests, multi-layer perceptrons (MLP), support vector machines (SVM), decision trees, XGBoost, and quadratic discriminant analysis (QDA), in both the "flat" configuration and as a part of the meta-classifier for a total of 12 different machine learning models. The results of these classifiers, along with several others, are shown in Table 3.2.

Both MLP and SVM are the highest performing classifiers and they achieve accuracy at level-3 between 63 and 64%. However, using just a single classifier instead of the hierarchy produces similar results at the activity level across virtually all models tested, which raises questions regarding the effectiveness of the hierarchical method. It's apparent that the hierarchical meta-classifier failed to offer significant advantages over "flat" learners when dealing with this pseudo free-living dataset. This may be due to the dataset being noisy, which exacerbates the problem of error propagation. Additionally, SVMs and MLP are shown to be the best classifiers. This is different from the work in Niazi et al. [25], which used random forest classifiers, but does align with many other studies which have revealed the high performance of SVMs and MLPs. Thus, with a pseudo free-living or noisy data set it may be beneficial to eschew complicated classifiers for simpler ones that generalize better, which is in line with Occam's razor.

Table 3.2: Evaluation of Classifiers

Classifier	Flat	Hierarchical
Random Forest	Accuracy: 58.03%	Level 1 Accuracy: 98.46% Level 2 Accuracy: 82.36% Level 3 Accuracy: 58.98%
MLP	Accuracy: 64.62%	Level 1 Accuracy: 98.36% Level 2 Accuracy: 84.57% Level 3 Accuracy: 63.17%
SVM	Accuracy: 63.85%	Level 1 Accuracy: 98.37% Level 2 Accuracy: 85.38% Level 3 Accuracy: 63.94%
Decision Tree	Accuracy: 50.77%	Level 1 Accuracy: 97.07% Level 2 Accuracy: 75.38% Level 3 Accuracy: 50.62%
XGBoost	Accuracy: 57.07%	Level 1 Accuracy: 98.51% Level 2 Accuracy: 81.63% Level 3 Accuracy: 58.70%
QDA	Accuracy: 47.93%	Level 1 Accuracy: 94.52% Level 2 Accuracy: 61.12% Level 3 Accuracy: 41.59%

A Wilcoxon signed-rank test was conducted to determine whether the MLP and SVM models had performed significantly better on the data than the hierarchical random forest. This is a paired statistical test that doesn't make assumptions about the distribution of the samples, which is beneficial in this case given the relatively small sample size. We are able to use this test because we performed leave-one-out inter-subject testing, which allows us to treat the participants as matched samples between the two classifiers. The computed test statistics and p-values are given in Table 3.3.

Table 3.3 Classifier Significance

	MLP	SVM
W-statistic	139	139.5
p-value	.04	.038

The results from the Wilcoxon signed-rank test show that both the MLP and SVM have p-values below 0.05, allowing us to conclude that both are significantly better classifiers than the hierarchical random forest used by Niazi et al. [25] on this dataset.

3.6 Intra-subject Testing

Niazi et al. [25] only used an intra-subject cross-validation strategy on his clean dataset, so an intra-subject test was performed on our pseudo free-living dataset in order for us to better compare the different results between the two datasets. Based on the classifier performance in section 3.5, an MLP was chosen as the classifier, and 10-fold cross-validation was performed on the data, with the resulting confusion matrix displayed in Figure 3.5.

The accuracy has increased from 64% to 86% when intra-subject recognition is employed. This is comparable to the results obtained by Niazi et al. [25] on their relatively clean dataset and demonstrates the high variability between different participants' activities. The two most confused activities, walking up and down an incline, have much higher accuracy here, which suggests that the variability between participants stems mainly from these two activities.

i. Level- 3

0	1	2	3	4	5	6	
3917	440	40	0	0	60	0	0
198	4450	76	0	100	79	0	1
38	40	10626	40	505	327	0	2
0	0	297	3177	100	0	0	3
0	20	2009	40	5605	172	0	4
0	40	913	60	614	6025	0	5
0	0	0	0	0	20	2959	6

0 – Sit, 1 – Stand, 2 – Walk at comfortable pace, 3 – Brisk walking, 4 – Up, 5 – Down, 6 – Jogging

LEVEL-3

Accuracy: 86.1%

Total Correctly Classified: 36759

Total instances: 42687

Figure 3.5: Intra-subject confusion matrix. Accuracy for MLP is greatly improved to 86.1%

3.7 Removing an Outlier

During the course of these experiments, an outlier participant was spotted and excluded from the data set. The accuracy for the individual participants is presented in Table 3.4.

It was determined that participant 16 is a statistically significant outlier. This could be the result of a faulty accelerometer, mislabeled data, or wildly inaccurate movement when performing the activities. Regardless, excluding this individual increased level-3 accuracy by 1.5% for the MLP, and resulted in improvements for every classifier, showing the benefits of having a large sample size so that accurate test results can be achieved. These results also reveal the large discrepancy between the classification accuracy for each participant. The standard deviation, with participant 16 excluded, is still at 11.2223%, demonstrating how inter-subject variability in movement poses one of the biggest challenges to human activity recognition.

3.8 Deep Learning

Increased computation power and deep learning have provided many exciting results in AI and machine learning, and these approaches have potential to be applied to the HAR problem as well [4, 19, 24]. One of the chief benefits of this approach is automatic feature learning.

Table 3.4: Accuracy for each individual test participant using MLP

1	2	3	4	5	6	7	8	9	10
74.2%	57.4%	63.4%	69.5%	73.6%	72.1%	72.3%	54.2%	43.3%	64.8%
11	12	13	14	15	16	17	18	19	20
52.7%	79.1%	40.5%	66.1%	59.8%	28.9%	45.7%	72.2%	68.8%	65.9%

Similar to the other learning schemes, standardization was performed on the raw data to prepare it for being fed into the neural network. A seven second window was used to segment the data, and the most frequent class was used as the label for each segment. Three-dimensional data in the form [total segments, input width, input channel] was generated, which was then reshaped so it had a height of 1 to feed into the network. The model itself consisted of one convolution layer, followed by max pooling and then another convolution layer, which was also followed by max pooling. That was then connected to a fully connected layer, and finally a Softmax layer. The architecture is given in Figure 3.6.

Previous experiments showed that around 250 epochs produced the best results without seriously overfitting the training data. A 7 second window is used due to resource and time constraints for the NN architecture, and the results are given below.

Table 3.5: Results of CNN performance on HAR data

Epoch Length	250
Training Accuracy	98.53%
Testing Accuracy	57.8%

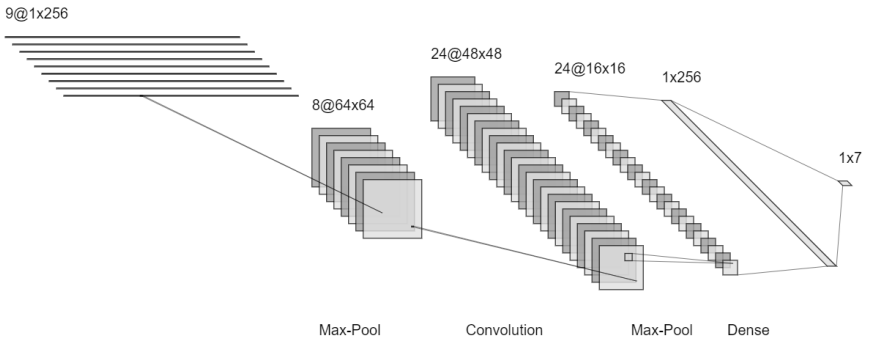

Figure 3.6: Pictorial representation of CNN architecture. 2 Convolutional layers and 2 Max-Pool layers with a fully connected dense layer before the output.

Due to time constraints, 20-fold cross validation on all participants could not be performed, so only a subset of 3 randomly selected participants were withheld from training and were used in the testing group. The resulting test accuracy is 57.8%, which is below that reached for other machine learning approaches, such as MLP and SVMs, and significantly below the training accuracy. However, this does show the power of learning features automatically as it almost reaches the same performance as random forests while being fed only the raw data, significantly simplifying the preprocessing steps.

3.9 Ensemble Methods

After examining many of the traditional learning models, ensemble methods were tried. The reasoning behind this is that they can help overcome bias that may be present in a single model, while the disadvantage of these approaches is that each learner acts as a black box, making it difficult to ascertain whether a particular model will be successful with other data sets or scenarios.

XGBoost and AdaBoost have become two popular boosting methods that use ensembles of relatively weak classifiers to improve overall performance [11]. XGBoost uses gradient boosted ensembles of decision trees to train and classify data, while AdaBoost can use any weak classifier as its base learner, and it utilizes a weighted some from their outputs to produce a prediction [11].

The Extra Trees classifier uses an ensemble of decision trees to create a predictive model. This classifier is slightly different than a random forest because it makes splitting decisions for each tree randomly, as well as randomly choosing subsets of features for each tree [26]. This

introduces an extra amount of uncertainty when the trees are constructed than with random forests, with the intent at reducing possible overfitting.

A voting classifier uses multiple base learners that each independently make a prediction, and each prediction is then tallied to determine what the final prediction should be. In this way, each lower level model casts a "vote", and in a hard-voting approach the class with the most votes wins. However, if soft-voting is used, then each classifier gives a probability along with its predicted class, so that classifiers that are very confident are given more weight in the voting.

Stacking classifiers use multiple base models to make predictions on the input data, which are then fed into another higher-level classifier which uses these predictions as inputs to make a final prediction for the output. The higher-level learners in a stacking classifier tend to be simple models, such as a decision tree, in order to reduce the chances of overfitting. Similar to the voting classifier, either the predicted classes or their corresponding probabilities can be fed into the higher-level learner.

Random forests are an ensemble method but were already tested in an earlier section. The results of the ensemble classifiers are given in Table 3.6.

Table 3.6: Evaluation of ensemble learning methods

Classifier	Full 52 Features
XGBoost	Level 1 Accuracy: 98.46% Level 2 Accuracy: 81.39% Level 3 Accuracy: 57.84%
AdaBoost	Level 1 Accuracy: 98.08% Level 2 Accuracy: 80.34% Level 3 Accuracy: 53.94%
ExtraTrees	Level 1 Accuracy: 98.41% Level 2 Accuracy: 82.02% Level 3 Accuracy: 59.71%
Voting Classifier	Level 1 Accuracy: 98.42% Level 2 Accuracy: 85.38% Level 3 Accuracy: 65.7%
Stacking	Flat only Level 3 Accuracy: 65.9%

Several new ensemble methods were attempted to see if any improvement in accuracy could be achieved. The worst performing method was AdaBoost which only had around 54% accuracy. It's generally accepted that boosting methods are more susceptible to overfitting noise, which helps explain its poor performance for this dataset. ExtraTrees classifier performed slightly better than a Random Forest classifier, while XGBoost performed slightly worse. Overall, the tree-based learners performed poorly and were outclassed by SVMs, MLPs and other ensemble methods. By far the best performing models were the voting and stacking classifiers, which achieved the best accuracies yet of 65.7%, and 65.9% respectively. The voting classifier used Random Forests, SVMs, and MLPs as its 3 base learners and used a soft voting approach to classification. These base classifiers were chosen because they employ different methods for classification, which allows them to compensate for each other's weaknesses.

The stacking classifier used the voting classifier, SVMs, and MLPs, and were chosen due to their high accuracy individually. This was much slower than the other approaches so only the flat classifier was tested but achieved promising results. Stacking classifiers are difficult to implement, and care must be taken to avoid feeding testing data into the training data set. Another downside for ensembles is that they act as a black box making it difficult to determine what the underlying patterns are or how they achieve their success. For instance, it is not clear why the voting classifier, SVM, and MLP configuration performed the best for the stacking classifier, since the voting one is itself an ensemble method. Nevertheless, it seems they can perform well in noisy data and help compensate for individual models. These results are again at odds with the study done by Niazi et al. [25] where they found ensemble methods, such as the

voting classifier, performed worse than the hierarchical approach, which again highlights how a pseudo free-living dataset needs different approaches and models than a clean one.

Chapter 4

Data and Parameter Analysis

4.1 Window Size

After researching several different machine learning models, a deeper look at the data's hyperparameters for HAR was undertaken. The first experiment deals with the effect of window size on level-3 accuracy (individual activity level) when using random forests, MLPs, and SVMs. These were chosen as the classifiers because the ensemble methods, such as voting classifier, were significantly slower and only offered marginal improvements in classification accuracy. They also feature prominently in the literature, especially for HAR. Studies have shown significant improvements in accuracy when altering window size [6], but very little work has been done on this topic using pseudo free-living data sets. Figure 4.1 shows how changing the window size can affect accuracy.

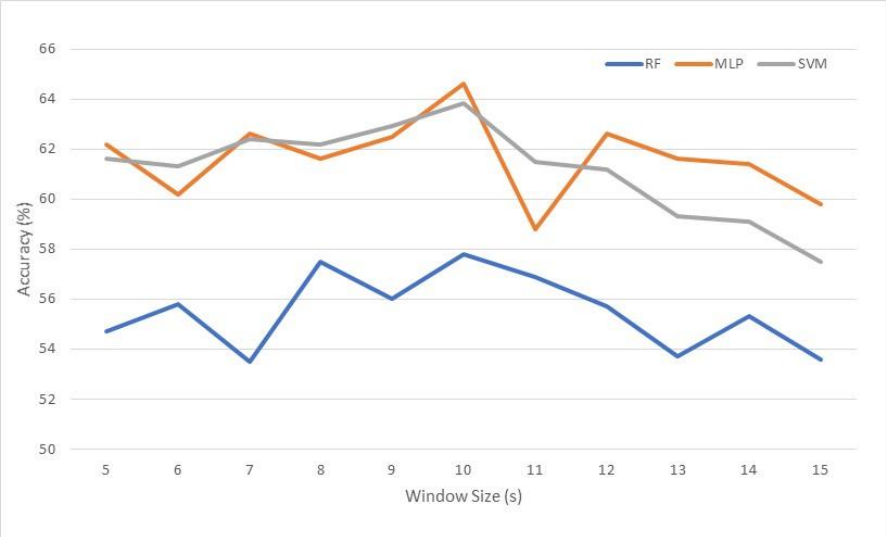

Figure 4.1: Activity classification accuracy as a function of window size. A peak occurs at a window size of 10 seconds

We see a peak classification accuracy at a 10 second window interval, while the second highest time is at the 7 second mark for MLPs and SVMs. The plots form a rough parabola, but it should be noted that there were only slight improvements in accuracy found at the 10 second mark, with an accuracy of around 64% compared to a 62% accuracy at the 7-second window size mark, showing that there were only slight variations at reasonable time spans for the window. These results also indicate that a relatively large window size of 10 seconds may help mitigate the effects of noise and large inter-subject differences, but when it gets too large the performance degrades. Smaller window sizes have the advantage of faster computation time, but apparently don't perform as well on this data set. This is in contrast to the work of Niazi et al., who observed significant differences in accuracy when the window size was altered [25], although we both reach the same conclusion that a 10 second window is optimal.

4.2 Window Overlap

Another big factor in feature extraction is whether to allow any window overlap. This has the potential to affect the accuracy because it produces more data, which could improve results. The experiment was run on the percent of overlap between windows. For instance, 0.2 overlap means the previous window shares 20% overlap of the data with the current window. The results of this experiment are given in Figure 4.2.

Random forests, MLPs, and SVMs were also used for this experiment. The window overlap percentage was tested because previously we only used non-overlapping windows. This technique has the advantage of producing more data, although processing time increases with as overlap increases.

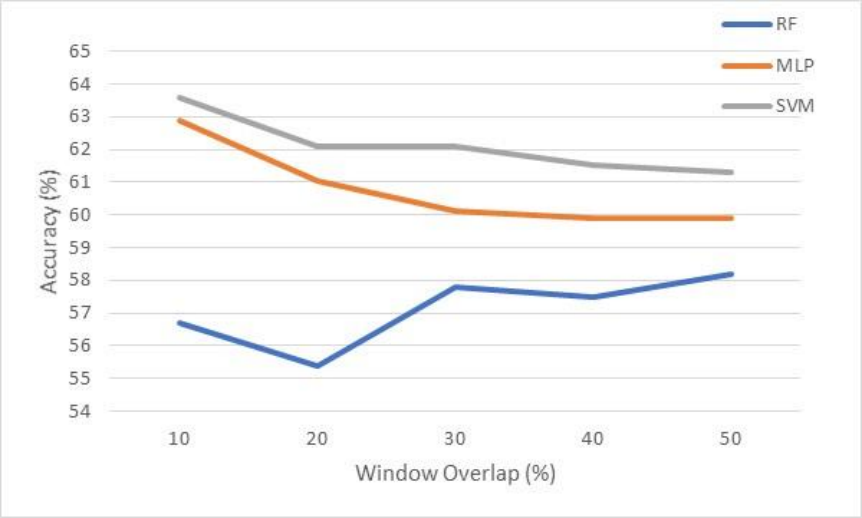

Figure 4.2: Activity Classification accuracy as a function of percent window overlap

There are only minimal changes in activity classification accuracy, with all falling in a range of 3 percentage points. A 10% overlap achieved the best results at 63.6% accuracy for the SVM, but overall, this indicates that there are no significant effects related to window overlap.

4.3 Learning Curve

The learning curve of a dataset is helpful in determining whether there is enough training data to learn useful information and have decent performance. Ideally, a logarithmic learning curve is desirable, which would mean that any more training data is unlikely to improve results and that you are not under-fitting the data.

The learning curve was generated for the HAR data as seen in Figure 4.3. This was accomplished by gradually increasing the number of participants that were included in the training set from 1 to 19 individuals. The participants in the training set were randomly selected, and this process was repeated 19 times for each particular training set size and the average was taken.

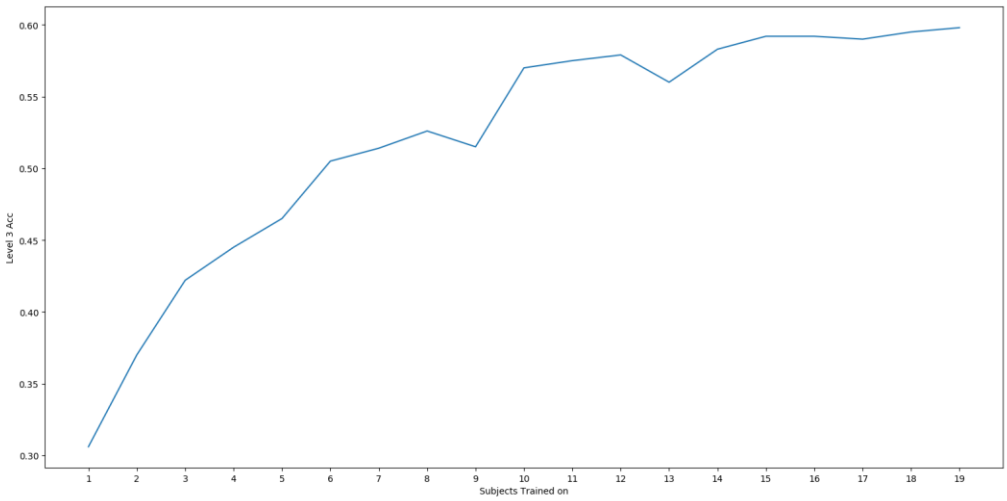

Figure 4.3: Learning curve. Displays the averaged accuracy of classifier as the number of participants used in training is changed

Initially the activity classification accuracy starts out at around 30%. The curve then follows a roughly logarithmic path until it levels out at around 60% accuracy, which occurs when 15 participants are in the training set. This indicates that adding more participants/data to the training set would not result in an increase in accuracy, as this is approximate to the accuracy seen when all 19 are used. Thus, the main limitation on the training accuracy appears to be the inter-subject variability present in the data, along with possible noise, instead of insufficient data. This also explains why altering window overlap didn't affect the performance at all. Increasing window overlap increases the amount of data present, but the learning curve suggests that enough data is present to avoid under-fitting.

4.4 Participant "Pre-training" Models

It appears that inter-subject variability is contributing to the poor performance in level-3 accuracy for the learning models. To alleviate this, experiments were ran that included increasingly large portions of the particular test participant's accelerometer data in the training set. This simulates the participant 'training' the accelerometer by providing some information

about what activity they perform. For instance, 19 participants are used to train an SVM model, while participant 1 will be used to test the learning model. We initially provide the first 10% of participant 1's data involving all 7 of the different activities to the training set. In theory, allowing some of participant 1's data to be included in training should improve accuracy. The goal is to remove some variability between participants while also maintaining a minimal, realistic amount of training data provided by the test participant. To this authors knowledge, this is the first such experiment performed using this method, and Figure 4.4 displays the results of the experiment.

Initially, just 10% of the participant's data is included for training to simulate a person 'training' an accelerometer before using it. An SVM model is used for classification, and the accuracy is improved at all levels from a base level of 61.7% accuracy. At 10% contribution, the accuracy increases to 64.5% accuracy. This jumps to 68.7% accuracy when the individual provides roughly 30% of their data in the training set. This jumps higher when 70% is provided, but this is unrealistic as the test set becomes too small to provide much information. Thus, we do see a noticeable increase in accuracy when the patient's data is included in the training set,

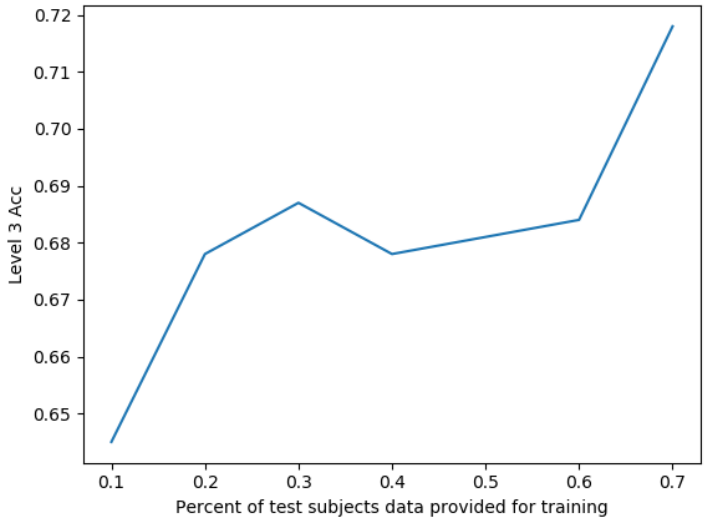

Figure 4.4: Accuracy of classifier as percent of intra-subject data is increased

especially when the data contributed is at 30%, before it eventually levels off. This is promising because it provides a way to reduce the inter-subject variability present in a pseudo free-living data set using a relatively small amount of data from the user in question, with over a 5% increase in base level accuracy seen.

4.5 Adding Data to Training

Anguita et al. [4] published a dataset to UCI that is publicly available and consists of accelerometer measurements obtained from a hip-worn sensor with several different classified activities. It may be beneficial to include their dataset with ours because their dataset is typically used as a benchmark in HAR studies. The theory is that high quality data may improve training, which in turn could lead to a more generalized model. The different possible activities include walking, walking upstairs, walking downstairs, sitting, standing, and laying, so samples labeled with "laying" are excluded from our dataset because we do not have an activity that matches it.

ii. **Level- 3**

0	1	2	3	4	5	6	
154	55	3	0	4	1	0	0
30	189	8	0	5	6	1	1
2	5	352	40	98	61	4	2
0	0	39	51	12	53	17	3
0	2	149	14	181	32	0	4
1	1	66	22	37	233	7	5
0	0	1	11	0	2	131	6

0 – Sit, 1 – Stand, 2 – Walk at comfortable pace, 3 – Brisk walking,
4 – Up, 5 – Down, 6 – Jogging

LEVEL-3

Accuracy: 62.1% Total Correctly Classified: 1291
Total instances: 2080

Figure 4.5: Confusion matrix for combined dataset. The UCI data was only added to the training set.

The UCI data was only included in training, and once again an SVM was used as the model in the hierarchical classifier, and the results are displayed in Figures 4.5 and 4.6.

The classifier performed worse when the UCI data is included in the training set for our own data, which isn't entirely unexpected because the learning curve showed that adding more data did not significantly improve accuracy after a certain point. The level-3 accuracy was reduced over 2% to 61% accuracy. Additionally, there may be other differences present in the data that have not been accounted for, which would undoubtedly deteriorate the results. From Figure 4.5, it is shown that the UCI reduced accuracy for the "walking up an incline" class, while most results for other classes were relatively unchanged when compared to results from our data only. This reveals that most of the difference between these two datasets occurs in the "walking up" activity. We can conclude that doping the training data with cleaner data from a well-studied dataset does not help the model generalize better.

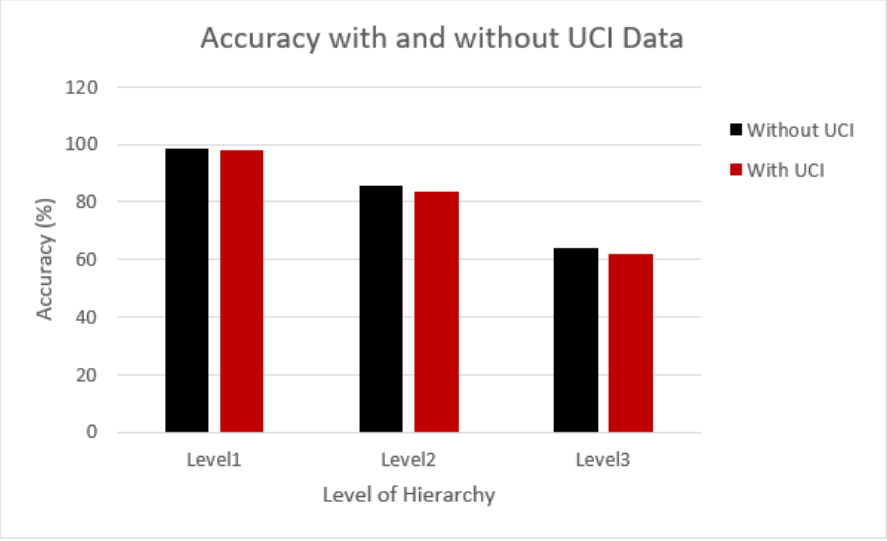

Figure 4.6: Accuracy with and without UCI data. Accuracy of the SVM classifier without UCI data included (black) and with the UCI data included in training (Red).

4.6 UCI Dataset Alone

In section 4.5, we studied the effects of adding the UCI dataset to our own, pseudo free-living dataset and found that results did not improve. However, it would still be informative to ascertain how well our current feature extraction techniques and models would perform on the UCI data alone, since it would allow us to see if the machine learning models were at fault. To this end, only the triaxial accelerometer data from the UCI data was used in this experiment. The same feature extraction techniques discussed previously were applied, and three machine learning models were tested on the data. These models were the random forest, MLP, and SVM learners, and the results are displayed in Figure 4.7 below. The results show that all three models performed significantly better on the much cleaner UCI data, with the random forest model having 91% accuracy, the MLP at 96.4% accuracy, and the SVM at 94.2% accuracy. These are significantly better than any model used on our realistic data and is comparable to the inter-subject accuracy obtained by Anguita et al. [4] in their original study.

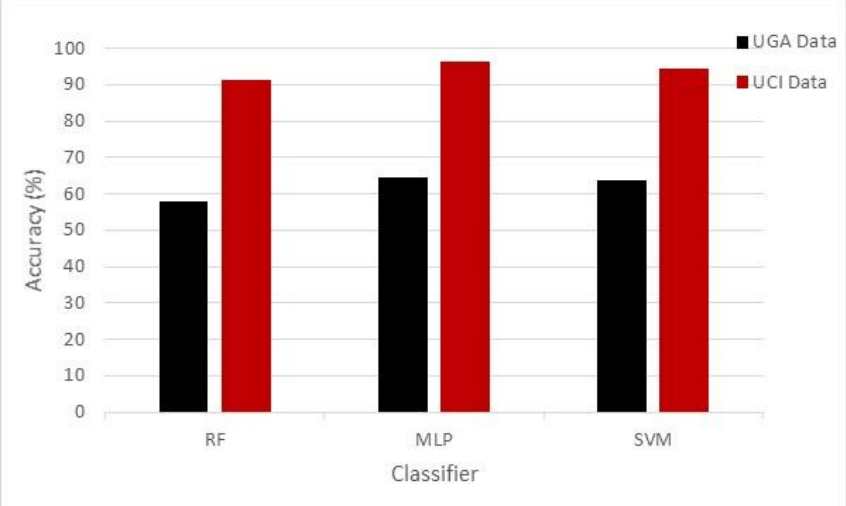

Figure 4.7: Classifier accuracy on two different HAR data sets. Accuracy of the RF, SVM, and MLP classifiers on the UGA data (black) and the UCI data (Red)

This reveals how approaches used on “clean” datasets can perform substantially different when used on a more realistic one collected in close to free-living conditions.

4.7 Simplifying Activities

The confusion matrices generated from these studies indicate that walking upstairs and walking downstairs are confused by virtually all the machine learning models. This implies that different participants are climbing these inclines in significantly different manners, so the models are having a difficult time generalizing to new individuals. However, if the participant or researcher is uninterested in the specific terrain an individual is walking across (i.e. an incline) or if stairs are reasonably expected to be rarely encountered, then it would make sense to label those instances as walking. Hypothetically, we would be able to achieve much better results by just considering those 2 classes as part of the “walking comfortably” class. To test this, the activities of walking upstairs and walking downstairs were folded into the “comfortably walking” activity so that no samples were discarded, just relabeled. Figure 4.8 displays the confusion matrix for this, using an SVM as the classifier. The accuracy obtained from this is 85.9%, which is comparable to many HAR studies in the literature.

i. **Level- 3**

0	1	2	3	4	5	6	
164	46	6	0	0	0	1	0
33	188	16	0	0	0	2	1
4	7	1234	52	0	0	1	2
0	0	96	63	0	0	13	3
0	0	0	0	0	0	0	4
0	0	0	0	0	0	0	5
0	10	1	5	0	0	138	6

0 – Sit, 1 – Stand, 2 – Walk at comfortable pace, 3 – Brisk walking, 4 – Up, 5 – Down, 6 – Jogging

LEVEL-3

Accuracy: 85.9%

Total Correctly Classified: 1787

Total instances: 2080

Figure 4.8: Confusion matrix with simplified activity classes. Confusion matrix from SVM classifier when Walking Up and Down classes are folded into the Walking Comfortably class. Accuracy is 85.9%.

i. **Level- 3**

0	1	2	3	4	5	6	
165	48	4	0	0	0	1	0
33	196	8	0	0	0	2	1
3	9	482	64	0	0	4	2
0	0	71	86	0	0	15	3
0	0	0	0	0	0	0	4
0	0	0	0	0	0	0	5
0	1	1	5	0	0	138	6

0 – Sit, 1 – Stand, 2 – Walk at comfortable pace, 3 – Brisk walking, 4 – Up, 5 – Down, 6 – Jogging

LEVEL-3

Accuracy: 79.9%

Total Correctly Classified: 1336

Total instances: 1067

Figure 4.9: Confusion matrix when activity classes are excluded. Displays confusion matrix from SVM classifier when Walking Up and Walking Down classes are excluded from the data. Accuracy is 80%.

These results show that most of the inter-subject variability arises from the walking upstairs and downstairs activities. This makes sense because there is a wide array of factors that can influence how someone climbs stairs.

The assumption that walking up and down an incline can be grouped with a comfortable walking pace may be overbroad. Therefore, a test was performed where these samples were removed from the dataset, so that only 5 activities remained. The results are shown in Figure 4.9. The accuracy from this experiment is at 80%, indicating that much of the inter-subject variability has been removed when walking up and down an incline is not included.

When the Up/Down classes were included as a single group, SVM classification accuracy increased to 67% accuracy (not shown). This confirms what is observed in the confusion matrix, that classifiers struggle to distinguish between terrain types, and that the walking on incline tends to be confused with walking at a comfortable pace.

Ch. 5 Conclusion and Future Directions

Using realistic data for the HAR problem presents many challenges in achieving high performance for machine learning models. The goal of this research is to create a robust model and methodology that is practical, has high performance, and can generalize classification to new individuals when using this type of dataset. It's been observed that the hierarchical meta-classifier fails to offer significant advantages over "flat" learners. This is most likely due to a noisy data set where error propagation is a significant downside for hierarchical approaches. Occam's razor suggests that the simpler classifiers, such as SVMs, are better able to generalize over noisy data and offer the best performance.

Feature selection also doesn't offer significant improvements in performance, despite some correlation among the attributes, although RFE offers a slight improvement by retaining 42 features. This is likely because the most of the feature selection approaches don't offer significant improvements over the machine learning models' inherent feature selection capabilities.

SVMs and MLPs perform the best out of the base learners on the data with 7 classes of activities, achieving an accuracy of around 63%. Ensemble methods perform the best at of all learners with both the voting and stacking classifiers achieving a level-3 accuracy of around 66%. Ensembles of learners may be able to better cope with noise in the data and compensate for weaknesses present in the individual models. However, it was found to greatly increase training and computation time, and they also acted as black boxes that would likely perform poorly if used on other data sets. Due to this, SVMs and MLPs are considered the best performers.

Analyzing the data and experimenting with the HAR hyperparameters produced several viable approaches for decreasing inter-subject variability and boosting performance. It was found that a window size of 10 seconds was optimal for this dataset and slightly increased accuracy. Conversely, window overlap and sampling size didn't produce any significant changes in the data. An interesting approach explored in this paper was including limited intra-subject data to improve accuracy. This yielded an increase of around 5% accuracy when 30% of the intra-subject data was provided. This demonstrated a practical, realistic way in which users can produce more reliable results for HAR. An attempt was made to further increase accuracy by including the UCI dataset in the training samples. The goal was to provide "cleaner" data for the models. Ultimately, this experiment showed that this approach will not work, and classification accuracy actually decreased when using this method.

Finally, it was observed that walking up and down an incline posed the greatest difficulty for the classifiers. If we are using HAR for caloric intake calculations or for general public health, it may be reasonable to assume that amount spent walking up an incline and down an incline will be irrelevant or unimportant. If we make this assumption, then we can lump this data into the "walking comfortably" class. By doing this, the number of classes decreased to 5 and overall classification accuracy increased to 85.9%. This classification accuracy is comparable to intra-subject accuracy found in the literature. When we simply excluded samples labeled as "walking up an incline" or "down an incline", then the classification accuracy increased to 80%. This indicates that much of the inter-subject noise is from these 2 classes.

Future directions include increasing the number of participants present. This could help improve robustness in the algorithms. Furthermore, exploring more deep learning methods can be useful. Convolutional neural networks and LSTMs are promising approaches to the HAR

problem. They benefit from having automatic feature extraction, which can help achieve higher performance than standard methods.

References

[1] Actigraph. (2019, November 15). retrieved from https://actigraphcorp.force.com/support/s/

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

[3] Kerem Altun and Billur Barshan. Human Activity Recognition Using Inertial/Magnetic Sensor Units. In *Human Behavior Understanding*. Springer: Berlin, Germany, pages 38-51, 2010.

[4] D. Anguita, A. Ghio, L. Oneto, J.-L. Parra, Xavier.and Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones, in: *European Symposium On Artificial Neural Networks, Computational Intelligence and Machine Learning ESANN*, 2013.

[5] Attal, F., Mohammed, S., and Dedabrishvili, M. Physical Human Activity Recognition Using Wearable Sensors, *Sensors*, 15(12): pages 31314-41338, 2015.

[6] O. Banos, J.-M. Galvez, M. Damas, H. Pomares, and I. Rojas. Window size impact in human activity recognition, *Sensors*, vol. 14, no. 4, pages 6474–6499, 2014.

[7] Bao, L., and Intille, S. S. Activity recognition from user-annotated acceleration data. In *Proceedings of the 2nd International Conference on Pervasive Computing*, pages 1–17, 2004.

[8] A. Bayat, M. Pomplun, and D.A. Tran. A Study on Human Activity Recognition Using Accelerometer Data from Smartphones, *Procedia Computer Science*, vol. 34, pages 450-457, 2014.

[9] Casale, Pierluigi and Pujol, Oriol and Radeva, Petia. Human activity recognition from accelerometer data using a wearable device, *Pattern Recognition and Image Analysis*, 289, Springer, 2011.

[10] C. Chen, R. Jafari, and N. Kehtarnavaz. A survey of depth and inertial sensor fusion for human action recognition, *Multimedia Tools and Applications*, doi: 10.1007/s11042-015-3177-1, available online, 2015.

[11] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY: Association for Computing Machinery, pages 785-784, 2016.

[12] González, S., Sedano, J., Villar, J. R., Corchado, E., Herrero, Á, and Baruque, B. Features and models for human activity recognition. *Neurocomputing*, 167, pages 52–60, 2015.

[13] P. Gupta and T. Dallas. "Feature selection and activity recognition system using a single triaxial accelerometer," *Biomedical Engineering, IEEE Transactions* on, vol. 61, pages 1780–1786, 2014.

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

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

[16] N. Y. Hammerla, S. Halloran, and T. Plotz. Deep, convolutional, and recurrent models for human activity recognition using wearables, in *ACM IJCAI*, New York, 2016.

[17] Z. He and L. Jin. Activity recognition from acceleration data based on discrete cosine transform and svm, *In Systems, Man and Cybernetics. SMC 2009. IEEE International Conference on*, pages 5041–5044, 2009.

[18] A. Khan, N. Hammerla, S. Mellor, and T. Plötz. Optimising sampling rates for accelerometer-based human activity recognition, *Pattern Recognit. Lett.*, vol. 73, pages 33–40, 2016.

[19] A. M. Khan, Y.-K. Lee, S. Y. Lee, and T.-S. Kim. A triaxial accelerometer-based physical-activity recognition via augmented-signal features and a hierarchical recognizer, *Information Technology in Biomedicine*, vol. 14, no. 5, pages 1166-1172, 2010.

[20] Kwapisz JR, Weiss GM, and Moore SA. Activity recognition using cell phone accelerometers; *Proceedings of the International workshop on Knowledge Discovery from Sensor Data*, pages 10–18, 2010.

[21] W. Liu, Z. J. Zha, Y. Wang, K. Lu, and D. Tao. p-laplacian regularized sparse coding for human activity recognition, *IEEE Transactions on Industrial Electronics*, 63 (8) pages 5120-5129, 2016.

[22] Y. Lu, Y. Wei, L. Liu, J. Zhong, L. Sun, and Y. Liu. Towards unsupervised physical activity recognition using smartphone accelerometers, *Multimedia Tools and Appl*., pages 1–19, 2016.

[23] U. Maurer, A. Smailagic, D. P. Siewiorek, and M. Deisher. Activity Recognition and monitoring using multiple sensors on different body positions, in *Proc. Workshop BSN*, Cambridge, MA, USA, pages 113–116, 2006.

[24] A. Murad and J.-Y. Pyun. Deep recurrent neural networks for human activity recognition, *Sensors*, vol. 17, no. 11, page 2556, 2017.

[25] Niazi, Anzah Hayat Khan. *A study in Human Activity Recognition: Hierarchical Classification and Statistical Analysis*. Diss. University of Georgia, 2016

[26] Pedregosa et al. Scikit-learn: Machine Learning in Python, JMLR 12, pages 2825-2830, 2011.

[27] J.-L. Reyes-Ortiz, L. Oneto, A. Samà, X. Parra, and D. A. Anguita. Transition-aware human activity recognition using smartphones, *Neurocomputing*, vol. 171, pp. 754–767, 2016.

[28] C.A. Ronao and S. Cho. Human activity recognition with smartphone sensors using deep learning neural networks. *Expert Systems with Applications*, pages 235-244, 2016.

[29] M. Shoaib, S. Bosch, O. D. Incel, H. Scholten, and P. J. M. Havinga. Complex human activity recognition using smartphone and wrist-worn motion sensors, *Sensors*, vol. 16, no. 4, page 426, 2016.

[30] Tapia EM, Intille SS, Haskell W, Larson K, Wright J, King A, and Friedman R. 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 1-4, 2007.

[31] A. Wang, G. Chen, J. Yang, S. Zhao, and C.-Y. Chang. A comparative study on human activity recognition using inertial sensors in a smartphone, *IEEE Sensors Journal*, vol. 16, no. 11, pages 4566–4578, 2016.

[32] Wei Wang, Alex X. Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. Understanding and Modeling of WiFi Signal Based Human Activity Recognition, *In Proc. ACM MobiCom*, 2015