Bachelor Thesis

Czech
Technical
University
in Prague

F3

Faculty of Electrical Engineering
Department of Computer Science

Use of data mining for analysis of human physical activity by accelerometer generated data

Tomáš Nagy

Supervisor: Ing. Pavel Náplava, Ph.D.
Study Programme: Software Engineering and Technology
August 2020

BACHELOR'S THESIS ASSIGNMENT

I. Personal and study details

Student's name: **Nagy Tomáš** Personal ID number: **453033**

Faculty / Institute: **Faculty of Electrical Engineering**

Department / Institute: **Department of Computer Science**

Study program: **Software Engineering and Technology**

II. Bachelor's thesis details

Bachelor's thesis title in English:

Use of data mining for analysis of human physical activity by accelerometer generated data

Bachelor's thesis title in Czech:

Využití data miningu pro analýzu fyzické aktivity člověka pomocí dat generovaných akcelerometrem

Guidelines:

Analyse how data from the accelerometer can be used to identify the movement habits of individuals. Focus primarily on finding human behavioural patterns.
Follow these steps:
1) Describe the accelerometer device and its usage.
2) Analyse and describe the data that can be acquired and evaluated from the accelerometer for measuring human activities.
3) Describe algorithms that are generally used for data mining and select at least one that can be used for human physical activity classification.
4) Apply the selected algorithm(s) to the dataset available at [1] as follows:
a) preprocess data from the dataset to the form that can be used by the selected algorithm(s);
b) apply the selected algorithm(s) to the preprocessed dataset and try to find correlations, clusters or other dependencies between different input parameters;
c) analyse obtained results
5) Based on the results try to find a practical application for encouraging a healthy lifestyle among individuals.

Bibliography / sources:

[1] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra and Jorge L.Reyes-Ortiz. A Public Domain Dataset for Human Activity Recognition Using Smartphones. 21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2013. Bruges, Belgium 24-26 April 2013.
Dostupné na:
http://archive.ics.uci.edu/ml/datasets/human+activity+recognition+using+smartpho
ne
[2] Charu C. Aggarwal, Managing and Mining Sensor Data, Springer Publishing Company, Incorporated, 2013
[3] A. Yassine, S. Singh and A. Alamri, "Mining Human Activity Patterns From Smart Home Big Data for Health Care Applications," in IEEE Access, vol. 5, pp. 13131-13141, 2017.

Name and workplace of bachelor's thesis supervisor:

Ing. Pavel Náplava, Ph.D., Department of Economics, Management and Humanities, FEE

Name and workplace of second bachelor's thesis supervisor or consultant:

Date of bachelor's thesis assignment: **14.02.2020** Deadline for bachelor thesis submission: **14.08.2020**

Assignment valid until: **30.09.2021**

Ing. Pavel Náplava, Ph.D. Supervisor's signature	Head of department's signature	prof. Mgr. Petr Páta, Ph.D. Dean's signature

III. Assignment receipt

The student acknowledges that the bachelor's thesis is an individual work. The student must produce his thesis without the assistance of others, with the exception of provided consultations. Within the bachelor's thesis, the author must state the names of consultants and include a list of references.

. _____
Date of assignment receipt

_____
Student's signature

Acknowledgements

Firstly, I would like to thank Ing. Pavel Náplava, Ph.D. for being my supervisor. His support, constant positive attitude, and constructive feedback helped me during the realization of this thesis.
Secondly, my gratitude belongs to my family and friends who shared their support, either morally, physically or financially. Thank you.

Declaration

I hereby declare that I have written this work independently and quoted all the sources of information used in accordance with the methodological instructions on ethical principles for writing an academic paper. Moreover, I state that this work has neither been submitted nor accepted for any other degree.

In Prague, 14. August 2020

Abstract

The rise of computing and sensing power has allowed us to implement various applications based on the ability to gather and analyse data in order to make our lives easier and smarter. For instance, applying Human Activity Recognition (HAR) serves such a purpose. An application based on HAR enables the recognition of a particular activity performed by a user. Building a HAR system is a complex task. A mistake in its fundamentals, like using an unreliable data source or utilising an incorrect technique, can result in inaccuracies within the classification of the performed activities. In this study, we examined five supervised machine learning algorithms to confirm whether they can predict the set of predefined physical activities with a desirable accuracy on a chosen pre-processed dataset. Based on that, a hypothesis was set. We evaluated the performance of the implemented algorithms. The result was that the Support Vector Classifier renders the most accurate classification results, reaching 91.24 %. Random Forest Classifier, Multinomial Logistic Regression, K-nearest Neighbors provided results with approximately 90 % of classification accuracy. The weakest results were obtained by Decision Tree Classifier. The hypothesis was proven.

Keywords: HAR, accelerometer, Machine Learning, Classification

Abstrakt

Nárast výpočtovej a snímacej sily nám umožňuje implementovať rôzne aplikácie založené na schopnosti zhromažďovať a analyzovať dáta, aby sa náš život stal ľahším a inteligentnejším. Pre tento účel slúžia aplikácie na báze rozpoznávania ľudskej aktivity (HAR). Aplikácia založená HAR umožňuje rozpoznať konkrétnu vykonanú činnosť uživateľom. Vybudovanie systému HAR je komplexná úloha. Chyba pri budovaní jeho základocv, ako napríklad použitie nespoľahlivého zdroja dát alebo použitie nesprávnej techniky, môže viesť k nepresnosti klasifikácie vykonaných činností. V tejto štúdii sme skúmali päť algoritmov supervizovaného strojového učenia, aby sme potvrdili, či môžu s požadovanou presnosťou predpovedať množinu preddefinovaných fyzických aktivít na vopred vybratom predspracovanom súbore dát. Na základe toho bola stanovená hypotéza. Hodnotili sme výkon implementovaných algoritmov. Výsledkom bolo, že klasifikátor podporných vektorov vykonáva najpresnejšie výsledky klasifikácie a dosahuje 91,24 % úspešnosti. Náhodný les, multinomická logistická regresia a algoritmus k-najbližších susedov dosiahli 90 % presnosť klasifikácie. Najslabšie výsledky boli získané pomocou klasifikátora rozhodovacích stromov s presnosťou 83 %. Hypotéza bola potvrdená.

Klíčová slova: HAR, Akcelerometer, Strojové Učenie, Klasifikácia

Contents

Part II
Practical part

Figures

Tables

Part I

Theoretical Part

Introduction

Motivation

The last decade has brought a significant improvement in smartphone technologies. The massive computing and sensing power have allowed us to implement numerous applications based on the ability to acquire and analyse data. Thanks to these technological innovations, various opportunities appeared in the research. One of them is Human Activity Recognition (HAR). The aim of the HAR system is the recognition of human activity patterns from a dataset using Machine Learning techniques. Various approaches exist to tackle a HAR problem. As the data source, some of them use video or static images, and others use sensors.[CS]

In this thesis, we focus on a dataset gathered from the accelerometer embedded in the smartphone. The sensor itself is characterised by low power consumption and high accuracy. It is considered as one of the most suitable electronic devices for HAR.[SS17]

The accelerometer provides motion recognition for a wide range of daily activities such as standing, walking, sitting, laying. Being able to identify the pattern can be priceless information. HAR is transforming the landscape of people's daily habits by contributing to a wide range of applications as health

and fitness tracking, elder care and automated monitoring. For instance, in a health care based application, it is possible to use the outcome as the core to build an overview of the patient movements. In 2015, a life insurance company created a program based on the activity tracking to motivate its patients to live a healthier way. The company logs the activity from wearable devices and provides statistics to its users. Depending on the client's achievement, it offers different benefits. [Bar]

We can claim, that HAR has become a relevant field which nowadays serves a wide range of applications.

During my bachelor studies, I had an opportunity to do an exchange program at the University of Buenos Aires. There happened to be my first encounter with the Artificial Intelligence (AI) field. Later, I became curious in data processing and data mining using different Machine Learning techniques. I attended several extra courses connected to the field.

While I was doing the research for the Semestral Project, I was introduced to the HAR field, in which I subsequently became interested in. I considered it as an important current topic for its possible applications, especially in the healthcare. I decided to do my Bachelor Thesis based on this relevant research area. Thanks to applications based on human activity identification, we will be able to help to prevent, treat and manage different diseases and provide for the physical and mental well-being of our citizens.

Research Challenges and Hypothesis

There are several challenges, and difficulties found even in a simple way of constructing a HAR system.

Firstly, selecting the right sensor or combination of sensors or picking the attributes and metrics to be measured can bring us several problems.

Secondly, choosing suitable tools and techniques to capture the differences between a set of daily activities is also challenging.

As the implementation of a complete HAR system is a complex task, there a pre-processed HAR dataset was chosen.

In this thesis, with the supervisor, we decided to focus more on the analysis and implementation part of the HAR system.

In order to confirm that the chosen accelerometer generated dataset could be used for classification of predefined physical activities, we will apply five different supervised Machine Learning techniques. According to several scientific papers, the prediction accuracy is on avarage 90 %, depending on the applied algorithm. [LL13] [BA09] Therefore, we state the following hypothesis:

- **H1** - At least one of the applied supervised Machine Learning algorithms, modelled on the chosen HAR dataset is capable of classifying a predefined set of daily activities with more than 90% of accuracy.

Objective and Milestones

The main objective of this work is to prove or disprove the above stated hypothesis based on our analysis and implementation of the HAR system. From the principal objective, the following milestones are determined:

- Analysis of opportunities for creating a human activity recognition classifier
- Assessment of the classifier Machine Learning algorithms
- Selection of suitable algorithms to accomplish the main objective of the work
- Implementation of physical activity recognition from a static dataset gathered by a triaxial accelerometer embedded into a smartphone
- Validation and evaluation of the results
- Finding usage and practical application based on results

Methodology

In order to accomplish the objective and milestones, the methodology follows these steps:

- **Step 1 - Literature review and related work** - This step involves information gathering from different works related to HAR.
- **Step 2 - Analysis of the collected information** - The set of information collected in Step 1 is analysed in order to be able to design a HAR system.
- **Step 3 - Development of the method** - The third step contains problem identification with a suitable design and model implementation based on the dataset.
- **Step 4 - Result analysis** - Here, the method's results are evaluated, and conclusions are drawn. By step 4, the hypothesis is proven or disproven.
- **Step 5 - Discussion and limitation** - We discuss the opportunities and possible applications of the system. Also, we describe the limitations of our implementation.

Structure

The thesis consists of two main parts, structured as follows:

- **Theoretical Part**
  - **Introduction** presents a brief introduction to the Human Activity Recognition (HAR) concept and its application; it sets the hypothesis and main objective of the work with its milestones; it shortly describes the followed methodology during this work.

- **Chapter 1** describes the general overview of HAR including accelerometer. Afterwards, it explains the concept of Machine Learning, including suitable algorithms candidates to build a system recognising daily activity patterns. In the end of Chapter 1, the utilised evaluation techniques are presented.

- **Practical Part**
  - **Chapter 2** introduces the dataset with a preprocessing procedure. Then, it demonstrates exploratory data analysis and the implementation of classification methods.
  - **Chapter 3** evaluates the obtained results with a performance measure. Based on results the hypothesis of the thesis is proven or disproven.
  - **Chapter 4** summarises the results and draws possible applications and the future opportunities.
  - **Conclusion** summarizes the study by reiterating the problem definition and our results.

Chapter 1

Chapter 1 provides a complete review of the literature about human activity recognition and the state-of-the-art technique. We followed databases such as ResearchGate, ACM Digital Library and IEEE Xplore. These sites provide a wide range of high-quality published scientific papers written by experts in the research field.

The first part of this chapter gives the groundwork for setting all the conditions for HAR. It begins with an introduction to the human activity recognition research area. HAR problem is defined, thereafter typical workflow to solve it is presented. Next, design issues are detailed. Lastly, the utilized sensor, accelerometer, is introduced. For a better understanding, a graphical layout of the first part of Chapter 1 is provided in Figure 1.1.

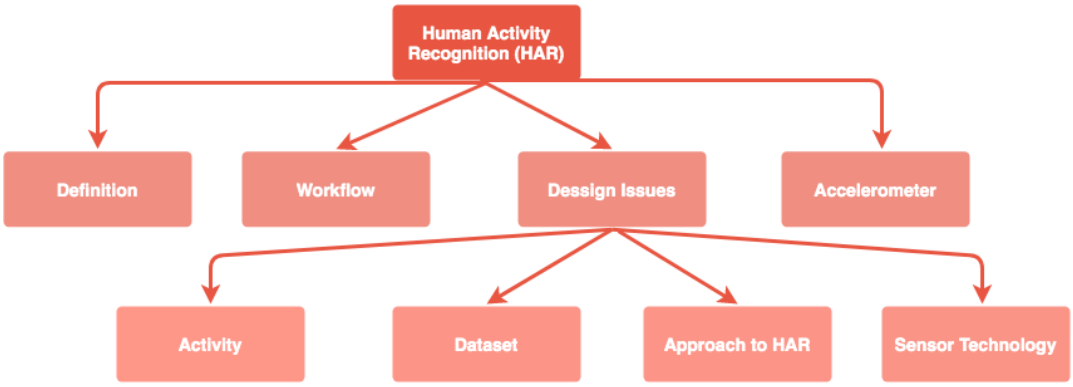

Figure 1.1: Chapter 1 - graphical layout of the first part

The second part presents a general overview of the state-of-the-art technology used for recognising physical activity. Firstly, there is a short introduction to data mining. Then, the theory of the Machine Learning area in the HAR context is discussed, including both supervised and unsupervised learning methods along with its applications. Futhermore, each Machine Learning method used in the practical part will be introduced.

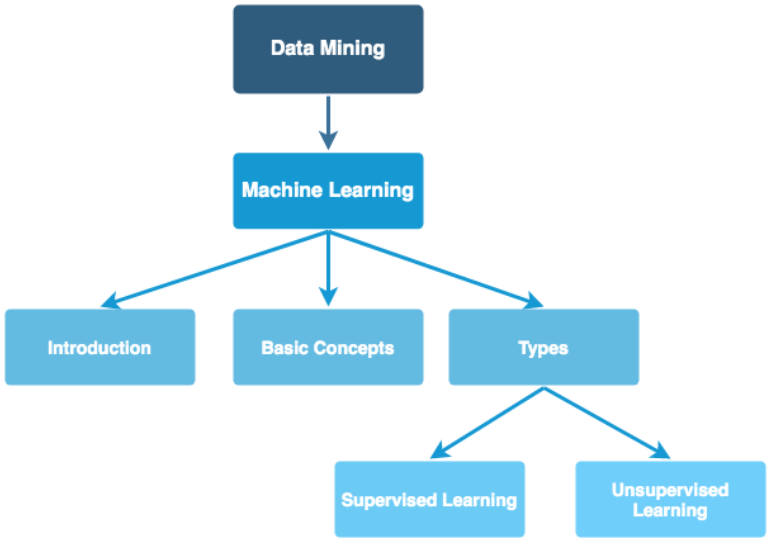

Figure 1.2: Chapter 1 - graphical layout of the second part

The third part of the first chapter describes the model evaluation metrics, which will help us to evaluate the success of the implemented Machine Learning techniques.

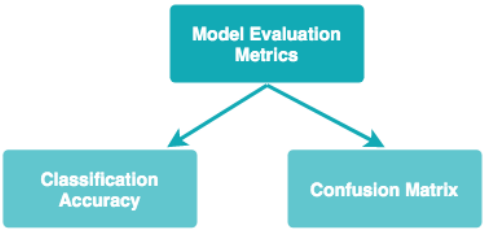

Figure 1.3: Chapter 1 - graphical layout of the third part

1.1 Human Activity Recognition

In the past decades, there has been an outstanding development of microelectronics and computer systems. Nowadays, thanks to these innovations sensors and mobile phones have high computational power, a small size, and low costs. The impact of such devices has become significant in people's daily life.

That was the beginning of Ubiquitous Sensing, a research area with the main objective of mining knowledge from the device-generated data. Notably, identification of human activities has become a key task in research, particularly in medicine, army, and health.[PLB10]

For instance, medical background examples could be patients with obesity, diabetes, or heart disease, who are usually required to follow a defined exercise routine as part of the treatment. Consequently, identifying activities such as walking or jogging contributes to building a general overview of the patient's activities for the caregiver. Similarly, patients with mental pathologies like dementia can be monitored to detect irregularities in behaviour patterns with the purpose to avoid unwanted consequences. [YYP08]

1.1.1 Human Activity Recognition Problem

After the short introduction in the HAR topic, we continue with a definition of the HAR problem.

Based on the paper [SS17], we are motivated to use wearable sensors in HAR. The measured attributes are usually related to the user's movement (e.g., accelerometers), physiological signals (e.g., heart rate) or environmental variables (e.g., temperature). These data are indexed over the time dimension. The following definition of the har problem is cited from the paper written by Lara and Labrador:

Definition 1.1. "Given a set W = $\{W_0, ..., W_{m-1}\}$ of m equally sized time windows, totally or partially labeled, and such that each W_i contains a set of time series $S_i = \{S_{i,0}, ..., S_{i,k-1}\}$ from each of the k measured attributes, and a set $A = \{a_0, ..., a_{n-1}\}$ of activity labels (e.g., sitting, walking, etc.). The goal is to find a mapping function $f : S_i \rightarrow A$ that can be evaluated for all possible values of S_i, such that f(S_i) is as similar as possible to the actual activity performed during W_i."

According to the definition the time series are divided into fixed-length time windows. The goal of the HAR problem is to find a function which assigns the correct activity to each time window W_i.

[LL13]

1.1.2 Solution to the HAR Problem

To solve the human activity recognition problem, we need to follow several steps. By accomplishing these steps we build a HAR system.

First, we need to gather the data from the perfomed activities, then pre-process it. Afterwards, we can use the data to implement a predictive model for identifying the activities.

Bulling et al. [BBS14] describes the typical workflow of creating a

human activity recognition system as follows:

- **Data Collection** - The data is acquired from a wearable sensor or an external device like cameras during the performed activity.
- **Data Pre-Processing** - To assign to each data point a corresponding activity, the data needs to be labelled. The data pre-processing also includes noise removal, resampling.
- **Data Segmentation** - The data set is divided into smaller segments, called windows. Each window corresponds to a particular activity.
- **Feature Generation** - and Selection: Here, new features are derived from each window, which reflects the characteristics of the data.
- **Classification** - To train a classification algorithm features are used. It enables us to distinguish between different activities.

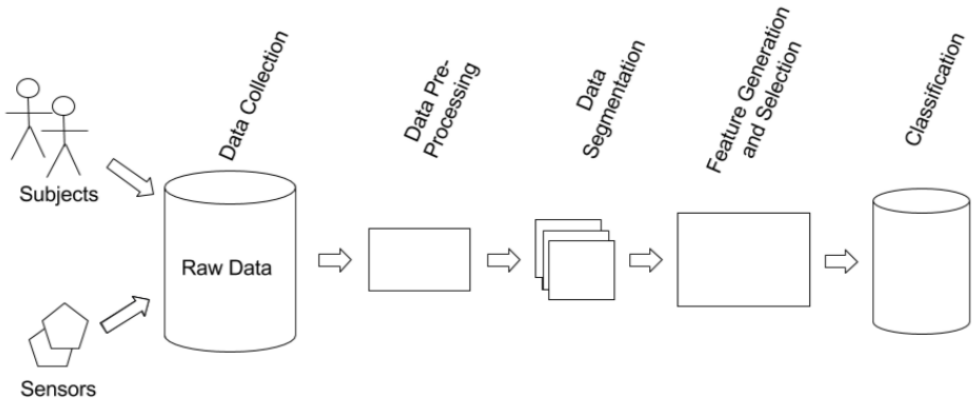

Figure 1.4: The process of creating a HAR system [HT16]

After introducing the process of creating a HAR system, we can claim, that it is a complicated set of tasks. Furthermore, according to various papers, to achieve high predictive accuracy results for HAR systems, we need

to use a well-preprocessed dataset. [SS17] [GD14] As we mentioned in the Introduction, in this thesis, we focus more on the analysis and implementation part of the HAR system. For these reasons, we will utilize a pre-processed dataset.

1.2 Design Issues

The recognition of human activities is a classification problem. To be able to construct a classifier model, first we need to clarify several factors, which have a significant impact on the outcome's success. [Kha11]

In the following part, the main problems and challenges concerning human activities, sensors, data and approach to create a har system, which are stated.

1.2.1 Activity

1.2.1.1 Number of Activities

The HAR system is capable of recognising a wide range of activities. It's less challenging to to identify a small number of activity patterns than a larger one. The reason is that the classifier has to differentiate among a broader range of activities as the number of activities increases. [Kha11]

1.2.1.2 Types of Activities

As mentioned in the paper written by Khan et al. [Kha11], the set of postures and activities can be divided into static and dynamic.
Activities such as sitting, standing, laying belong to static ones and walking,

running to the dynamic ones. These basic human activities were chosen based on the utilised dataset and possible applications of our work.
Some static and dynamic activities are more difficult to discriminate because their patterns can overlap in the feature space. For instance, postures, like sitting and standing or walking upstairs and walking downstairs, are more challenging to distinguish owing their possible overlapping pattern.[Kha11]

1.2.1.3 Conditions During the Performed Activity

Data can be collected in the laboratory or under free-living conditions. The dataset generated in the laboratory usually follows some protocol, which forces the participating subject to accomplish the activity at the constant speed and duration. In real conditions, people may act differently and in less restricted ways, which can affect accuracy in a negative way. [Kha11]

1.2.2 Data

1.2.2.1 Data Quantity

Data collected from a small number of people might not be sufficient to provide flexible activity patterns recognition of a new user. Lara et al. [LL13] shows there should be collected data from people with different age, gender, height, physical conditions, to be able to provide high accuracy.

1.2.3 Approach to Build HAR

All human beings have different characters, because of age, gender, weight, height, lifestyle or physical abilities. For this reason, each individual

has a unique moving pattern, which implies that the same activity possesses various representations. An independent subject activity recognition with a high-accuracy is hard to achieve because each set of activities has a high variability of performance. Papers suggest that recognition models should be able to generalise as much as possible concerning the final user and the execution context. According to the previous works, there are three approaches, which help to achieve generalisable activity recognition: subject-independent, subject-dependent, and hybrid. Chen et al. [CS] called these approaches rest-to-one, one-to-one, and all-to-one. The graphical representation of these HAR approaches is in Figure 1.5.

1.2.3.1 Subject-independent(Impersonal)

The subject-indepedent approach does not use the end-user data to develop the activity recognition model. It creates a flexible single activity recognition model that generalises the diversity between users and provides high-accurate results once a new user is classified.

1.2.3.2 Subject-dependent (Personal)

The subject-depedent approach does use the end-user data to develop the activity recognition model. The model generalises the real context well, and can also capture peculiarities. The disadvantage is that it must be implemented for each end-user.

1.2.3.3 Hybrid

The hybrid approach uses the end-user data and the data of the other users for recognition model development. The motivation behind this approach is to easily recognise the performed activity with a higher-accuracy. [BBSB10]

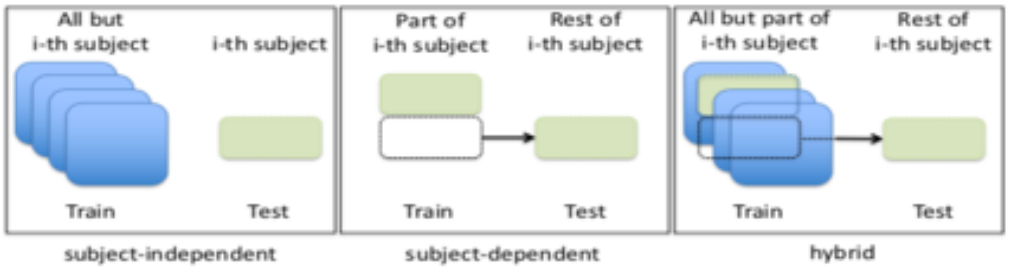

Figure 1.5: Three different approaches to achieve activity recognition [BBSB10]

1.2.4 Sensor Technology

1.2.4.1 Type of Sensors

There are various ways to gather data for a single user activity recognition. A popular method to collect data from an activity can be implemented by using sensors. Mobile phones and other wearable devices incorporate different types of sensors which include accelerometer, GPS, Gyroscope, temperature and blood pressure sensors.[KWM11]

Some process information about the environment, others about the user's locomotion. Every sensor does not provide sufficient information for identification of a set of activities. The most used sensor, which provides sufficient information to be able to recognise basic activity patterns (walking, lying, jogging) is the accelerometer.

Accelerometers do not require high power, and they are an integral part of today's phones and other wearable devices; also, they are cheap. Multiple studies used the accelerometer as a motion sensor, achieving excellent results. Under different evaluation methodologies, the human recognition accuracy of the primary activities with an accelerometer can be up to 98%. [LL13]

1.2.4.2 Sensor Location

There are two ways to collect data for a HAR classifier model, depending on the sensor location:

- Data Collection External sensors - these tools are placed at fixed positions in the environment (cameras, household appliances, environmental sensors)
- Body-worn sensors - the sensor is attached to the user or embedded in a device such as a mobile phone or wearable device like smartwatch, fitness tracker.

[Kha11]

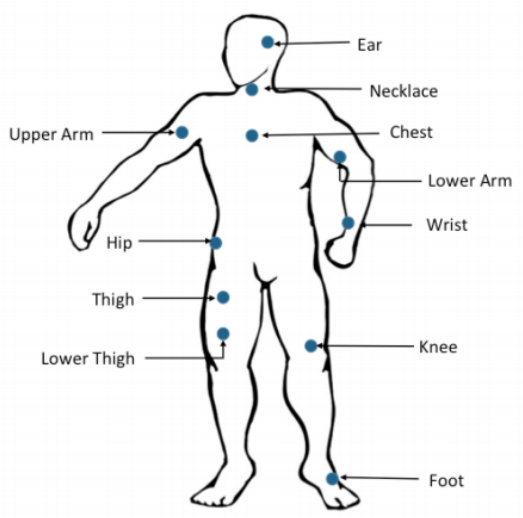

Figure 1.6: Illustration of body-worn sensors placement [Vel17]

To choose body-worn sensor, like accelerometer has several benefits:

the activity is continually measured, and it is independent of the environmental conditions (e.g., light, sound) and geographic location. The dataset used for the thesis, contains data gathered by smartphone's accelerometer.

1.2.4.3 Sensor Location on the Body

The position of a smartphone or other wearable devices affects the sensor-generated data. For instance, reading data differs when the user is walking, wearing the phone in the pocket or holding it in hands. [STJ14] Gupta & Dallas (2014) created a HAR system using a single sensor placed at the waist. They received a classification accuracy of 98%. [GD14] Bonomi et al. [GD14] have received an accuracy of 93% while detecting for similar activities and sensor placement. Consequently, according to the papers, placing the the sensor on the waist can be an ideal position. In Figure 1.6 the possible sensor locations can be seen.

1.2.4.4 Number of Sensors

Using data from multiple sensors at the same time, for instance, an Accelerometer with Gyroscope or Magnetic field sensor provides an absolute orientation in space. However, using multiple sensors can challenge the battery capacity of the electronic device. For example, real-time HAR systems require continuous sensing. In order to reduce battery usage, activity recognition with high accuracy can be done using a single triaxial accelerometer sensor in a smartphone. [LZYG13]

1.3 Accelerometer

As we mentioned earlier, today, accelerometers are among the most used sensors that are capable of identifying daily activities.

We examined studies, where the accelerometer was used in combination with another sensor achieving excellent results at HAR. [Kha11] There were also studies, in which the accelemetor was the only sensor used for HAR. Those studies confirmed that it is sufficient to use only accelerometer gathered data to achieve a accurate classification results.[HT16] [Kha11] Hence, for the purpose of the thesis the used data will come from an accelerometer sensor embedded in a smartphone.[STJ14] We can find it in almost every smartphone or other wearable smart devices (e.g., smartwatch).

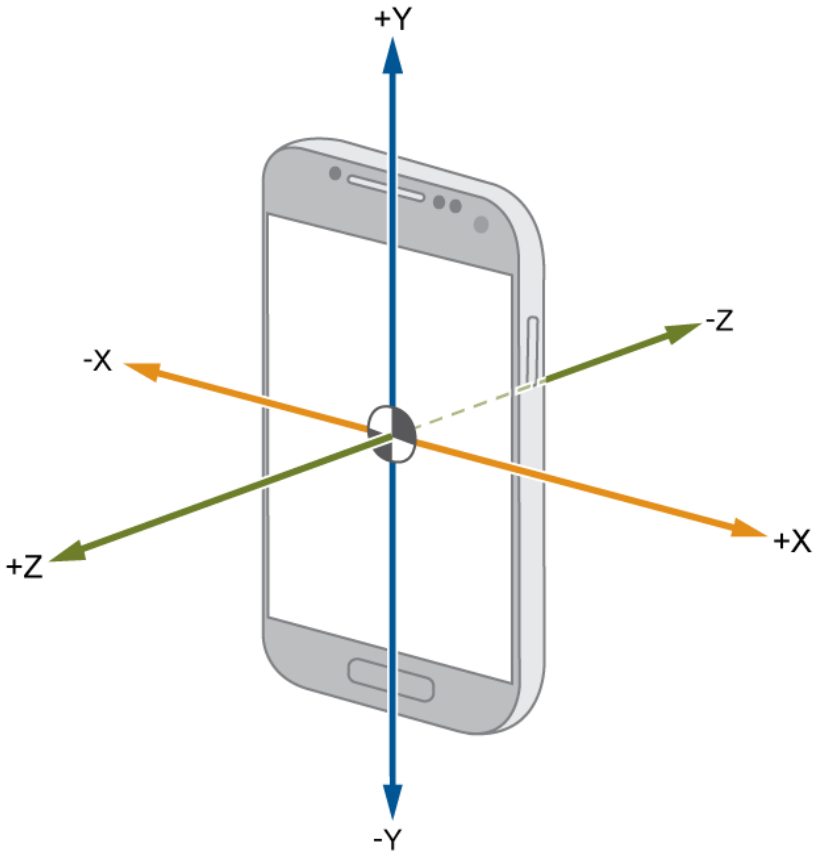

Figure 1.7: Illustration - three-axis of accelerometer embedded in a smartphone [Mat]

Next, we detail how this electronic device works. An accelerometer is an electromechanical device that meaures static and dynamic acceleration

forces. When it is embedded in a smartphone or other wearable devices, it measures the acceleration of the object, shown in Figure 1.7. Triaxial accelerometers are perhaps the most broadly used sensors, which extract 3-axis data (X, Y, and Z). These axes relative to the device remain constant. The phone's lateral movement is described by the X-axis and the perpendicular one by the Y-axis. The Z-axis represents the movement in and out of the plane defined by the X and Y axes. For example, if we put the device face-up on the floor, the Z-axis measures the acceleration of Earth gravity. It will output 9.81 in m/s2. The X and Y axes are vertical to the acceleration of Earth gravity. They will both output 0.00 in m/s2. [Mat]

1.3.1 Description of Performed Activities

In our thesis, the utilised dataset contains six activities: walking, sitting, lying, standing, walking upstairs and walking downstairs. Most people in daily life perform them. Thus, it is essential to differ the acceleration pattern and angular velocity for these activities.

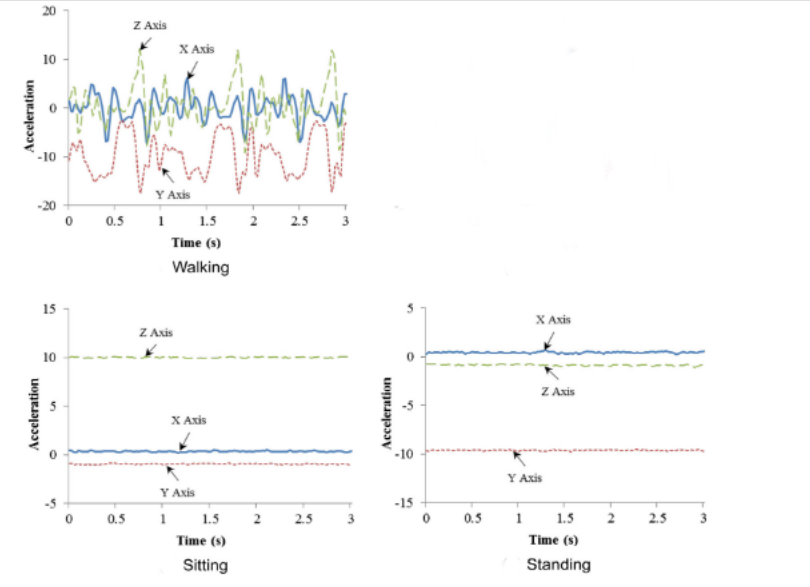

Figure 1.8: Acceleration of 3 activities [KKB14]

Figure 1.8 describes the acceleration of all three axes. Walking X, Y, and Z values vary significantly, because of the dynamic nature of the movement, whereas, values for static activities, like sitting, lying and standing, are almost constant. An interesting fact: the Y values for walking and standing have the lowest accelerations, and for sitting and lying, the Z values have the most significant accelerations. The reason behind this is that the force of gravity plays its role in influencing the entire acceleration in the direction of the centre of the Earth. The direction of gravity mainly corresponds to the Y axis for walking and standing. On the other hand, for sitting and lying, it mainly corresponds to the Z axis. [KKB14]

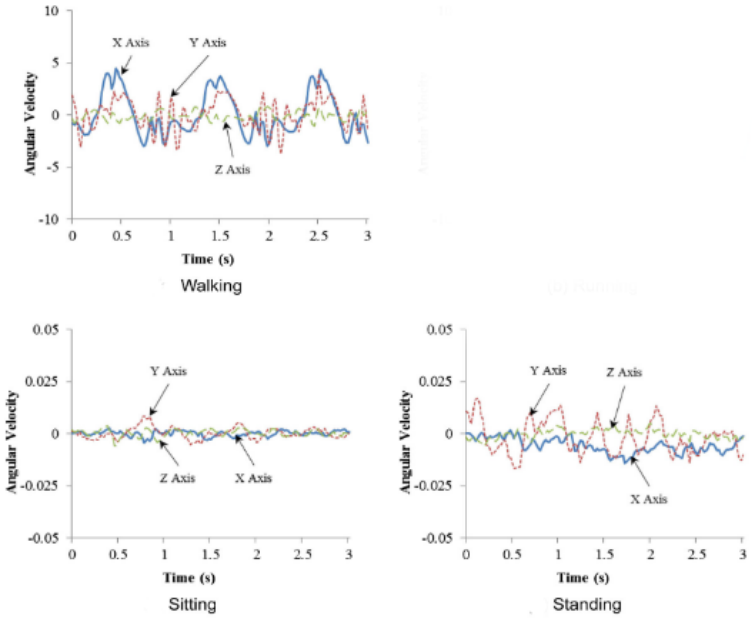

Figure 1.9: Angular velocity signals of 3 performed activities [KKB14]

Figure 1.9 describes the acceleration of all three axes. Each axis value markedly changes for walking. Differently, for sitting, standing and lying, the axis values are minor. However, values for standing change significantly, because it is a relatively unstable activity compared to the other static activities. As the angular velocity pattern show similarities within all the static activities and their magnitudes are small, it is not suitable for distinguishing them.[KKB14]

Figure 1.10 shows, how the triaxial acceleration signals look during the two performed activities. Red colour reflects the X-axis, blue the Y-axis and Z-axis. We can see, that walking and walking upstairs differ in the X-axis pattern.

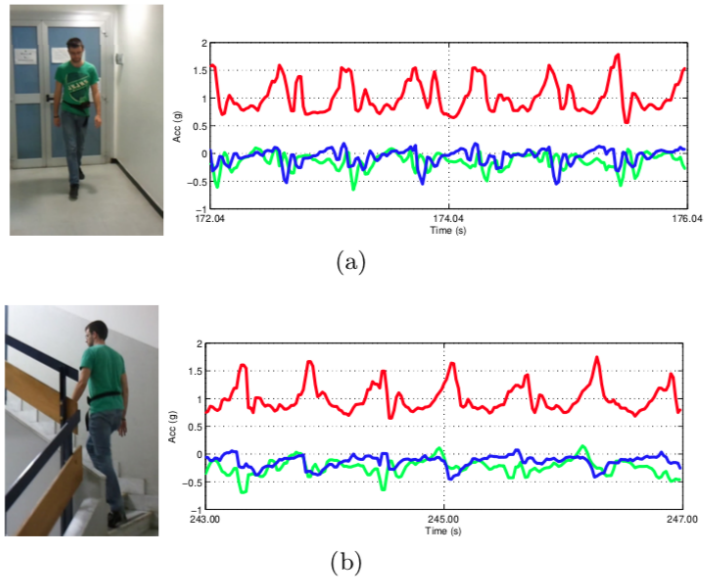

Figure 1.10: Acceleration of 3 activities [KKB14]

X	Y	Z	activity
0.2571	-0.0232	0.0146	standing
0.3424	-0.0419	-0.1229	walking
0.2820	-0.0579	-0.1198	sitting

Table 1.1: A simplified sample of the 3 axial accelerometer generated data

1.4 Data Mining

Data mining is a process involving the collection and selection of data, the pre-processing of data, data analysis, modelling including the visualisation

of results, interpretation of findings, and the application of knowledge.

Data mining involves tasks of descriptive and predictive nature. The descriptive ones generate a knowledge base in the form of models that identify patterns or relationships and correlations in data. Descriptive data mining tasks are often exploratory. The objective of the predictive data mining task is to predict the value of a particular attribute based on the values of other attributes. The predicted attribute is known as target (dependent variable), while the attributes used for making the predictions known as independent variables.

To be able to prove or disprove the given hypothesis, we build models for classification. It is considered as a predictive data mining task. To realise a predictive data mining task for HAR, we will apply data modelling techniques based on Machine Learning algorithms.

1.4.1 Introduction to Machine Learning

With the rapid growth of information technologies, a massive amount of data is generated from different sensors, devices or IoT technologies. The collected data provides a rich user context. The current computational power of machines allows generating that knowledge from the massive amount of data. Here Machine Learning (ML) comes into play. We can think about ML as a set of algorithms that finds patterns and relationships in the dataset. In this thesis, the dataset is a collection of sensor readings provided by an accelerometer, and the generated knowledge comes from the relationships between sensor readings.

ML is a subfield of Artificial Intelligence. It has significant applications in medicine, economics, finance, natural and technical sciences, ecology and many others. Methods like data analysis, data mining, text classification and text mining, recognition of speech, handwriting, images, etc., are all using ML. [cit07]

Tom Mitchell defines ML as follows:

Definition 1.2. "A computer program is said to learn from experience E with

respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E." [Ger17]

A shorter definition of machine learning is written by Kelleher et al. (2015):

Definition 1.3. "It's an automated process that extracts patterns from the data". [Kel15]

The ML algorithm infers the properties on a given dataset. That information makes it possible to identify or make predictions from a dataset that the algorithm has never seen before. It is achievable because almost all nonrandom data contains patterns which allow a machine to generalise them. For instance, a machine learning algorithm can classify future data points into walking and standing groups after being trained on a collection of sample accelerometer data marked as walking or standing.

ML can be divided into three groups depending on the type of learning:

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

In the theoretical part, we are going to introduce the Supervised and Unsupervised learning. Here, reinforcement learning is not detailed. In the implementation part, we will apply only Supervised Learning techniques.

1.4.2 Basic Concepts

Before we go deeper into supervised and unsupervised learning, let's define some terminology.

1.4.2.1 Instance (feature vector), Attributes (features)

We have a set of activities defined in Table 1.1. Each row represents an instance, the table represents a set of instances. Instance, also called a feature vector holds the information about each activity. It is characterized by a predetermined set of features or attributes (table's columns). An attribute or feature is a property or characteristic of an instance.

1.4.2.2 Nominal and Numerical Attributes (features)

Attributes can have numerical or nominal values. Numerical attributes can take integer or real numbers. Nominal, also, called categorical attributes, can take a collection of information that is divided into a group. In the table 1.1 X,Y,Z are numerical attributes, and activity is a nominal (categorical) attribute. In this thesis, we will work with numerical features and one categorical feature, the activity, which describes the target.

1.4.2.3 Learning Algorithm, Execution Algorithm, the Classification Model

The learning algorithm generates new knowledge from the set of input data. The execution algorithm uses the generated knowledge for solving new problems. The generated knowledge is known as the model. Since the HAR models describe categorical labels (daily activities), it can be called a classification model or classifier.

1.4.3 Supervised Learning

Supervised learning algorithms work with a labelled training dataset, including the desired output. The learning algorithm constructs the model representing the input, output and function parameter relationships. With other words, we try to find out how the labelled variable Y is affected by the features $x_0, x_1, ..., x_{p-1}$. Hence, we are looking for a functional relationship $Y \approx f(x_0, x_1, ..., x_{p-1})$.

1.4.3.1 Training and Testing Dataset

It is common practice to divide the dataset into two sets. The training set serves to build the model, and the testing set helps to check the success of the model. The success of the model is measured by testing the generated model with an independent set of data; in our case, it is the testing dataset. During this process, the correct classifications are known but are hidden to the classifier. The model's accuracy is calculated as the number correctly classified cases divided by the total number of cases.

1.4.3.2 Classification and Regression

Supervised learning methods can be divided into classification and regression problems. If the attribute to be predicted is nominal, it is called classification. In the other case, it is called regression. Human activity recognition is considered a classification problem.

1.4.3.3 Supervised Learning Algorithms

There are several algorithms, which can be a suitable candidate to solve the classification problem:

- Decision Tree Classifier
- Random Forest Classifier
- Logistic regression
- Support Vector Classifier
- Neural networks
- Bayesian networks

According to different surveys, supervised learning is suitable to solve the classification problems in the HAR context:

- Another example was presented by Bourobou et al. [BS] using artificial neural networks and K-pattern clustering to identify and predict user activities in smart environments.
- In the survey by Lara et [LL13] human activity recognition activity is mainly carried out with the support of ML techniques, k- Nearest Neighbor, and Decision Tree.
- Chawla and Wagner argued that due to the high performance of classifier algorithms Decision Tree, Support Vector Machine, Artificial Neural Network, K-Nearest Neighbour could be used for real-time human activity recognition. [OL18]
- Fleury et al. [FVN10] proposed a smart home system on health care using the SVM algorithm. The system classifies daily living activities based on the data from different sensors.

Figure 1.11 demonstrates different Supervised Learning algorithms, including their pros and cons and possible application. It helps us to justify the choice to select the right algorithm for the right problems to be able to prove the hypothesis.

Category	Type	Algorithms	Pros	Cons	Applicability in SBs
Supervised Learning	Classification	Neural networks	Requires little statistical training; Can detect complex non-linear relationships	computational burden; Prone to Overfitting; Picking the correct topology is difficult; Training can take a long time and a lot of data	Used for classification, control and automated home appliances, next step/action prediction.
		SVM	Can avoid overfitting using the regularization; expert knowledge using appropriate kernels	Computationally expensive; Slow; Choice of kernel models and parameters sensitive to overfitting	Classification and regression problems in SBs such as activity recognitions, human tracking, energy efficiency services
		Bayesian networks	Very simple representation does not allow for rich hypotheses	You should train a large training set to use it well.	Energy management system and human activity recognition.
		Decision trees	Non-parametric algorithm that is easy to interpret and explain.	Can easily overfit	Patient monitoring, healthcare services, awareness and notification services.
		Hidden Markov	Flexible generalization of sequence profiles; Can handle variations in record structure	Requires training using annotated data; Many unstructured parameters	Daily living activities recognition classification
		Deep Learning	Enables learning of features rather than hand tuning; Reduce the need for feature engineering	Requires a very large amount of labeled data, computationally really expensive, and extremely hard to tune.	modeling occupant's behavior, and in human voice recognition and monitoring systems; Context-aware SB services.
	Regression	Orthogonal matching pursuit	Fast	Can go seriously wrong if there are severe outliers or influential cases	For regression problem such as energy efficiency services in SBs.
		clustered-based	Straightforward to understand and explain, and can be regularized to avoid overfitting.	It is not flexible enough to capture complex patterns	Gesture recognition.
	Ensemble methods	N/A	Increased model accuracy through averaging as the number of models increases.	Difficulties in interpreting decisions; Large computational requirements.	Human activity recognition and Energy efficiency services.
	Time series	N/A	Can model temporal relationships; Applicable to settings where traditional between-subject designs are impossible or difficult to implement	Model identification is difficult; Traditional measures may be inappropriate for TS designs; Generalizability cannot be inferred from a single study.	Occupant comfort services and energy efficiency services in SBs.

Figure 1.11: Supervised learning algorithms with its pros, cons and application context

Based on our study, five ML algorithms were chosen suitable for the classification problem, which will be used to train the models in the implementation part. Next, these algorithms are described from the high-level perspective.

1.4.3.4 Decision Tree Classifier (DTC)

A decision tree is an arrangement of data located in a tree structure. It represents several possible decision paths and an outcome for each path. Depending on whether the output produces categorical or numerical values, classification and regression trees are distinguished. In the following part, the classification trees are discussed. The classification decision tree consists of decision nodes and leaf nodes, which return the classification result.

Entropy and the information are two essential components, which help to build the decision tree. There is a set S of data. Each data is labelled with one of a finite number of classes $c_1, ..., c_n$. Entropy measures how the set of attributes are ordered. If most of the data points belong to a single class, the level of uncertainty is low, which means the entropy is low. On the other hand, there would be a higher entropy. Information gain is the second principal component. It helps to select which feature to choose to reduce the uncertainty of the set. Information gain is defined by entropy minus the weighted sum of entropies. The tree for learning from a set of labelled data uses the ID3 algorithm, which operates in the following manner: Let us be given some labelled data, and a list of attributes $F_1, ..., F_n$ to consider branching on. The algorithm uses the information gain to split the set, starting first building the root of the tree with the attribute, by the highest information gain. After, it goes all the way down splitting the data up into different groups based on the information gain of attributes. The algorithm finishes in 2 cases: or it runs out of attributes or the data at the node have the same class of our interest. A constructed tree can classify new entries.

The main benefit of the decision trees is that they do not require data preprocessing and normalization. It has however a weakness. It is inclined to overfitting a model when the depth of the tree is not limited to a particular level.

1.4.3.5 Random Forest Classifier (RFC)

Random Forest Classifier is an ensemble learning method that combines several weak learners in order to produce a more robust classification model.

It is created by combining the output of different decision trees that have already been trained using the bootstrap method (bagging process). The general idea behind the bootstrap that we first create from the input training dataset D, n different datasets $D_1, ..., D_n$. They have the same size as D dataset. We use repetitive selection. We will teach the decision tree on a Di dataset usually with the depth two or three. Let us denote these trees $T_1, ..., T_n$. Through each data point, we run all the trees $T_1, ..., T_n$ and from each of them we save the decision (output). The final decision to determine the output depends on the sum of the output tested by all the trees.

Random forests can be used for both classification and regression problems. One of the advantages is that random forests are fast to train but are relatively slower when making predictions. The main disadvantage of the random forests is that a large number of trees makes the model slow.

1.4.3.6 Multinomial Logistic Regression (MLR)

Multinomial logistic regression is used for predicting nominal dependent variables based on using the linear combination of multiple independent variables. It allows us to predict more than two categories of the dependent variable. The algorithm uses the maximum likelihood estimation to evaluate the probability of categorical membership.

1.4.3.7 Support Vector Classifier (SVC)

Support Vector Classifier or SVC is a ML algorithm whereby a model categorises data around a hyperplane. Intuitively, a hyperplane can be a “line” that separates and classifies a set of data. If the data point lies further from the hyperplane, the confidence of the correctly classified data point is rising. Support Vector Classifier is used for both classification problems.[KDN]

1.4.3.8 K-nearest Neighbors (KNN)

According to the literature, k-nearest neighbour is one of the easiest supervised learning algorithms. [K.17] Let us be given a scatter plot and distance function. The distance function allows us to compute the distance between two random points on the plot. To be able to train the model, it is required to have already classified data. If there is a new data point to classify, the algorithm looks at k nearest points on a scatter plot and decides which class it belongs to. It is important to choose an appropriate k value. For this purpose is recommended to run tests for different k values. It needs to be small enough not to gather insignificant neighbours, but large enough to provide enough data points for a valid sample to be taken. KNN is a simple concept, but it can be extremely powerful. Despite its simplicity, it has its drawbacks: the curse of dimensionality. [K.17]

1.4.4 Unsupervised Learning

HAR has proceeded through the use of supervised learning techniques in recent decades. However, there are some cases when unsupervised learning method could be utilised. Unsupervised learning algorithms do not work with a labelled training dataset. Compared with the supervised learning, there is no training set with labels compared to the instances. [Ger17]

From the unsupervised learning types, clustering is the most popular. Here, the algorithm analyses the similarities between the input samples and classifies them into different clusters. [cit07] In the HAR context, unsupervised learning algorithms can be used to recognise various activities, when it's challenging to have labels (output) for input data, for instance:

- The paper [AGO+13] presents an unsupervised learning method approach to recognise the number of perfomed activities.

In the practical part, we work with a labelled dataset. The Unsupervised

Learning technique for building a HAR will not be used.

1.5 Model Evaluation Metrics

1.5.1 Performance Measures

Performance measure includes the last steps in a ML project. When we train the model, there are different methods to learn performance measures, which evaluate how effectively a particular classifier operates. They help us to evaluate the correctness, efficiency and usefulness of the design and the modelling process. The following two methods are going to be used to measure the performance of the model: Classification accuracy and Confusion matrix.

1.5.1.1 Classification Accuracy

Classification accuracy compares the ratio of the number of correct predictions to the total number of input samples:

$$Accuracy = \frac{Number of Correct Predictions}{Total number of Predictions}$$

The main disadvantage of the method is that it is working well only when we have the same or almost the same number of samples belonging to each class. So, the misclassification of minor class samples is high. For instance, consider 80% of samples from class A and 20% from a different class B. We can easily reach high accuracy by simply predicting all samples to be of the class A. [MIS]

1.5.1.2 Confusion Matrix (CM)

Confusion matrix, as the name says, returns the output in the form of a matrix, which describes the complete performance of the classifier. The main idea is to count the number of times instances of class A classified as class B. In a confusion matrix, each row describes an actual class, while each column represents a predicted class. [Ger17] For instance, to know how many times the classifier confused STANDING with SITTING, we take a look at the STANDING row and SITTING column.

	Predicted: SITTING	Predicted: STANDING
Actual: SITTING	34 (TP)	10 (FN)
Actual: STANDING	6 (FP)	56 (TN)

Table 1.2: Example of a Confusion Matrix

Let us assume we have samples belonging to two classes: SITTING and STANDING. Furthermore, we have a model which classifies the data point to one of the classes. An activity, for instance, SITTING can be classified SITTING or STANDING. To understand from the matrix how the model has performed, we define four terms:

- True Positives (TP): The case in which we predict SITTING and the actual value is also SITTING. It is the number of positive records correctly predicted as positive by the model.
- True Negatives (TN): The case in which we predict STANDING and the actual value is STANDING. It is the number of negative records correctly predicted as negative by the model.
- False Positives (FP): The case in which we predict SITTING and the actual value is STANDING. It is the number of negative records incorrectly predicted as positive by the model.
- False Negatives (FN): The case in which we predict STANDING and the actual value is SITTING. It is the number of positive records incorrectly predicted as negative by the model.

Accuracy is meant as the total number of correct predictions proposed by the model which includes the positive and negative predictions.
It is calculated as:
$Accuracy = \frac{TruePositives+TrueNegatives}{TotalNumberOfSamples}$
[MIS]

Summary

Chapter 1 pointed out that thanks to today's technological development, HAR has become one of the essential studies with an increasing number of practical applications.

Thanks to the theory, we analysed the opportunities for creating a human activity recognition classifier including supervised Machine Learning algoirthms. Based on that we selected five algorithms to confirm the hypothesis of the thesis.

Part II

Practical part

Chapter 2

In the previous part, we got familiar with the theoretical background of the thesis. Chapter 2 describes the methodology for building human activity recognition classifiers, which will help us to evaluate the hypothesis set in the Introduction.

The methodology is a sequence of events, described as separate sections within this chapter. It begins with the description of the dataset. According to this dataset, the HAR system will be designed. Further, we perform an exploratory data analysis, where the utilised data is inspected and visualised. Then comes the preprocessing step, where we prepare the dataset for the classification algorithms. The last step includes modelling. In it, we will demonstrate how classification algorithms are used to build models from the preprocessed data.
The evaluation of the constructed models will be discussed in Chapter 3.

2.1 The Utilised Dataset

In this section, the dataset used in this work is detailed. We begin with the basic description of the dataset. Next, we shortly introduce how the data preprocessing and feature extraction were done.

2.1.1 Data Source

A dataset called "A Public Domain Dataset or Human Activity Recognition Using Smartphones" is used to accomplish the objective of the thesis. The experiment has been carried out with a group of 30 volunteers between the age of 19-48 years. Each person followed a protocol to perform six activities (walking, walking upstairs, walking downstairs, sitting, standing and laying) wearing a smartphone on the waist.

During the experiment, the data was simultaneously gathered from 2 sensors, an accelerometer and a gyroscope embedded in a smartphone. There were captured 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz. Although the dataset contains data both from accelerometer and gyroscope gathered data, for the purpose of this thesis, we'll use only data coming from the accelerometer. [DARO13]

2.1.2 Data Preprocessing and Feature Extraction

As a first step, the collected signals went through noise reduction with a median filter and a 3rd order low-pass Butterworth filter with a 20 Hz cutoff frequency. Next, the time signals were sampled in fixed sliding windows of 2.56 second with 50%.(128 readings/window) After was used another Butterworth low-pass filter was used to separate gravitational and body motion components. Then, from each window, a vector of features was calculated by the time and frequency domain variables. In the end, the authors split the dataset into training and testing sets with the distribution

of 70:30. [DARO13]

Raw Signal	Definition
tBodyAcc-XYZ	Body acceleration in time
tGravityAcc-XYZ	Gravity acceleration in time
tBodyAccJerk-XYZ	Jerk in body acceleration in time
tBodyAccMag	Magnitude of body acceleration in time
tGravityAccMag	Magnitude of gravity acceleration in time
tBodyAccJerkMag	Magnitude of jerk in body acceleration in time
fBodyAcc-XYZ	Body acceleration in frequency
fBodyAccJerk-XYZ	Jerk in body acceleration in frequency
fBodyAccMag	Magnitude of body acceleration in frequency
fBodyAccJerkMag	Magnitude of jerk in body acceleration in frequency

Table 2.1: Description of raw signals from HAR experiment [DARO13]

The 'XYZ' denotes the three-axis directions X, Y, Z for each of the tri-axial signals; t indicates time-domain variables; f denotes frequency domain variables. [DARO13]

Descriptive	Definition
mean()	Average value
std()	Standard deviation
mad()	Median absolute deviation
max()	Maximum value
min()	Minimum value
sma()	Signal magnitude area
energy()	Energy value
iqr()	Interquartile range
entropy()	Signal entropy value
arCoeff()	Autoregression coefficient
correlation()	Correlation coefficient
maxInds()	Index of the largest magnitude frequency component
meanFreq()	Weighted average of the frequency component
skewness()	Skewness of the frequency domain signal
kurtosis()	Kurtosis of the frequency domain signal
bandsEnergy()	Energy of the frequency within the FFT of each window
angle()	The angle between the vectors

Table 2.2: Description of derived variables from raw signals [DARO13]

Feature extraction helps to achieve a better quality of the dataset which raises the performance of the classifier models. During this process, new features have been derived from the initial obtained tri-axial signals from the accelerometer.

First, the total acceleration signals were split into tBodyAcc-XYZ and tGravityAcc-XYZ.

Next, the Jerk signals tBodyAccJerk-XYZ were derived from the raw signals. Jerk signals reflect the rate of change in acceleration over time.

Then, Euclidean norm is used to calculate the magnitude of each of the signals. It results in features as tBodyAccMag, tGravityAccMag, tBodyAccJerkMag.

Additionally, a Fast Fourier Transform was applied to produce the features as fBodyAcc-XYZ, fBodyAccJerk-XYZ, fBodyAccJerkMag.[DARO13]

2.2 Design of the HAR system based on the dataset

2.2.1 The Problem Setting

The main tasks of the practical part are doing an exploratory data analysis and developing models that are capable of recognising multiple daily activities. After evaluating the success of the developed models, the hypothesis will be proven or disproven.

In order to build a generalisable activity recognition classifier, a subject-independent (impersonal) approach is selected. It means, that we try to create a model based on the training dataset that generalises the diversity between users. Thanks to this ability, the model is able to classify the predefined six daily activities of a new user. During the implementation, five supervised Machine Learning techniques are applied on the dataset to solve the classification problem. Our utilised dataset went through data-prepocessing and feature extraction. Therefore, they're partially prepared for the implementation. The realisation of the practical part will be detailed in the following sections.

2.2.2 Set of Activities

Activity name	ID	Type
WALKING	1	dynamic
WALKING_UPSTAIRS	2	dynamic
WALKING_DOWNSTAIRS	3	dynamic
SITTING	4	static
STANDING	5	static
LAYING	6	static

Table 2.3: Table of the perfomed activities

2.2.3 Technology

To handle the practical part of the thesis Python 3.7.6 was used. The experimentation and data visualisation happened in Jupyter notebooks. The written program has dependencies to some standard packages.

To collect and transform data efficiently, we use the Pandas package. It offers the DataFrame structure for storing heterogeneous tabular data and also provides an efficient operations with the data.

Further, we utilise numpy 1.18.1 and scipy 1.4.1 packages to perform statistical computations. For applying ML algorithms, we use implementations provided by the scikit-learn version.

Finally, all the plots presented in this thesis are generated using matplotlib 3.1.3 and seaborn 0.10.0.

The experiment was done on a Macbook Pro 2010 with a processor 2,4 GHz Intel Core 2 Duo and RAM 8 GB 1067 MHz DDR3. We present the technical configuration, as the training and testing time of the models will be given in the appendix.

2.3 Exploratory Data Analysis

Exploratory data analysis is a method to analyse and investigate a set of data with the purpose to summarise its main characteristics, discover patterns or anomalies with the help of statistics and graphical representation. [Pat]

In order to achieve the exploratory data analysis, first, the data is collected and placed into an appropriate format using the Pandas package.

2.3.1 Data Characteristics

The first element of the performed data investigation is the number of the instances and its dimension. The training and testing dataset contains 10299 instances and 561 features.

Here, we discover, that the dataset contains features coming from accelerometer and gyroscope sensors, too. Since we defined data genereted only by the accelerometer, we drop all the features coming from the gyroscope. After the elimination of unneed features, the dataset contains 10299 instances and 351 features. In table 2.4 we can a see 10 features out of 351 used for building the model.

Feature ID	Feature
1	tBodyAcc-mean()-X
2	tBodyAcc-mean()-Y
3	tBodyAcc-mean()-Z
4	tBodyAcc-std()-X
5	tBodyAcc-std()-Y
6	tBodyAcc-std()-Z
7	tBodyAcc-mad()-X
8	tBodyAcc-mad()-Y
9	tBodyAcc-mad()-Z
10	tBodyAcc-max()-X

Table 2.4: Part of the features used for training the model

Now, histograms from the training and testing dataset are created to understand the data distribution of the activity and the subject. In figures 2.1 - 2.4 it is observed, that in the training and testing datasets, each activity is represented by sufficient data points. Although, there are fluctuations in the activity counts, they are almost equally distributed.

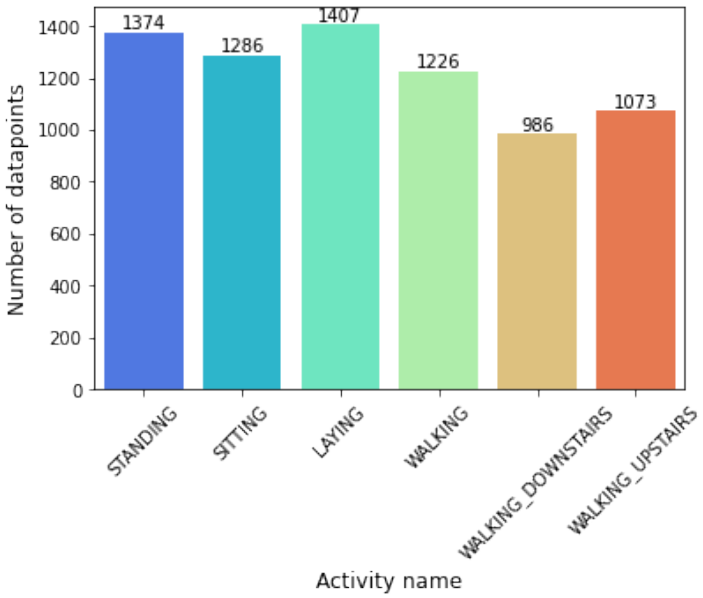

Figure 2.1: Histogram - count of the target variable (training set)

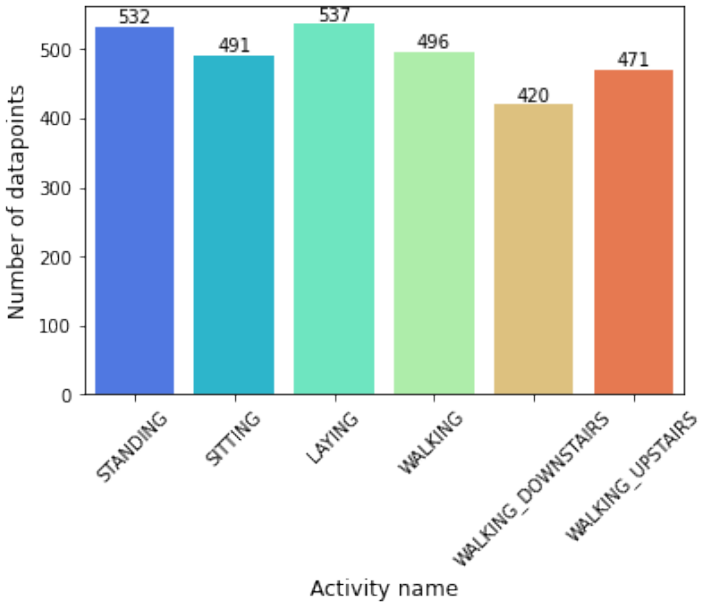

Figure 2.2: Histogram - count of the target variable (testing set)

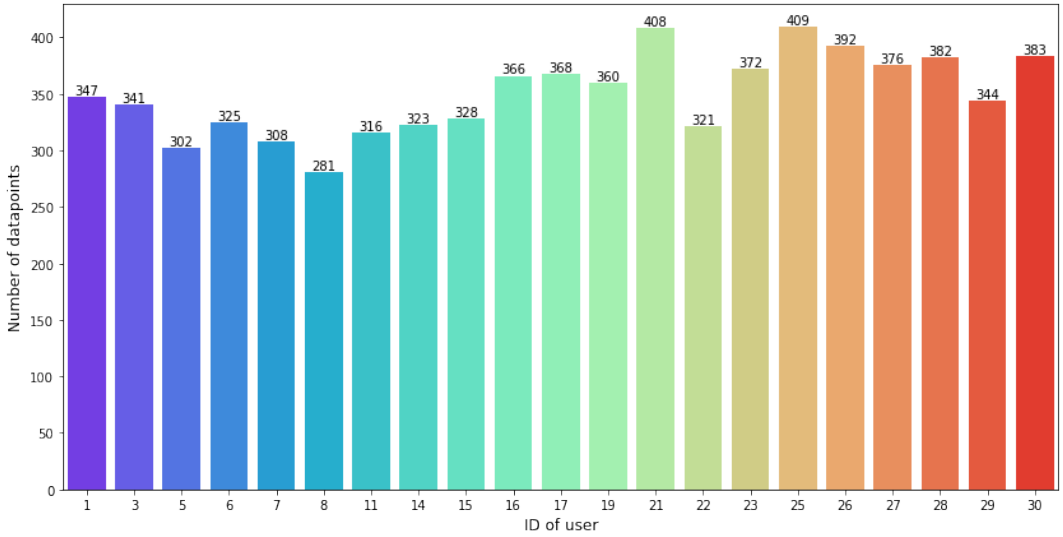

Figure 2.3: Histogram - Activities perfomed by subjects

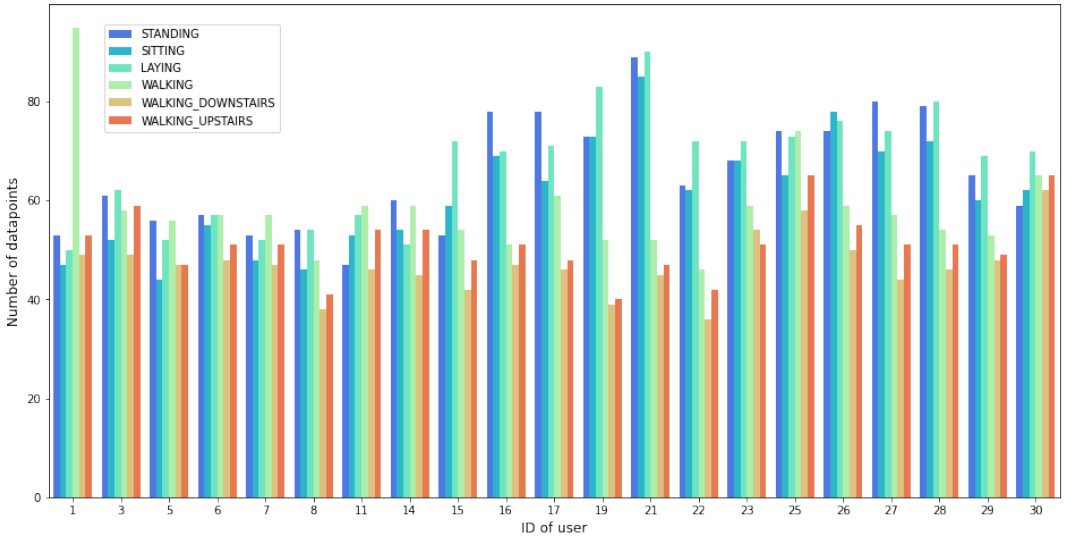

Figure 2.4: Histogram - Particular activities performed by subjects

2.3.2 Pattern Discovery (Activity Exploration)

The dataset is geared towards classifying six activities performed by 30 participants. In the following part, we investigate whether data points corresponding to the activities are separable.

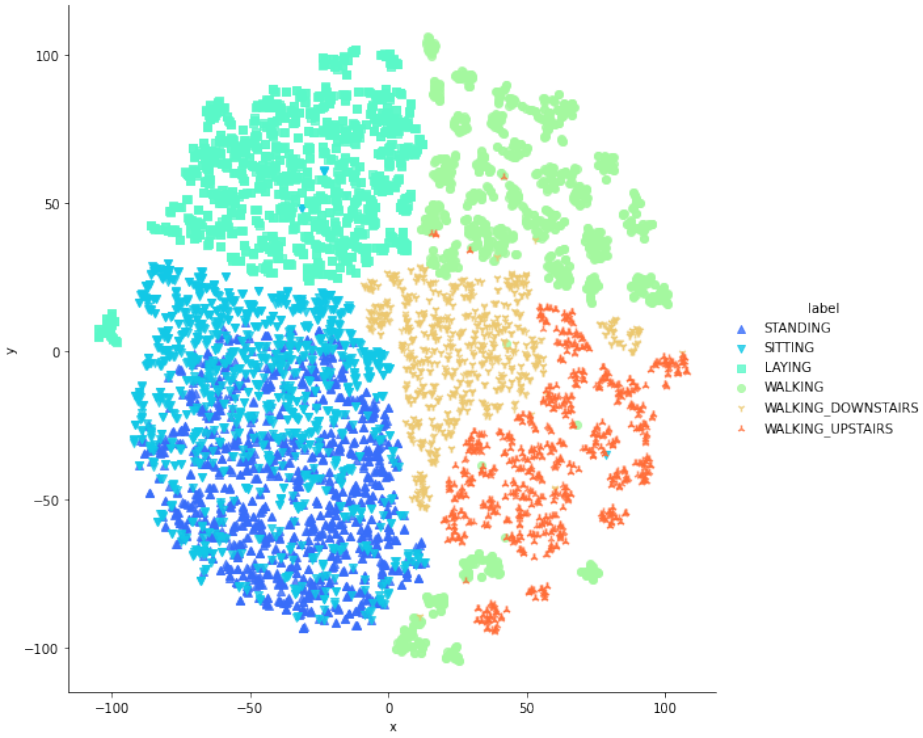

Figure 2.5: Visualisation of particular activities in a two dimensional space

2.3.2.1 t-distributed Stochastic Neighbor Embedding

Since the training and testing datasets have a high-dimension (contain several features), we need to find an appropriate tool to be able to visualise the diversity of the data points in a low-dimensional space. t-distributed Stochastic Neighbor Embedding (t-SNE) is a machine-learning algorithm that

is capable of visualising high-dimensional data. Shortly, in the first step, the algorithm creates a probability distribution that represents the relationships between various neighbouring points. In the second step, it tries to recreate a lower-dimensional space that follows the probability distribution in the best way. [sld]

After reducing the high-dimensional data-points, we visualise the results in the two-dimensional graph, represented in Figure 2.5. We see, that most of the activities represented by data points create clusters. We can claim, that they are mostly separable. Exceptions are the activities STANDING and SITTING. We see that the data points of these activities lie very close to each other, often overlap in the reduced two-dimensional space.

2.3.2.2 Differentiating Static and Dynamic Activities

For differentiating static and dynamic activities we will use the feature BodyAccMag-mean. Probability density function (PDF) helps us to set a condition to separate static and dynamic activities. The rendered graph, located in Figure 2.6 shows that the static activities are situated below the value - 0.5 and the dynamic activities above -0.5.

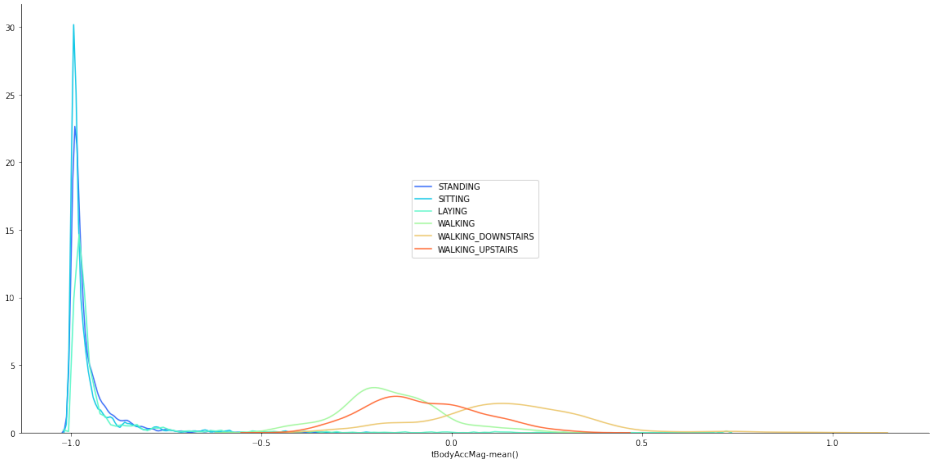

Figure 2.6: Analysis of the tBodyAccMag-mean feature

Boxplots can also be used to differentiate static and dynamic activities using the Body Acceleration Magnitude feature. In Figure 2.7 the rendered graph is located. We see, that static activities are situated below the value -0.8 and the dynamic activities above -0.6.

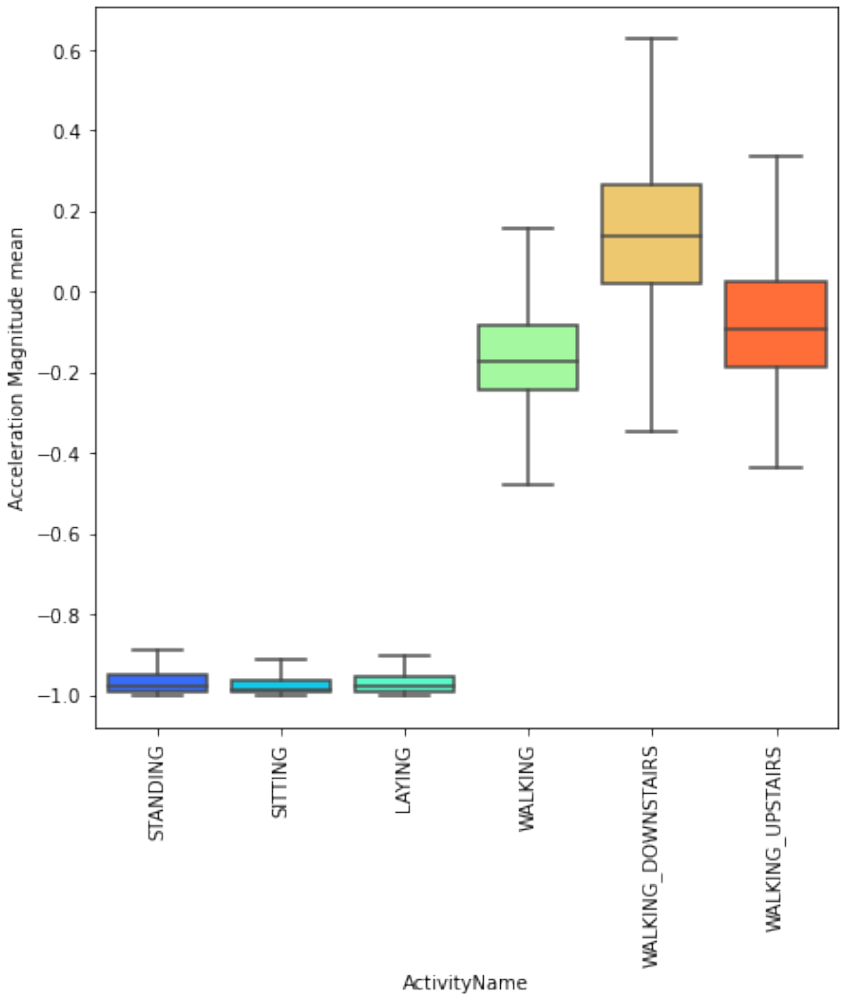

Figure 2.7: Boxplots - tBodyAccMag-mean feature

2.4 Preprocessing Step

2.4.1 Dimension Reduction

The dataset contains 351 features. To work with a high dimensional dataset brings difficulties not just at visualizing, but it can be computationally expensive and complicated at model training, too.

There are different techniques to reduce a large number of attributes. During the implementation process we tried out one of them, Principal Component Analysis (PCA). Looking at the technique on high-level, it takes a higher dimensional data space and transforms the data points to a lower-dimensional space.

After applying PCA in our datasets with variance 0.9, we discover, that 62 principal components should be sufficient to train the model with a similar accuracy level. We train the Logistic Regression model with a reduced feature number. In the implementation, we observe that the required modelling time is decreasing because we work with fewer features. On the other hand, the model's accuracy is not improving because there is some information loss that harms the model's performance. We stop examining PCA on other models.

It follows that in our case, it is better to avoid dimension reduction and work with the original dataset containing 348 features. The reasons are the following: we receive a higher accuracy with the original amount of features and the dataset contains only 10299 data points, which doesn't yet require high computational power.

2.4.2 Modelling

2.4.2.1 Hyperparameters

To train a model requires finding its optimal parameters in order to gain the highest possible accuracy. These parameters are called hyperparameters.

They are important, because they control the behaviour of the training algorithm and have a significant impact on the performance of the model. To set the right parameters for the training process, we use the cross-validation technique combining with the GridSearchCV methodology. Thanks to this technique, multiple values are tried out with the purpose to find out the optimal hyper-parameter space depending on the cross-validation score.

The process consists of 2 main parts:

- Step 1 - Modelling - Here, the model is trained using cross-validation using the traiing dataset (70% of the total records).
- Step 2 - Performance measure - Here, the model is evaluated using the trained model created in Step 1. The model uses the testing dataset, which consists of the remaining 30% of the total records.

The following ML algorithms were implemented:

- Decision tree
- Random forests
- Logistic regression
- Support Vector Classifier
- KNN

Chapter 3

Evaluation

This chapter contains the evaluation of the implementation results obtained the different generated models listed in the second chapter. The outcomes are examined individually, containing the classification accuracy and an analysis confusion matrix. As we mentioned earlier, the models were trained using 70% of the HAR data set. The rest of the data was used for testing the performance of each model.

3.1 Evaluation of the Models

3.1.1 Decision Tree

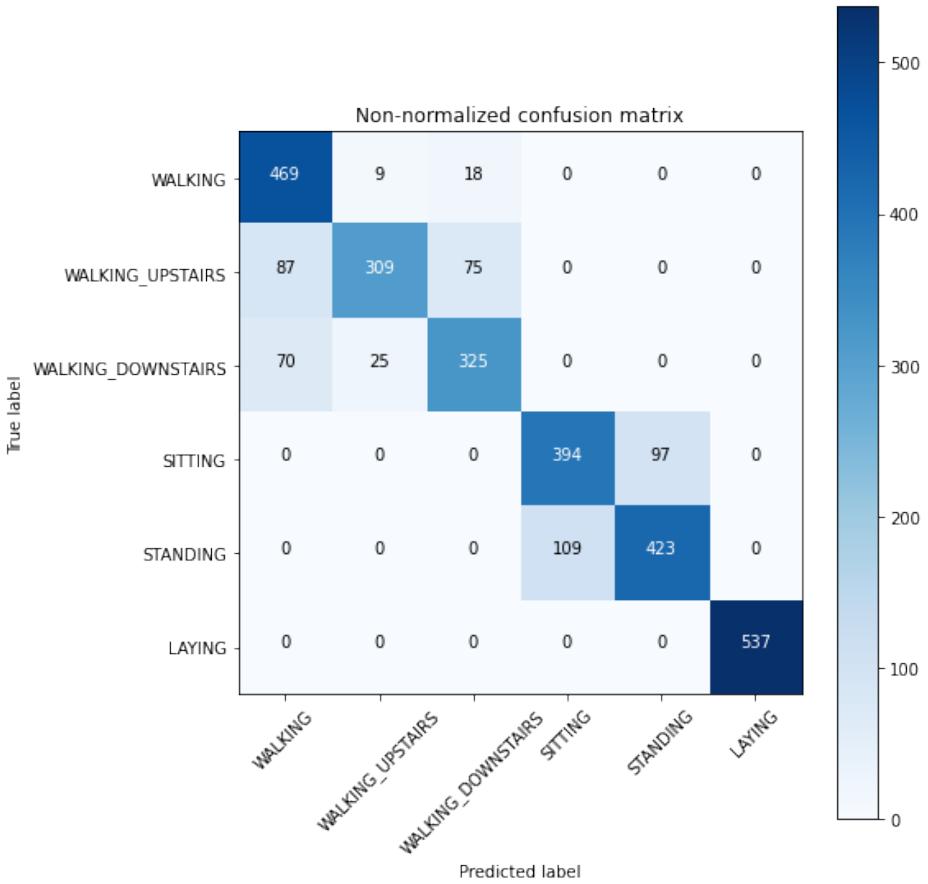

Figure 3.1: Confusion Matrix For Decision tree

The first implemented model is the decision tree. After evaluating the classification accuracy using the testing dataset, we reached 83.33% accuracy. When we look at the results summarised by confusion matrix, it can be observed in Figure 3.1 that the following model suffered the most in

differentiating the following activities: SITTING AND STANDING WALKING_UPSTAIRS and WALKING_DOWNSTAIRS One of the reasons that the decision tree fails at predicting the above-mentioned activities stems from the section design issues in Chapter 1. We mentioned, that some tasks such as sitting and standing or walking upstairs and walking downstairs could be more challenging for ML techniques because their pattern can easily overlap.

3.1.2 Random Forests

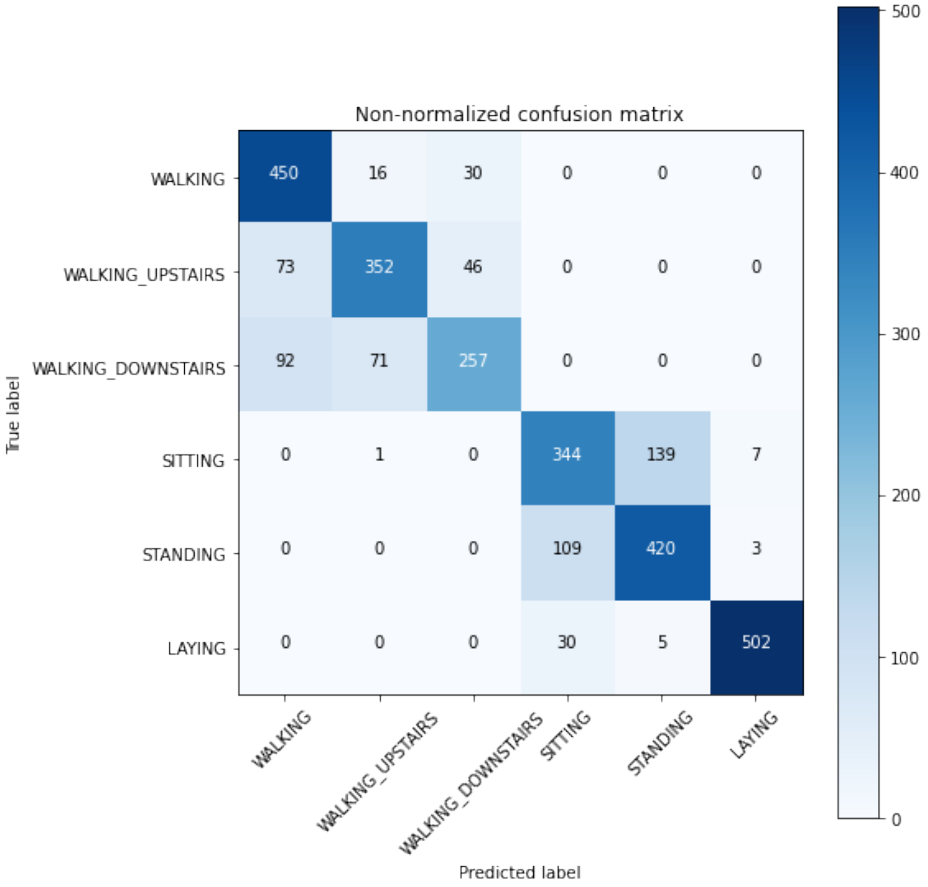

Figure 3.2: Confusion Matrix For Random Forests

The next implemented model is the random forests classifier. As we mentioned in the theoretical part, the algorithm is based on a set of decision trees with depth two or three, which work together to bring a better performance. Comparing random forests classifier to the decision tree, we can claim that this model has brought a higher accuracy. The classification accuracy of the random forests classifier is 88.05%, and thus 5% higher than the result from the model based on decision trees. However, the training time for this model required more time than the decision model. It took approximately 20 minutes to build up the model. When we analyse the confusion matrix, we see above that random forests classifier has brought a better performance in each classified activity. The model is failing at most at activities SITTING AND STANDING and WALKING_UPSTAIRS and WALKING_DOWNSTAIRS.

3.1.3 Multinomial Logistic Regression

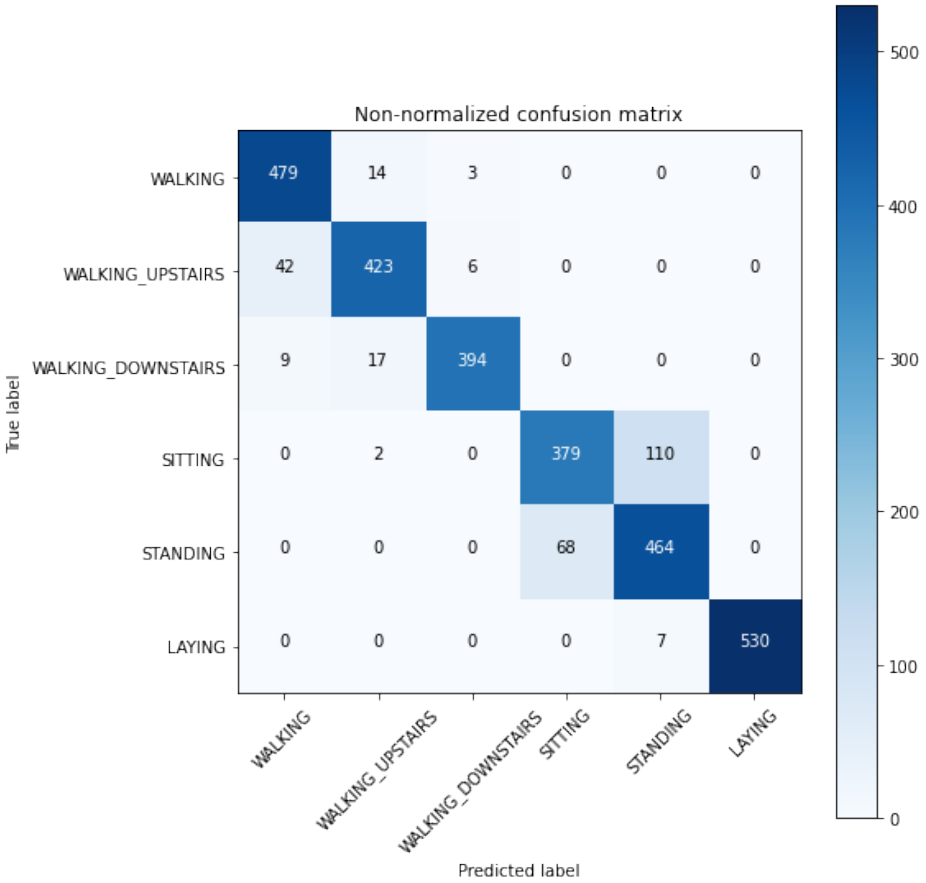

Figure 3.3: Confusion Matrix For Logistic Regression

The multinomial logistic regression model is one of the best performing models. When we compare the logistic regression model to the previously built models, we conclude that the model outperforms the previously obtained accuracy results. The model reached 90.56% in classification accuracy. After examining the confusion matrix, we observed that there is low cluster misclassification by the model but some amount of misclassification occurred for the stationary activities, concretely, for SITTING and STANDING.

3.1.4 SVC

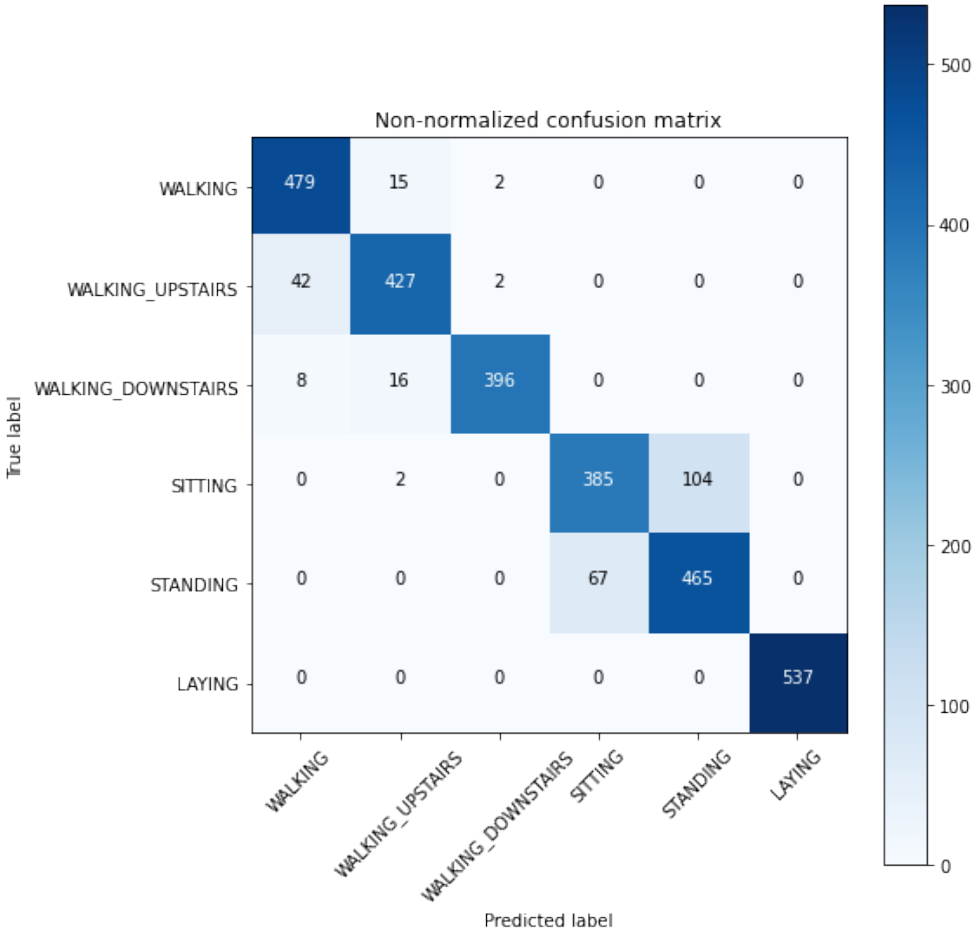

Figure 3.4: Confusion Matrix For SVC

Among the performed models, Support Vector Classifier has provided the best scores. This model has reached 91,24% in classification accuracy in total. However, the training time required approximately 20 minutes, similarly to the Random Forests Classifier. Figure 3.4 presents the confusion matrix of the SVC model. It can be seen that most of the activities have been classified with high accuracy. On top of that, the model classifies laying activity with 100% accuracy. However, standing and sitting activities were

still slightly misclassified.

3.1.5 KNN

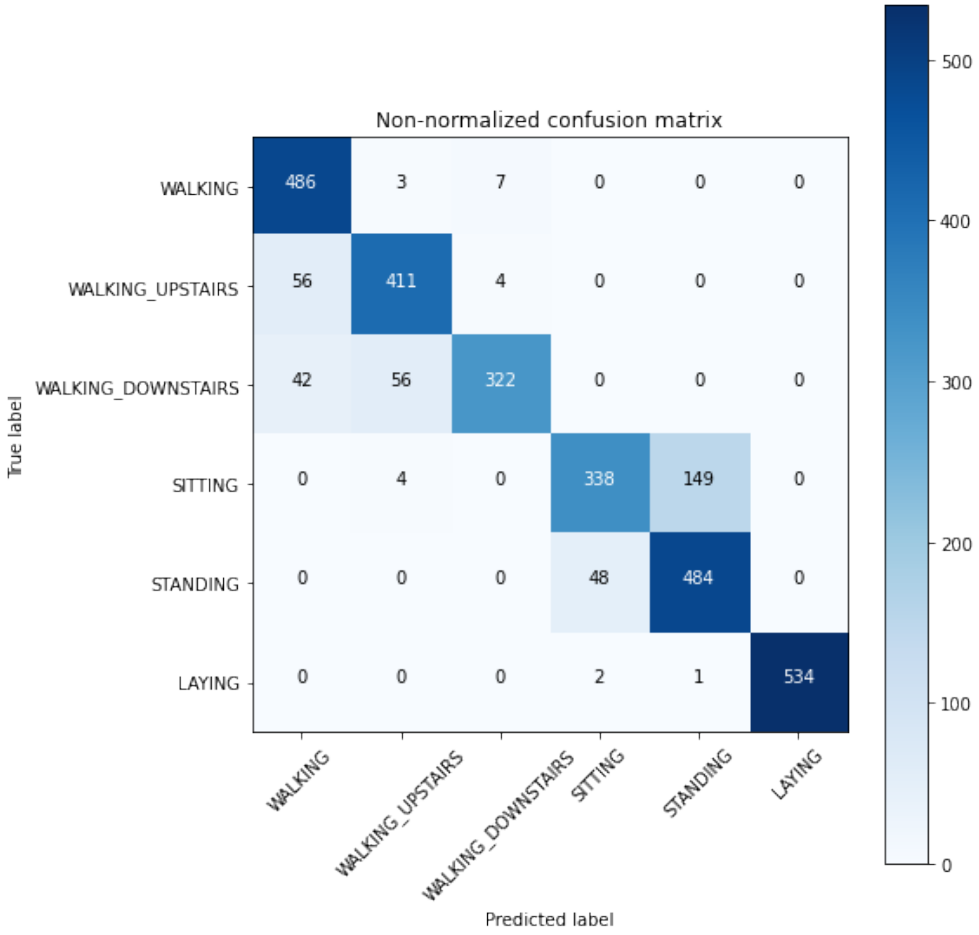

Figure 3.5: Confusion Matrix For KNN

The final individual algorithm was the K nearest neighbors algorithm. KNN reached 87.38% classification accuracy. The algorithm has lower misclassifications levels, mostly failing at static activities SITTING and STANDING.

3.2 Evaluation of the Hypothesis

The hypothesis set in the Introduction is the following:

- H1- At least one of the applied supervised Machine Learning algorithm, modelled on the chosen HAR dataset is capable of classifying a predefined set of daily activities with more than 90% of the accuracy.

The previous sections demonstrated the evaluation of each Machine Learning method, including classification accuracy and confusion matrix. The research hypothesis requires at least one ML method to achieve more than 90 % accuracy. To be able to prove or disprove the hypothesis, we use the results of the classification accuracy for each implemented method tabulated in table 3.1.

Model	Classification accuracy
Decision Tree	83.33%
Random Forests	88.05%
Logistic Regression	90.56%
Support Vector Classifier	91,24%
KNN	87.38%

Table 3.1: Classification accuracy of each implemented algorithm

Table 3.1 shows, that two Machine Learning methods, Logistic Regression and Support Vector Classifier outperform the predefined 90% of accuracy. These 2 models meet the hypothesis statement. Logistic Regression achieved 90.56% and Support Vector Classifier 91.24% of accuracy. Thus, our hypothesis is proven.

3.3 Strengths and Limitations of the Models

3.3.1 Strengths

3.3.1.1 Capable of differentiating static and dynamic activities

According to results obtained from the confusion matrix, we found out that all created models are capable of differentiating static and dynamic activities with negligible number of mistakes.
Decision tree and random forests regression work with 100% of accuracy at differentiating static and dynamic activities.
LR and SVC make 2 mistakes at classifying walking upstairs instead of standing and KNN 4.

3.3.1.2 High Classification Accuracy

One of the strengths of the implementation is the high accuracy performance achieved by LR and SVC models. The evaluation part indicated that those models resulted in a high precision classification of the activity patterns, and it signifies that these obtained results could be further replicated.

3.3.2 Limitations

3.3.2.1 Classification Problems

The first limitation of the implementation part is that most of the models classify certain activity patterns with low accuracy. As we presented

the confusion matrix of each model, we observed that some models failed in differentiating WALKING_DOWNSTAIRS and WALKING_UPSTARIS activities. Furthermore, we observed that every implemented model fails at predicting SITTING and STANDING activities. According to the theory, these activities have similar pattern representation. Thereby, they can overlap easily in the feature space.

3.3.2.2 Number of utilized ML techniques

The second limitation is connected to the number of utilized ML techniques. For the thesis purpose, five ML models were implemented. Two of them fulfilled the hypothesis statement. Therefore, we succeeded in verifying the hypothesis. On the other hand, in the theoretical part, we mentioned other methods, which are also convenient for our classification problem, but we did not implement them. If the obtained results are used in further work (i.e. in an application), we cannot be sure that we would choose an optimal model (the one with the highest accuracy). In order to claim that we have found the optimal model, we should implement every recommended method.

Chapter 4

In the previous chapter, we evaluated the results. We proved that there is a supervised Machine Learning algorithm, modelled on the HAR dataset, that is capable of classifying the predefined six daily activities with more than 90% of accuracy. The beginning of the fourth chapter demonstrates the opportunities and possible applications, where the HAR system could be used. Afterwards, a review of the current study is summarized. It iterates the objective of the research with all the important steps and makes conclusions.

4.1 Application

The implementation part of the thesis has shown, that a well-preprocessed dataset brings us promising results in classification of daily activities even using a single sensor. Now, we will discuss practical applications of the HAR system.

4.1.1 Medical Application

A well-functioning healthcare system is required to maintain the health and prevent sickness of citizens. The motivation behind such a system is to prevent, treat and manage diseases and provide the physical and mental well-being of a person. There exist lifestyle-diseases, which can lead to certain chronic diseases such as diabetes, stroke, high blood cholesterol, hypertension or cardiac failure. [oMUCoQoHCiA01] One of the lifestyle-diseases, which has seen a rapid increase in the Czech Republic over the past decades, is obesity. According to the statistics made in 2018, in the Czech Republic, 47% of men and 33% of women are slightly overweight. Obesity affects almost 20% of men and 18% of women. If we compare the results to previous years, obesity has an increasing tendency among individuals. [CZS] To treat or prevent lifestyle-disease, a combination of a healthy exercise routine and a balanced diet is recommended. In the current circumstances, our health care system needs innovations. The paper [Bra] suggests to not only focus on treating people but also advising and guiding them about how to deal with and prevent chronic medical conditions. One of the possible solutions is telemonitoring, which involves monitoring the patients. A device used for telemonitoring, for instance, a smartwatch not just tracks the patient's mobility, but can also record heart rate and blood pressure. The core of mobility detection is based on the HAR system. It details the amount of time spent in dynamic activities or static activities. The obtained data can help healthcare workers monitor the patient physical condition, make better decisions for recommendation or allocation of a particular treatment.[CNSL]

4.1.2 Personal Lifelog

Another possible application of the HAR system is personal lifelog. Personal lifelog is a set of data containing an individual's daily activities. The data helps to understand the user's life interactions. By mining the logged data, the application can offer a detailed lifestyle summary or report about the sleeping habits. Most of the recognized activities provide users with medically useful information, such as step counts for walking, jogging, going up-stairs/down-stairs, or data on walking distance and duration. [CNSL]

4.2 Limitations of Results for the Real-life Scenario

As we described in the third chapter, in the utilised HAR dataset, each person had to follow a protocol to perform six activities wearing a smartphone on the waist. In a real-life scenario, it could be challenging to achieve similar results with our generated models as actions can draw different patterns in the feature space, and the smart device location also can be different.
Here, we discuss which limitations we need to face at real-life scenario implementation and which suggestions we would have.

4.2.1 Generated Dataset From Real-life Scenario

The HAR dataset is a model dataset. It contains data coming from performing six activities under laboratory conditions. However, training models which use such dataset for practice are not sufficient. Models should be built upon a datasets, where data is also collected from a real-life scenario.

4.2.2 Flexible Sensor Location

The generated dataset used for the thesis' purpose was acquired from a mobile phone's accelerometer located in the waist. The sensor location can vary from user to user. Hence, the location of the sensor embedded in a smartphone or other smart device needs to be more flexible for a real-life scenario.

4.2.2.1 Multiple sensors, better accuracy

In the implementation, only the accelerometer's generated data and their derivation were used for building classifying models. As shown in work [KDN], accelerometer combined with gyroscope can bring us more accurate results. Therefore, it's another opportunity to use data from 2 sensors to carry out more reliable models for a real-life application.

4.2.2.2 Real-time Data Processing

In this thesis, we showed that a dataset generated from the accelerometer can be used for creating a HAR system. For this purpose, using a static dataset was sufficient. However, for a real-life scenario, we need to process the data in real-time, which requires us to create a more complex system.

Conclusion

The aim of this thesis was to confirm or disprove the stated hypothesis through an experiment, which assumes, that at least one of the applied supervised Machine Learning algorithms, modelled on the accelerometer based HAR dataset is capable of classifying a predefined set of daily activities with more than 90 percent of accuracy. The work focuses on the study of Machine Learning techniques in the HAR research area. We picked 5 supervised Machine Learning algorithms for our experiment. In the implementation part we focused on exploratory data analysis, modelling and evaluating the learning methods. In the exploratory data analysis we supposed, that the selected dataset is most likely suitable for modelling. We observed, that each activity is represented by sufficient data points and that they are almost equally distributed. Next, using t-SNE technique on the dataset, we noticed, that in the two-dimensional space most of the activities are separable from each other. Additionally, visualisation of the feature representing the body acceleration magnitude showed us clearly, that the static and dynamic activities are differentiable. We built our models from the training dataset using k-fold cross-validation technique for hyperparameters tuning. After evaluating the success of these models, we noticed, that 2 of the 5 models, Logistic Regression and Support Vector Classifier at classifying outperform the set accuracy at classifying. Logistic Regression achieved 90.65 % and SVC 91.24% of accuracy. By this step the hypothesis of the thesis was proven.

The study proved, that the data coming from the accelerometer embedded in a smartphone can be used for differentiating the predefined 6 activities with more than 90% of accuracy. Thanks to that, we can monitor the movement of the individuals. An application based on this feature can be used mostly for personal life logging or by healthcare providers for therapeutical and prophylactic purposes.

Thanks to the study, I got familiar with designing a HAR system. Furthermore, I learnt to use several libraries connected to Data Processing and Machine Learning.

Appendices

Appendix A

Acronyms

HAR Human Activity Recognition. vii, viii, x, xi, 3–7, 10, 12–15, 18–20, 24, 26, 28, 32, 33, 39, 41, 43, 53, 60, 63–67

ML Machine Learning. 3–5, 7, 10, 11, 24, 25, 28, 29, 31, 33, 35, 43, 60, 63, 67

Appendix B

Bibliography

[AGO+13] D. Anguita, A. Ghio, L. Oneto, X. Parra, and Jorge Luis Reyes-Ortiz, *Energy efficient smartphone-based activity recognition using fixed-point arithmetic*, J. UCS **19** (2013), 1295–1314.

[BA09] Yin B Westerterp KR. Bonomi AG, Goris AH, *Detection of type, duration, and intensity of physical activity using an accelerometer.*

[Bar] Suzanne Barlyn, *Strap on the fitbit: John hancock to sell only interactive life insurance*, [cit. 2020-08-08]. https://www.reuters.com/article/us-manulife-financi-john-hancock-lifeins/strap-on-the-fitbit-john-hancock-to-sell-only-interactive-life-insurance-idUSKCN1LZ1WL.

[BBS14] Andreas Bulling, Ulf Blanke, and Bernt Schiele, *A tutorial on human activity recognition using body-worn inertial sensors*, ACM Comput. Surv. (2014).

[BBSB10] Martin Berchtold, Matthias Budde, H. R. Schmidtke, and Michael Beigl, *An extensible modular recognition*

concept that makes activity recognition practical, 400–409.

[Bra] A. Braun., *Proactive vs. reactive: Shifting paradigms in health care provision*, [cit. 2020-07-25].

[BS] Yoo Y. Bourobou ST, *User activity recognition in smart homes using pattern clustering applied to temporal ann algorithm*, [cit. 2020-08-03], url = doi:10.3390/s150511953.

[cit07] *Machine learning and data mining: Introduction to principles and algorithms*, Horwood Publishing Limited, 2007.

[CNSL] P. Barralon N. Noury D. Lyons C. N. Scanaill, S. Carew and G. M. Lyons.

[CS] Y. Chen and C. Shen, *Performance analysis of smartphone-sensor behavior for human activity recognition*, [IEEE Access, vol. 5, pp. 3095-3110, 2017].

[CZS] CZSO, *Prumerny cech trpi mirnou nadvahou*, [cit. 2020-08-1].

[DARO13] Luca Oneto Xavier Parra Davide Anguita, Alessandro Ghio and Jorge L. Reyes-Ortiz, *A public domain dataset for human activity recognition using smartphones. 21th european symposium on artificial neural networks.*

[FVN10] A. Fleury, M. Vacher, and N. Noury, *Svm-based multimodal classification of activities of daily living in health smart homes: Sensors, algorithms, and first experimental results*, IEEE Transactions on Information Technology in Biomedicine (2010), 274–283.

[GD14] P. Gupta and T. Dallas, *Feature selection and activity recognition system using a single triaxial accelerometer*, IEEE Transactions on Biomedical Engineering (2014), 1780–1786.

[Ger17] Aurelien Geron, *Hands-on machine learning with scikit-learn and tensorflow : concepts, tools, and techniques to build intelligent systems*, O'Reilly Media, 2017.

[HT16] Hans-Olav Hessen and Astrid Johnsen Tessem, *Human activity recognition with two body-worn accelerometer sensors.*

[K.17] Frank K., *Hands-on data science and python machine learning*, Packt, 2017.

[KDN] KDNuggets, *Support vector machines: A simple explanation*, [cit. 2020-07-15].

[Kel15] Mac Namee B. D'Arcy A.en Kelleher, J. D., *Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies*, MIT Press, 2015.

[Kha11] Adil Mehmood Khan, *Human activity recognition using a single tri-axial accelerometer.*

[KKB14] Yongjin Kwon, Kyuchang Kang, and Changseok Bae, *Unsupervised learning for human activity recognition using smartphone sensors*, Expert Systems with Applications **41** (2014), no. 14, 6067 – 6074.

[KWM11] Jennifer R. Kwapisz, Gary M. Weiss, and Samuel A. Moore, *Activity recognition using cell phone accelerometers*, 74—82.

[LL13] O. D. Lara and M. A. Labrador, *A survey on human activity recognition using wearable sensors*, IEEE Communications Surveys Tutorials (2013), 1192–1209.

[LZYG13] Yunji Liang, Xingshe Zhou, Zhiwen Yu, and Bin Guo, *Energy-efficient motion related activity recognition on mobile devices for pervasive healthcare*, Mobile Networks and Applications (2013).

[Mat] MathWorks, *Accelerometer*, [cit. 2020-08-01]. https://www.mathworks.com/help/supportpkg/android/ref/accelerometer.html.

[MIS] A. MISHRA, *A. metrics to evaluate your machine learning algorithm*, [cit. 2020-08-02].

[OL18] Godwin Ogbuabor and Robert La, *Human activity recognition for healthcare using smartphones*, 41–46.

[oMUCoQoHCiA01] Institute of Medicine (US) Committee on Quality of Health Care in America., *Crossing the quality chasm: A new health system for the 21st century*, 2001.

[Pat] Prasad Patil, *What is exploratory data analysis?*

[PLB10] A. J. Perez, M. A. Labrador, and S. J. Barbeau, *G-sense: a scalable architecture for global sensing and monitoring*, IEEE Network (2010), 57–64.

[sld] scikit-learn documentation, *sklearn.manifold.tsne*.

[SS17] T. Sztyler and H. Stuckenschmidt, *Online personalization of cross-subjects based activity recognition models on wearable devices*, 2017.

[STJ14] X. Su, H. Tong, and P. Ji, *Activity recognition with smartphone sensors*, Tsinghua Science and Technology (2014), 235–249.

[Vel17] Haritha Vellampalli, *Physical human activity recognition using machine learning algorithms*.

[YYP08] J. Yin, Q. Yang, and J. J. Pan, *Sensor-based abnormal human-activity detection*, IEEE Transactions on Knowledge and Data Engineering (2008), 1082–1090.