NTNU

Norwegian University of
Science and Technology

Data Analytics for HUNT: Recognition of Physical Activity on Sensor Data Streams

Øyvind Reinsve

Master of Science in Computer Science
Submission date: June 2018
Supervisor: Kerstin Bach, IDI

Norwegian University of Science and Technology
Department of Computer Science

Abstract

Human Activity Recognition (HAR) is the field of recognizing activities by analyzing measurements of a subject's movement and environment. A major application of HAR systems is medical research. The Nord-Trøndelag Health Study is one of the largest health studies in the world, containing health data on 120 000 subjects. The fourth version of the study (HUNT4) commenced in the fall of 2017, where activity data for the first time is collected through physical measurements and not by questionnaires. Subjects are asked to wear two accelerometers for a week to record their activities. To analyze this data an effective and accurate HAR system is needed. With the large amounts of data, a manual analysis is not feasible. Prior studies have developed promising HAR systems, classifying activities with a high degree of accuracy (Hessen and Tessem [2016], Vågeskar [2017]). This thesis aims to make improvements to the HAR system presented in Vågeskar [2017] by increasing the efficiency of the system and adding a sensor no-wear time classifier.

Three goals were defined for this thesis: Goal 1 was to explore the state of the art machine learning methods and datasets that are commonly used in HAR research. This was to be explored in a systematic literature review. Goal 2 was to increase the effectiveness of the HAR system presented in Vågeskar [2017] while maintaining the accuracy of 94 percent, based on the results of the specialization project preceding this thesis (Reinsve [2017]) which indicated that the 138 features used to train the HAR classifier in Vågeskar [2017] could be significantly reduced while maintaining the accuracy. Goal 3 was to develop a classifier that was able to detect instances of sensor no-wear time (SNT) by classifying the configuration of sensors attached to a subject at any given time.

In this thesis, a systematic literature review on machine learning methods and publicly available datasets used in HAR research, is presented. The feature importances for the 138 different features were presented. It was shown that when tested on the TFL dataset, a model with the 5 most important features was sufficient in order to achieve an accuracy of 90.0 percent, while a model using the 27 most important features was capable of reaching 94.0 percent accuracy. By calculating only the most important features, an increase in effectiveness of 5.9 times for the feature calculation step of the HAR system was achieved using 27 features. With 5 features a speedup of 23 times was achieved. The SNT classifier achieved an accuracy of 95.6 percent using 2 minute windows and a random forest classifier, when tested on the SNT dataset.

Sammendrag

Helseundersøkelsen i Nord-Trøndelag (HUNT) er den største helseundersøkelsen i Norge. Siden studiet ble unnfanget i 1984, har det blitt samlet inn helseinformasjon om 120,000 personer. Hittil har det vært gjennomført tre utgaver av studiet: HUNT1, HUNT2 og HUNT3[1]. Informasjon som høyde, vekt, blodtrykk, puls, hørsel og oksygenopptak har blit samlet inn. Høsten 2017 ble den fjerde utgaven av studiet satt i gang (HUNT4). Dette blir det første av HUNT studiene til å samle inn fysiske målinger av aktivitetsnivået til deltakerene. Det er forventet at ca. 58,000 vil ta del i studiet, og at ca. 50,000 av disse vil gjennomføre de frivillige aktivitetsmålingene. Deltakerene som ønsker å utføre målingene får utdelt to akselerometere som de blir bedt om gå med i syv dager. For å håndtere de store datamengdene trengs det et system som automatisk analyserer dataene. Dette forskiningområdet omtales som "Human Activity Recongnition" (HAR). Gjennom arbeidet til Hessen and Tessem [2016] og Vågeskar [2017] ble det utviklet et fullt fungerede HAR system for dette formålet. I denne oppgaven bygges det videre på dette arbeidet for å gjøre dette systemet bedre rustet til å takle utfordringene ved storskala-bruk. Målet er å øke effektiviten av systemet presentert i Vågeskar [2017] samt utvikle et system for deteksjon av det som kalles "sensor no-wear time".

Tre konkrete mål ble definert for oppgaven: å undersøke hvilke maskinlæringsmetoder og alment tilgjengelige datasett som blir brukt i forskning innenfor HAR fagfeltet, forbedre effektiviteten av HAR systemet presentert i Vågeskar [2017], og utvikle et system for å detektere tilfeller av "no-wear time", der sensorene ikke er montert på personen mens målingene pågår.

En systematisk litteraturstudie ble utført for å finne hvilke maskinlæringsmodeller som blir brukt i forskning på HAR, samt hvilke HAR datasett som er alment tilgjengelige. Informasjon om hvilke "features" som var mest betydningfulle for HAR systemet brukt i Vågeskar [2017] ble hentet ut. Det ble vist at en modell med de 5 viktigste features'ene var tilstrekkelig for å opnå en nøyaktighet på 90 prosent. Med 27 features oppnådde modellen 94.0 prosent nøyaktighet. Med 5 features økte effektiviteten av systemet betraktelig; utregninger ble gjort 23 ganger raskere enn for modellen med 138 features. Med 27 features var systemet 5.9 ganger raskere. Modellen for detektering av sensor no-wear time oppnådde en nøyaktighet på 95.6 prosent.

[1] *https://www.ntnu.no/hunt/om*

Preface

This master's thesis was written in the spring of 2018 for the Department of Computer Science (IDI) at the Norwegian University of Science and Technology (NTNU). The subject and scope of this thesis was defined by my supervisor, Kerstin Bach, who has been instrumental in guiding me through the process and helping with technical advice and support.

There has been several other people contributing to this project. I would like to thank Eirik Vågeskar for helping me better understand the code he developed for his masters thesis, which this project builds heavily on. I would also like to thank Paul Jarle Mork, Professor at the Department of Public Health and Nursing (NTNU), for creating the sensor no-wear time dataset on short notice, with the help of Vegar Rangul and Atle Kongsvold, both at the HUNT research center. This dataset was instrumental in helping me develop a sensor no-wear time detection system. I would also like to thank both Paul and Atle for taking the time to answer questions that I had.

Table of Contents

List of Tables

List of Figures

Abbreviations

ANN	artificial neural net
ARC	Activity Recognition Chain
AX3	Axivity AX3 sensor
CSV	Comma-Separated Values
CVA	cross validation
CV	computer vision
DT	decision trees
FI	feature importance
HAR	human activity recognition
lb	lower back
ML	machine learning
NWT	no-wear time
RF	random forests
RFC	random forest classifier
SNT	sensor no-wear time
SVM	support vector machine
th	thigh
TFL	Trondheim Free Living

Chapter 1

Introduction

This chapter introduces the background and motivation for the thesis. Furthermore, it defines the goals and the structure of the thesis.

1.1 Background and Motivation

Human Activity Recognition (HAR) is the field of recognizing activities by analyzing measurements of a subject's movement and environment. HAR systems have applications in health care services, sports, security (video surveillance etc), and personal health tracking. Another application is medical research. Gaining an understanding of the activity level of a population can help us understand the development of life style deceases, heart conditions, or mortality. This information can aid policy makers and others in taking proactive steps to improve the quality of lives of people on a substantial scale.

In Norway, the Nord-Trøndelag health study is the largest collection of health data on a population. The study spans decades, going back as far as 1984 (Hunt Research Center [2017]). In total, approximately 120,000 people, all from the province of Nord-Trøndelag, have participated in the HUNT studies. The study have collected measurements on height, weight, blood pressure, heart rate, hearing,and oxygen uptake, among others [1]. Traditionally, the HUNT studies have collected data on physical activity (PA) levels through questionnaires. Unfortunately, questionnaires about PA are bound to be unreliable and do not necessarily reflect an objective view of a person's activity level. Measuring the activity level directly with sensors would produce more accurate and objective results.

In the fall of 2017, the fourth version of the study, HUNT4, commenced. The study is a collaboration between the HUNT research center, the Norwegian University of Science and Technology (NTNU), the Norwegian ministry of Health and Care services (HOD),

[1] *https://www.ntnu.no/hunt/hunt3*

and others. It is the first HUNT study to collect PA data through direct measurement. It is expected that like in the HUNT3, approximately 58,000 people will take part in the study. Almost 30 000 have already participated [2]. Subjects are given the choice to wear two accelerometers for a period of seven days. It is expected that approximately 50,000 of the participants will take part in wearing the accelerometers. Additionally, about 10% of the subjects are offered to wear a heart rate monitor for two out of the seven days. The accelerometers, which will be the core domain of this thesis, will register every movement the subjects make throughout the day, even while they are sleeping. The data will be used to gather a better view of the overall activity level in the population. The subjects will also be given a precise summary of the activities they performed during the period they wore the sensors, giving them the opportunity to make improvements in their daily routines.

To be able to analyze the large amounts of data collected through the HUNT4 study, an accurate and effective HAR system is needed. Due to the large amount of participants, a manual analysis of the data is not feasible. This thesis builds on the foundation of two prior master theses: Hessen and Tessem [2016] and, Vågeskar [2017], both from the Department of Computer Science (IDI) at the Norwegian University of Science and Technology (NTNU). Their work served as prestudies to the HUNT4 study, and resulted in a fully functioning HAR system. This project aim to make improvements to the effectiveness of the HAR system presented in Vågeskar [2017], and to introduce a system that is able to detect instances of sensor no-wear time. No-wear time is when a sensor is detached from the subject during the recording when it was intended to be attached. The random forest classifier (RFC) developed in Vågeskar [2017] delivered great accuracy for the Trondheim free-living (TFL) dataset: 94,2%. However, as shown in the specialization thesis proceeding this master (Reinsve [2017]), the use of 138 input features to the RFC made the system slower than necessary while maintaining a similar accuracy. A significant number of the features presented in Vågeskar [2017] was not necessary to maintain the accuracy of the RFC. Reinsve [2017] also showed that the feature calculation step of the HAR system was the most time consuming by a significant margin. Therefore, a significant increase in effectiveness could be made with a corresponding reduction of features used in the model. This thesis aim to present an improved version of the HAR system presented in Vågeskar [2017], using less features in order to increase the effectiveness of the feature calculations step in the HAR system. This will reduce time it takes to predict the activities for the subjects participating in the HUNT4 study, as well as providing insight into which features are the most important to create a robust model.

In addition to increasing the effectiveness of the HAR system, this thesis aims to create a system that is capable of detecting sensor no-wear time. It is expected that for a number of the participants in the HUNT4 study, one or both sensors will fall off at some point during the week long recordings. Some of these participants will try to reattach the senors, while others will not. The sensors might also be removed for any particular reason by the participants themselves. This introduces inaccurate measurements into the HUNT4 data. The current HAR system (Vågeskar [2017]) uses a dual-sensor model for all predictions.

[2] As of May 2018

For example, if a participant wore both sensors for three days, and one sensor for the last four days, the dual-sensor model would produce inaccurate prediction results. Therefore, in order to maintain a similar degree of accuracy for this type of data, as with properly recorded data, a system that can detect when a sensor is detached from the subject would be useful. Such a system can then inform the HAR system to use either the dual-sensor model or a sensor specific single-sensor model for predictions, when necessary. To train a system for this task, a dataset with labeled instances of no-wear time was needed. A new dataset was created for this thesis for this purpose; the sensor no-wear time (SNT) dataset will be used to train a separate system to detect instances of no-wear time in recordings. The dataset consists of recordings from two subjects wearing two accelerometers each, one on the back and one on the thigh. Both subjects performed two separate protocols each. The protocols described the sequences of which sensor to put on, or take off, and for how long to wear them. This resulted in a total of four different recordings. This dataset will hopefully provide the necessary data in order to train the model used by the no-wear time detection system. With such a system in place, it is desired that the subject recordings from the HUNT4 containing no-wear time can be used; and at the same time achieve approximately the same degree of accuracy as the for the data not containing no-wear time. Put together, the improvements in effectiveness and the detection of no-wear time should result in HAR system that is better prepared for large scale usage such as with the HUNT4 study.

1.2 Goals and Research Questions

The following goals and research questions were defined for this thesis:

- **Goal 1:** Explore the state of the art in HAR systems that use machine learning
  - **Research Question 1:** What is the state of art machine learning methods used in HAR systems and how are they performing in terms of accuracy?
  - **Research Question 2:** Which publicly available datasets have been used in HAR research??
- **Goal 2:** Reduce the number of features used in the HAR classifier model to improve the speed of predicting data while maintaining a high degree of accuracy
  - **Research Question 1**: What features are the most important when used with a random forest classifier for the TFL dataset?
  - **Research Question 2**: How does the feature importances vary with sensor placement?
  - **Research Question 3**: How many and which features are sufficient to achieve 90% and 94% accuracy?
- **Goal 3:** Create a HAR system that is able to detect sensor recordings with instances of no-wear time

  - **Research Question 1**: What are the instances and reasons where data is missing or not recorded as intended in the HUNT4 data?
  - **Research Question 2**: To what degree is missing records and no-wear time data present in the HUNT4 data?
  - **Research Question 3**: What is a reasonable approach to detect no-wear time of sensors, and what accuracy can it achieve when implemented?

1.3 Thesis Structure

- **Chapter 2: Background Theory** details the background theory necessary to understand the concepts covered in this thesis.
- **Chapter 3: Systematic Literature Review** presents the systematic literature review (SLR). It includes state of the art machine learning methods used in HAR research and the datasets that are commonly used in HAR research.
- **Chapter 4: Datasets** describes the two datasets used in this thesis: the TFL and SNT datasets.
- **Chapter 5: Methodology** details the HAR system's design used for the experiments in this work.
- **Chapter 6: Experiments** presents the the experiments of this thesis, covering the setup, results, and motivation for each of the experiments.
- **Chapter 7: Discussion** discusses the results of the SLR and the experiments.
- **Chapter 8: Conclusion** summarizes and evaluates the work presented in this thesis and discusses possible areas of interest for further work on the topic.

Chapter 2

Background Theory

This chapter present background theory on subjects necessary to understand the concepts and experiments explored in this thesis. Section 2.1 briefly introduces human activity recognition, section 2.4 presents machine learning, section 2.5 introduces decision trees, section 2.6 explains ensemble methods in general and specifically the random forests method. Section 2.6 briefly introduces the performance metrics used to evaluate the classifiers in this thesis, while section 2.8 describe features and list the features used for the HAR classifier in this thesis. Lastly, section 2.3 explains sensors in the context of HAR.

2.1 Human Activity Recognition

Human Activity Recognition (HAR) is the field of recognizing the activities performed by a a person or a group of people by analyzing measurements of a subject's movement and environment. HAR has been a research field since the earliest works were done in the late '90s (Lara and Labrador [2013]). HAR systems have a wide range

2.1.1 The HAR problem definition

A comprehensive and detailed definition of HAR is given in Definition 2.1.1

Definition 2.1.1. ***HAR problem:*** *Given a set $S = S_0, ..., S_{k-1}$ of k time series, each one from a particular measured attribute, and all defined within time interval $I = [t_\alpha, t_\omega]$, the goal is to find a temporal partition $\langle I_0, ..., I_{r-1} \rangle$ of I, based on the data in S, and a set of labels representing the activity performed during each interval I_j (e.g., sitting, walking, etc.). This implies that time intervals I_j are consecutive, non-empty, non-overlapping, and such that $\bigcup_{j=0}^{r-1} I_j = I$,* ***Lara and Labrador [2013]***

2.1.2 Structure of HAR systems

A HAR system can be defined in terms of a sequence of operations performed to recognize activities from wearable sensors. In Bulling et al. [2014] this sequence of operations is presented as the Activity Recognition Chain (ARC); a sequence of five operations. A simplified overview of the the ARC is presented in figure 2.1.

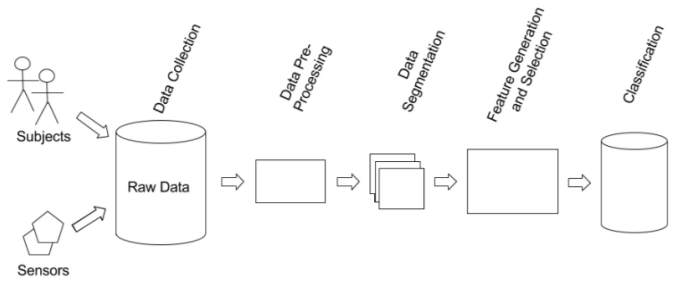


Figure 2.1: A simplified overview of the Activity Recognition Chain presented in Bulling et al. [2014]. Figure from Hessen and Tessem [2016]

The domain of this thesis is primarily the feature generation and selection step, and the classification step. The three steps: data collection, data pre-processing and data segmentation are described in detail in Hessen and Tessem [2016]. Relevant parts from these three step are presented below in section The feature generation and selection step, hereby referred to as the feature calculation step is discribed in the context of machine learning in 2.4.

2.2 Data collection

2.3 Sensors

A sensor is a device that detects or measures a device which detects or measures a physical property and records, indicates, or otherwise responds to it [1].

A range of different sensors have been used in HAR systems Lara and Labrador [2013], including accelerometers, microphones, light sensors, GPS, and others. They can be categorized in terms of what attributes they measure: environmental, acceleration, location and physiological signals (Lara and Labrador [2013]).

- **Environmental attributes:** Examples of environmental attributes include temperature, light and audio levels, humidity, etc. These attribute provide context information about the surrounding of the subject.

[1] *http://www.oed.com/view/Entry/176005?rskey=5Tz8tB&result=1&isAdvanced=false#eid*

- **Acceleration:** Accelerometers are widely used for HAR purposes. They are inexpensive,are found in many of the devices we carry around on a daily basis, and they require low power Lara and Labrador [2013]. Triaxial accelereometers are among the most widely used for recognizing activities such as walking, running, etc (Lara and Labrador [2013]). Systems using accelerometers have reported high recognition accuracies, achieving up 90 percent Lara and Labrador [2013].

- **Location:** The most common location based sensor is probably the Global Positioning System (GPS). Most phones come equipped with a GPS inside, making it a convenient sensor to use. Drawbacks of using GPS includes poor indoors performance and privacy issues.

- **Physiological signals:** There is a wide range of physiological signals: heart rate, blood oxygen.

2.4 Machine Learning

Machine learning is the science of creating computer programs that improve with experience. The use of machine learning has exploded in the last decade; dramatically improving web searches (Google etc), speech recognition systems (Apple's Siri etc), and brought us closer to the self-driving car reality. Machine learning is also a profound tool in the context of HAR. Analyzing the vast amounts of data collected through the HUNT4 study would simply not be possible without the help of machines. In this section, the basics of machine learning are introduced.

2.4.1 The Learning Problem

For the computer to be able to learn, a proper definition of learning is necessary. Definition 2.4.1 gives a definition for learning in the context of computer programs.

Definition 2.4.1. ***Learning*** *A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. [Mitchell, 1997, p. 2]*

Many learning problems can be described by the way of this definition. Consider the task of recognizing words in handwriting. The task T is then to recognize and classify words in pictures of handwritten words. The performance measure P could be the percentage of word that are correctly classified, also called the accuracy. The experience E would be a database with pictures of handwritten words, that are labeled with their correct classification. In the context of HAR the task T is to recognize activities based on some measurements of a subjects activity, with the help of experience E usually provided as a labeled set of training data. The performance measure can be several different metrics: accuracy, precision, recall, and more. The metrics used in this thesis are presented in section 2.7 One might also be interested in the amount of time the system uses; the efficiency of the system.

2.4.2 Types of Learning

There are three main types of learning in machine learning. They are defined in terms of feedback that is available to learn from. [Russel and Norvig, 2010, p. 693-695] describes the three types of learning:

Supervised learning: The agent observes input-output pairs to learn a function that maps from the input data to the output. An example of supervised learning is an agent tasked with braking a car. As input the agent receives a series of perceptions in the form of images or other sensor data. The output is provided by a teacher who says ”brake”. The agent will then try to learn what factors determines the need to brake.

Unsupervised learning: No feedback is provided to the learning agent. The agent learns patterns in the input. An example of unsupervised learning is clustering.

Reinforcement learning: With this type of learning the agent learns from reinforcements, which comes in the form of rewards or punishments. A prime example of an agent utilizing reinforcement learning would be a chess-playing agent. The agent receives no reward or punishment until the game is over. If the agent wins and receives two points, the reward indicate that the agent did something right. It does not however inform the agent of which moves it performed, prior to winning the game, that were good (or bad moves) This is up to the agent itself to decide.

These distinctions might not always as distinct. There is also semi-supervised learning which blurs the line between unsupervised and supervised learning. With semi-supervised learning the learning agent is given only a few labeled training examples.

2.4.3 Supervised Learning

The HAR problem is typically solved with supervised learning. A definition of the supervised learning task is defined in definition 2.4.2.

Definition 2.4.2. ***Supervised learning*** *Given a* ***training set*** *of N example input-output pairs* $(x_1, y_{1)}, (x_2, y_2), ..., (x_N, y_N)),)$ *where each* y_j *was generated by an unknown function y = f(x), discover a function h that approximates the true function f. [Russel and Norvig, 2010, p.695]*

2.4.4 Machine Learning tasks with Respects to Output

ML tasks tasks can be categorized in terms of the output they provide. Classification, regression, and clustering. The three methods are described in [Russel and Norvig, 2010, p. 694-696]
Classification: when the output, y, is one of a finite set values. An example would be the finite set of values (*sunny, cloudy, rainy*) if trying to predict the weather of tomorrow, as an example. If y can only take on two different values, it is called binary classification.

Regression: when the output, y, is a number. An example would be to predict tomorrow's temperate. The probability of predicting the exact number is actually zero, hence regression is finding a conditional expectation or average value of y.
Clustering: detects potentially useful clusters in the input data.

2.4.5 Classification

The HAR problem is a classification task. As seen in Definition 2.1.1, we wish to find a temporal partition of the time interval, where each partition gets assigned an activity label. The activity is selected from a set of activities.

2.5 Decision Trees

The random forest (RF) model is used for both the HAR and SNT systems in this thesis. RF are ensembles of decision trees (DT). Hence, understanding DT are important to grasp the RF model. A short introduction to DT is presented in this section.

According to [Mitchell, 1997, p. 52], decision tree learning is one of the most widely used and practical methods for inductive inference. Decision trees is a method for approximating discrete-valued functions that is robust to noisy data ([Mitchell, 1997, p. 52]). Decision trees can be used to solve a wide range of ML tasks: classification, regression, clustering, ranking and probability estimation (Flach [2012]).

Decision tree is a simple tree based method. DTs have been used extensively for AR purposes, but the reported accuracies have generally not been on par with other methods such as random forest, artificial neural nets, support vector machines and hidden Markov models

2.6 Ensemble methods and Random Forests

An ensemble of classifiers is a set of classifiers where the individual decisions of the included classifiers are combined in some way (Dietterich [2000]). The classifier ensemble is generally much more accurate than the individual classifiers in the ensemble.

Dietterich [2000]

2.6.1 Bootstrapping

Bootstrap is a method for creating different random samples of datasets. It creates random subset of data by uniformly sampling the data with replacement. Because samples are taken with replacement, bootstrap samples will general contain duplicates [Flach, 2012, 331]. This is not a disadvantage however, the difference between the bootstrap samples introduces diversity to the models in the ensemble.

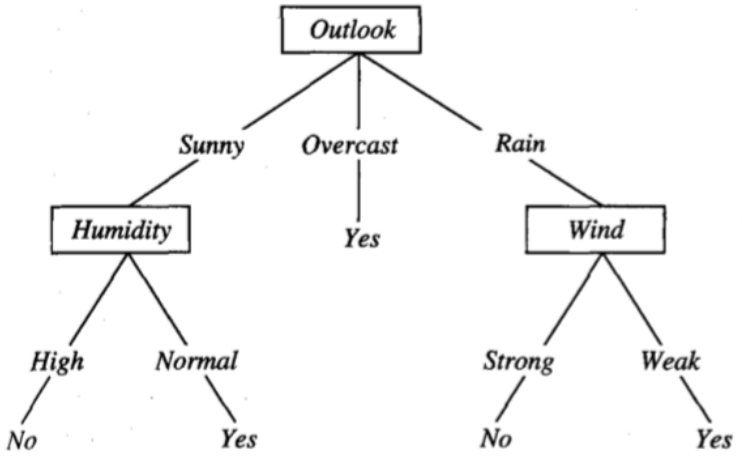


Figure 2.2: An illustration of a decision tree for the concept *"PlayTennis"*. Each node in the tree corresponds to a test of an attribute of an instance. Examples are classified by sorting the examples through the tree and testing the attribute values of the examples on the internal nodes in the tree. An instance with the attribute values *(Sunny, No, Strong)* is classified as *Yes*. Figure from [Mitchell, 1997, p. 52]

2.6.2 Bagging

Bagging is short for "bootstrap aggregating. It is a simple, yet effective ensemble method that created models on different bootstrap samples of the original dataset. Figure 2.8 presents the basic bagging algorithm.
The bagging algorithm is relatively simple. It takes a dataset, an ensemble size, and a learning algorithm as input.

2.6.3 Random Forests

Random forest is an ensemble classifier that was properly introduced first in Breiman [2001]. The definition of random forests shown in 2.6.1 was presented in Breiman [2001].

Definition 2.6.1. ***Random Forest*** *A random forest is a classifier consisting of a collection of tree-structured classifiers* $\{h(\boldsymbol{x}, \theta_{k)}, k = 1, ...\}$ *where* θ_k *are independent identically distributed random vector and each tree casts a unit vote for the most popular class at input* $\boldsymbol{x}$*. (Breiman [2001]).*

One of the advantages of RF is that it can be used for both regression and classifications problems.

The RF introduces additional randomness over the bagging method. When growing the trees, nodes are split on the best feature available in a random subset of the remaining features at that point. This helps bring great diversity in the trees that are created, in turn

```
Input  : data D; set of features F.
Output : feature tree T with labelled leaves.
if Homogeneous(D) then return Label(D) ;          // Homogeneous, Label: see text
S ←BestSplit(D, F) ;                              // e.g., BestSplit-Class (Algorithm 5.2)
split D into subsets D_i according to the literals in S;
for each i do
    if D_i ≠ ∅ then T_i ←GrowTree(D_i, F) else T_i is a leaf labelled with Label(D);
end
return a tree whose root is labelled with S and whose children are T_i
```

Figure 2.3: The GrowTree algorithm presented in [Flach, 2012, p. 132]. Combined with the algorithm in figure 2.4 they make up an decision tree learner.

```
Input  : data D; set of features F.
Output : feature f to split on.
I_min ←1;
for each f ∈ F do
    split D into subsets D_1,...,D_l according to the values v_j of f;
    if Imp({D_1,...,D_l}) < I_min then
        I_min ←Imp({D_1,...,D_l});
        f_best ←f;
    end
end
return f_best
```

Figure 2.4: The BestSplit-Class algorithm presented in [Flach, 2012, p. 137]

generally resulting in better models.

Input : data set D; ensemble size T; learning algorithm $\mathscr{A}$.
Output : ensemble of models whose predictions are to be combined by voting or averaging.

1 **for** $t = 1$ to T **do**
2 build a bootstrap sample D_t from D by sampling $|D|$ data points with replacement;
3 run $\mathscr{A}$ on D_t to produce a model M_t;
4 **end**
5 **return** $\{M_t | 1 \leq t \leq T\}$

Figure 2.5: The bagging algorithm. The algorithm train an ensemble of models from bootstrap samples of the original dataset. Figure from [Flach, 2012, p. 332]

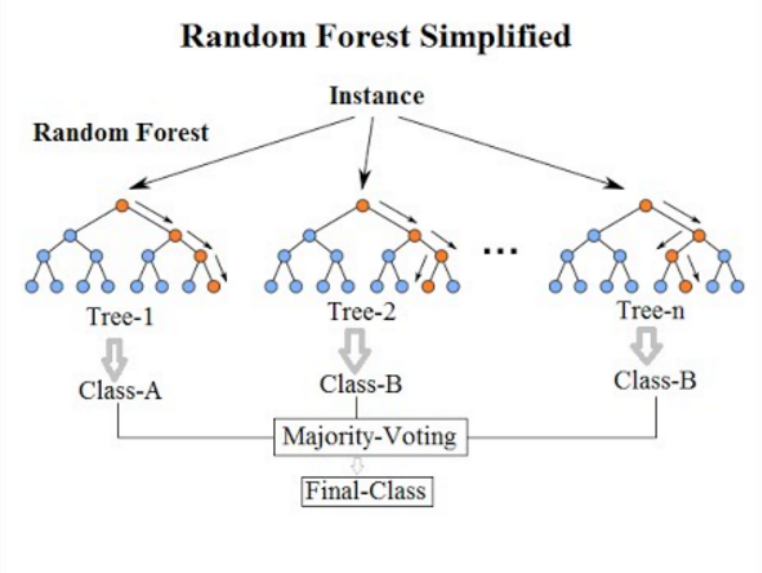


Figure 2.6: Illustration of the Random Forest method. The model create n instances of decision trees, each with a random subset of the data. The predicted class is the majority vote of all the trees. [2]

[2] *https://cdn-images-1.medium.com/max/1600/1*i0o8mjFfCn-uD79-F1Cqkw.png*

Input : data set D; ensemble size T; subspace dimension d.
Output : ensemble of tree models whose predictions are to be combined by voting or averaging.

1 **for** $t = 1$ to T **do**
2 build a bootstrap sample D_t from D by sampling $|D|$ data points with replacement;
3 select d features at random and reduce dimensionality of D_t accordingly;
4 train a tree model M_t on D_t without pruning;
5 **end**
6 **return** $\{M_t | 1 \leq t \leq T\}$

Figure 2.8: The random forest algorithm. Figure from [Flach, 2012, p. 333]

2.7 Quality metrics

Quality metrics are metrics the quality of classifications are evaluated on. Many different metrics are used in the field of ML. The quality metrics used to evaluate the two classifiers in this thesis are listen in table 2.1.

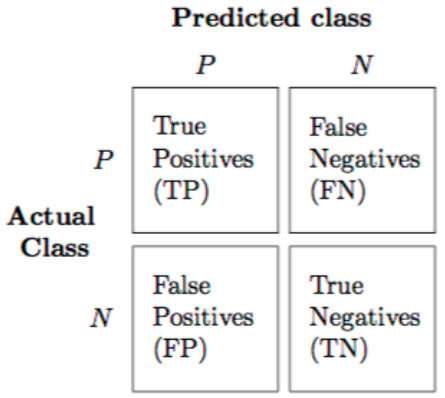


Figure 2.9: A confusion matrix

Metric	Description
$Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$	Percentage of correctly classified samples
$Precision = \frac{TP}{TP+FP}$	Proportion of predicted positives that are positives
$Recall = \frac{TP}{TP+FN}$	Proportion of positives that are correctly classified as positives
$Specificity = \frac{TN}{TN+FP}$	Proportion of negatives that re correctly classified as negatives
$F_1 = 2 \times \frac{Recall \times Precision}{Recall+Precision}$	The weighted average of precision and recall

Table 2.1: Quality metrics

2.8 Features

A model is only as good as its features. Features can be though of as a kind of measurement of that can be performed on any instance. More specifically they are functions that map from an instance space to a domain of the feature; the feature values.

2.8.1 Time Domain Features

Time domain features are statistical features that are extracted from a signal without first transforming it to some other domain. Time domain features are efficient to extract, with a computational complexity of O(n) where n is the number of samples. The time domain features used in this thesis is shown in table 2.10. They are the same time domain features used in Vågeskar [2017] For a comprehensive explanation on time domain features see Vågeskar [2017].

2.8.2 Frequency Domain Features

Frequency features that require the signal to be transferred to the frequency domain to be extracted. Frequency features are more computationally demanding to extract than time domain features. The signal has to undergo frequency domain transform. The frequency domain features used in this thesis are shown in figure 2.11. These are the same frequency domain features as used in Vågeskar [2017]. For a comprehensive explanation of frequency domain features see Vågeskar [2017].

Name	Definition	Description
Mean	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	Arithmetic mean of values for an axis.
Standard deviation	$s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$	Root of the uncorrected variance (the average squared distance from mean).
Skewness	$b_x = \frac{\frac{1}{n}\sum_{i=1}^{n}(x - \bar{x})^3}{s_x^3}$	How "skewed" the distribution of values are around the mean.
Magnitude maximum, mean, and standard deviation	$m_i = \sqrt{x_i^2 + y_i^2 + z_i^2}$ $\bar{m}, s_m, \max(m)$	The maximum, mean, and standard deviation of the magnitude of the signal.
Zero crossing rate	$zcr_x = \frac{\sum_{i=2}^{n} \lvert sgn(x_i) - sgn(x_{i-1})\rvert}{2(n-1)}$	Number of times the signal's value changes from negative to positive and vice versa.
Mean crossing rate	$mcr_x = \frac{\sum_{i=2}^{n} \lvert sgn(x_i - \bar{x}) - sgn(x_{i-1} - \bar{x})\rvert}{2(n-1)}$	Like zcr, but number of times the mean is crossed.
Root mean square	$rms_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$	The root of the mean of the squared values.
Energy	$E_x = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad E_{total} = \frac{1}{3N}(E_x + E_y + E_z)$	A measure of the signal's strength.
Median	$\tilde{x}_{odd} = x_{\frac{n+1}{2}}, \quad \tilde{x}_{even} = 1/2(x_{\frac{n}{2}} + x_{\frac{n}{2}+1})$	Value(s) separating the sorted values into two equal halves.
Range	$\max(x) - \min(x)$	Difference between maximum and minimum of a sequence
Interquartile range	$iqr_x = Q_{3_x} - Q_{1_x}$	A quarter of the values in the sorted sequence x are below or equal to Q_{1_x}, and three quarters below or equal to Q_{3_x}
Correlation	$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$	Pearson's product-moment coefficient. The degree of linear dependence between two series.
Hadamard product mean, standard deviation, and maximum	$xy = x \circ y, \quad xyz = x \circ y \circ z$ $\overline{xy}, s_{xy}, \max(xy); \quad \overline{xyz}, s_{xyz}, \max(xyz)$	The Hadamard product is the element-wise multiplication of the entries in a vector, resulting in a vector of equal length.

Figure 2.10: Time domain features. Figure from Vågeskar [2017]

Name	Definition	Description
Mean amplitude	$\bar{a} = \frac{1}{k}\sum_{j=0}^{k} a_j$	The arithmetic mean of the amplitudes.
Amplitude standard deviation	$s_a = \sqrt{\frac{1}{k}\sum_{j=0}^{k}(a_j - \bar{a})^2}$	The root of the uncorrected variance for all the amplitudes.
Maximum amplitude	$\max(a)$	The maximum amplitude.
Median amplitude	$\tilde{a}_{odd} = a_{\frac{k+1}{2}}$ $\tilde{a}_{even} = {}^1/_2(a_{\frac{k}{2}} + a_{\frac{k}{2}+1})$	The value which separates the sorted amplitudes into two equally sized halves. If the number of amplitudes is even, it is the arithmetic mean of the two values which separate the values.
Spectral centroid	$sc_a = \frac{\sum_{j=0}^{k} a_j \times f_j}{\sum_{j=0}^{k} a_j}$	Analogous to the center of mass of the frequencies if one regards the amplitude a_j as analogous to volume and the frequency f_j as analogous to density.
Dominant frequency	$f_{\mathrm{argmax}_j\, a}$	The frequency with the maximum amplitude.
Spectral entropy	$p_j = \frac{a_j^2}{\sum_{j=0}^{k} a_j^2}$ $H = -\sum_{j=0}^{k} p_j * \log p_j$	The disorder in the spectrum.

Figure 2.11: Frequency domain features. Figure from Vågeskar [2017]

Chapter 3

Systematic Literature Review

This chapter present the systematic literature review (SLR).

3.1 Introduction

A SLR is a way of systematically identifying and assessing literature on a topic, narrowing down the set of identified work, and synthesize the research to answer a set of research questions. There are multiple definitions of systematic reviews available, two of them are presented here in Definition 3.1.1 and 3.1.2.

Definition 3.1.1. ***Systematic review*** *"A review of a clearly formulated question that uses systematic and explicit methods to identify, select, and critically appraise relevant research, and to collect and analyse data from studies that are included in the review. Statistical methods (meta-analysis) may or may not be used to analyse and summarise the results of the included studies" (Siddaway [2014]).*

A shorter definition is shown in 3.1.2.

Definition 3.1.2. ***Systematic review*** *A systematic review attempts "to identify, appraise and synthesize all the empirical evidence that meets pre-specified eligibility criteria to answer a given research question"* [1]

The goal of this SLR is to answer the two research questions of goal 1, presented in section 1.2.
The SLR presented in this chapter follows the guidelines of the SLR process described in Siddaway [2014]. Some of the stages in the process are done less comprehensively than described, these are presented as limitations of the review, in section 7.1.2. Additionally, inspiration and ideas found in Prestmo [2017] were included in the SLR.

[1] *http://www.cochranelibrary.com/about/about-cochrane-systematic-reviews.html*

3.2 Systematic Literature Review

Siddaway [2014] describes a five stage SLR process: scoping, planning, identification (searching), screening, and eligibility. These stages are carried out successively in this section. Each stage is briefly explained to the reader. The results of the research synthesis and analysis is presented in section 3.3, followed by a summary and discussion in chapter 7.

3.2.1 Scoping

Scoping is the first step of the SLR process. The goal is to formulate research questions and clarify whether the planned SLR have been conducted by others in the field. The goals and research questions for this thesis are presented in section 1.2. Goal 1, and its research questions RQ1 and RQ2, are to be answered by this SLR.

- **Goal 1:** Explore the state of the art in HAR systems that use machine learning.
  - **Research Question 1:** What is the state of art machine learning methods used in HAR systems and how are they performing in terms of accuracy?
  - **Research Question 2:** Which publicly available datasets have been used in HAR research?

It is important to emphasize that the focus of this thesis, and hence the SLR, lies on activity recognition in the context of inertial sensors, particularly accelerometers, which are used in the experimental part of this thesis. Hence, vision-based methods are not considered relevant in the context of this literature review.

Search for similar systematic literature reviews
Prior to conducting a SLR, it is necessary to research whether similar SLR have been conducted by others, to avoid doing duplicate and unnecessary work. The two RQs listed above are not directly related to each other, and as such, the search for similar SLRs was carried out with to similar, but separate searches; one for each of two research questions. The results of these searches are shown below. Google Scholar and Google was used as search engines. When searching for articles, articles published prior to the last ten years were filtered out from the search results, as they where seen as less relevant than more recent research. 6 search string were created for both of the searches, based on the terms found in table 3.2.

- **Research Question 1:** For each of the 6 search string, the top 20 most relevant articles were examined by screening the title and/or abstract. Based on the findings, there seems to be very few proper systematic literature reviews on HAR methods and their performance. However, there are multiple surveys and articles covering the topic in their related work/literature review sections, summarizing the topic well. These seem to go a long way in covering machine learning methods in HAR, and their performance. Therefore, the systematic literature review part, related to RQ1, will be focused on extracting information from these studies, in addtion to searching for studies more recent that may not be included in those papers.

- **Research Question 2:** The same procedure as for RQ1 was carried out for RQ2. Based on the findings of this search, no SLR was found on the topic. Very few papers in general seemed to cover the topics of available dataset.

The articles that presented relevant surveys or summaries on the topics of the two topics were included in the SLR. These articles are presented in table 3.1.

Study ID	Title	Authors
S1	Activity recognition using inertial sensing for healthcare, well-being and sports applications: A survey	Akin Avci et al.
S2	A Survey on Human Activity Recognition using Wearable Sensors	Oscar D. Lara et.al
S3	A Tutorial on Human Activity Recognition Using Body-Worn Inertial Sensors	Andreas Bulling et al.
S4	A Study on Human Activity Recognition Using Accelerometer Data from Smartphone	Akram Bayat et al.
S5	Activity identification using body-mounted sensorsfi?!a review of classification techniques	Stephen J Preece et al.
S6	A Survey on Activity Detection and Classification Using Wearable Sensors	Maria Cornacchia et al.
S7	UniMiB SHAR: A Dataset for Human Activity Recoignition Using Acceleration Data from Smartphones	Daniela Micucci et al.
S8	Accelerometry analysis of physical activity and sedentary behavior in older adults: a systematic review and data analysis	E. Gorman et al.

Table 3.1: The articles that were included after searchin for similar SLRs.

3.2.2 Planning

The planning phase of the SLR consists of three phases: breaking down the research questions into search questions and search terms, formulating inclusion and exclusion criterias, and creating a comprehensive record keeping system.

Search questions
Six search questions (SQs) were created to break down the research questions into more specific and detailed questions. This approach was taken from Prestmo [2017]. Separately, the SQs corresponds to parts of a RQ, collectively they represent the research questions.

SQ1 What machine learning methods are applied in HAR research?

SQ2 What accuracy do these methods achieve?

SQ3 Which ML methods achieves the highest degree of accuracy?

SQ4 What are the publicly available datasets used in HAR research?

SQ5 What devices or sensors are used to collect these dataset?

SQ6 How many subjects are typically used to collect data in these dataset?

SQ1-SQ3 are related to RQ1, and SQ4-SQ6 are related to RQ2. Search terms were created based on the search, and research questions. Additionally, some words that were closely related or synonymous with words in the list was added to increase the likelihood of finding relevant articles.

Search Terms
A total of 13 search terms were created. The terms were divided into 3 groups with related terms and synonyms (table 3.2). An article was classified as relevant if it contained at least one of the terms from group 1, and one from either group 2, 3, or both. Based on this definition, a total of 19 search strings were created.

	Group 1	**Group 2**	**Group 3**
Term 1	Human Activity Recognition	Machine learning	Data set(s)
Term 2	Activity Recognition	Machine recognition	Dataset(s)
Term 3	Human activities	Artificial Intelligence	Data
Term 4	Daily living	Algorithms	Accelerometer
Term 5		Random forest	

Table 3.2: The search terms of the SLR, divided into four groups of related terms.

Inclusion and exclusion criteria
Inclusion and exclusion criteria makes it possible to clearly define the boundaries of the SLR. They are used as criteria for evaluating the articles (section 3.2.4 and 3.2.5) in three steps: title, abstract, and full-text screening. In defining the criteria, a range of aspects were considered: Field of study, topic, research design, time frame, and data. Relevant studies are the ones that meet the inclusion criteria, but not the exclusion criteria.

The defined inclusion criteria:

IC1 The field of study of the article is computer science

IC2 The main topic of the article is activity recognition

IC3 The article is published after 2007

IC4 The study focuses on machine learning methods and/or datasets used in HAR

IC5 The HAR dataset(s) used in the study is collected with an accelerometer

IC6 Empirical results are presented in the study

The defined exclusion criteria:

EC1 The focus of the article is on movement analysis and/or gesture recognition

EC2 The focus of the article is on activity recognition using computer vision methods

EC3 The set of activities differs significantly from the set used in this thesis

EC4 The study uses fewer than 10 subjects for data collection

Record keeping system Two record keeping systems were used for this SLR. To keep track of searches, Microsoft OneNote was used. Records were kept for all searches in a systematic fashion. To keep track of articles, tables were created for each step of the screening process, listing all of the articles for each step. The articles selected in the last filtering step (eligibility) were added to EndNote X8, for reference management.

3.2.3 Identification

The identification step of the SLR is concerned with finding all relevant works that addresses the research questions. This process starts with selecting electronic databases to conduct searches on. It is favorably if two or more databases are used in the search process. This stage also serves as a refinement step, where the inclusion and exclusion criteria, as well as the search questions and search terms, can be refined iteratively if the search results are not satisfactory. The search terms and criteria defined in Section 3.2.2 went through multiple steps of refinement to improve the results of searches.

Selecting the right electronic databases to conduct the searches on is important to increase the likelihood of finding a wide range of relevant and high quality articles. There are many academic databases and search engines to chose from; each with their own advantages and disadvantages. For this SLR, Google Scholar was selected as the main search engine. It is an academic search engine estimated to contain about 160 million

documents (Ordua-Malea et al. [2014]).

Additional searches
To reduce the likelihood of missing out on highly relevant studies, additional searches were performed. Two steps were taken: (i) Search the standard Google search engine with some of the search strings, (ii) Examine the reference section of works identified.

(i) Using Google Scholar exclusively, might lead to some relevant studies from being discovered . Not necessarily because they are not in the database, but because for some reason, they did not show up in one of the searches. By searching the standard Google search engine, more articles could potentially be found. This approach yielded favorable results; several highly relevant articles were found by searching Google with some of search strings created for the Google Scholar search. A PRISMA flow diagram for this SLR is presented in figure 3.1. The PRISMA flow diagram illustrates how many articles (records) that were identified through the different search processes, and the number of articles filtered out through the different screenings.

(ii) The reference section of the works that were included after the title and abstract screening in section 3.2.4 were screened for additional relevant studies. This step helped find some articles for the SLR, and some articles for the background theory in Chapter 2.
Searching
Google Scholar was searched with all of the 19 search string that were created. Google was searched with 6 of the search string. Each of top 20 articles for each of the search string, on both databases were screened (Section 3.2.4).

Database	Search strings	Hits
Google Scholar	19	686,287
Google	6	1,000,000 +

Table 3.3: The number of search results returned, for all the search strings on each of the databases

3.2.4 Screening

This stage is concerned with sifting the title and abstracts of the articles identified in the identification step of the SLR, and exporting the articles to the reference manager (EndNote X8). The screening procedure was divided into two stages to find articles more effectively: title screening and abstract screening.

Title screening
A total of 517 articles were identified through the search process. These were all evaluated on the basis of their title relative to the inclusion and exclusion criteria defined for this SLR. For the title screening, IC1, IC2, IC3 and IC4, as well as EC1, and EC2 were the criteria for inclusion and exclusion respectively. A total of 39 articles passed the title screening, while 478 were excluded.

Abstract screening
The 39 articles that passed title screenings were accessed to read the abstract section of the article. In this step all the inclusion criteria and EC1-EC3 were used. A total of 17 article passed the abstract screening, excluding 22 of the articles that passed the title screening.

3.2.5 Eligibility

In the eligibility stage of SLR process, the goal is to reduce the set of works identified, to the final set to be be used as the resources for the synthesize and analysis. This is to be accomplished by screening the full-text version of articles and extracting relevant information to be analyzed later in the process. The inclusion and exclusion criteria are the metric by which the articles are evaluated by. For this screening, all of the inclusion and exclusion criteria were used to determine the relevance of articles. Out of the 17 articles that passed the abstract screening, 14 articles passed the full-text screening and were included in the SLR. These 14 articles include the 8 articles that were included in the search for other systematic literature reviews.

In total, 517 articles were examined. 39 out of the 517 articles passed the title screening. 17 out of the 39 articles that passed the title screening, passed the abstract screening. 14 articles were included in the study after passing the full-text screening step. The final set of included articles is presented in table 3.4.

3.3 Results

In this section the search questions are answered on the basis of the findings from the SLR.

SQ1 What machine learning methods are applied in HAR research?

MFA bin Abdullah et al. [2012] presents a comprehensive list of machine learning models used in HAR research. It is evident that most of the popular classifiers have and are being used in HAR research and systems. Based on this sample of studies, Decision Trees appear to be used frequently inn research. This is true for other classifiers such as the Support Vector Machine (SVM), Naive Bayes, and k-Nearest Neighbour as well.

Study ID	Title	Authors
S1	Activity recognition using inertial sensing for healthcare, well-being and sports applications: A survey	Akin Avci et al.
S2	A Survey on Human Activity Recognition using Wearable Sensors	Oscar D. Lara et.al
S3	A Tutorial on Human Activity Recognition Using Body-Worn Inertial Sensors	Andreas Bulling et al.
S4	A Study on Human Activity Recognition Using Accelerometer Data from Smartphone	Akram Bayat et al.
S5	Activity identification using body-mounted sensorsfi?!a review of classification techniques	Stephen J Preece et al.
S6	A Survey on Activity Detection and Classification Using Wearable Sensors	Maria Cornacchia et al.
S7	UniMiB SHAR: A Dataset for Human Activity Recoignition Using Acceleration Data from Smartphones	Daniela Micucci et al.
S8	Accelerometry analysis of physical activity and sedentary behavior in older adults: a systematic review and data analysis	E. Gorman et al.
S9	Human Activity Recognition: Using Sensor Data of Smartphones and Smartwatches	Andreas Dengel et al.
S10	Centinela: A human activity recognition system based on acceleration and vital sign dat	Oscar D. Lara et al.
S11	Activity Recognition using Cell Phone Accelerometers	Jennifer R. Kwapisz et al.
S12	Simple and Complex Activity Recognition through Smart Phones	Stefan Dernbach et. al
S13	Classification Algorithms in Human Activity Recognition using Smartphones	Mohd Fikri Azli bin Abdullah et al.
S14	A comparison study of classifier algorithms for mobile-phonefis accelerometer based activity recognition	Media Anugerah Ayu et al

Table 3.4: The final list of selected works for the SLR.

Year	Ref	Classification Algorithms
2011	[39]	Hidden Markov Models
2011	[37]	Multiclass Logistic Regression
2011	[40]	Transfer learning Embedded Decision Tree
2011	[51]	Hidden Markov Chain
2011	[38]	Decision Tree, Naïve Bayes, Random Forest, Logistics Regression, RBF Network, Support Vector Machine
2011	[53]	Smoothed Single-layer Hidden Markov Models
2011	[34]	Decision Tree
2011	[36]	Hidden Markov Model
2010	[43]	Support Vector Machine
2010	[42]	Decision Tree, Logistic Regression and Multilayer Neural Networks
2010	[44]	Recurrent Fuzzy Inference Systems
2010	[45]	k-Nearest Neighbour, Direct Density, Class Local Outlier Factor, Local Classification Factor
2010	[52]	Artificial Neural Network
2010	[41]	Active Learning, Self-learning, and Co-learning
2010	[47]	Decision Tree, Bayesian Network, Naïve Bayes, k-Nearest Neighbour, Support Vector Machine
2010	[46]	Decision Tree, k-Nearest Neighbours and Sequential Minimal Optimization
2010	[27]	Gaussian Mixture Model (GMM) and Support Vector Machine
2009	[54]	K-Nearest Neighbour
2009	[55]	Artificial Neural Network
2009	[49]	Decision Trees, Naive Bayes, k-Nearest Neighbour and the Support Vector Machine
2008	[48]	k-Nearest Neighbour, Support Vector Machine, Adaptive Minimum-distance
2008	[50]	Naïve Bayes

Figure 3.2: A comprehensive list of the machine learning models used in HAR research. Figure is taken from bin Abdullah et al. [2012] and have been modified.

SQ2 What accuracy do these methods achieve?

When assessing the accuracy of machine learning models across studies it is important to consider that most studies use different datasets, features, parameters, etc. Hence, an absolute comparison between models across multiple datasets is unreliable. Therefore, a comparison between models tested on the same dataset is more representative of the differences in performance between models. Such a comparison is presented in Ismail

et al. [2012], where 45 algorithms were tested on a dataset collected specifically for the study, using an HTC Mini mobile phone. The results of the most accurate out of the 45 classifiers are shown in figure 3.3.

Classifier Category	Algorithm		Training Accuracy		Testing Accuracy	
	Shirt Pocket	Hand Palm	Shirt Pocket	Hand Palm	Shirt Pocket	Hand Palm
Bayes	NaiveBayes NaiveBayesSimple NaiveBayesUpdateable	NaiveBayes NaiveBayesSimple NaiveBayesUpdateable	99.06%	99.00%	95.00%	99.00%
Functions	SimpleLogistic	SMO	98.75%	99.75%	93.75%	100.00%
Lazy	IB1 Ibk	IB1 IB*k*	99.06%	100.00%	92.50%	100.00%
Meta	RotationForest	ClassificationViaRegression RandomCommittee	99.38%	99.25%	95.00%	98.00%
Misc	VFI	VFI	99.06%	96.50%	91.25%	92.00%
Rules	DTNB	NNge	100.00%	99.25%	90.00%	98.00%
Trees	LMT	FT	98.44%	99.50%	93.75%	100.00%

Figure 3.3: A comparison of the accuracy achieved by different machine learning methods trained and tested on the same dataset. Figure is taken from Ismail et al. [2012]

Preece et al. [2009] presents a comprehensive overview of studies that compare different classifiers, and their results. This presents the opportunity to look at how two given models can perform better or worse than the other depending on the study. This is why it is difficult to compare models without training and testing them on the same datasets; one classifier might be better for a given set of features and data, and then perform worse when used in conjunction with other data and features. Figure 3.4 lists the studies and the results of the comparisons.

Publication (number of subjects)	Activities (number of activities)	Accelerometer placements	Inter-subject classification accuracy
Bao and Intille (2004) (20 subjects)	Walking, sitting, cycling, running, vacuuming, folding laundry and more (20)	Shank, thigh, upper arm, wrist and hip	Decision tree (84%) kNN (83%) Naive Bayes (52%)
Parkka *et al* (2006) (16 subjects)	Lying, sitting, walking, Nordic walking, rowing, cycling and more (8)	Chest and wrist	Decision tree (86%) Hierarchical (82%) Neural network (82%)
Maurer *et al* (2006) (6 subjects)	Sitting, standing, walking, ascending/descending stairs and running (6)	Wrist	Decision tree (87%) Naive Bayes[a] (<87%) kNN[a] (<87%)
Pirttikangas *et al* (2006) (13 subjects)	Typing, watching TV, drinking, walking upstairs, cycling and more (17)	Both wrists, thigh and necklace	Neural network (93%) kNN (90%)
Ermes *et al* (2008) (12 subjects)	Lying, sitting, walking, Nordic walking, rowing, playing football and more (9)	Hip and wrist	Neural network (87%) Hierarchical (83%) Decision tree (60%)
Ravi *et al* (2005) (2 subjects)	Standing, running, sit-ups, vacuuming, brushing teeth, walking and more (8)	Waist	Naive Bayes (64%) SVM (63%) Decision trees (57%) kNN (50%)
Lester *et al* (2005) (2 subjects)	Walking, driving, jogging, ascending/descending in an escalator and more (10)	Shoulder	Naive Bayes (67%) HMM (47%) HMM and binary classifiers (95%)
Allen *et al* (2006) (6 subjects)	Sitting, standing, lying, walking and four postural transitions (8)	Waist	Gaussian mixture model[b] (91%) Hierarchical (71%)

[a] No data were presented on classification accuracy.
[b] Some subject-specific training was used for this classifier.

Figure 3.4: An overview of studies comparing different classifiers. Figure is taken from Preece et al. [2009]

An overview of the performance of different classifiers across multiple studies is presented in Avci et al.. This overview is shown in figure 3.5. The results indicate that the Decision Tree classifier performs consistently well in the three studies that are listed. The results for Nearest Neighbour shows that the variations in the accuracy of a classifier can be significant, achieving a 91 percent accuracy in one study and 49.7 percent in another.

Classification Method		Ref.	Sensor Placement	Accuracy
Threshold-based		[34]	rear hip, neck, wrists, knees, and lower legs	NA
		[78]	waist	NA
		[40]	waist	89.14%
		[79]	waist	NA
Pattern Recognition	Decision Tables	[55]	wrist, upper-arm, waist, thigh, ankle	46.75%
		[56]	waist	46.67%
	Decision Trees	[55]	wrist, upper-arm, waist, thigh, ankle	84.26%
		[39]	waist	90.8%
		[80]	thorax	80%
	Nearest Neighbor	[56]	waist	49.67%
		[60]	wrist, belt, necklace, trousers pocket, shirt pocket, bag	NA
		[58]	wrist	91%
	Naïve Bayes	[34]	rear hip, neck, wrists, knees, and lower legs	NA
		[9]	knee	NA
		[81]	NA	83.97%
	SVM	[56]	waist	73.33%
		[57]	ext. objects	84.28%
		[44]	NA	87.36%
	HMMs	[82]	shoulder, waist, wrist	90%
		[83]	wrist, elbow	72%
		[63]	10 for right arm, 9 for left	87.36%
	GMMs	[84]	hip	88.76%
		[85]	waist	76.6%
Artificial Neural Networks		[86]	chest	84%
		[42]	wrist	95

Figure 3.5: An overview of the performance of different classifiers across multiple studies. Figure is taken from Avci et al.

Another study testing multiple classifiers on the same dataset is Bayat et al. [2014]. The study test 6 different classifiers. The Multilayer perceptron classifier achieves the highest accuracy of the 6 models when the smartphone is placed both in the hand and in the pocket of the subject. The random forest classifier comes in at a strong second place if averaging the results of the two experiments, edging out the SVM by a slight margin.

Classifier	Accuracy (in-hand)	Accuracy (in-pocket)
Multilayer Perceptron	89.48%	89.72%
SVM	88.76%	72.27%
Random Forest	87.55%	85.15%
LMT	85.89%	85.04%
Simple Logistic	85.41%	85.05%
Logit Boost	82.54%	82.24%

Figure 3.6: An overview of the accuracies of 6 different classifiers commonly used for HAR purposed, tested on the same dataset. Figure is taken from Bayat et al. [2014]

SQ3 What ML method have the highest accuracy?

There might not be one true answer to this questions. It depends on a multitude of factors, as discussed in SQ2. Through the research identified and included in this SLR we can however present some suggestion to the answer of this question. Looking at figure 3.4 we can see that the decision tree classifier performs better than both k-Nearest Neighbour (kNN) and Naive Bayes in both the Bao and Intille study, as well as the Maurer study. This might indicate that the decision tree classifier is a better classifier for HAR purpose than the two. The same than be said about the result in 3.5 However, this is presumable not always the case. As seen in the Ravi study in figure 3.4, the Naive Bayes classifier performs significantly better than the decision tree classifier. The highest accuracy achieved across all the studies comparing different classifiers is 95 percent. This was achieved in the Lester study using binary classifiers. This does however not indicate that binary classifiers are the most accurate classifiers. In figure 3.5, Artificial Neural Network (ANN) delivered the highest accuracy of all the classifiers with 95 percent. This does not mean that it is the best classifier for HAR purposes. To evaluate such a claim, the other classifiers would have to be tested on the same data, and in the same way.

SQ4 What publicly available datasets are used in HAR research?

Datasets are either publicly available or not. Using publicly available dataset can be efficient at minimizing the bias of the system by separating the development of the system from the creation of the dataset. Little research was found that gives an overview of the datasets that are publicly available. In fact, through this SLR only one such article was found: Micucci et al. [2017]. The article presents the publicly available datasets containing samples from smartphone sensors. The overview is displayed in figure 3.7).

Dataset	Year	ADLs	Falls	Nr. of Subjects	Gender		Age	Height	Weight
					Female	Male	(Years)	(cm)	(Kg)
DMPSBFD [24]	2015	yes	yes	5	-	-	-	-	-
Gravity [23]	2016	yes	yes	2	-	-	26–32 29 ± 4.2	170–185 178 ± 10.6	63–80 71.5 ± 12
MobiFall [11]	2014	yes	yes	24	7	17	22–47 27 ± 5	160–189 175 ± 7	50–103 76.4 ± 14.5
MobiAct [25]	2016	yes	yes	57	15	42	20–47 25 ± 4	160–193 175 ± 4	50–120 76.6 ± 14.4
RealWorld (HAR) [26]	2016	yes	no	16	7	8	16–62 32 ± 12	163–183 173 ± 7	48–95 74.1 ± 13.3
Shoaib PA [27]	2013	yes	no	4	0	4	25–30 -	-	-
Shoaib SA [28]	2014	yes	no	10	0	10	25–30 -	-	-
tFall [29]	2013	yes	yes	10	7	3	20–42 31 ± 9	161–184 173 ± 1	54–98 69.2 ± 13.1
UCI HAR [30]	2012	yes	no	30	-	-	19–48 -	-	-
UCI HAPT [31]	2015	yes	no	30	-	-	19–48 -	-	-
UCI UIWADS [32]	2013	yes	no	22	-	-	- -	-	-
UMA Fall [33]	2016	yes	yes	17	6	11	14–55 27 ± 10	155–195 172 ± 9	50–93 69.9 ± 12.3
WISDM [34]	2012	yes	no	29	-	-	-	-	-
UniMiB SHAR	2016	yes	yes	30	24	6	18–60 27 ± 11	160–190 169 ± 7	50–82 64.4 ± 9.7

Figure 3.7: Publicly available datasets containing samples from smartphone sensors. Figure taken from Micucci et al. [2017]

Sensor	Location	Applications
Microphone	EOB	Speaker recognition, localisation by ambient sounds, activity detection, object self-localisation [Amft et al. 2005; Clarkson et al. 2000; Lu et al. 2009]
Accelerometers or gyroscopes	EOB	Detection of body movement patterns, object use, ambient infrastructure [Godfrey et al. 2008; Westeyn et al. 2005; Huynh and Schiele 2005; Blanke and Schiele 2009; Bächlin et al. 2009]
Magnetometer	—B	Orientation of the body [Lee and Mase 2002] or relative position sensing of body parts [Pirkl et al. 2008]
Inertial measurement units	-OB	Absolute orientation, multiple sensors for body model reconstruction [Blanke et al. 2011; Ogris et al. 2008; Stiefmeier et al. 2007; Zinnen et al. 2009a; Blanke and Schiele 2010; Bulling et al. 2012]
Capacitative sensing	—B	Breathing, fluid intake [Cheng et al. 2010]
Pressure sensor	—B	Vertical motion, e.g. in elevator or staircase [Lester et al. 2005]
Light sensor (visible, IR, UV)	—B	Localisation of windows, lamps, light tubes [Maurer et al. 2006; van Laerhoven and Cakmakci 2000]
Skin temperature	—B	Health state (e.g. fever) [Anliker et al. 2004]
Galvanic skin response	—B	Measure of skin conductivity to infer emotional states or levels of arousal [Pentland 2004]
Environment temperature	E—	Discrimination of outdoor vs. indoor settings
Oximetry	—B	Blood oxygen: Detection of sleep apnoea [Oliver and Flores-Mangas 2007]
ECG	—B	Electrocardiography: Monitoring of physical activity and health state
EOG	—B	Electrooculography: Analysis of eye movements and recognition of cognitive processes [Bulling et al. 2011, 2012; Bulling and Roggen 2011; Bulling et al. 2009]
EMG	—B	Electromyography: Detection of muscle activation [Kang et al. 1995]
EEG, fNIR	—B	Electroencephalography and functional near-infrared spectroscopy: Measure of brain activity
Strain, stress	—B	User's breathing (respiration belt), movement (strain sensors in clothes) [Lukowicz et al. 2006; Mattmann et al. 2007; Morris and Paradiso 2002]
UWB	E—	Ultra wide band: User localisation [Ogris et al. 2008]
GPS	E-B	Global positioning system: User localisation, activities at locations, prediction of future locations [Liao et al. 2005; Krumm and Horvitz 2006]
Camera	E-B	Localisation, body model reconstruction [Clarkson et al. 2000]
Reed switches	EO-	Use of objects and ambient infrastructure [van Kasteren et al. 2008]
RFID	EO-	Radio-frequency identification: Use of objects and ambient infrastructure [Philipose et al. 2004; Stikic et al. 2008; Wang et al. 2007; Buettner et al. 2009]
Proximity	E-B	motion detection, tracking, localisation [Schindler et al. 2006]; behaviour analysis [Wren et al. 2007]; obstacle avoidance [Cassinelli et al. 2006]

Figure 3.8: Commonly used sensor for HAR and their applications Bulling et al. [2014]

The table lists 14 different datasets, but as stated in the article, the MobiAct and UCI HAPT are updated versions of MobiFall and UCI HAR. Hence, there are 12 unique publicly available datasets in the list.

Readers interested in a survey on video datasets for activity recognition should explore Chaquet et al. [2013].

SQ4 What devices or sensors are used to collect these dataset?

The device type used to collect the datasets in figure 3.7 is smartphones. The accelerometer embedded in the phones were used.

SQ6 How many subjects are typically used to collect data in these dataset?

The number of subjects used for collecting data varies greatly as seen in figure 3.7. When excluding the MobiFall and UCI HAR datasets, the average number of subjects across the 12 studies is 19 subjects. This is close to the number of subjects in the TFL dataset, in which 16 out of the 22 subjects are used in this thesis. The dataset with the fewest number of subjects is the Gravity dataset, using only two subject for data collection. MobiAct is the largest of the datasets in terms of number of subjects. It uses 57 subjects, almost twice as many as the datasets with the second largest number of subjects: the UCI HAPT and the UniMiB SHAR datasets. Both datasets uses a total of 30 subjects for data collection.

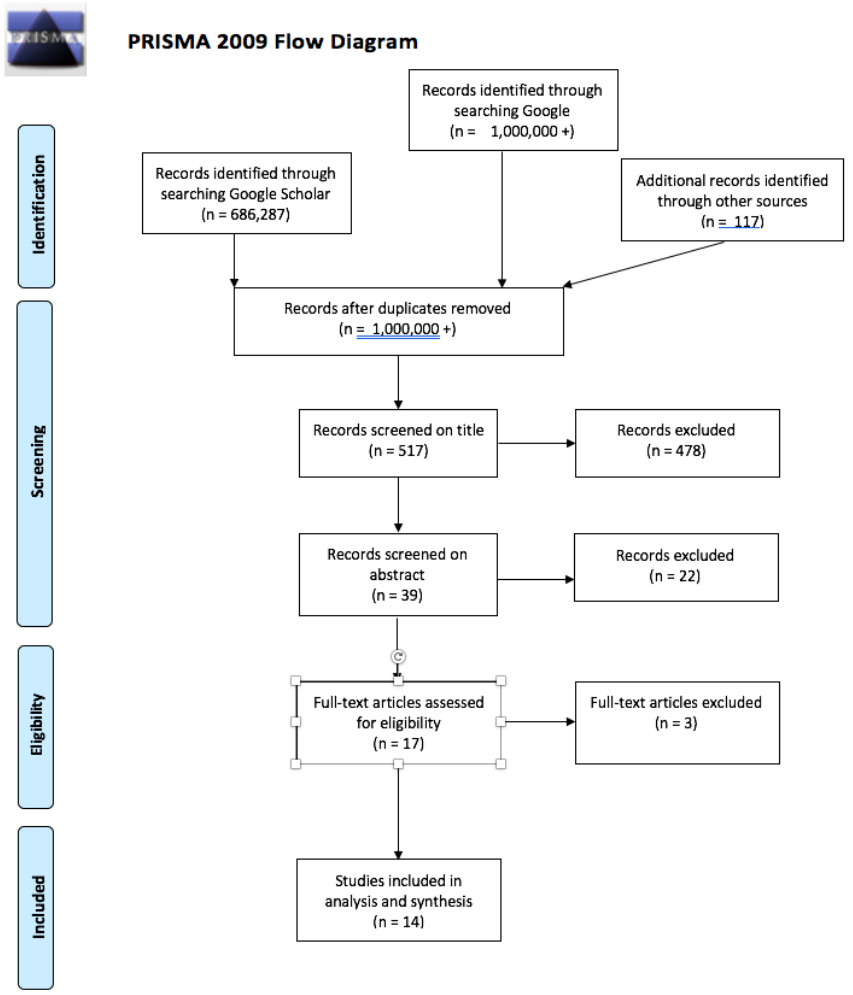


Figure 3.1: PRISMA flow diagram for the SLR, detailing the number of articles at each step of the process

Chapter 4

Datasets

This chapter presents the two datasets used in the experiments presented in chapter 6, the Trondheim free living (TFL) and the sensor no-wear time (SNT) datasets.

4.1 The Trondheim Free Living Dataset (TFL)

The Trondheim Free Living (TFL) dataset consists of measurements on 22 subjects in an out-of-lab environment. This dataset was first used in Larsen and Vågeskar [2016], then Vågeskar [2017] and later in Reinsve [2017]. This section presents the equipment used to collect data, the setup and the collection process.

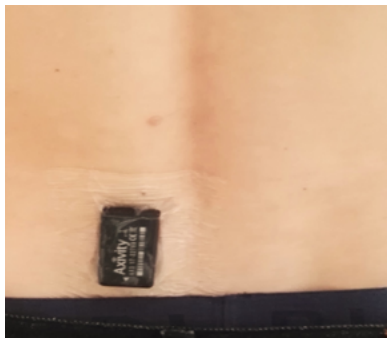
4.1.1 Equipment and Setup

The sensor used for measuring the performed activities was the Axivity AX3 accelerometer, pictured in figure 4.1. It measures proper acceleration in three axis. The sensors weighs only 11 grams and its dimensions are: 23 x 32.5 x 7.6 mm[1]. The recording frequency of the sensors were set to 100 Hz. The data was collected using two different setups: Subject 001 through subject 005 wore two sensors: one on the thigh, and and one on the upper back. Subject 006 through subject 022 wore a total of four sensors, one on the thigh, upper back, lower back, and left or right wrist. The lower back and thigh placements are illustrated in figure 4.1.1.

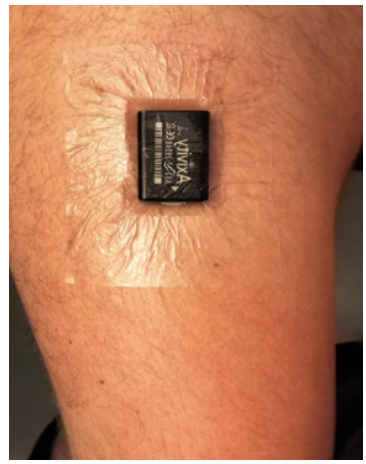
A Go Pro camera was used to capture video of the activities carried out in the out-of-lab environment. It was placed around the chest of the subjects and was pointed downwards towards the feet.

[1] *https://axivity.com/product/ax3*

Figure 4.1: The Axivity AX3 accelerometer



(a) Lower back



(b) Thigh

Figure 4.2: The sensor placements on the subjects

4.1.2 Data collection process

The recording were started at St. Olavs hospital in Trondheim, where the subjects were fitted with the two accelerometers and the chest mounted camera. The subjects wore the camera and accelerometers for approximately two hours while conducting a set of activities. The camera recordings served as the basis for labeling the data for the first two hours. The labelling of the activity data was performed manually at the hospital on a frame-by-frame basis.

Figure 4.3: The chest-mounted camera

4.1.3 Subjects

The subject group consisted of 22 subjects, all them adults. 15 of them male and 7 female. The average age of the subjects was 40 years. The subject IDs range from 001 to 022, excluding ID 007 which was removed because of a sensor malfunction. Figure 4.4 presents a bar chart with the activity distribution over all the subjects with ID in the range 006-022 (excluding subject 007). In Appendix B a bar chart of the activity distribution for each subject in the dataset is presented for reference. Only subject 006 through subject 022, excluding subject 007, was used in this thesis because of the difference in back sensor placement compared to the remaining subjects.

4.2 The Sensor No-wear Time Dataset (SNT)

The Sensor No-wear Time (SNT) dataset was created for the purpose of this thesis, to explore the possibility of using temperature readings from the AX3 to detect instances were a sensor is taken off (and possibly put back on again) during the recording of data.

4.2.1 Equipment and Setup

The data was collected using the same AX3 accelerometer as used to collect the TFL data. The AX3 includes a on-board temperature sensor, measuring the internal electronics board temperature. This sensor is the Microchip MCP9700 [2], depicted in figure 4.5. The MCP9700 is capable of accurately measuring temperatures between -40C and +150C[3]The temperature is used to calibrate the accelerometer signal of the AX3, and is not a direct measurement of skin temperature. It was however expected that the temperature

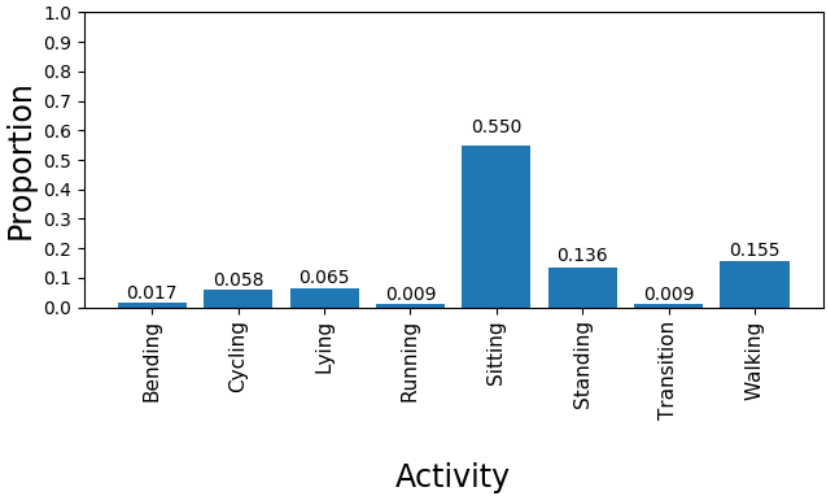


Figure 4.4: The activity distribution across all subjects

measured by the sensors would strongly correlate with the skin temperature of the subject wearing it. The sensor was placed on the lower back and on the thigh in the same positions as shown in figure 4.1.1.

4.2.2 Data collection process

The data was collected at HUNT research center in Trondheim, Norway in April of 2018. The data was collected using two different protocols: P1 and P2. Both subjects did two recordings, one for each protocol. This resulted in four recordings in total, each with a sensor recording for both the back and thigh sensor.

Protocol 1
This protocol consisted of 4 steps:

[3] *https://www.microchip.com/wwwproducts/en/en022289*

Figure 4.5: The Microchip MCP9700 temperature sensor

Step 1 Put the two sensor on, one on the thigh and one on the back, and do three claps (with the hands). Wear both sensors for about an hour while being somewhat active.

Step 2 Take off the back sensor and do three claps. Let the back sensor lie on a table for about an hour.

Step 3 Put on the back sensor, take off the thigh sensor, and do three claps. Let the thigh sensor lie on a table for about an hour.

Step 4 Take off both sensor, do three claps.

Protocol 2
This protocol consisted of 6 steps:

Step 1 Put the two sensor on, one on the thigh and one on the back, and do three claps (with the hands). Wear both sensors for about an hour while being somewhat active.

Step 2 Take off the back sensor and do three claps. Let the back sensor lie on a table for about an hour.

Step 3 Do three claps, put the back sensor back on, with the opposite device orientation. No defined duration before applying next step.

Step 4 Take off the thigh sensor and do three claps. Let the thigh sensor lie on a table for about an hour.

Step 5 Put the thigh sensor back on, with the opposite device orientation. No defined duration before applying next step.

Step 6 Take off both sensors, do three claps.

4.2.3 Subjects

The SNT dataset contains recordings from two male subjects, both adult. The data was collected by Vegar Rangul and Atle Kongsvold at the HUNT resarch center. The dataset is in its current form just a proof-of-concept. The dataset is expected to be extended in the future.

4.3 Annotation process

With two protocols and two subjects, four sets of recording were made. For each of the four recordings, a text file with the timestamps for the changes in sensor configurations were created. The temperature readings were extracted with the cwa-convert[4]program (provided by Axivity); which converts a cwa file into a csv. Based on the timestamps provided and the temperature readings (CSVs, a label file was created. Three labels were defined: one for both sensors on (A for all sensors), one for thigh sensor on (T for thigh), and one for back sensor on (B for back). The only instances where both sensors are off are at the end of the recordings where they were taken off the subject. A label for no sensor on was experimented with, but not included for the final experiment in this thesis. The reason for this is discussed in Section 7.2.3.

[4] *https://github.com/digitalinteraction/openmovement/blob/master/Software/AX3/cwa-convert/c/README.md*

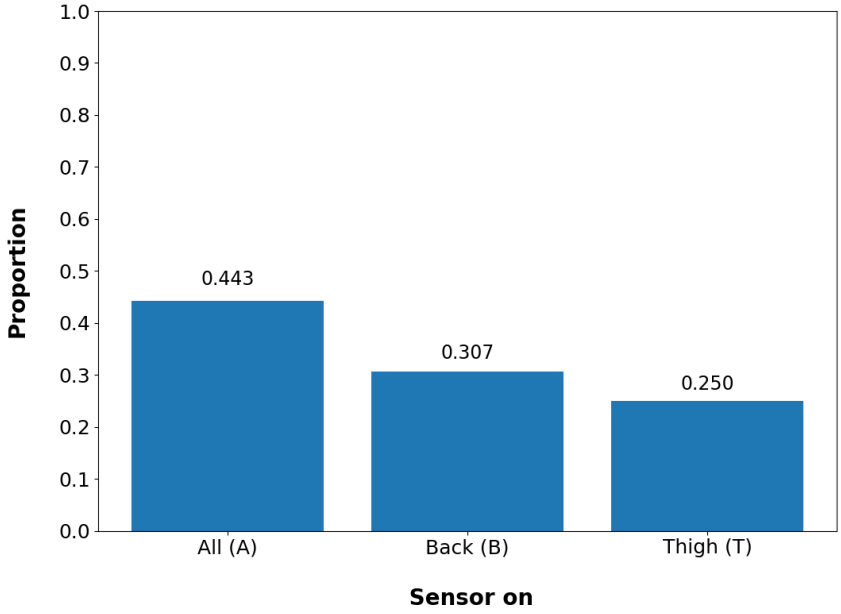


Figure 4.6: Distribution of sensor configurations for the entire SNT dataset.

Chapter 5

Methodology

This chapter presents the methodology used for the experiments found in chapter 6. Two of the three experiments performed in this thesis use the HAR system presented in Vågeskar [2017]. Hence, the explanation will be similar to the methodology presented in Vågeskar [2017] albeit less detailed. For a more detailed and comprehensive presentation of the design of the HAR system see Vågeskar [2017]. The same methods are also used for most aspects of the SNT classifier, hence both systems will be discussed in each of the steps. The explanation in this chapter will follow the steps of the the Activity Recognition Chain presented in section 2.1.2.

5.1 Data Acquisition

The TFL and SNT datasets used in this work are both presented in chapter 4. This section presents the process of how the raw data is process in order to be used bu the HAR system in this work. For a detailed description of the process see Vågeskar [2017]

5.1.1 Sensor Synchronization

The AX3 sensor captures data in the AX3 continuous wave accelerometery (CWA format). For the TFL dataset the Axivitys OMConvert [1]software was used to convert the CWA files into WAV files. Further, for the purpose of synchronizing the recording of the two sensors, the Timesync software package was used. Timesync takes two WAV files as input and synchronizes them based on the magnitudes of the signals. A master and slave file can be specified by the script's user. The output of the Timesync software is a seven column CSV with three columns for each of the synchronized sensors signals. The columns contain the acceleration along the X, Y, and Z axes in g_0s in floating point numbers Vågeskar [2017]. In addition to the synchronizing of the senors, a synchronizing between the video annotations and the sensors have to be performed. This process is described in its entirety in Vågeskar [2017]. This process was already completed for the TFL dataset before being applied in this thesis. This was not the case for the SNT dataset,

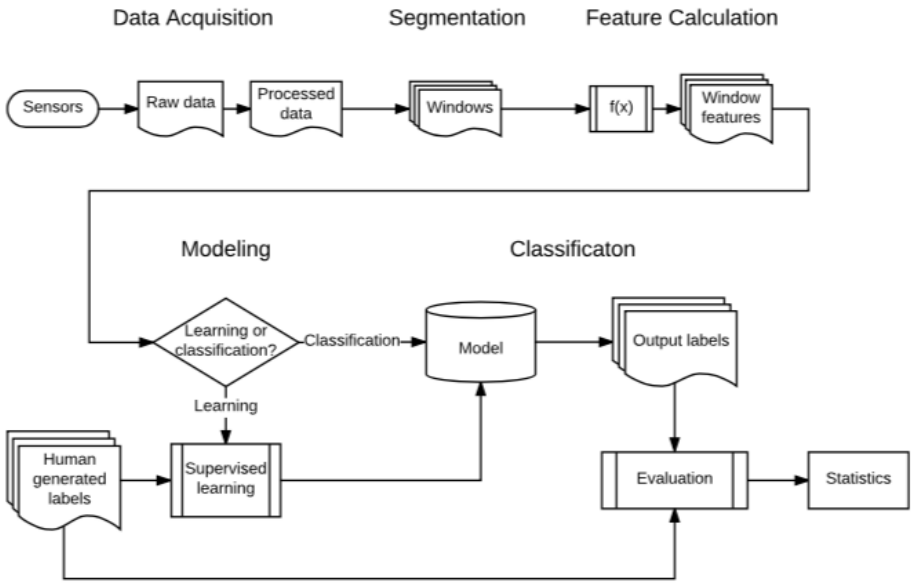


Figure 5.1: The Activity Recognition Chain. Figure from Vågeskar [2017]

where the files were provided in the CWA format. To extract the temperature readings of the sensor Axivity's cwa-convert program was used. This extracted the temperature readings of the two sensors and converted the two CWA files into two comma-separated values (CSV) files with two columns: the time stamp of each temperature sample, and the corresponding temperature in ADC units. The two temperature recordings were not synchronized because of difficulties with getting temperature readings of the sensor when using the OMConvert and Timesync software packages. While labeling the SNT dataset, the effect of not having synchronized the sensors was evident. It was discovered that the time stamps between the two sensors and their recordings could be off by several minutes for samples later in the recording. To deal with this the following labeling approach was used:

5.2 Segmentation of the data

The segmentation of the data in the CSVs into windows of a given length, was done for both the TFL and the SNT dataset. For the TFL dataset a window length of 3.0 seconds was used. This length was chosen based on the the evaluation on window lengths in Vågeskar [2017], and for the results of this thesis' experiments related to increasing the effectiveness of the HAR system in Vågeskar [2017], to be comparable with each other. For the SNT dataset multiple window lengths were experimented with. The SNT data was

[1] *https://github.com/digitalinteraction/openmovement/tree/master/Software/AX3/omconvert*

collected with a sampling frequency of 50 Hz for the acceleration sensor. A temperature reading was extracted for every 120 acceleration sample. This meant that, on average, one temperature reading was extracted for every 2.4 second. The rate of change for the the temperature measurements is obviously much slower than that of the acceleration measurements. Therefore the window length had to be substantially longer than that used for the HAR system with the TFL data. Window lengths of 1, 2, 3, and 4 minutes were experimented with. A window length of 30 second was also experimented with- A window length of 2 minutes consistently performed better than all the others. With 2 minute windows, each window contains 50 temperature readings.

5.3 Relabeling and removal of activities

Prior to modelling and classification some windows with certain activities were either relabeled or removed. The removed activities are the same ones that were removed in both Hessen and Tessem [2016] and Vågeskar [2017]. These were removed for the same reasons as in them. The 4 removed activities are: shuffling, transitions, undefined activity, and non-vigorous activity.

Activity	Relabeled as	Justification
Ascending stairs	Walking	Ascending and descending stairs was often misclassified as walking. After bending, stairwalking was found to be the RF classifier's least accurate activity when compared to the CNN classifier in Hessen and Tessem [2016, table 6.5]. Recognizing this activity for stroke patients and other patients with gait impairment has previously been shown to be problematic by Lonini et al. [2016, figure 3] and Capela et al. [2016, table 5].
Descending stairs	Walking	Same as ascending stairs.
Picking	Bending	Hessen and Tessem found this to often be confused with bending because of high interclass similarity: Picking occurs when the subject places, touches, or picks up an object below knee height, which must occur in the middle of a bending activity. The activity is also the only manipulative gesture in the activity set, while all other activities are related to either locomotion or posture.

Figure 5.2: Activities that were relabeled. Figure from Vågeskar [2017]

5.4 Feature calculation

The features used by the HAR classifier were listed in figure 2.10 and 2.11. The features used in the SNT classifier are listed in table5.1. For both classifiers, features were

Activity	Justification
Shuffling	Shuffling overlaps with two other activities: It is either a short walking bout or standing with some leg movements. This makes it difficult to recognize, and as it overlaps with two other activities, it is not a candidate for relabeling.
Transition	Transition is a movement between activities. Windows with this label show little similarity to each other.
Undefined activity	This activity contains activities which cannot be identified from the video or which occur before sensors and camera has been attached. Can contain a multitude of different activities, and recognizing it is therefore hard.
Non-vigorous activity	Activities which are recognizable, but do not classify according to the definitions are labeled as non-vigorous activity.

Figure 5.3: Activities that were relabeled. Figure from Vågeskar [2017]

extracted separately from each sensor. For the HAR classifier all of the 20 time and frequency domain features were calculated. These unique features ”produced” at total of 138 distinct features inputted to be used in the modelling and classification. The 4 features for each of the sensors used in the SNT modelling and classification were also extracted separately for each of the two sensors.

In some instance the feature calculations of the HAR classifier would result in undefined or infinite values. The affected features were the spectral entropy, spectral centroid and, correlation features. The system developed in **?** dealt with these instances by replacing the values with 0. This was never the case with the feature calculations of the SNT classifier. None of the features in 5.1 produced such values.

The feature calculation in the HAR systems also deals with sensors that are wrongly attached on subjects (Vågeskar [2017]). This is the case when a sensor is attached upside down, making the values along one or more of the senor’s axes be opposite of the expected. The systems deals with this by removing the sign off all sensor values. This is applied only to the final output of the feature calculations for both training and testing. This system significantly increases the accuracy on subjects wearing sensors that are oriented wrongly (Vågeskar [2017])

Name	Definition	Description
Max	max(x)	The highest value in a sequence.
Min	min(x)	The lowest value in a sequence.
Max-min delta	max(x) - min(x)	The difference between the maximum and minimum value in a sequence.
First-last delta	last(x) - first(x)	The difference between the last and first value in a sequence

Table 5.1: The features used in the SNT system. The features are time domain features calculated from the temperature readings of the sensors.

5.5 Random Forest Classifier as the choice of classifier

The RFC are used for all three experiment in chapter 6. The random forest classifier was the obvious choice for the two experiment related to increasing the effectiveness of the HAR system in Vågeskar [2017], which used this classifier. For the SNT experiment it was decided that a RFC was to be used for simplicity; using similar classifiers made it easier and quicker to implement and test the SNT classifier.

Scikit-learn's random forest classifier implementation was used for the implementation of both the HAR and SNT classifiers. [2]

5.5.1 Parameters

A few parameters of the RFC were set differently than their default values. For the HAR classifier the following parameters were set:

- **Number of trees (n_estimators): 50.** This number of trees was set to match the settings of the RFC in Vågeskar [2017]. This number of trees was selected in Vågeskar [2017] because they provided a good balance between accuracy and training time. Using significantly more threes reportedly resulted in very slight changes in accuracy while taking significantly longer to train.
- **Class weight (class_weight): "balanced".** This parameter determines the class weights of the classes. The "balanced" option gives the classes equal weight. This is the results in the same effect as oversampling the dataset until an equal class distribution is achieved Vågeskar [2017].

For the RFC classifier used for the purpose of SNT detection and sensor configuration classification, the following parameters were set:

- **Number of trees(n_estimators): 100.** Training time for the SNT system was significantly shorter than for the HAR system. Therefore, the number of trees making up the model were increased. This resulted in a slight increase in accuracy over using 50 tees.
- **Class weight (class_weight): "balanced".** This was used on the basis of that of the HAR classifier. The sensor configurations in the SNT dataset are not evenly distributed as show in in figure 4.6.

[2] *http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html*

5.6 Feature importances

To evaluate the the importance of the features used in the random forest model, scikit-learns *"feature_importances_"* attribute of the RFC was used[3]. The feature importance measures the relative importance of each feature on the prediction. Information on the implementation of the feature importance calculation in scikit-learn was hard to find. It is indicated that the framework use the gini importance, which is defined as total decrease in node impurity weighted by the probability of reaching that node, averaged over all the trees of the ensemble[4]. The higher the feature importance of a feature, the more important the feature.

5.7 Training the models

A total of 138 different RFC was built using different number of features. The number of features included in the model is called the feature count of the model. The model with feature count 1 was trained on data containing exclusively the feature with the highest feature importance. The model with feature count 2 was trained with the two most important features, etc. This process was repeated for all 138 different feature counts. Segmentation and feature calculation of the data was done once with all the 138 features included, but before training each of the models, the corresponding n features with the highest feature importance of each feature count was extracted from this set. This was to reduce the testing time of the system.

5.8 Testing with subject-wise and record-wise cross-validation

To assess the models ability to generalize beyond the observed training data, the testing of the models were done with subject- and record-wise cross-validation. This means that for each record or subject in the dataset the following was performed: exclude the recording or subject from being included in the training data, use this to test the model. Use all the remanding records or subjects to train the model. When finished training the model, run the predictions on the excluded data and measure the performance of the classifier. Then, after completing this process, average the performance values over all the tests based on the number of predictions in each of the tests.

[3] *http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html*
[4] *https://medium.com/theartificialimpostor/featureimportancemeasuresfortreemodelsparti47f187c1a2c3*

Chapter 6

Experiments

This chapter presents the experiments that were performed. The goal of the experiments was to increase the effectiveness of the HAR system presented in Vågeskar [2017] and to develop a system capable of detecting sensor no-wear time by classifying the sensor configuration of the sensors attached to a subject.

6.1 Selecting features based on feature importances and model accuracy

The RFC model presented in Vågeskar [2017] used 13 time domain features and 7 frequency domain features. The 20 unique features were represented as 69 input features to the RFC, for each of the two sensors. In total, 138 features were used to both train the model and predict. The reason for there being more input features to the classifier than unique features is that some features are calculated for all three axis of the acceleration signal, for both sensors. This can result in 6 features being created from one unique feature. From this point onward ”features” refers to the 138 features inputted to the model and ”unique features” to the 20 time and frequency domain features, unless otherwise stated. In Reinsve [2017] it was shown that many of 138 features were not necessary to maintain the accuracy presented in Vågeskar [2017]. It also showed that the feature calculation step in the activity recognition chain was the most time consuming of the steps. Therefore, reducing the number of unique features to calculate would result in significant time saving for the HAR system when viewed as a whole. One key factor had to be considered while doing so: maintain a similar degree of accuracy as that of 94.2 percent Vågeskar [2017]. This experiment consists of three steps: extracting the feature importances of all the 138 features, finding which of the unique features the 138 features and their indexes map to, and lastly, evaluating the accuracy of 138 models trained on 1-138 of the most important features. The goal of the experiment was to find out how many, and which features were sufficient to achieve an accuracy of 90 and 94 percent.

6.1.1 Setup

This experiment uses the HAR system presented in Vågeskar [2017] with the TFL dataset. Scikit-learn's feature importances[1]was used to extract the feature importances for the 138 features. 138 ordinary RFC classifiers are trained on the TFL dataset using different number of features. The first model is trained on the single most important feature only, as determined by the feature importances score. The second model is trained exclusively on the two most important features. This process continued up until a model trained on all the 138 features was created. All of the classifiers were trained on all the subjects in the TFL dataset using subject-wise cross validation. The testing was repeated 3 times and the scores were averaged.

6.1.2 Results

Figure 6.1 shows the feature importances for the 69 features of the lower back sensor sorted by their importance. The x-axis is the ranking of the feature in terms of feature importance. The feature with the highest feature importance is at index 1. The x-axis value does not reflect the index of the feature when passed to the RFC, nor which of the unique features it corresponds to. Figure 6.2 present the feature importances for the thigh sensor features.

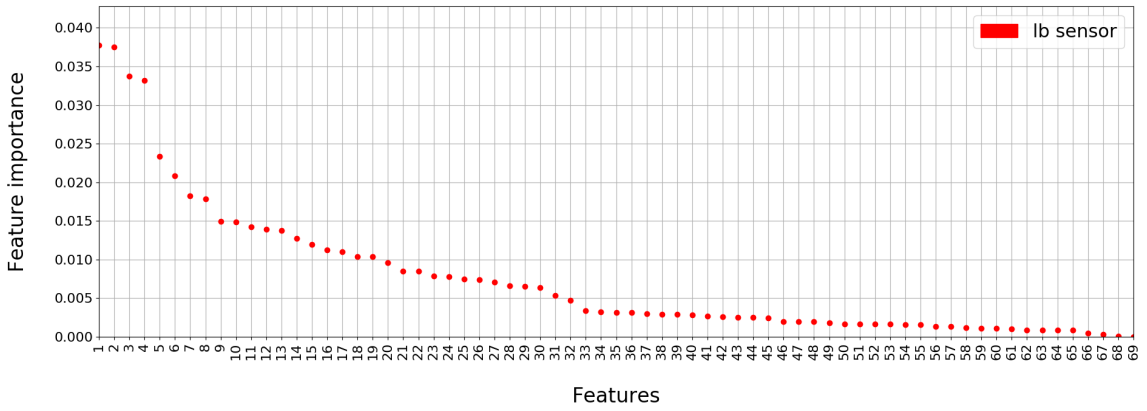


Figure 6.1: Feature importances of the features derived from the lower back accelerometer signal.

[1] *http://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html*

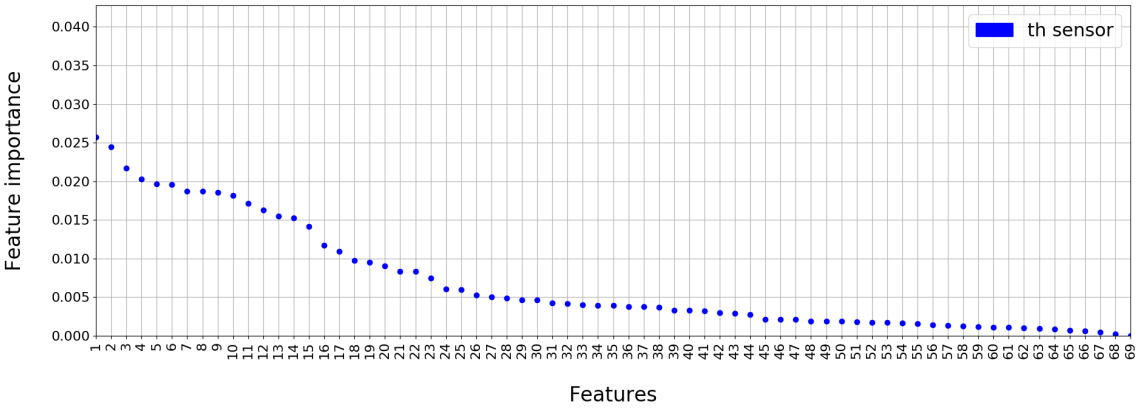


Figure 6.2: Feature importances of the features derived from the thigh accelerometer signal.

In figure 6.3 the feature importances of both sensors are show on top of each other. The four most important features calculated from the lower back sensor are all more important than any of the features derived from the thigh sensor.

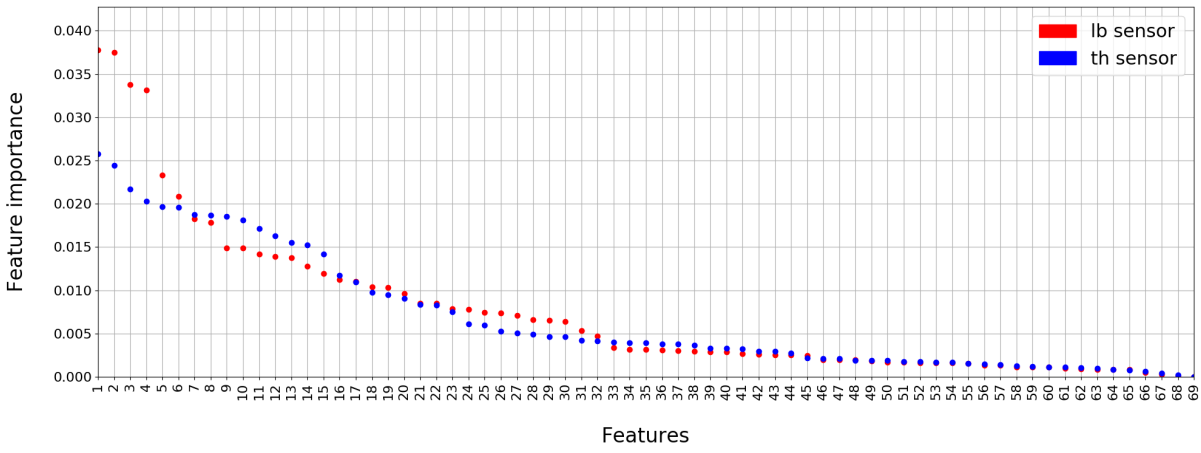


Figure 6.3: Feature importances of features for both sensors, presented on top of each other for better comparison. The x-axis values indicate the index of the feature when sorted based on feature importance. This is not the index of the feature when passed to the RFC.

The accuracies of the 138 different models trained with different number of the most

important features are shown in figure 6.4.

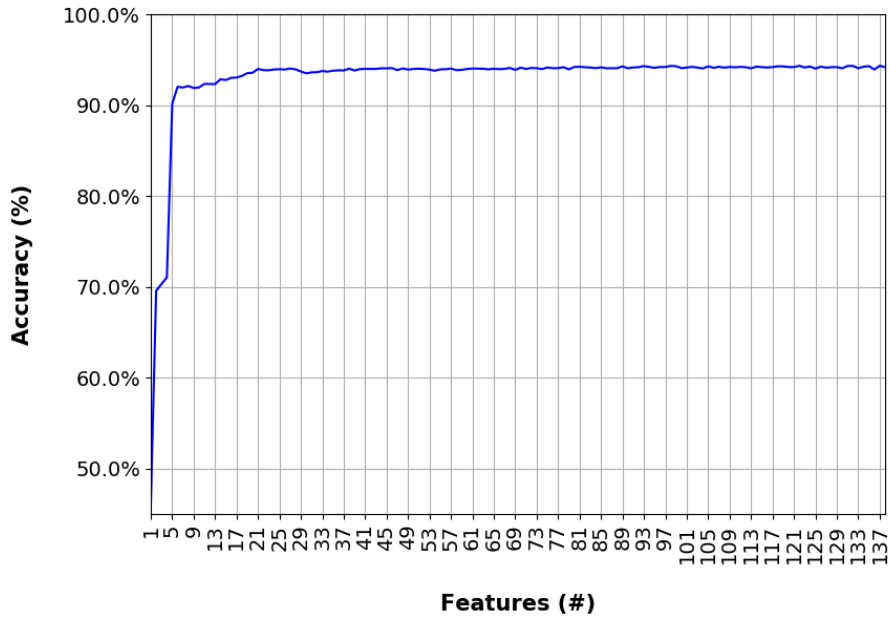


Figure 6.4: The overall accuracy for subject-wise cross validation for all the feature count combinations from 1 to 138 features. The accuracy increases rapidly when adding the 5 most important features. The model with the 5 most important features is is the first model to reach an accuracy of 90 percent.

Figure 6.5 shows the confusion matrix for the default model using all 138 features. This model has an accuracy of 94.1 percent.

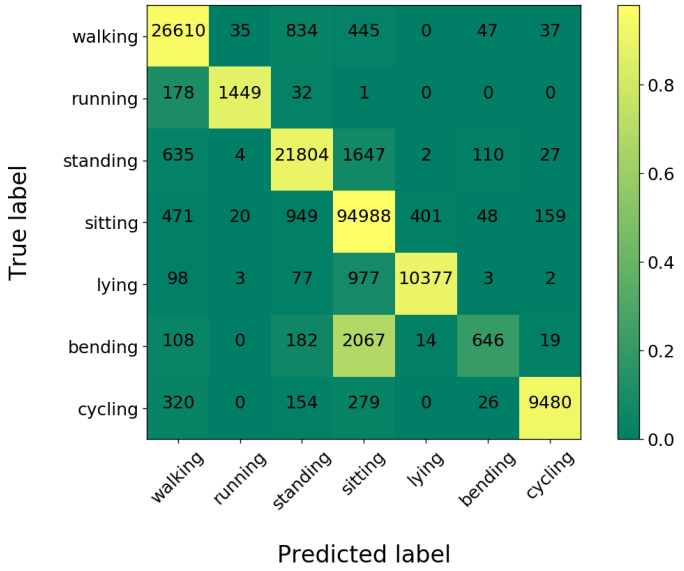


Figure 6.5: Confusion matrix for the RFC trained with all the 138 features.

The confusion matrices for the RFCs with 5 and 27 of the most important features are shown in figure 6.6 and 6.7 respectively.

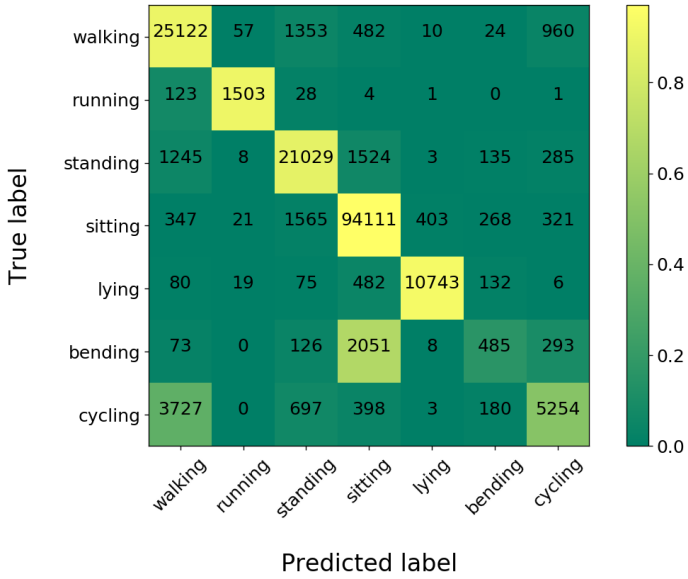


Figure 6.6: Confusion matrix for the RFC trained with the 5 most important features.

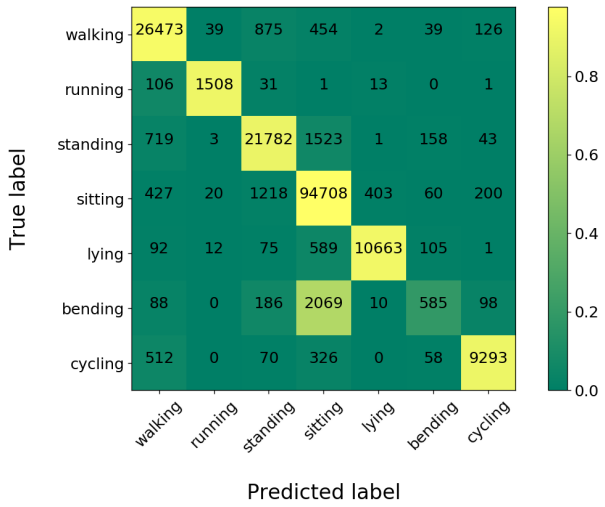


Figure 6.7: Confusion matrix for the RFC trained with the 27 most important features

Figure 6.8 present the accuracies of the models with 1 to 10 most important features. The accuracy rises sharply until adding the 6th most important feature. The classifier with 5 features obtains 90.0 percent accuracy. The classifier using the 6 most important features achieve 92.0 percent accuracy.

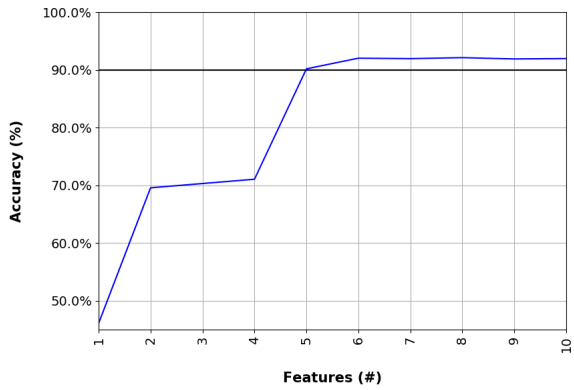


Figure 6.8: The overall accuracy for subject-wise cross validation for all the feature count combinations from 1 to 10 features. Adding additional features after the 5 most important features produce incremental changes in the accuracy. The model trained on 5 features is the first model to reach an accuracy 90 percent. It is worth noting that when adding the 6th most important feature to the model, an accuracy of 92 percent was achieved

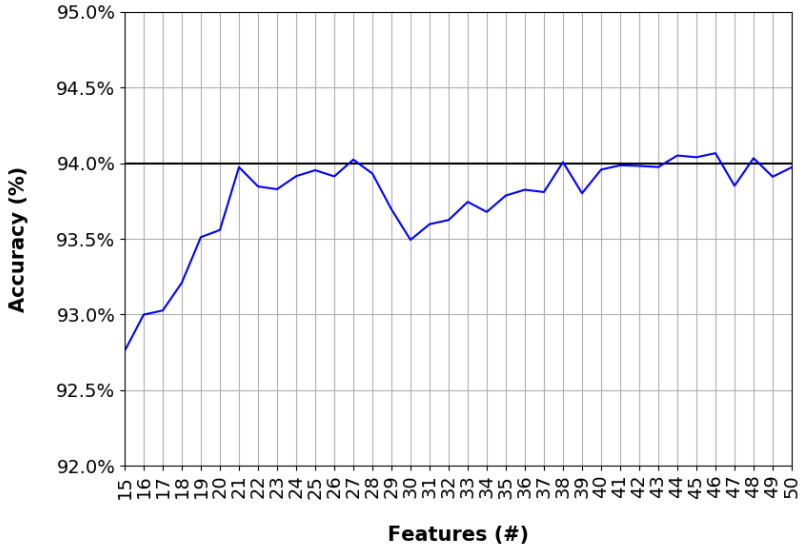


Figure 6.9: The overall accuracy for subject-wise cross validation for all the feature count combinations from 15 to 50 features. The model trained on the 27 most important features is the first to achieve 94 percent accuracy. However, this accuracy is not achieved consistently until the 40 to 50 most important features are used

In figure 6.9 the accuracies of the models with the 15 to 50 most important features are illustrated. The model using 27 features delivers an accuracy of 94.0 percent. The accuracy stabilizes at this level of accuracy around the 40-50 features range.

	5 features				**27 features**			
	Accuracy				**Accuracy**			
	0.902				0.940			
Activity	**Precision**	**Specificity**	**Recall**	F_1	**Precision**	**Specificity**	**Recall**	F_1
Bending	0.407	0.996	0.157	0.227	0.659	0.998	0.195	0.301
Cycling	0.741	0.989	0.521	0.612	0.952	0.997	0.904	0.927
Lying	0.962	0.997	0.930	0.945	0.962	0.997	0.933	0.947
Running	0.931	0.999	0.901	0.915	0.949	1.000	0.908	0.928
Sitting	0.950	0.937	0.972	0.961	0.950	0.937	0.978	0.964
Standing	0.848	0.975	0.869	0.856	0.904	0.985	0.900	0.902
Walking	0.821	0.963	0.896	0.857	0.933	0.987	0.945	0.939

Table 6.1: Quality metrics table for the performance of the activity recognition classifiers using 5 and 27 features, tested with subject-wise cross validation on the TFL dataset

	138 features			
	Accuracy			
	0.942			
Activity	**Precision**	**Specificity**	**Recall**	F_1
Bending	0.741	0.999	0.209	0.326
Cycling	0.974	0.998	0.925	0.949
Lying	0.962	0.997	0.907	0.934
Running	0.958	1.000	0.870	0.912
Sitting	0.947	0.933	0.979	0.963
Standing	0.908	0.985	0.902	0.905
Walking	0.936	0.988	0.951	0.944

Table 6.2: Quality metrics table for performance of the activity recognition classifier with 138 features, tested with subject-wise cross-validation on the TFL dataset.

6.2 Increasing efficiency of the HAR system by calculating fewer features

This experiment builds on the findings of the experiment in section 6.1. Having shown that a significant number of features can be excluded in the feature calculation step while maintaining the accuracy, the next step was to assess the potential increase in effectiveness of the HAR system. The motivation behind increasing the efficiency is obvious; the HUNT 4 dataset is expected to consist of about 50 000 subjects with PA measurements. Predicting the activities of all the subjects would take considerable time with the current number of features. Table 6.3 shows the time statistics of the different steps of the HAR system for three different data lengths, using all of the 138 features.

Number of subjects	**Dataset**	**Time (m)**	**Model training (s)**	**Windows extraction**	**Feature calculation**	**Prediction**
1	TFL	139	13.32	0.94s	21.97s	0.31s
1	HUNT4	8,640	13.32	58s	23m	19s
50,000	HUNT4	432,000,000	13.32	34d	790d	11d

Table 6.3: Time statistics for the HAR system using all 138 features with a sampling frequency of 100 Hz and a window length of 3 second. Times are stated in s for seconds, m for minutes, and d for days. Training the model is done once, and remains the same regardless of the number of subjects to be analyzed. Window extraction, feature calculation, and prediction scale linearly with the data length. Feature calculation take the most time of all the steps. Calculating the features for the 50,000 subjects that are expected to participate in the HUNT4 study wearing the accelerometers would take 790 days running on a single thread on a reasonably fast computer. Subject 006 was used as the testing subject. All the other times are extrapolated from the the time it took for the system to classify the activities of this subject.

6.2.1 Setup

This experiment used the same dataset as in the experiment in 6.1: the TFL dataset. 138 models were again created, using 1 to 138 features, starting with the most important feature. The time it took to complete the model training, window extraction, feature calculation, and prediction steps in the HAR system were measured. Subject 006 in the TFL dataset was used as test subject for measuring the time. The testing was repeated three times and the results were averaged over all three runs.

6.2.2 Results

Figure 6.10 show the time spent executing the four step of the HAR system for subject 006.

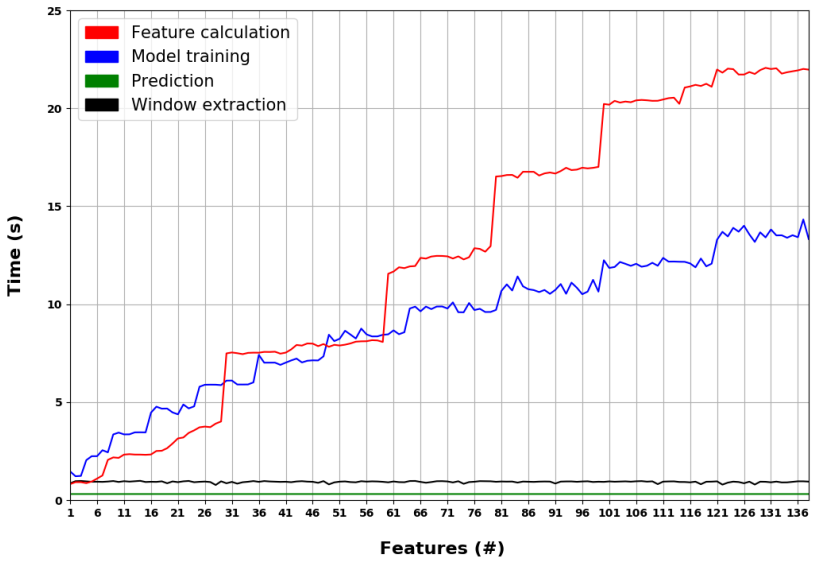


Figure 6.10: The time usage for the 4 steps in the HAR system for predicting the activities of subject 006 in the TFL dataset. The time for each step and feature count is averaged over three iterations.

Figure 6.11 combine the accuracy plot in figure 6.4 with the time statistics for the feature calculation step in figure 6.10.

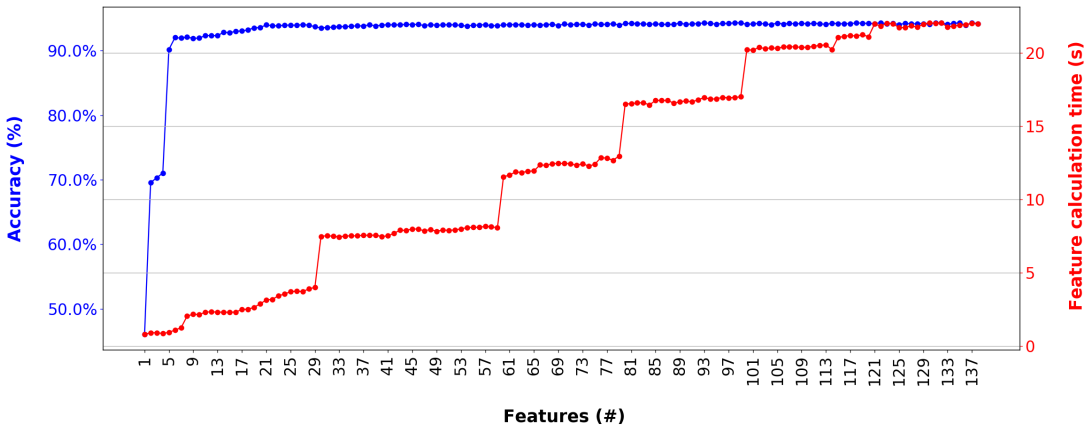


Figure 6.11: Combined plot with both accuracy and feature calculation time for models with feature count 1 to 138. The accuracy is a result of a subject-wise cross validation on the 16 subjects used in the TFL dataset (subject 006 and subjects 008-022). The feature calculation time is extracted from the predicting the activities of subject 006 in the TFL dataset.

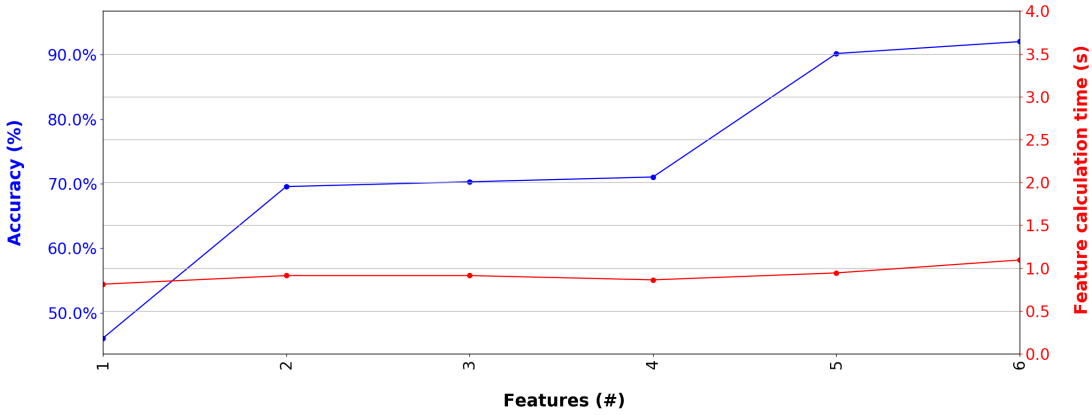


Figure 6.12: Combined plot with both accuracy and feature calculation time for models with feature count 1 to 138. The accuracy is a result of a subject-wise cross validation on the 16 subjects used in the TFL dataset (subject 006 and subjects 008-022). The feature calculation time is extracted from the predicting the activities of subject 006 in the TFL dataset.

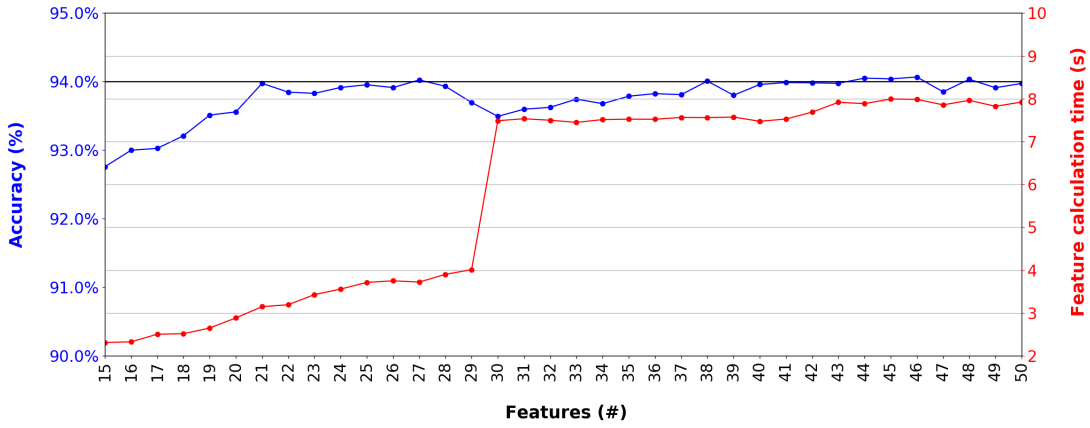


Figure 6.13: Combined plot with both accuracy and feature calculation time for models with feature count 1 to 138. The accuracy is a result of a subject-wise cross validation on the 16 subjects used in the TFL dataset (subject 006 and subjects 008-022). The feature calculation time is extracted from the predicting the activities of subject 006 in the TFL dataset.

The results of this experiment and the experiment in 6.1 was put together to get a better understanding of the optimal accuracy and.

6.3 No-wear time detection and sensor configuration classification by the use of sensor temperature readings

There are generally three instances where data is not recorded or obtained as intended: missing recordings, records with errors, and records containing sensor no-wear time samples where either one or both of the sensors have been detached during recording. Missing records are the easiest to deal with. If both records are missing, the classifier is unable to make any predictions. If either the back or thigh recording is missing, the HAR system can use sensor-specific models for classification. To produce an estimate for the number of subject with missing records, 4304 unique subject records from the HUNT4 data were looked at. Figure 6.14 presents the results. There are multiple reasons for why sensor recordings are missing: some sensor stop working during recordings, for some unknown reason, making it impossible to extract the collected data. Some sensors have even started burning when they were connected to a computer, after having completed a number of subject recording. This has happened because of the glue around the micro-USB has caught fire when plugged in to charge.

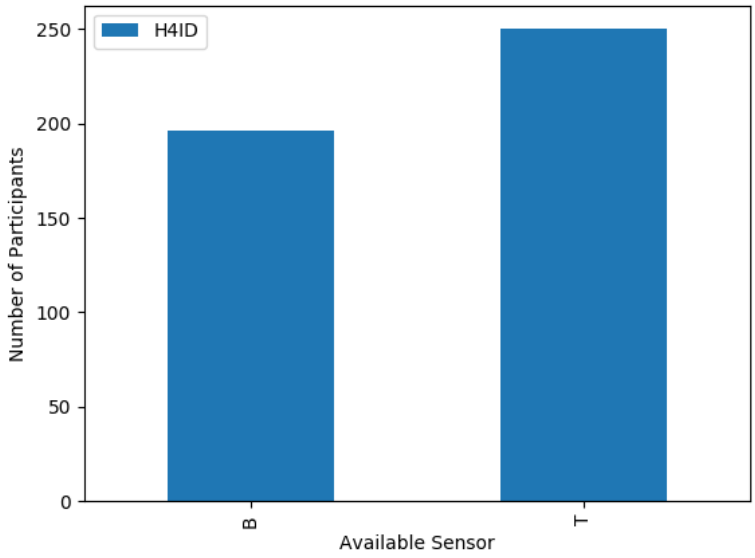


Figure 6.14: The frequency of records missing for the lower back sensor (B) and thigh sensor (T) in a sample size of 4304 subjects in the HUNT4 dataset. Out of the 4304 subjects, 446 subjects have one missing sensor recording. 196 are missing the lower back sensor, and 250 are missing the thigh sensor.

In addition to missing sensor records, there are instances where the recording contain obvious error. For example, a sensor might have a malfunction during the recording, but it was possible to extract the data from it. These instances are the hardest to deal with and often have to be manually examined before being either discarded or modified. Lastly, there's the instances with sensor no-wear time. Sensor have been reported to have fallen off during recordings of data for the HUNT4 study. Some of the participants in the study might also decide to take one or both of the sensors off for any particular reason. Some subjects might try to reattach the sensor(s), others might not. To get an idea of the extent of sensor no-wear time in the HUNT4 data collected thus far, a subset of the data was looked at. For each of the subjects wearing accelerometers in the HUNT4 study, an activity report is generated by the HAR system. Based on these reports we might be able to understand the frequency of no-wear time in the HUNT4 dataset. Figure 6.15 and 6.16 show the activity report of two subjects that have participated i the HUNT4 study. The subject in figure 6.15 looks to have been very active, with a lot of short walks during the daytime (green), while the subject in figure 6.16 seem to have been sitting more still (light blue), with some longer walks in the middle of the day. Both figure 6.15 and 6.16 looks like the way they are expected to look with no, or very little, sensor no-wear time. This means that both sensors have been attached for most or all of the duration of the recording. Figure 6.17 and 6.18 however, illustrate how recordings containing sensor no-wear time are expected to look. The sensors attached to a subject during a recording

is from now one referred to as the sensor configuration of the subject. The goal of this experiment is to develop a sensor configuration classifier, which is tasked with assigning one of the three labels (A, B, T) to windows of data from the recordings. ”A” indicates that both sensors are attached to the subject, ”B” indicated that the back sensor (only) is attached to the subject, and ”T” which indicates that the thigh sensor (only) is attached to the subject. No-wear time is predicted by the classifier for any window not assigned label A.

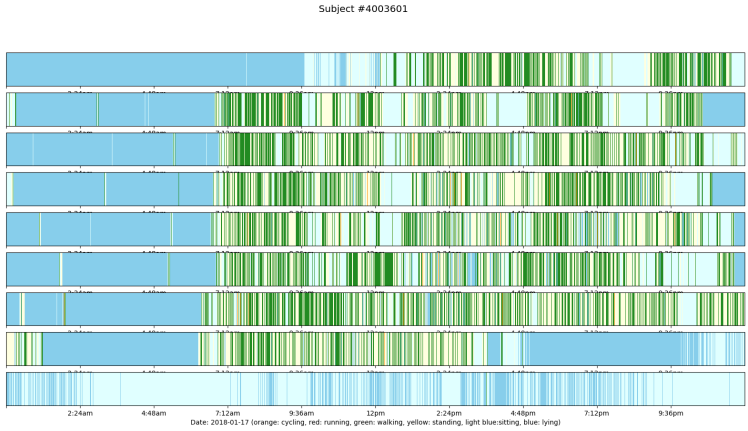


Figure 6.15: An example activity report for a subject in the HUNT4 dataset (probably) not containing no-wear time, except for day 9 and the evening of day 8.

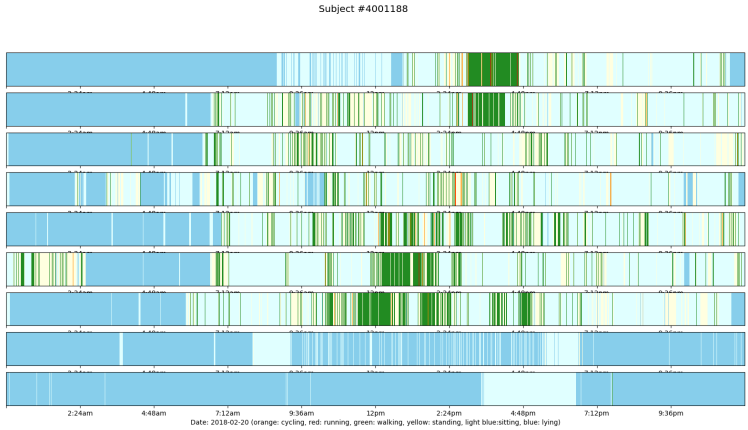


Figure 6.16: An example activity report for a subject in the HUNT4 dataset (not likely) containing no-wear time for the first 7 days. The last two days of the recording the sensors are clearly not attached, hence there is sensor no-wear time.

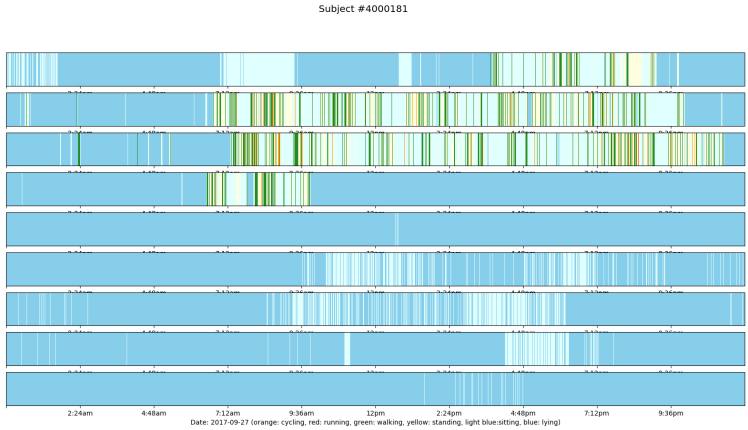


Figure 6.17: An example activity report for a subject in the HUNT4 dataset (very likely) containing no-wear time. The no-wear time looks to start on the morning of the 4th day of recording

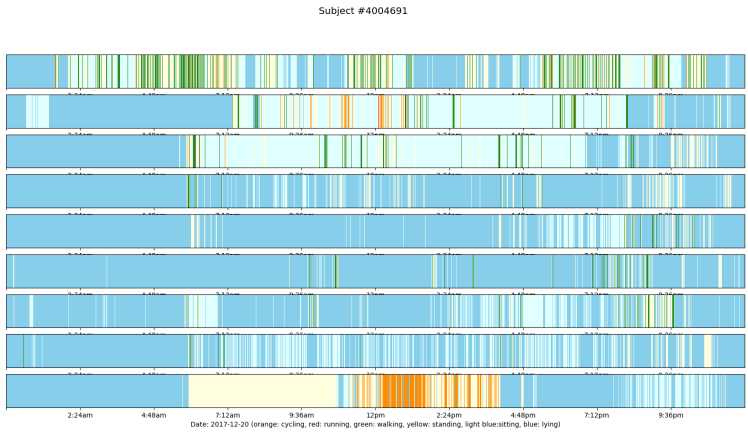


Figure 6.18: An example activity report from a subject in the HUNT4 dataset (very likely) containing no-wear time for day 4 to 9.

In looking for a reasonable way to detect no-wear time of sensors, the idea of utilizing the temperature sensor in the AX3 accelerometer came up. The idea behind this is that the skin temperature on both the back and thigh should be fairly similar. So if both sensors are attached, they temperature difference between the two will be insignificant. And if one of the senors are detached, the detached sensor will experience a significantly larger change in temperature than the sensor that is attached.

6.3.1 Setup

The dataset used for this experiment was the SNT dataset presented in section 4.2. The 5 features listed in 5.1 were extracted for both the sensors separately before being combined. An ordinary RFC was used with the parameters defined as in section 5.5.1. The testing was performed with a subject-wise cross-validation of the four recordings in the SNT dataset, holding out one of the recordings while training on the other three recordings, and repeating this process for each recording in the dataset.

6.3.2 Results

The SNT classifier achieved an accuracy of 95.6 percent accuracy on average when classifying the four recordings in the SNT dataset. A confusion matrix of the results is presented in figure 6.19. Table 6.4 shows the performance of the classifier in terms of the quality metrics introduced in 2.7.

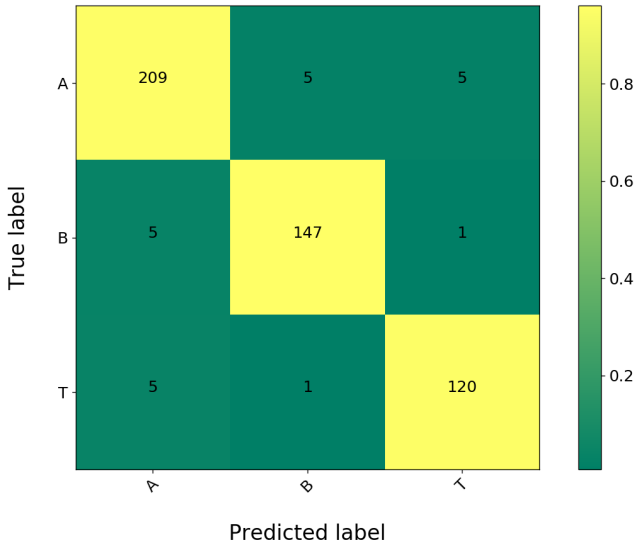


Figure 6.19: Confusion matrix for the record-wise cross-validation test on the SNT dataset

Sensor on the subject	Precision	Specificity	Recall	F_1
All (A)	0.954	0.964	0.954	0.954
Back (B)	0.960	0.983	0.961	0.961
Thigh (T)	0.952	0.984	0.952	0.952

Table 6.4: Quality metrics table for the results of subject-wise cross-validation testing on the SNT dataset with 2 minute windows.

Chapter 7

Discussion

7.1 Systematic Literature Review

7.1.1 Strength and limitations of the literature

The literature used for the literature review were all of acceptable quality. Most of the papers were fairly recent, and presented goals, results, and discussions clearly. A significant number of articles were found covering different machine learning methods in HAR. This made it possible to extract information about classifiers across papers, and that have been applied to different datasets, as well as comparisons using the same training and testing data.

A limitation of the literature was the lack of articles found on the topic of publicly available datasets for accelerometer based HAR. There was however substantial amounts of literature and surveys on the topic of video-based datasets for HAR purposes. These were not included in because of the difference in domains. There should however be created similar survey for accelerometer based HAR systems. This would benefit researchers by making it easier to select datasets and implement new models.

7.1.2 Limitations of the review

There are multiple limitations in how this SLR was carried out. Using Google Scholar as the only academic search engine in the process, might have limited the range of documents that were found. Furthermore, in the process of creating search questions and terms, there is no telling what the optimal combination of search terms are. Therefore, it is impossible to guarantee that all relevant studies in the respective databases were included in the search results. Furthermore,the filtering of the articles, through multiple steps, was performed by a single person; the author of this thesis. Therefore, the authors bias will be reflected in the works that were selected as the basis for the research synthesis. This creates room for error and misjudgments, which would likely be less prevalent

with more people involved in the process.

Because of time constraints, only the top 20 articles in the search results were examined in any form. Clearly there will always be time constraints, but there might have been articles of interest outside of the top 20 relevant articles returned by the search engine.

7.2 Experiments discussion

The three experiments performed in chapter 6 were aiming to reach goal 2 and 3 of this thesis: increasing the effectiveness of the HAR system presented in Vågeskar [2017], and developing a SNT classifier able to detect instances of sensor no-wear time by classifying sensor-configurations for recordings in the SNT dataset. The two experiments in section 6.1 and 6.2 both aimed to achieve goal 2. The experiment in section 6.2 builds on the results of the experiment in section 6.1, therefore they must bee seen in conjunction.

7.2.1 Selecting features based on feature importances and model accuracy

The first step towards achieving goal 2 of the thesis was performed in section 6.1. The feature importances for the lower back and thigh sensors were presented in figure 6.1 and 6.2, respectively. Looking at the feature importances it is clear that many of features are less important than others. The 4 features with the highest feature importance in figure 6.1 have significantly higher importances than the rest of the features. This is less pronounced in figure 6.1, which depicts the feature importances for the thigh sensor. There, a more linear decrease in feature importance towards the right of the figure is evident. Figure 6.3 show the feature importance fort both sensor on top of each other. It is interesting to see that the four features with the highest feature importances are all from the lower back sensor. This indicates that the this sensor is more important when classifying the 7 activities with the TFL dataset. In fact, as seen in the accuracy plot in figure 6.8, the model using the 4 most important features (meaning that only features from the back sensor was used) achieved an accuracy slightly above 70 percent. Moving beyond the four most important features, the difference in feature importances between the two sensors decrease. In figure 6.4 the accuracy for all the 138 models that were trained is show. It is clear, and maybe surprisingly so, that very few of the 138 features are necessary to achieve an accuracy in the 90 to 94 percent range. The accuracy seen to stabilize after the 37 or so most important features are included in the model. The classifier with 5 features is the ”smallest” of the models to reach an accuracy of 90 percent, achieving an accuracy of 90.2 percent. The model with 27 features is the first to reach an accuracy of 94 percent. Compared to the accuracy of 94.2 percent for the model incorporating all of the 138 features, this is a surprising result.

7.2.2 Increasing efficiency of the HAR system by calculating fewer features

Based on the promising result presented in section 6.1, it was decided to go ahead with modifying the feature calculation part of the HAR system. The results of this change are convincing. Figure 6.10 presents the time usage of the system for 4 of the steps in the HAR system related to predicting the activities of subjects: training the model, extracting windows, calculating features, and running the predictions. All 138 models are shown in the figure. The time it takes to train the different models increases with the number of features in the model. This is to be expected. It is important to note that the feature calculation step of the training process was not included in this statistic. This is a big limitation of the experiment, but can be somewhat neglected based on the fact that training the model is done once only. The time used by the RFC to classify the activities of subject 006 in the TFL dataset remains consistently low relative to time spent on the other parts of the system. The same holds true for the window extraction part of the system. It uses about the same time regardless of the number of features in the model. This is exactly as anticipated because the window extraction extracts the exact same windows for all the models. Therefore an increase in effectiveness is to be achieved primarily by a reduction in the number of features that is calculated for each window.

Figure 6.11 combines the results of the experiment in 6.1 with the feature calculation graph in figure 6.10. This figure presents an informative overview over the trade-offs to be made when balancing between accuracy and the overall efficiency of the HAR system. It is evident that using more features than the 29 of the features results in a significant decrease in effectiveness and only a very slight increase in accuracy. For the model with 138 features, the feature calculation took 21.97 seconds on average when tested on subject 006 in the TFL dataset. When using the 5 most important features, the feature calculation step of the HAR system took 0.94 seconds on average. This is 4.3 percent of the time it took for the model with 138 features, or 23 times faster. With the 27 most important features the feature calculation step took 3.72 seconds, 16.9 percent as much as the model with 138 features. This is a speedup of 5.9 times This model achieved an accuracy of 94.0 percent, 0.2 less than the full model. Hence, the choice between the model with 27 features and 138 features should be simple. When it comes to choosing the optimal number of features for both efficiency and accuracy the choice depends on what the user of the system prefers. If time is not an issue, the model with the highest accuracy should be chosen. If time is of more importance, the user must decide themselves what accuracy (or performance in terms of the other quality metrics) is acceptable. If 90 percent accuracy is acceptable figure 6.12 indicate that using 6 features is the best choice. It uses marginally more time than that of the model with 5 features, but increases the accuracy from 90 percent to 92 percent. If the the highest accuracy is desired figure 6.13 show that using about 40 or more features is sufficient. The accuracy stabilizes fairly well around that mark. If a certain emphasis is put on the the effectiveness, 21 to 29 features achieve about the same accuracy while being twice as effective. Based on these result it is clear that an increase in the effectiveness of the HAR system is achieved while maintaining the high accuracy of the system. It is up to the user of the system to make the desired compromise between the performance and effectiveness of the system.

7.2.3 No-wear time detection by the use of sensor temperature readings

The SNT classifier achieves an accuracy of 95.6 percent when tested with record-wise cross validation on the SNT dataset. It is able to successfully classify the sensor configurations with a high degree of certainty, in effect detecting instances of no-wear time (A or B). These results were achieved using 4 simple time domain features derived from temperature readings and a window size of two minutes. The performance of this system using these 4 simple features is surprisingly good.

There is however several crucial limitations of the system that limit the system from being used in real world applications such as the HUNT4 study. Most notable is the lack of the sensor configuration " N (no sensor)". Including this label in the system was in fact experimented with in the initial stages of the developing the experiment. However the accuracy of this system was poor, about ten percent lower than that of the system using A, B, and T. This might be due to the fact that the system only considers temperature data, and that separating between the A and N configuration using this data is hard. The temperature difference between the two sensors in configuration A is minimal. Unfortunately, the same will most often likely be true for the two sensor in configuration N. Additionally, the temperature changes of the sensors in configuration A are often very slow, meaning that the temperature is very stable. The same can be said for the N configuration. This led to several misclassifications between the two. It is believed that incorporating features derived from the acceleration signals of the accelerometer will alleviate this problem, because the "profile" of the acceleration signals should differ significantly between the two configurations. Lastly, the fact that the sensor were not synchronized before being labelled is another limitation of the experiment. The impact of this aspect on the results of the experiment has not been investigated.

In spite of these limitations, the potential of the classifier is clear. An additional benefit of having a sensor configuration classifier is the potential for a significant increase in efficiency of the HAR system used with the HUNT4 study (Vågeskar [2017]). Even though the subjects were the sensors for about 7 days on average, the recordings often last 8 or 9 days. This is because the sensors are transported back around before and after the recording are performed. This means that anywhere between 1 to 3 days of the recordings have the sensor configuration N. The current HAR system used for classification in the HUNT4 study classifies this period the same way it does for the rest of the data. In other words, the system performs a significant number of unnecessary and irrelevant calculation, slowing down the process, and in turn, the HUNT4 study analysis. This problem can be fixed with the SNT detection system. Removing the 1-3 last days of the recordings automatically with the help of the SNT classifier will result in significant time savings. This is in addition the the already significant timesaving achieved by the reduction in features.

Chapter 8

Conclusion and Future Work

This chapter presents the conclusion of this thesis along with potential areas of interest for future work.

8.1 Conclusion

Three goals were defined for this thesis in section 1.2. Goal 1 was to explore the state of the art machine learning methods and datasets that are commonly used in HAR research. This was presented in the systematic literature review in chapter 3. Goal 2 was to increase the effectiveness of the HAR system presented in Vågeskar [2017] by building on the results of the specialization project preceding this thesis (Reinsve [2017]). This indicated that the number of features used to train the random forest classifier could be significantly reduced while maintaining the accuracy. Goal 3 was to develop a classifier that was able to detect instances of sensor no-wear time and classify the the configuration of sensors attached on a subject. Goal 2 and goal 3 were explored through the three experiments presented in chapter 6. The results of these experiments were presented in section 6.1, 6.2, and 6.3.

All three goals of this thesis were reached. This work has resulted in one improvement and one addition to the HAR system that was presented in Vågeskar [2017]. Firstly, the effectiveness of the feature calculation has been significantly increased while maintaining the accuracy of the model. Secondly, a functioning sensor no-wear time classifier has been developed.

The effectiveness of the HAR system presented in Vågeskar [2017] was increased by modifying the feature calculation step of the activity classifier. This system classifies 7 different activities: walking, running, standing, sitting, lying, and bending. The increase in efficiency was achieved first by determining which features that were the most important for the RFC. The features were evaluated on the basis of their feature importance value, provided by the scikit-learn ML framework. The original system (Vågeskar [2017])

used 20 different time and frequency domain features. These features produced a total of 138 input features (69 features for each of the two sensors) to the RFC. When tested with a subject-wise cross-validation of the 16 subjects from the TFL dataset, the accuracy of this system proved to be 94.2 percent, equaling that of the same system in Vågeskar [2017]. It was shown that in order to reach an accuracy of 90.0 percent, a classifier trained on the 5 most important features was sufficient, resulting in 90.2 percent accuracy. By including the 6th most important feature in the model, the accuracy was raised to 92.0 percent. When further trained with the 27 most important features, the RFC achieved an accuracy of 94.0 percent. Based on these results, a significant increase in effectiveness was achieved by making changes to feature calculation part of the HAR system. This made it possible to calculate only the n most important features when classifying activities of subjects, while excluding the others. Then, models with corresponding features were used to make predictions. When using the 5 most important features, the feature calculation step of the HAR system took 0.94 seconds on average. This was about 23x (4.3 percent) faster than that of the model using 138 features, which took 21.97 seconds (subject 006 TFL, 134 minutes). With the 27 most important features the feature calculation step took 3.72 seconds, a 5.9x (16.9 percent) speedup compared to using 138 features, and at the same time achieving 94.0 percent accuracy.

A sensor no-wear time classifier was developed to detect instances of sensor no-wear time in recordings and classifying the sensor configuration worn by a subject. The system used temperature readings extracted from the on-board temperature sensor on the two AX3 accelerometers used. A random forest classifier was trained on the SNT dataset, which was created for this thesis. The system was used to recognize three sensor configurations: all sensors (A), back sensor (B), and thigh sensor (T). The configuration labels indicated which sensors were attached to the subject at any given time. The classifier obtained an accuracy of 95.6 percent using 4 features for each of the two temperature sensors when tested with subject-wise cross-validation on the 4 recordings in the SNT dataset.

In conclusion, the feature importances together with the accuracy and time statistics obtained in this work, provide useful insight into what features are the most important and necessary for the system to perform well overall. This insight make it possible to make decisions on models to use while balancing the trade-off between accuracy and effectiveness. The increase in effectiveness should prove useful for processing the HUNT4 data, which will be a time consuming process. The SNT classifier will make it possible to handle no-wear time instances for the 50,000 subjects expected to make up the HUNT4 dataset, and will aid the feature calculation and classification step of the HAR system to use the correct set of models for each of the subjects. This should increase the accuracy for the HUNT4 predictions.

8.2 Future Work

This section presents areas of interest for future work, identified throughout the work on this thesis.

8.2.1 Further optimisations on feature calculation

The feature importances that were presented in this work made it possible to identify which features that were important in order to successfully classify activities. To be able to use any combinations of features for training the RFCs, the mapping between the 20 unique time and frequency domain features had to be found. The indexes of the 138 features were mapped to the unique features producing them. With this mapping, it was possible to reduce the number of features that were calculated for a subject's windows of activities by reducing the number of the unique features that were calculated. This enabled a speedup of the HAR system in the feature calculation step. A limitation of this solution is that most of the unique features produced more than 1 input feature to the classifier. Hence, a unique feature were calculated in its entirety as soon as one of its produced features was selected for inclusion. This resulted in the non-linear growth of the time used on feature calculation as show in figure 6.10. Although this still resulted in a significant time saving, further work can be done in refining this process. A prime example of this is the current implementation of the 7 frequency domain features. They are calculated in the same Python method in the feature calculation module. This means that as soon as the 1 out of the 21 features this feature produces is selected for inclusion, all of them are calculated even though only one is used. This applies to many of the 20 real features. This is very ineffective. With modifications to the feature calculation approach used in this thesis, additional efficiency improvements can be achieved using the same approach and knowledge deduced in this work.

8.2.2 Analyze the impact on accuracy by adding more activities

For the experiments in this thesis, 7 out of the 18 activities defined for the TFL dataset were used. Adding more of the activities should affect the accuracy of the classifier. This affect should be explored in particular relation to the efficiency improvements presented in this work. With the 7 activities the model with the 5 and 27 most important features both performed relatively well. This might not be the case when increasing the set of activities to be classified. As a result, achieving the same results might not be possible with the substantial reduction in features presented in this thesis. Experiments should be conducted to examine this effect.

8.2.3 More comprehensive SNT classifier

The system developed in this thesis for detecting instances of no-wear time only considers the temperature readings of the sensor. For this very reason, it is difficult for the system to differentiate between a sensor configuration where either both sensors are attached to the subject, or both are detached. Making more features available to the classifier by including features derived from acceleration measurements of the sensor should help the encountering those configurations. This is based on the assumption that two accelerometers attached to a person should experience significantly more acceleration forces than two sensors that are detached. This might not be true in all cases however. The accelerometers can be left in a car that is driving around, or left in a cloth drier etc. Regardless, it is believed that when combining features derived from both the acceleration and tempera-

ture recording of the sensor, the system should be capable at dealing well with instances were both sensors are detached. This approach should also increase the accuracy of the predictions for the A, B, and T classes, as compared to using temperature-only features. Another benefit of including acceleration features would be the increase in "time resolution". The current system uses windows of two minute lengths for classification. This might be decreased significantly depending on the contribution of the acceleration features. Another limiting factor for the application of the sensor no-wear time classifier to real world data, is that it was trained exclusively on data collected inside of buildings. The temperature inside a building is usually relatively stable. When moving between inside and outside environments however, one might experience substantial temperature changes in a short period of time. This is especially true for locations in cold climates. For example, during winter in Norway the temperature might be around $23\,^{\circ}\mathrm{C}$ inside a given building, while the outside temperature might be as low as $-23\,^{\circ}\mathrm{C}$. While such dramatic temperature changes might not significantly affect the system, the extent of the effect cannot be determined until a dataset capturing these variables have been created and tested on. This aspect should also benefit from the addition of acceleration features.

8.2.4 Switching of models based on SNT sensor configuration-classifications

As shown in section 6.3, it is expected that a number of recordings in the HUNT4 study will contain instances of sensor no-wear time. The SNT classifier was effective in detecting instances with the sensor configurations A, B, and T. With further improvements to this system, as described above, it is believed that it will be better at classifying instances where both sensors are detached. With such a system in place an implementation can be integrated into the HAR system to be used for analyzing the HUNT4 data, as a step before the classifications step. With the HAR system now being able to recognize which sensors are attached to a subject at any given time, the system can change the models it uses for prediction on the go. One way to implement the model switching would be to separate subjects in the dataset into two groups: no-wear time "detected", and "not detected". What the threshold should be before a subject is put in the no-wear time "detected" category, is to be determined. For example, one might define that a subject with 5% or more samples classified differently than "A" (All sensors on) should be put in the no-wear time "detected" category. After dividing the subjects into groups, the system might iterate over the subjects in the "detected" category and classify the sensor configurations for each of the windows in the subject's recording using the SNT classifier. Then, slight modifications to the (activity) prediction step of the HAR system need to be implemented. The SNT system should pass on a flag to the prediction step, indicating whether the subject was placed in the "detected" group and is to use multiple models for prediction. The prediction step will use the sensor configuration predictions to select the corresponding models for each window and subject it predicts. For instance, if a subject wore both sensors for 2 days and then wore only the back sensor for 5 days, the system would use the dual-sensor model for the first two days and the single-sensor back model for the 5 days. Using the SNT classifier as developed in this work, models can be selected and applied on window lengths of 2 minutes.

8.2.5 A more comprehensive SNT dataset

The SNT dataset created for this thesis consists of four recording collected from two subjects performing both the P1 and P2 protocols. The dataset was sufficient to develop a proof of concept SNT classifier. However, it would be beneficial if a larger dataset with more subjects and variations in recording environments was made. This would also increase the changes of identifying edge cases where the system performs poorly. Such a dataset would provide a more robust framework for the continued development of the SNT classifier. This is necessary to aspire confidence in the ability of the system to function as intended if applied in large scale environments such as the HUNT4 study.

8.2.6 Video Annotation improvements for Labeling Training Data

The annotation process for the TFL dataset is both tedious and time consuming. The annotators have to examine the videos recorded of the subjects on a frame-by-frame basis and label them with the ground-truth activities. With an average video length of over two hours for each subject in the TFL dataset, labeling all the videos accurately would be very laborious. Speeding up this process would make it easier to label more, and longer, videos. This could in turn inspire more datasets to be created. A way to do this could be to utilize the classifier(s) that are already available and have been trained on other datasets, or a subset of the dataset, to process the accelerometer recordings of new data and predict the activities. Then, these predictions could be used as a temporary ground truth assignment in the video labeling tool. With all the frames labeled, the job of the annotator would be to quickly go through the video and to verify that segments of the video, of greater length than a single frame, match the corresponding segments of labeling performed by the classifier. If a mismatch occurs, the annotator should be able to manually adjust the timeframe of the activities and /or relabel them. The change from manually selecting activities on a frame-by-frame basis, to verifying already labeled segments, should significantly speed up the process of labeling training data.

Bibliography

A. Avci, S. Bosch, M. Marin-Perianu, R. Marin-Perianu, and P. Havinga. Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey. In *Architecture of computing systems (ARCS), 2010 23rd international conference on*, pages 1–10. VDE. ISBN 380073222X.

A. Bayat, M. Pomplun, and D. A. Tran. A study on human activity recognition using accelerometer data from smartphones. *Procedia Computer Science*, 34:450–457, 2014. ISSN 1877-0509.

M. F. A. bin Abdullah, A. F. P. Negara, M. S. Sayeed, D.-J. Choi, and K. S. Muthu. Classification algorithms in human activity recognition using smartphones. *International Journal of Computer and Information Engineering*, 6:77–84, 2012.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. ISSN 0885-6125.

A. Bulling, U. Blanke, and B. Schiele. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys (CSUR)*, 46(3):33, 2014. ISSN 0360-0300.

J. M. Chaquet, E. J. Carmona, and A. Fernndez-Caballero. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*, 117 (6):633–659, 2013. ISSN 1077-3142.

T. G. Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

P. Flach. *Machine Learning The Art and Science of Algorithms that Make Sense of Data*. Cambridge University Press, New York, 2012. ISBN 978-1-107-09639-4.

H.-O. Hessen and A. J. Tessem. Human activity recognition with two body-worn accelerometer sensors. *Master's thesis, Trondheim*, 2016.

N. Hunt Research Center. About hunt. December 2017. URL `https://www.ntnu.edu/hunt/about-hunt`.

S. A. Ismail, A. F. A. Matin, and T. Mantoro. A comparison study of classifier algorithms for mobile-phone's accelerometer based activity recognition. *Procedia Engineering*, 41: 224–229, 2012. ISSN 1877-7058.

O. D. Lara and M. A. Labrador. A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys and Tutorials*, 15(3):1192–1209, 2013.

F. G. Larsen and E. Vågeskar. Investigating the performance of a human activity recognition system on out-of-lab data. *Technical report, Trondheim*, 2016.

D. Micucci, M. Mobilio, and P. Napoletano. Unimib shar: A dataset for human activity recognition using acceleration data from smartphones. *Applied Sciences*, 7(10):1101, 2017.

T. M. Mitchell. *Machine Learning*. McGraw-Hill series in computer science. McGraw-Hill, New York, 1997. ISBN 0070428077. URL `Publisherdescriptionhttp://www.loc.gov/catdir/description/mh022/97007692.htmlTableofcontentshttp://www.loc.gov/catdir/toc/mh022/97007692.html`.

E. Ordua-Malea, J. M. Aylln, A. Martn-Martn, and E. D. Lpez-Czar. About the size of google scholar: playing the numbers. *arXiv preprint arXiv:1407.6239*, 2014.

S. J. Preece, J. Y. Goulermas, L. P. Kenney, D. Howard, K. Meijer, and R. Crompton. Activity identification using body-mounted sensorsfi?!a review of classification techniques. *Physiological measurement*, 30(4):R1, 2009. ISSN 0967-3334.

T. Prestmo. Decision support in patient-centered health care. *Master's thesis, Trondheim*, 2017.

Ø. Reinsve. Improving the activity recognition and speed of prediction in an existing human activity recognition system. *Project thesis, Trondheim*, 2017.

S. J. Russel and P. Norvig. Artificial intelligence: A modern approach. 2010.

A. Siddaway. What is a systematic literature review and how do i do one. *University of Stirling*, (I):1, 2014.

E. Vågeskar. Activity recognition for stroke patients. *Master's thesis, Trondheim*, 2017.

Appendices

Appendix A

Activity definitions

Activity	Definition
Walking	Locomotion towards a destination with one stride or more, (one step with both feet, where one foot is placed at the other side of the other). Walking could occur in all directions. Walking along a curved line is allowed.
Running	Locomotion towards a destination, with at least two steps where both feet leave the ground during each stride. **For chest mounted camera:** Running can be inferred when trunk moves forward in a constant upward-downward motion with at least two steps. Running along a curved line is allowed.
Shuffling	Stepping in place by non-cyclical and non-directional movement of the feet. Includes turning on the spot with feet movement not as part of walking bout. **For chest mounted camera:** Without being able to see the feet, if movement of the upper body and surroundings indicate non-directional feet movement, shuffling can be inferred.
Stairs, ascending or descending	**Start:** Heel-off of the foot that will land on the first step of the stairs. **End:** When the heel-strike of the last foot is placed on flat ground. (If both feet rest on the same step with no feet movement, standing should be inferred.)
Standing	Upright, feet supporting the person's body weight, with no feet movement, otherwise this could be shuffling/walking. Movement of upper body and arms is allowed until forward tilt and arm movement occurs below knee height. Then this should be inferred as bending. **For chest mounted camera:** If feet position is equal before and after upper body movement, standing can be inferred. Without being able to see the feet, if upper body and surroundings indicate no feet movement, standing can be inferred.
Sitting	When the person's buttocks is on the seat of the chair, bed or floor. Sitting can include some movement in the upper body and legs; this should not be tagged as a separate transition. Adjustment of sitting position is allowed.
Lying	The person lies down. Adjustment after lying down is allowed if it does not lead to a change between the prone, supine, right and left lying positons. Movement of arms and head is allowed.
Transition	Transitioning between any of the activities listed here.
Bending	While standing/sitting, bending towards something below knee-height is tagged as bending. Steps can occur during bending.
Picking	Refers to picking/placing/touching an object from below knee height. Picking occurs when the trunk is at its lowest and the person has touched/placed/picked an object. When the person starts to rise the trunk, picking finishes, and bending begins. Adjustment of position while picking is allowed.
Undefined activity	Until all the sensors are attached, or final adjustment made to position the video camera can be tagged as undefined. All postures/movements that can not be clearly identified due to blocking of the camera/view should be tagged as undefined.
Cycling (sitting)	Pedaling while the buttocks is placed at the seat. Cycling starts on first pedaling and finishes when pedaling ends. **Not pedaling:** Sitting without pedaling should be tagged separate as sitting.
Non-vigorous activity	All non-cyclic movements that are recognizable, but do not classify according to the definitions. Can occur in all directions. Can be crawling, rolling, cleaning the floor etc.

Figure A.1: Description of the activities used for labeling the TFL dataset

Appendix B

Distribution of activities for the TFL data set

This chapter presents bar plots of the activity distribution for all subjects used in this project from TFL data set. These are not all the activities in the data set, but the ones that the classifier was trained on.

(a) Distribution over all subjects. Time: 1757 minutes

(b) Subject 006. Time: 134 minutes

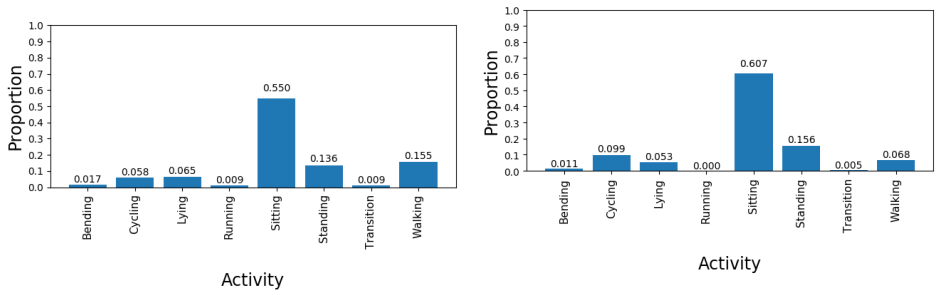
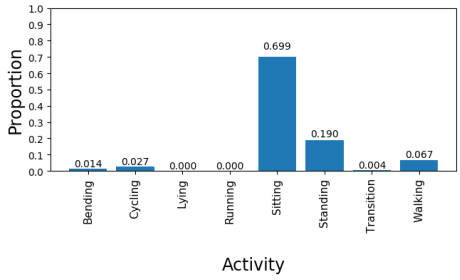
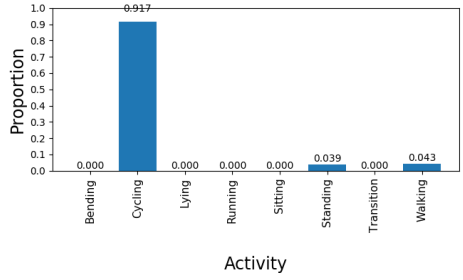


Figure B.1: Activity distribution

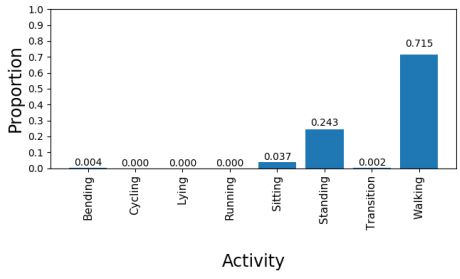
(c) Subject 008. Time: 114 minutes



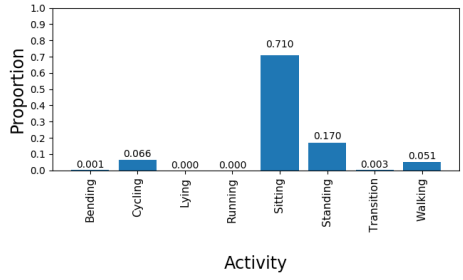
(d) Subject 009. Time: 51 minutes



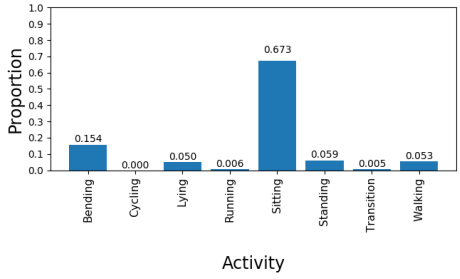
(e) Subject 010. Time: 86 minutes



(f) Subject 011. Time: 115 minutes



(g) Subject 012. Time: 126 minutes



(h) Subject 013. Time: 112 minutes

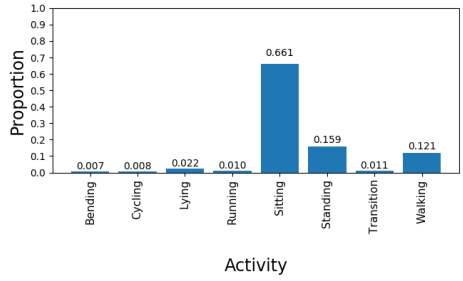
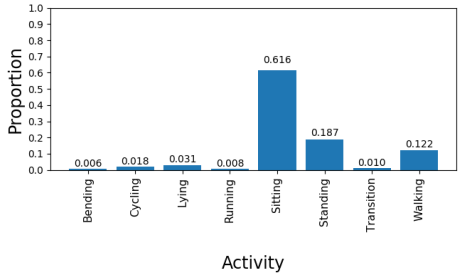
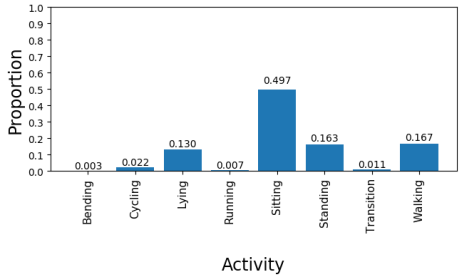


Figure B.1: Activity distributions

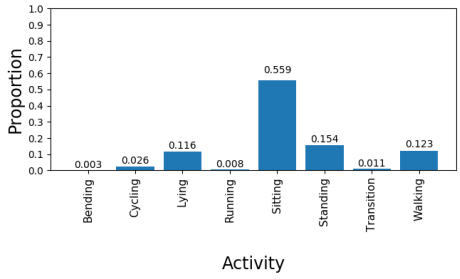
(i) Subject 014. Time: 100 minutes



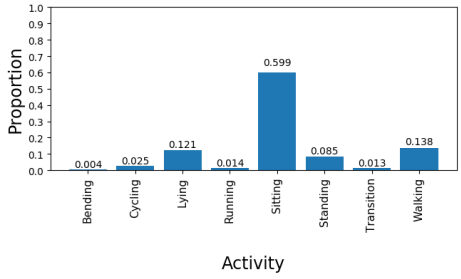
(j) Subject 015. Time: 137 minutes



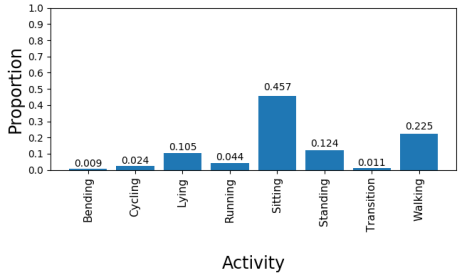
(k) Subject 016. Time: 114 minutes



(l) Subject 017. Time: 121 minutes



(m) Subject 018. Time: 104 minutes



(n) Subject 019. Time: 105 minutes

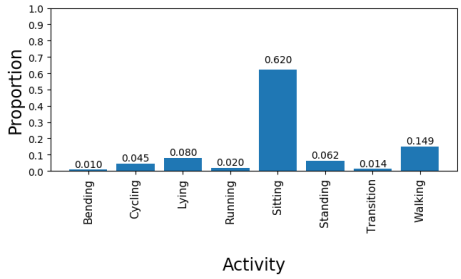
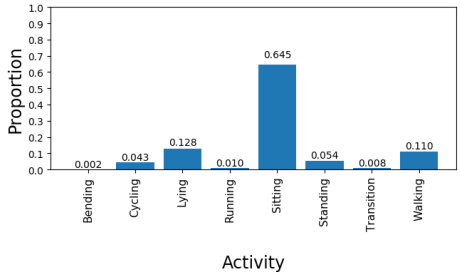
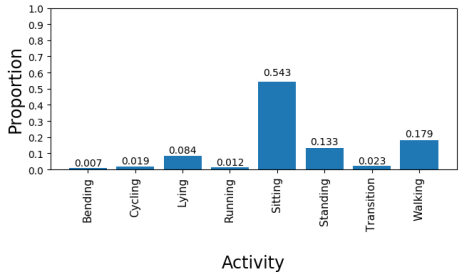


Figure B.1: Activity distributions

(o) Subject 020. Time: 122 minutes



(p) Subject 021. Time: 107 minutes



(q) Subject 022. Time: 109 minutes

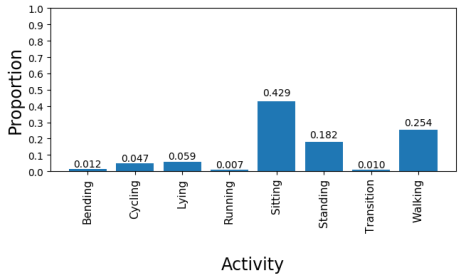


Figure B.1: Activity distributions

Appendix C

Distribution of sensor configurations for the SNT dataset

(a) Distribution over all recordings. Time: 999 minutes

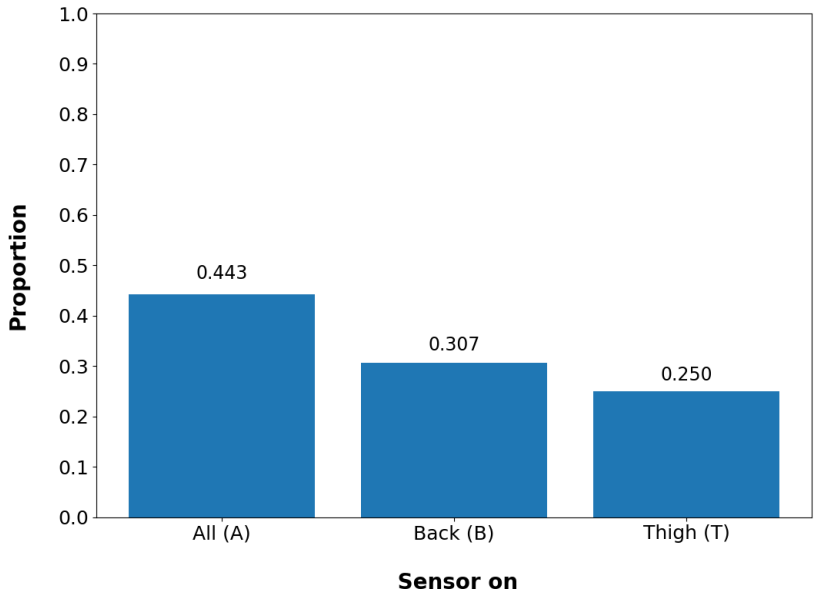


Figure C.1: Sensor configuration distribution

(b) Subject 1 and subject 2 (Protocol 1) . Time: 259 minutes

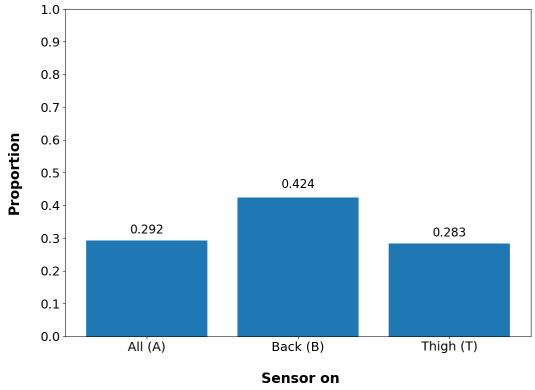


Figure C.1: Sensor configuration distribution

(c) Subject 1 and Subject 2 (Protocol 2) . Time: 568 minutes

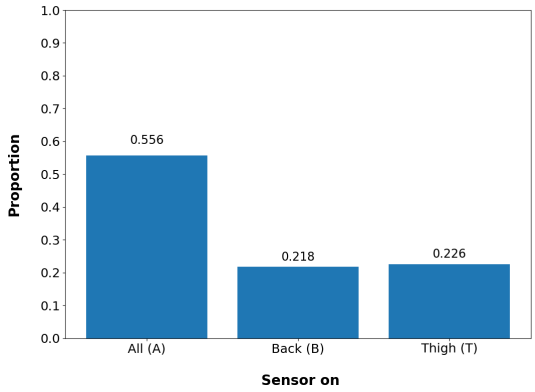
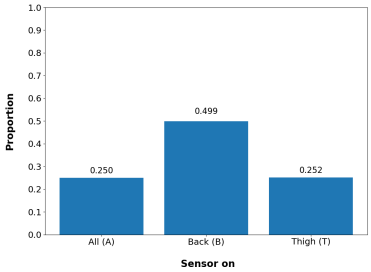
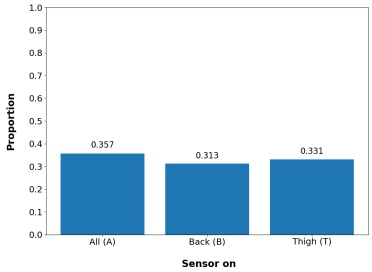


Figure C.1: Sensor configuration distribution

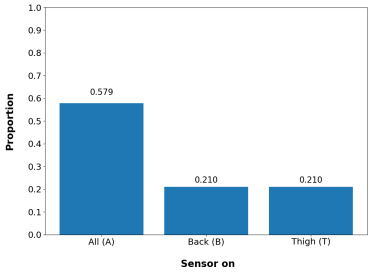
(d) Subject 1 (Protocol 1) . Time: 259 minutes



(e) Subject 2 (Protocol 1). Time: 172 minutes



(f) Subject 1 (Protocol 2). Time: 292 minutes



(g) Subject 2 (Protocol 2). Time: 277 minutes

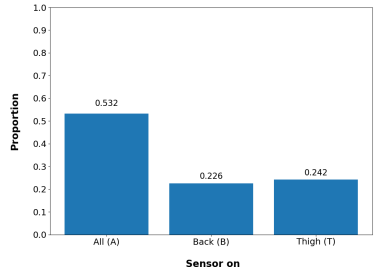


Figure C.1: Sensor configuration distribution