Machine Learning for the Prevention of Injuries in the Construction Industry

Sarah Johannesson and Johanna Ögren

DIVISION OF INNOVATION ENGINEERING | DEPARTMENT OF DESIGN SCIENCES
FACULTY OF ENGINEERING LTH | LUND UNIVERSITY
2020

MASTER THESIS

Machine Learning for the Prevention of Injuries in the Construction Industry

Sarah Johannesson and Johanna Ögren

LUND
UNIVERSITY

Machine Learning for the Prevention of Injuries in the Construction Industry

Published by
Department of Design Sciences
Faculty of Engineering LTH, Lund University
P.O. Box 118, SE-221 00 Lund, Sweden

Subject: Innovation Engineering (INTM01)
Division: Innovation Engineering
Supervisor: Emil Åkesson
Co-supervisor: Emma Fitzgerald
Examiner: Anders Warell

Abstract

The Swedish construction industry is subject to a high rate of occupational injuries, where overload factors are a significant cause. Through Human Activity Recognition, movement data can be collected and analyzed, enabling the identification of harmful movement patterns with the use of machine learning. This study aims to describe the environmental barriers and stakeholder attitudes towards a smart construction helmet which enables this kind of data collection, while evaluating the performance of the supervised machine learning algorithm Random Forest when applying it to movement data. It asks whether collecting movement data violates the privacy of construction workers, or if there are other significant aspects to consider in the adoption process.

Based on a literature review on the Swedish construction industry, digitalization and privacy, interviews were conducted with stakeholders within five relevant roles to gather their attitudes towards the smart helmet. Furthermore, a group of eleven subjects participated in the collection of movement data which was further analyzed with the Random Forest algorithm. Analysis of the interview responses demonstrated a positive attitude from all stakeholders, where technology resistance was an obstacle, while privacy was a less emphasized issue. The movement data analysis showed significant recognition skills after using reviewed methods to manipulate the data. However, the collected dataset was not satisfactory to alone show these results but was complemented by an external dataset. The results indicate that the construction industry may be ready for a smart helmet if the presented gains outweigh the technology resistance and the added weight of the IoT-device. Further research is however needed to develop the recognition skills to analyze more detailed movement data.

Keywords: machine learning, Random Forest, Human Activity Recognition, construction industry, digitalization, privacy

Sammanfattning

Den svenska byggbranschen är hårt drabbad av arbetsskador, där belastningsskador är en övervägande orsak. Genom Human Activity Recognition kan rörelsedata samlas in och analyseras, vilket möjliggör identifiering av skadliga rörelsemönster med maskininlärning, ett växande vetenskapsområde. Studien syftar att beskriva de omvärldsfaktorer och olika intressenters attityd gentemot en smart bygghjälm som möjliggör den här typen av datainsamling. Parallellt utvärderas prestandan hos den vägledda maskininlärningsalgoritmen Random Forest, när den appliceras på rörelsedata. Studien ifrågasätter om insamling av rörelsedata är en inskränkning på den personliga integriteten hos byggarbetare, eller om det finns andra viktiga aspekter som bör tas hänsyn till i implementeringen av hjälmen.

Baserat på en litteraturstudie rörande den svenska byggbranschen, digitalisering och personlig integritet, utfördes intervjuer med intressenter inom fem relevanta roller, för att sammanfatta deras attityder gentemot den smarta hjälmen. Vidare deltog en grupp av elva representanter i insamlingen av rörelsedata, som senare analyserades med Random Forest-algoritmen. Analys av intervjusvaren visade på en positiv attityd bland samtliga intressenter, där teknologiskt motstånd var ett uttryckt hinder, medan den personliga integriteten var ett mindre betonat problem. Analysen av rörelsedata visade signifikant igenkänning av olika rörelser efter användning av granskade metoder för att manipulera data. Emellertid var det insamlade datasetet inte tillräckligt för att på egen hand visa dessa resultat, utan var kompletterat med ett externt dataset. Resultaten indikerar att byggbranschen kan vara redo för en smart hjälm, under förutsättning att de presenterade vinsterna väger upp för det teknologiska motståndet och den adderade vikten till hjälmen som den teknologiska applikationen medför. Vidare forskning på området krävs dock för att utveckla igenkänningen och möjliggöra analys av mer detaljerade rörelsedata.

Nyckelord: maskininlärning, Random Forest, Human Activity Recognition, byggbranschen, digitalisering, personlig integritet

Acknowledgements

The production of this thesis has been an interesting process, giving us the opportunity to combine our previous knowledge while deep diving into two areas that were completely new to us - the science of machine learning and the construction industry. While we many times have encountered challenges to try our patience, this process has also been very inspiring and rewarding, which we will carry with us in future projects.

First, we would like to thank Cybercom and all employees involved in developing the smart helmet for the opportunity to work with this project and providing us with the tools needed to finish it. We would especially like to thank our supervisor at Cybercom, Dennis Zikovic for his continuous support and management whenever this project took new turns.

Secondly, we would like to thank our supervisors at LTH, Emil Åkesson, and Emma Fitzgerald for guiding us through this project. Taking part of their input and knowledge did always generate giving discussions to bring us further and improve our thesis. We would also like to thank our opponents Ellen Peber and Erik Wästfelt for providing us with new perspectives and constructive feedback, which helped us fine-tune this thesis.

Finally, we would like to express our gratitude to all the experts and company representatives that we have received insights from through the conducted interviews, as well as to the research subjects that participated in the smart helmet data collection.

Lund, May 2020
Sarah Johannesson and Johanna Ögren

Table of Contents

List of Acronyms and Abbreviations

AI	Artificial Intelligence
CAGR	Compound Annual Growth Rate
FFT	Fast Fourier Transformation
GDPR	The General Data Protection Regulation
HAR	Human Activity Recognition
IoT	Internet of Things
IT	Information Technologies
JSON	JavaScript Object Notation
KNN	K-nearest Neighbor
LDA	Linear Discriminant Analysis
MCC	Matthew Correlation Coefficient
RQ	Research Question
SVM	Support Vector Machines
UCI	University of California Irvine

1 Introduction

The introductory chapter aims to firstly present the problem definition based on a short underlying background, which will be further explained in Chapter 2. This will be followed by the research question and the limitations that are considered in this thesis. Ultimately, the contributions that the thesis will bring to academic research will be presented.

1.1 Problem Definition

The Swedish construction industry is subject to the highest rate of occupational injuries and illnesses among all industries in Sweden. Construction workers are often likely to operate on construction sites under constantly changing conditions, creating an unsafe environment and an increased risk of accidents (AFA Försäkring, 2017). Among the reported injuries, 17 percent are caused by movements leading to overload and/or overuse injuries. Furthermore, in a long-term perspective, as much as 47 percent of occupational injuries are caused by overload factors (Samuelson, 2018).

Meanwhile, the science of machine learning is experiencing explosive growth, with its market's expected compound annual growth rate (CAGR) of 43.7 percent between the years of 2020 and 2030 (Prescient & Strategic Intelligence, 2020). Human Activity Recognition (HAR) has been an active field of research for over two decades and builds on detecting body movements with the help of data collected from either external or wearable sensors (Lara & Labrador, 2013). With the help of stronger machine learning capabilities, this is an area of research that has expanded lately.

The possibility to detect harmful movement patterns could prevent them from causing long-term and irreversible injuries. This could in turn enable a healthier work-life for construction workers as well as resulting in reduced healthcare costs. However, due to the General Data Protection Regulation (GDPR) it is important to consider what type of personal information is permitted to collect and store regarding employees, and how this is handled.

With this idea in mind, the purpose of this research is to evaluate the possibility of developing a smart construction helmet (later referred to as the smart helmet), which

by applying machine learning on movement data can prevent these kinds of injuries, without violating the construction workers' integrity rights.

1.2 Research Question

This project aims to answer the following research question (RQ):

RQ 1 Is Random Forest applicable for the smart helmet and what are the stakeholder attitudes towards the smart helmet?

> ***RQ 1.1*** What are the environmental barriers in the Swedish construction industry for the smart helmet, with regards to the stakeholder attitudes?
>
> ***RQ 1.2*** Can Random Forest make good classification from the movement data collected from the smart helmet?

1.3 Limitations

To enable giving a clear answer in the pre-study, the scope has been narrowed with two conditions: only studying the Swedish construction industry, and only considering how the smart helmet would be implemented with regards to Swedish regulations. This, as both the industry structure and the regulatory framework of each country are believed to have an impact on the implementation environment and process.

Due to the outbreak of Covid-19 in the spring of 2020, the data collection was not executed according to plans, since the variety of research subjects could not be used as first expected. Therefore, the research subject attributes presented in the methodology chapter are not as widely spread as had been wished for. Furthermore, during the data collection unforeseen connection errors to the Azure server were encountered. This resulted in a less efficient data collection, and hence less amount of data due to the time limitation.

1.4 Contributions

This project contributes to the current research on smart helmets within the construction industry by evaluating both the social and technological aspects of its implementation. From the social perspective, it provides a review of the construction industry and its stakeholder attitudes, and the possible implications of introducing a smart helmet with a certain focus on the privacy of users. From the

technological perspective, it contributes with an evaluation of the Random Forest machine learning algorithm when applied on an activity dataset, inspired by movements in the construction industry, collected through the smart helmet. This, as a first step in research towards identifying and preventing harmful movement patterns among construction workers. Furthermore, an implementation code for preprocessing data according to the used methods will be provided. In addition to this, the explorations may not only be limited to the construction industry but could also provide guidance for projects related to other industries.

1.5 Disposition

The background to this thesis will first be presented, elaborating on the construction industry structure, its digitalization initiatives, and how privacy is related to this. Moreover, the fundamentals of HAR and machine learning will be explained with the purpose to give the reader an understanding of the subject, to better follow the analysis further on. Further, the methodology used for this thesis will be presented, describing how the work process was conducted and the method chosen to answer the research questions from two perspectives: the environment of the smart helmet, and the implementation of the machine learning algorithm. Subsequently, the methodology will be followed by the interview and observation results of the pre-study, to create a practical base for the problem to be better understood. The results of the body movement analysis will then be presented to describe the performance of the Random Forest algorithm. The results will then further be evaluated and discussed to answer the purpose of the thesis. Lastly, a conclusion will present the reflections from this work process.

2 Background

The theory chapter is divided into five parts. The first two parts will explain the prerequisites of the pre-study, focusing on the social aspects. First, the relevant fundamentals of the construction industry are presented. This includes an explanation of the Swedish industry structure and its issue with occupational injuries, together with the general digitalization progress of the industry. Second, it will be reviewed how privacy and personal integrity rights can be handled in an increasingly digital business and working environment. The following parts will focus on the prerequisites from a technical perspective. This will include the definitions and technology behind human activity recognition, as well as related research on the subject. Furthermore, the science behind machine learning will be explained, with a certain focus on supervised learning and the Random Forest algorithm which will be evaluated in the report.

2.1 The Construction Industry

The Swedish construction industry consisted in 2019 of 327,000 employees, which represents 6.4 percent of the engaged workforce in Sweden, and entails an increase of 152,000 employees since the turn of the millennium (Byggföretagen, 2020b; Statistiska Centralbyrån, 2020). Approximately 10 percent of these are employed at one of the three largest construction firms: PEAB, Skanska, and NCC (Byggföretagen, 2020a). Out of 107,582 companies within the construction industry, only 662 companies had more than 50 employees in 2019 (Byggföretagen, 2020b). Hence, most construction companies are small enterprises, often active as subcontractors to larger firms.

2.1.1 **Industry Regulations**

In Sweden the right of association gives all employees the right to join a trade union of their choice, which in each sector determines collective agreements to which companies within that sector need to comply with. The trade unions and employer organizations are responsible for the fulfillment of these regulations towards employees and employers, while the government does not interfere (Unionen, 2020). Considering the construction industry, Byggnads is a trade union with over

100,000 members, supporting around 80 different occupational groups, meaning that approximately one-third of all employees in the construction industry are members (Byggnads, 2019).

It is the employer together with several involved project stakeholders that are responsible for safe work conditions in construction projects. In each project there are two work environment coordinators responsible for the work environment during the planning phase and the construction phase respectively (Arbetsmiljöverket, 2019). Alongside, AML 1977:1160 is a general law for all Swedish industries, providing legal regulations to prevent health issues and accidents in the work environment. Furthermore, AFS 2001:1 ensures that the employer continuously evaluates their business and activities for the same purpose. If the employer itself does not have the sufficient resources and competence, the same regulation states that the employer must engage an occupational health service to provide health care.

2.1.2 Injuries in the Construction Industry

Despite the regulations and safety measurements previously presented, the construction industry is subject to the highest rate of occupational injuries and illnesses among all industries in Sweden. Construction workers are often likely to operate on construction sites under constantly changing conditions, creating an unsafe environment and an increased risk of accidents (AFA Försäkring, 2017). Apart from the construction site changes due to the project process, the workforce on the site will also change (Kines, et al., 2011). This may increase the risk of accidents, as safety information on the worksite is not guaranteed to reach all construction worker (Stergiou-Kita, et al., 2015).

Although the yearly number of reported injuries has decreased in recent years, the frequency of reported injuries was still 11.5 per 1,000 construction workers in 2018. Among these, 17 percent of the reported injuries are caused by movements leading to overload and/or overuse injuries. Furthermore, from a long-term perspective, as much as 47 percent of occupational injuries are caused by overload factors (Samuelson, 2018). Generally, overload factors most commonly originate from repetitive activities, heavy lifting, transferring of objects, and inconvenient body positioning (Arbetsmiljöverket, 2018). More specifically, approximately 30 percent of construction workers state that they lift more than 15 kg at least once per day, while the percentage is less than half in other industries (Arbetsmiljöverket, 2015).

The construction industry did until 1993 have an occupational health service specifically assigned to them, called “Bygghälsan” (“The Construction Health”). The years following the termination of Bygghälsan, expert knowledge among occupational health services were considered deficient by employers and trade unions, as construction companies were forced to employ more general health services. Furthermore, the knowledge about health and safety-related issues varies

between large and small construction companies (Byggvärlden, 2015). This issue is also intensified as it is more difficult for smaller companies to employ these types of services compared to larger companies, due to limited resources and that the occupational health services often prioritize larger customers in time and resources (Åström Paulsson, et al., 2014; Johansson, 2016).

However, nearly 20 years after the termination of Bygghälsan new initiatives started to sprout. A collaboration between Luleå University of Technology, Development Fund of the Swedish Construction Industry (SBUF) and the insurance company AFA Försäkring evaluated the health services provided in the construction industry, by interviewing the trade association Byggföretagen, trade unions, local managers, safety officers and construction workers (Byggvärlden, 2015).

In the following years, NCC and Peab, both listed as the top three largest Swedish construction firms signed an agreement with the occupational health service Feelgood (Feelgood, 2008; Feelgood, 2019). The objective would be to increase the companies' work with long-term health among employees, through a preventative approach. The introductory intervention was presented in a report from Linköping University. The intervention included a test group of 123 construction workers, all considered as high-risk with regards to their physical condition. For one year, half of the participants would go through regular physical check-ups and get feedback to improve their condition, while the other half would represent the control group. Results could be noticed both regarding the physical health of the participants, as well as through gains in efficiency thanks to decreased sick leave among workers (Bernfort, et al., 2013).

2.1.3 **Digitalization of the Construction Industry**

While digitalization already has reached many industries as of today, the construction industry is experiencing a delay due to technology resistance (Oliver Wyman, 2018). A Swedish report from Tillväxtverket (2018) investigates the digitalization in Swedish industries, defined as the usage of information technologies (IT) in firms. Internationally, Sweden performs well in overall digitalization of society, the knowledge and usage of technology. However, the digitalization trend differs among the sizes of firms and industries. Businesses within the service sector are generally more digitalized than those within the industrial sector, and among the latter, the construction industry is considered the least digitalized industry in Sweden.

Simultaneously, the need for a disruption is growing to increase the efficiency of the industry according to a report from McKinsey & Company (2016). The authors state that in an international context projects often take 20 percent longer to finish than initially planned for and end up at 80 percent over budget. Furthermore, the labor productivity of the construction industry compared to the total economy has declined in the German and UK market since 1995.

A report from Oliver Wyman (2018) shows that certain trends are pushing towards digitalization within the construction industry more than others. As clients of construction companies are influenced by other more rapidly changing industries, they demand the buildings and infrastructure to fit usage expectations and connectedness. Simultaneously, the costs of sensors, hardware, and software are decreasing, while their efficiency is increasing, making technologies more available. The accelerating technology adoption is creating new technology-related jobs as well as increasing the opportunities for startups within the industry. Furthermore, digitalization can help reduce the environmental impact of construction projects. This in turn, is needed to fill governmental requirements and regulations that can be seen particularly in the Nordic countries and the UK. The requirements on data capacity and cybersecurity in buildings and infrastructures will also increase, with a sustainable approach towards the GDPR. Although construction industry stakeholders are still hesitant about new technologies, they may be required to develop new strategies for the digital age to reach continuous success in the future.

The industry must also lay the groundwork for these types of initiatives to thrive as discussed by McKinsey & Company (2016). The report shows that projects within the construction industry are often extremely diverse. On top of that, smaller construction firms with varying sophistication levels often function as subcontractors, creating a complexity between actors that must be managed. There is a demand for better processes for project planning, incentives for risk-sharing and innovation in contracts, performance management, and evaluating up-front investments compared to their long-term benefits. The latter is significant, with R&D expenditures in construction of less than 1 percent of revenues, versus up to 4.5 percent in the auto and aerospace sectors.

Furthermore, it is shown that firms are more likely to digitalize within areas related to increased efficiency and streamlining of processes, while it is less common to apply technology within business development. It is therefore considered important for companies to also identify digitalization possibilities to enhance the value proposition and customer benefits (Tillväxtverket, 2018).

However, some initiatives related to construction safety with the help of new technologies have been seen lately. Brown (2020) states one example, where the Boston-based construction firm Suffolk has introduced drones to capture images of construction sites. Using a machine learning algorithm trained with ten years of accident data and safety hazard images, the objective is to predict where accidents will happen. Furthermore, considering not only the construction industry but workwear in general, Yang, et al. (2018) introduces a wearable system with textile electrodes, motion sensors, and real-time data processing. The system is used to conduct risk assessments in various types of activities with various levels of workload, by collecting data on the heart rate and leg motion. Eight research subjects were used, whereof two from the construction industry, to collect data and

provide a basis for the prevention of musculoskeletal disorders and cardiovascular disorders based on physical workload.

2.2 Privacy Consequences from Digitalization

The word integrity originates from Latin, meaning whole or complete, and relates to each person's inalienable intrinsic value. The idea of personal integrity rights is that a person shall not be violated, neither physically nor mentally. Physical integrity relates to the human body, to which no actions should be made without the owner's consent. Mental integrity analogously relates to the human mind, such as values, ideas, opinions, and desires, which shall not be subject to encroachment (Statens Medicin-Etiska Råd, 2020). In Sweden these rights are explained by the Fundamental Law on Freedom of Expression, one of four fundamental laws in The Swedish Constitution stated by the government (Sveriges Riksdag, 2016). Related to integrity, and more generally used in an international context is the word privacy, which according to the Cambridge Dictionary is defined as "the right that someone has to keep their personal life or personal information secret or known only to a small group of people". Both integrity rights and privacy has become a topic of discussion lately, as digitalization progress and an increasing amount of personal data is collected from users of digital devices.

Johansson Stålnacke & Pettersson (2016) evaluates through their research the view on personal data and integrity rights in increasingly digital services from two perspectives: the receiver often in the role of a product developer, and the sender as the product user. The report states that although the both parts are initially skeptical to share personal data in private use of devices, the receiver expresses more trust after the purpose is explained, due to their bigger knowledge regarding data storage and processing. Furthermore, a difference could be noticed in what type of data was considered most sensitive to share. Both parts considered health data highly sensitive, while location information was more sensitive to senders than to receivers. The authors argue that future developments will result in more complex applications and related information sharing, requiring a more trustful and transparent relationship between the sender and the receiver.

Johansson Stålnacke & Pettersson (2016) further argues that although the research on Internet of Things (IoT) usage has expanded lately, little is said about how IoT influence people's integrity and sense of security. With an increasing amount of collected personal data thanks to IoT, it is important for companies to ensure integrity quality, and build strong and lasting relationships with customers. As proposed by Weinberg, et al. (2015) privacy by design is a method in this process, where the integrity of customers is considered throughout the development process. Cavoukian (2010) presents seven principles on which privacy by design builds on, related to how privacy should be considered:

(1) proactive rather than reactive;
(2) default condition rather than optional;
(3) embedded in the design rather than an add-on;
(4) not to interfere with the product functionality;
(5) to be protected, security is applied throughout the entire system where sensitive data may travel;
(6) procedures are transparent to be trustworthy in delivering on privacy-related objectives;
(7) respectful with regards to users' interests and being, empowering them to manage their data through actions related to consent, accuracy, access, and compliance.

Brill (2014) discusses the increasing amount of data collected and stored in society through IoT devices, along with its involved stakeholders, opportunities, and risks. There is a big potential in this evolvement of data collection for solving social challenges, and breakthroughs in healthcare are already seen thanks to wearable devices used to measure movements, sleep, and other health aspects. However, stakeholders notice risks related to the privacy of individuals, as personal information becomes affluent and more easily available. With various combinations of offline and online personal data, it is possible to create alarmingly accurate consumer profiles, letting companies track and advertise towards their customers in a more precise way. Brill (2014) presents three practices to help device and service providers cope with these privacy issues. First, as previously mentioned, privacy by design is recommended to promote the privacy of consumers and the ethical aspect throughout organizations and processes. Second, the importance of the de-identification of personal data is stressed. Third, it must be recognized that effective transparency is fundamental for privacy protection, easily explaining to consumers what nature of data their devices collect and transmit.

Apart from general initiatives to protect the privacy of users in the digital environment, the regulatory framework has also been updated in recent years. GDPR 2016/679 was implemented in May 2018, with the purpose to create a uniform basis for the protection of personal data throughout the EU, enabling a freer flow of data within Europe (Datainspektionen, 2020). GDPR superseded the Data Protection Directive 95/46/EC which was implemented by several countries within the EU in 1998, also with the purpose to protect personal data. However, as presented by IDG (2019), there are some essential differences between the two regulations. Firstly, GDPR states that a company cannot own any personal data of individuals. Data can only be lent from the individuals to the company and must be disposed of as soon as the two parties are no longer involved. Secondly, companies must clarify to the owner the purpose with which they are collecting personal data, before collecting it.

Processing of personal data is only lawful if it follows at least one of six legal bases as presented in Chapter 2 (Art. 6 §1) of GDPR 2016/679 (2016):

(1) the data subject has given consent to the processing of his or her personal data for one or more specific purposes;
(2) processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract;
(3) processing is necessary for compliance with a legal obligation to which the controller is subject;
(4) processing is necessary in order to protect the vital interests of the data subject or of another natural person;
(5) processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller;
(6) processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child.

2.3 Human Activity Recognition

The field of HAR has been active in research since the late 1990s and can provide distinct information on the activities and behaviors of people (Lara & Labrador, 2013). Today it is of interest for several industries, and the technology is used in for example medical, military and security applications to give feedback to the user on its movements (Jia, 2009; Yin, et al., 2008).

2.3.1 **The HAR System**

Two different methods for collecting and recognizing activity data can be seen. Either, applications may use external sensors such as cameras or smart home devices, positioned in the environment of interest. The second approach is using wearable sensors, which are positioned on the body to collect the user's movement, environmental variables, or physiological signals (Lara & Labrador, 2013).

With external sensors several difficulties can be seen, especially if using cameras. First, as the method implicates continuous monitoring and recording by a camera, it may violate the privacy of the user. Second, capturing the specific user's movements from various angles and within the photographic reach can be difficult and limiting. Third, video processing demands heavy computing which becomes both complex and expensive, limiting a scalable real-time system. These difficulties can motivate the use of wearable sensors instead of external sensors in HAR systems (Lara & Labrador, 2013).

The HAR system with wearable sensors is built up by four general architectural parts: (1) the wearable sensors measuring the attributes of interest, (2) the integration devices communicating with the sensors, possibly preprocessing the data and sending it to a server, (3) communication protocol over which the sensors can communicate with other applications, and (4) storage and inference where the data may be stored, monitored or visualized. The architecture is illustrated in Figure 2.1 (Lara & Labrador, 2013).

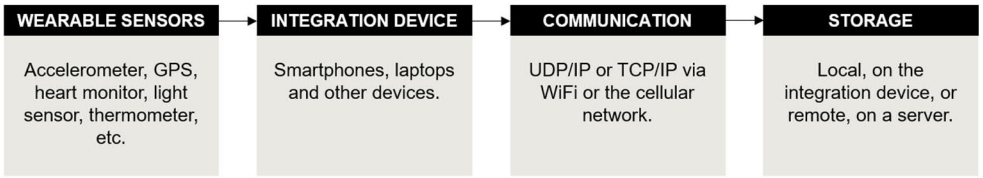

Figure 2.1. The general structure of a HAR system (Lara & Labrador, 2013).

2.3.2 **The HAR Problem Definition**

The HAR problem, i.e. recognizing movements from time series of attribute values, such as acceleration values can be defined in several ways (Lara & Labrador, 2013). Lara & Labrador proposes a definition of the HAR problem to make it problem deterministically solvable, by limiting the combinations of attribute values and activities. This can be achieved if using attribute values recorded over some time with a constant sample frequency, which are then cut into equally sized sections with regards to the number of samples and time, so-called fixed-length time windows. Within each of the time windows it will thereby be easier to evaluate what activity is performed, since the window is relatively small compared to the time a person naturally performs one activity.

It is however argued by Lara & Labrador that this definition creates some errors in the model, as more than one activity might be performed within one single time window if letting the participant move freely and switches activities. Yet, it is assumed that the number of transition windows will be much smaller than the total number of time windows, making the error insignificant.

2.3.3 **System Design Issues**

When designing a HAR system Lara & Labrador (2013) identifies seven main issues to consider.

(1) The selection of attributes and their corresponding sensors, which can be categorized as environmental (e.g. temperature, humidity, audio level), acceleration, location, and physiological signals (e.g. heart rate, respiration rate, skin temperature), are critical for the result. In the selection phase the

relevance and combination of these attributes must be evaluated, to reach as high recognition accuracy and descriptiveness as possible in the future analysis of data. Related to the use of accelerometers, Maurer, et al. (2006) propose that there is no significant gain in accuracy above 20 Hz for ambulation activities, such as walking, running, or climbing stairs. Furthermore, Lara & Labrador (2013) stresses the importance of the accelerometer placement on the user, with regards to the motions that the system aims to recognize. Ravi, et al. (2005) have approached the problem by using motion sensors dedicated to different body parts, e.g. waist, wrist, chest, and thighs, with good classification performance as a result. Meanwhile, the system should not include more attributes and sensors than is needed, as this increases the system cost and energy expenditures through potential wireless connections, as well as it introduces obtrusiveness.

(2) Obtrusiveness should be avoided, meaning that the HAR system should not be noticeable for the user, neither through the number of sensors or the need to interact with them. However, with more sensors the collected data will be richer, why it is important to find a balance between both sides. The number of sensors needed will also depend on the type of activities. For example, Bao & Intille (2004) conclude in their study that only two accelerometers, either on the wrist and thigh or wrist and hip, are sufficient to recognize ambulation and other daily activities. Anguita, et al. (2013) argue that using smartphones for data collection may less obtrusive and invasive than solutions such as wearable sensors.

(3) The environment and the individuals who are part of the data collection must be considered. Collecting data in a controlled laboratory environment will give more accurate results than in the natural environment. Furthermore, to obtain comprehensive training data, individuals of various characteristics should be used to ensure the flexibility of the model.

(4) The recognition performance of the system depends on several aspects that must be considered, such as the activity set and the complexity of the activities, the training data quality and quantity, the feature extraction method, and the choice of learning algorithm based on the characteristics of the dataset.

(5) Energy consumption may not be too high, as HAR applications often rely on mobile devices with an energy constraint. Energy expenditures are caused by processing, communication and visualization tasks, and can be limited mainly by minimizing the amount of transmitted data, and by using short-range networks such as Wi-Fi or Bluetooth over the cellular network.

(6) Processing of data can be made either in the server or in the integration device, where the former provides a larger capacity for processing and storing, while the latter can reduce energy expenditures from data not having to be continuously transmitted as well as being more responsive from not depending on wireless communication.

(7) The level of flexibility of the system must be considered, and hence it is important to decide whether a general recognition model should be used for all users, or if the model should be adjusted to each user's characteristics. The choice may depend on factors such as the number and type of activities, and the variation of behavior among users.

2.3.4 **Research Topics and Future Developments**

Anguita, et al. (2013) brought out a study on HAR using smartphones, arguing that the device brings new research opportunities on the area, with users as a rich source of context information and the device a firsthand sensing tool. As later models of smartphones come with built-in sensors, they provide a flexible and affordable way of monitoring daily activities in an unobtrusive sense. In their study Anguita, et al. presents a dataset collected from smartphone accelerometers and gyroscopes, intending to recognize six different human activities. There were 30 research subjects included in the study, with ages ranging from 19 to 48 years, following a protocol of activities while wearing a smartphone on their waist. The collected data was thereafter further processed in both the time domain and in the frequency domain.

Lara & Labrador (2013) argues that there are several topics for future research which could create value when developing the technology of HAR systems further. Some of these are: (1) enabling the analysis of more complex behaviors and composite activities other than the more fundamental groups earlier presented, (2) enabling the identification of overlapping activities, such as walking while eating, (3) creating greater context awareness by not only classifying the activity, but also other attributes such as age and gender, (4) enabling recognition of collective activity patterns which gives the possibility to estimate exercise habits and health conditions of a target population. Anguita, et al. (2013) also mentions the issue of identifying non-dynamic activities as a possible topic for future research, as their study shows a significant misclassification overlap between e.g. standing and sitting.

2.4 The Fundamentals of Machine Learning

Machine learning is a part of a wider concept called artificial intelligence (AI) and includes the concept of deep learning with neural networks. It is a field that can synthesize the underlying relationship among data without being explicitly programmed. The goal with using machine learning is to estimate the outcome of a situation that is unknown to the computer (Khanna & Awad, 2015). A more precise and widely accepted definition about the concept of machine learning was defined by Mitchell (1997). He defined it as:

> *"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."* (pp. 2)

Machine learning is today applied in a wide range of applications and areas, such as in image and audio processing, determining diseases and social behavior analysis. Based on the underlying mappings between input data and the expected output value, different machine learning algorithms can be used (Jiang, et al., 2017). These can simply be categorized into three main groups: supervised learning, unsupervised learning and reinforcement learning.

As mentioned, this report will evaluate the Random Forest algorithm, which is a supervised learning algorithm. Therefore, the following parts of this chapter will focus on supervised learning, while all three groups of learning algorithms will be briefly explained in this section to give the reader an understanding of the differences between them. The choice of machine learning algorithm will be further explained and motivated in *2.4.1 Supervised Learning* and *2.6.1 Evaluation of Algorithms within Similar Research Projects*.

2.4.1 **Supervised Learning**

Supervised learning is when a model is using a labeled training dataset, i.e. samples which are tagged with one or more labels, to learn the link between the input data and the expected output values (Khanna & Awad, 2015). From this, a prediction model can be developed to forecast output values for a new dataset (Jiang, et al., 2017).

A high level of generalization and predictive power for new input datasets is desirable when working with supervised learning algorithms. Since the performance increases with the size and variance of the training dataset, supervised learning algorithms require a potentially expensive training process.

The majority of classification and regression algorithms are supervised, where classification tasks use categorical output variables, while regression tasks use numerical output variables (Medium, 2018; Jiang, et al., 2017). Some examples are linear regression models, K-nearest neighbor (KNN), support vector machines (SVM), Bayesian learning, and Random Forest (Khanna & Awad, 2015; Jiang, et al., 2017). The latter will be further explained in *2.5 The Random Forest Algorithm*, while the key characteristics of each are presented in *Appendix B.1 Supervised Machine Learning Algorithm Characteristics*.

2.4.2 **Unsupervised Learning**

The family of unsupervised learning techniques detect and group data that behaves similar to each other, without them being pre-specified and labeled as done in supervised learning. Therefore, the unsupervised learning technique does not need a specific training dataset. Instead it learns the underlying structure of the dataset, while rejecting unstructured noise. Unsupervised learning algorithms include most clustering and dimensionality reduction algorithms (Khanna & Awad, 2015).

2.4.3 **Reinforcement Learning**

Reinforcement learning is inspired by behavioral psychology, where the learning technique relies on a dynamic iterative learning and decision-making process (Jiang, et al., 2017). The learning methodology is built on feedback loops, where rewards and punishments are associated with a sequence of actions, as illustrated in Figure 2.2. A given set of experimental *actions* is performed by an *intelligent agent* which will result in observed responses to the *state* of the *environment*. Depending on the action, the agent may also be *rewarded*. The agent is motivated to maximize the cumulated reward to find the best possible behavior or path in a specific state (Khanna & Awad, 2015).

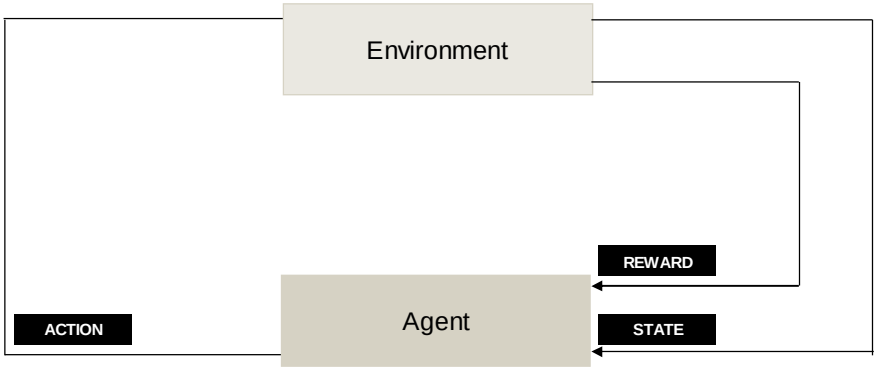

Figure 2.2. The fundamental parts of reinforcement learning: an agent takes actions in an environment, which is interpreted into reward and a representation of the state, which is fed back into the agent (Amiri, et al., 2018).

2.4.4 **Summary**

To summarize the subsections above, a family-tree of machine learning techniques, their type of training data, learning method, and use can be constructed as Table 2.1 shows.

Mitchell (1997) described a classification problem as when the algorithm needs to determine the category of the data, where the possible categories are included in a

dataset of categorical output variables. The aim with the collected dataset was to identify and group activities performed by the research subjects. With this in mind, it could be considered a classification problem, why it was found motivated to mainly study supervised learning algorithms and evaluation methods. Hence, the following sections will only concern supervised machine learning.

Table 2.1. Family-tree of machine learning techniques and their key characteristics, based on their training data, learning method and use.

	Supervised learning	***Unsupervised learning***	***Reinforcement learning***
Training data	Labeled training dataset	No training dataset	The agent generates its own data through interaction with the environment
Learning method	Learning from historical experience	Learning from structured patterns, by rejecting unstructured noise	Reward based learning
Use	Used for classification and regression algorithms	Used for clustering and dimensional reduction algorithms	Reward and recommendation algorithms

2.5 Evaluation of Supervised Learning Algorithms

Depending on the problem a machine learning model is trying to solve, different evaluation metrics are used. The metrics evaluates how well an algorithm performs, and different machine learning models can be compared if standard metrics are used (Caruana & Niculescu-Mizil, 2006). In the following subsection, a selection of different standard evaluation metrics for supervised learning algorithms will be presented.

2.5.1 **Confusion Matrix**

The confusion matrix is one of the most used evaluation metrics in supervised learning, and is a way to summarize the performance of a classification algorithm (Seref & Bostanci, 2018; Xu, et al., 2020). The confusion matrix is built up by two dimensions. One dimension is indexed with the actual class of an object and the other is indexed with the predicate class (Deng, et al., 2016). In Table 2.2, a standard form of a confusion matrix is presented. The algorithm reaches its optimal when it classifies as much of the dataset as possible as *true positives* and *true negatives*, i.e. the numbers along the diagonal are as high as possible, while the *false positives* and *false negatives* are as low as possible.

Table 2.2. The standard form of a confusion matrix with predicted classes presented horizontally and actual classes presented vertically (Chicco & Jurman, 2020).

	Predicate positive	***Predicate negative***
Actual positive	True positives, TP	False negatives, FN
Actual negative	False positives, FP	True negatives, TN

Note: True positives (TP) and true negatives (TN) are the correct predictions made by the classifier, while incorrect predictions are made with false negatives (FN) and false positives (FP)

By analyzing the number of true positives, false positives, false negatives, and true negatives in the confusion matrix several other evaluation metrics of classification performance can be defined. These are presented in the following subsections.

2.5.2 **Accuracy**

Accuracy is the proportion of the total number of predictions that were correctly identified among the total number of cases examined (Deng, et al., 2016).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.1}$$

2.5.3 **Precision**

Precision is a measure that shows the proportion of correctly identified instances that has been predicted in the positively identified set (Deng, et al., 2016). In other words, when the algorithm predicts the positive result, it is a measure of how often it is the correct prediction (Seref & Bostanci, 2018).

$$Precision = \frac{TP}{TP+FP} \tag{2.2}$$

2.5.4 **Recall**

Recall is a measure that distinguishes the proportion of the true positive predictions compared to the complete set of actual outcomes. It is an indicator of how complete the results are (Deng, et al., 2016).

$$Recall = \frac{TP}{TP+FN} \tag{2.3}$$

2.5.5 **F1-score**

The F1-score is a measure of a test's accuracy and defined as the harmonic mean of the precision and the recall of the algorithm predictions (Deng, et al., 2016). The ranges for the F1-score is [0,1], and as for accuracy, the minimum is reached when all positive samples are misclassified (TP = 0) and the maximum is reached for perfect classification (FN = FP = 0) (Chicco & Jurman, 2020).

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.4}$$

2.5.6 **Matthew Correlation Coefficient**

Matthew Correlation Coefficient (MCC) is a measure of the quality of the binary classification. Therefore, to get a high-quality score the classifier must make correct predictions in most of the positivity cases and negativity cases respectively. This, independently from their ratio in the overall dataset (Chicco & Jurman, 2020). The MCC ranges in [-1, 1]. The maximum is reached when perfect classification and respectively the minimum is achieved with perfect misclassification. The value of MCC = 0, indicates that the prediction was no better than a random flip of a fair coin (Boughorbel, et al., 2017).

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{2.5}$$

One of the advantages with MCC, according to Chicco & Jurman (2020), is that it can generate reliable results from an imbalanced dataset (the number of examples for each class label in the dataset is unbalanced). Further they explain that, F1-score and accuracy, can produce misleading results when applied to an imbalanced dataset, since these measurements fail to consider the rate between positive and negative elements. Thus, it is recommended by Chicco & Jurman (2020) to use MCC to evaluate the algorithm's performance.

2.6 The Random Forest Algorithm

Random Forest is a supervised machine learning algorithm that can perform both regression and classification tasks (Hastie, et al., 2009). The following subsection will firstly present a theoretical motivation to why the Random Forest algorithm was chosen for this project. Furthermore, as this thesis concerns a classification problem, the fundamentals of the Random Forest classification algorithm will be presented.

2.6.1 **Evaluation of Algorithms within Similar Research Projects**

Nowadays, an attractive research topic is HAR based on wearable sensor data, due to its applications in areas like healthcare and smart environments. Remarkable results have been presented from research using accelerometer and gyroscope data for HAR (Jordao, et al., 2018), similar to the setup of this study. Furthermore, multiple research studies have been made on the publicly available HAR dataset from the University of California Irvine (UCI) Machine Learning Repository presented in *2.3.4 Research Topics and Future Developments*, trying to evaluate which algorithm is the best classifier. As seen in Table 2.3, Random Forest outperformed all other algorithms' results with its score closes to the maximum of 1.0 in accuracy, and is according to the result of various studies the best classifier method.

Table 2.3. A comparison of the performance in accuracy between different supervised learning algorithms based on the UCI HAR dataset, presented in four research studies.

	Random Forest	***Decision Tree***	***SVM***	***KNN***	***LDA***	***Naive Bayes***	***Parallel Random Forest***
Parmar (u.d.)	**0.9987**	0.8999	0.9877	-	-	-	-
Dewi & Chen (2019)	**0.9857**	-	0.9796	0.9748	0.9823	-	-
Lavanya & Gayathri (2017)	**1.0**	0.9777	0.9555	0.8988	-	0.9555	-
BhanuJyothi, et al. (2017)	**0.9313**	-	-	0.8628	-	-	0.9298

In a similar study, five popular machine learning algorithms were used to train models for prediction of accident occurrence and severity in the construction industry. Once again, Random Forest provided the best performance regarding accuracy, as seen in Table 2.4. The dataset used in the study was collected from a constructor in Singapore and included 27 construction projects over seven years,

between 2010 and 2016. Furthermore, it consisted of 785 safety monthly inspection records, 418 accident cases along with their related monthly project-related attributes (Poh, et al., 2018).

With basis in the presented evaluation results and the previously presented background, the Random Forest algorithm was considered the most suitable for this research project.

Table 2.4. A comparison of the performance in accuracy between different machine learning algorithms from Poh, et al. (2018).

Random Forest	***Decision Tree***	***SVM***	***KNN***	***Linear regression***[a]
0.78	0.71	0.44	0.73	0.59

[a] Note: Linear regression is not a classification algorithm.

2.6.2 **The Decision Tree as a Building Block**

The building block of the algorithm is a large collection of de-correlated decision trees (Hastie, et al., 2009). A *decision tree* can be described as a series of true/false questions about the data that is leading to a predicted class for a classification problem (BhanuJyothi, et al., 2017). This means for each node top-down, the tree will find one feature that allows it to split the observations into a new classification, so that the resulting groups are as different from each other as possible while the members of each resulting subgroup are as similar to each other as possible (Medium, 2017).

One simple example that illustrates the logic of a decision tree is predicting the outside temperature of tomorrow for any city of choice. At each node the remaining observations will be classified after “season”, “historical average”, and “temperature of today”, resulting in a final prediction at the end of each subtree as illustrated in Figure 2.3 (Medium, 2017).

When Random Forest is used for classification, a prediction will be made from a majority vote from each tree’s class prediction, as seen in Figure 2.4 (Hastie, et al., 2009).

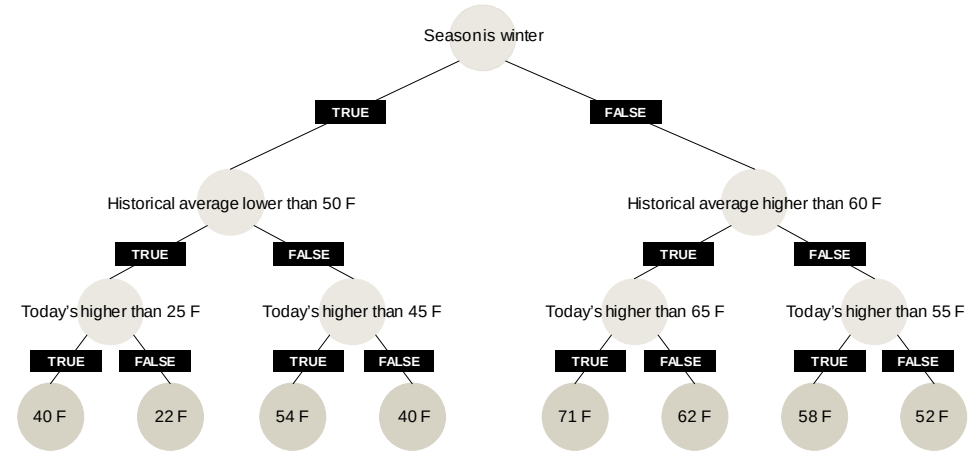

Figure 2.3. The logical reasoning of each decision tree in the Random Forest algorithm with a prediction of tomorrow's temperature (Medium, 2017).

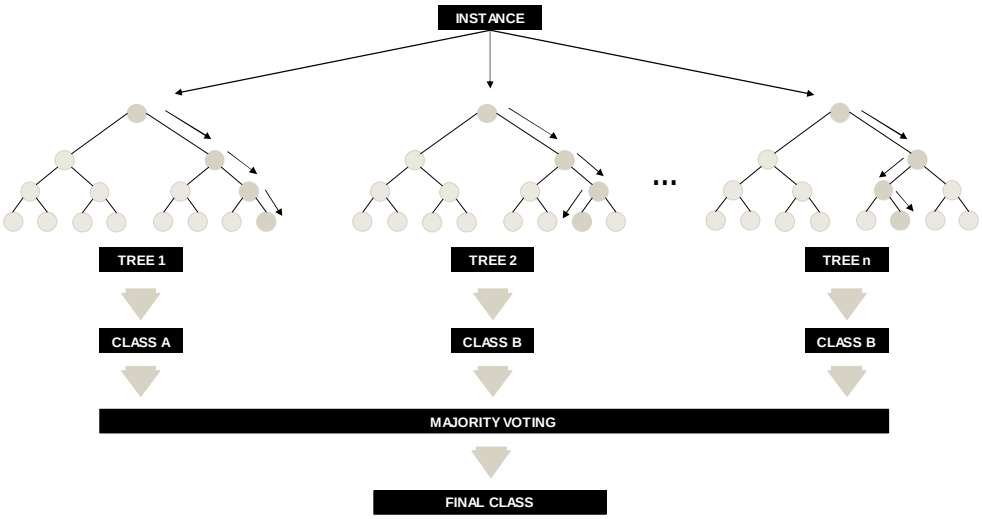

Figure 2.4. Structure of the Random Forest classification, with n trees between which a majority voting is performed (Medium, 2017).

According to Oshiro, et al. (2012) previous literature on Random Forest has very limited directive regarding the number of decision trees needed to compose a forest. The result of their research report and experiment of 29 datasets showed that the optimal range is 64 to 128 trees in a forest. Furthermore, according to the authors, no significant performance gain can be seen from increasing the number of trees higher than to the mentioned threshold, since this only increases the computational cost.

2.6.3 **Bias and Variance**

When designing supervised learning algorithms, both *bias* and *variance* are sources of error. Thus, they should both be minimized. As they are functions reacting in the opposite direction, a tradeoff must be made since minimizing the variance will increase the bias errors and vice versa (Geurts, 2002).

Supervised learning algorithms are said to suffer from bias error when wrong assumptions are made from the learning process. A high bias can result in missing relevant relations between features and target output since the algorithm is not satisfactory to solve the problem (James, 2003). Using the example in *2.6.3 The Decision Tree as a Building Block*, the bias would be caused if not enough questions were asked to give a credible prediction, and an unfounded conclusion is drawn.

A high variance can cause the algorithm to overfit (James, 2003). Overfitting is caused by the level of specificity in the tree, i.e. the depth of the tree in the example above, which may include noise to the training dataset, i.e. asking irrelevant questions that misleads the prediction (Towards Data Science, 2018; James, 2003). Hence, this source of error is a result of the sensitivity to small fluctuations in the training dataset (James, 2003).

The idea in Random Forest is to reduce the variance without increasing the bias. This is achieved in the tree-growing process by bagging (also known as bootstrap aggregation) and by splitting the nodes of each decision tree with a random subset of features along with the use of a committee prediction. Bagging refers to training a model on different datasets multiple times. In other words, each decision tree used in the algorithm learns from a different subset of data that is chosen at random with replacement. Each tree will therefore be unique. Thus, it reduces the correlation between the trees (Hastie, et al., 2009).

3 Methodology

This chapter aims to elaborate on questions related to the methodology of the thesis, such as the chosen research strategy and its validity. The methodology will be divided into two main parts according to the research questions. Firstly, the methodology for the pre-study will be presented. This will include the empirical and theoretical approach conducted, to explain the methods used to gather information. Furthermore, it will include an analytical approach for the findings made. Secondly, a more technical approach for the implementation of the machine learning algorithm will be presented, describing the technological prerequisites for the project, the data collection and selection process, and finally the data analysis with regards to the Random Forest algorithm.

3.1 Understanding the Environmental Barriers of the Smart Helmet

This section describes the methods used to understand the environmental barriers of an introduction of the smart helmet, to answer *RQ 1.1*. The theoretical framework and methods used will first be presented, followed by an explanation of how interviews and observations were conducted.

3.1.1 **Theoretical Framework**

When developing a new product, services, or internal process, design thinking is a human-centered approach for creative problem-solving. The approach has its focus on the human need behind any considered business need, and may reduce the risk associated with launching new ideas (IDEO, 2020).

Brown (2020) suggests that the first step in the design thinking process is to discover which constraints are important to the development of an innovation. These constraints can be organized into three overlapping criteria for successful ideas called 'The Three Lenses of Innovation': feasibility, desirability, and viability as presented in Figure 1.1. Feasibility describes what is functionally possible for an innovation within a foreseeable future. Desirability includes whether the innovation is of any value to people and organizations. Viability explains whether the

innovation will be valid for a sustainable business model. The idea is to find a balance between these three criteria, as an innovation will not manage in practice with only one or two of them fulfilled.

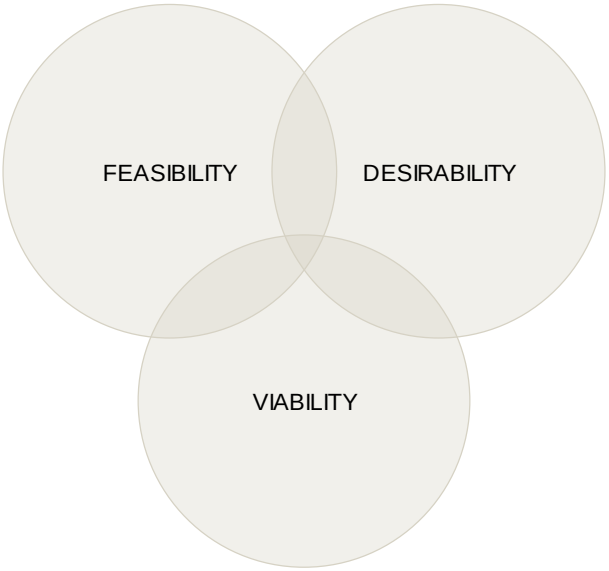

Figure 1.1 A Venn diagram illustrating the framework which this thesis' pre-study is built upon (IDEO, 2020).

To better understand the needs and constraints of the construction industry and the social environment in which the smart helmet would be applied, it was decided to conduct a pre-study as a first part of this thesis. In the pre-study the presented design thinking framework is adopted to better understand the environment in which the smart helmet would be implemented, together with the barriers that may be encountered. However, the framework is adjusted according to the scope of the thesis, to better fit the prerequisites of the thesis.

Within the feasibility criteria, the aim is to understand the technological aspect related to machine learning and human activity recognition, as well as to consider the social aspect regarding regulations, regulating units, and stakeholders on the labor market. Within the desirability criteria the receptiveness among construction workers in their role as users is evaluated, and their attitude towards privacy in the case of the smart helmet. Furthermore, the desirability is evaluated also from other stakeholder perspectives, such as the value construction companies and the trade union can see that this smart helmet would create. Finally, the viability criteria will only be briefly commented on, as it was assumed to follow the business model of existing construction helmets.

As a first step in the pre-study, the aim was to get an understanding of the fundamentals of the construction industry and the regulatory environment

surrounding it. Building this background to the problem situation would help in the following process of understanding which key stakeholders to contact for interviews, what relevant questions to ask them, as well as what theoretical areas to dig deeper into. Furthermore, the theoretical approach with a focus on machine learning was developed from an initial literature review to generally understand the science of machine learning and its underlying topics, as well as the related research made on human activity recognition. This is done to create a base to motivate the following approach and discussions on what machine learning algorithm was appropriate to evaluate for this thesis.

A preliminary mapping of theoretical areas and relevant contributions within those were made, where relevance was mainly assessed based on the number of citations on articles together with the applicability to the thesis situation. As this thesis is limited to only evaluating the Swedish construction industry, Swedish articles related to the regulatory and cultural environment was prioritized to get a more valid picture of the situation. When searching for more general subjects, not as influenced by the regulatory and cultural environment of Sweden, the citation method would be prioritized together with the specificity of the articles. However, when using non-academic articles as sources, the validity and reliability of the authors were assessed.

As stated by (Höst, et al., 2006) the most common methods when collecting empirical data are interviews, observations, and archival analysis. Among these, conducting interviews was considered the most appropriate method for collecting primary data, to cover several social areas within the time limitation. Observations were also chosen as a preliminary secondary strategy to collect data, depending on the outcome of the interviews and what information would be missing.

3.1.2 **Interviews**

Due to not having any existing network within the subjects covered in the pre-study, snowball sampling was found to be a reliable method to widen the social network and reach out to relevant stakeholders. Noy (2008) proposes a definition of snowball sampling as a method where the researcher gets access to interviewees through contact information provided by already known interviewees. It is argued that snowball sampling is specifically useful for two purposes: (1) to capture social knowledge dynamically, and (2) to understand the power relations between interviewees (Noy, 2008). This approach enabled capturing the greatest findings from the initial leads, while also being presented to dedicated experts within the research areas.

3.1.2.1 Interviewee Selection

When selecting interviewees, the aim was to gain a comprehensive understanding of the context in which the smart helmet would be applied, including stakeholders involved in the product's future adoption and use. As some background research

had been made before starting to conduct the interviews it was possible to define stakeholder attributes within the thesis' framework, i.e. the three lenses of innovation that were considered important for the purpose of the thesis.

As proposed by Mitchell, et al. (1997) it is possible to identify three stakeholder attributes: power, legitimacy, and urgency. Power can be defined as the ability of a stakeholder to impose its will in a relationship, while legitimacy is the perception of a stakeholder's actions being desirable, proper, or appropriate with regards to the social organization it is conducted in. Ultimately, urgency is proposed to be an attribute adding a dynamic dimension to the model, where time-sensitivity and criticality are included in the relationship (Mitchell, et al., 1997a).

Considering the previously presented stakeholder attributes, the thesis framework, and the initial literature reviews, a stakeholder analysis was conducted to identify appropriate interviewee positions for qualitative data collection as input to the work.

(1) **Employers** were defined as representatives at construction companies at white-collar positions. These were interviewed to get an apprehension of their attitude towards the smart helmet as customers, from both a feasibility, desirability and viability perspective. From a feasibility perspective, they were considered to have power in arguing for whether the implementation of the smart helmet would be organizationally possible. They were also asked to share any former experience from collecting and handling these types of personal data of employees. From a desirability perspective, they were thought to have power in telling whether the helmet would fulfill any needs of the employers and/or employees. Third, from a viability perspective, they were thought able to tell what the purchasing process would look like, and what level of willingness to pay would have to be met.
(2) **Employees** were defined as construction workers. These were interviewed to better understand their willingness to adopt the smart helmet, mainly from a desirability perspective. The aim was to investigate their acceptance towards letting the employer collect individual movement data in their daily work, as well as towards the technological adoption of the helmet. Another aim was to understand the power relationship between employees and employers in the situation. This, related to the requirements employers may set for employees in terms of demanding the use of the helmet for safety reasons, versus the requirements employees may set for employers in terms of not wanting to share personal data for privacy reasons.
(3) **Trade unions** were from the initial research understood to have moderate power in the implementation of a smart helmet, mainly from a feasibility perspective. They were therefore interviewed to better understand their influence on the implementation, and what obligations and/or possibilities they would have to protect the privacy of construction workers. Furthermore, they were thought to have an apprehension of the behavior among construction workers, both regarding injuries and their attitude towards technological devices.

(4) **Legal services** were considered to have high power with regards to the social aspects of the feasibility perspective. Since regulations set the framework for what is legal and therefore practically implementable, legal services were interviewed to give a better understanding and confirmation of the regulations concerned when implementing a smart helmet.
(5) **Technical experts** were considered to have high power in understanding the implementation of similar solutions to the helmet, and would therefore be able to give support to the feasibility perspective by explaining their experience on the area.

In Table 3.1 the interviews conducted are presented, describing which stakeholders were included at each meeting. Furthermore, in the right column the code with which they will be cited in *4 Interview Results* are presented to act as a reference for the reader.

Table 3.1 The interviews held with various stakeholders and the code with which they will be cited with in the interview results.

Code	Interview
TU	Interview with a representative with long experience from the trade union Byggnads.
NCC1	Group interview with four employers on various positions at the construction company NCC.
NCC2	Group interview with two employers and two employees at NCC with varying experience of the industry.
LEG	Interview with a representative from legal services, specialized on GDPR within business law.
TECH	Interview with a researcher on the topic of human activity recognition.

3.1.2.2 Interview Structure

Interviews were conducted with a semi-structured approach. As argued by Barriball & While (1994), semi-structured interviews are appropriate when there is a need to explore certain opinions and perceptions among the respondents, and if respondents have different professional, educational or personal backgrounds, meaning that a standardized interview form cannot be applied. Hence, as the empirical part of this study was aimed towards several different professional groups, a semi-structured approach was found adequate, and would allow adjusting the interview form for each of the respondents which in turn would create a more flowing conversation.

Before each interview, a personal interview form was sent to the respondent to give him or her an opportunity to read and understand the interview questions, as well as to prepare any needed research in advance. The interview form consisted of some social and technical background to give an understanding of the thesis project and its purpose. The form was then followed by a few introductory questions where the respondent would tell about their position and knowledge on the area. Apart from

providing some background information on the respondent, these types of questions may also make the respondent feel more confident and comfortable in the interview situation (Lekvall & Wahlbin, 2001). The interview was often followed by a couple of sections with different focuses, e.g. the privacy consequences of the helmet, or regarding the power position of the respondent in the potential implementation of the helmet. Questions were based on the conducted literature review. However, due to the exploratory approach and lack of knowledge in some areas, questions would be rather open and speculative to capture as many thoughts and opinions as possible. The interview forms can be forms in *Appendix A*.

3.1.2.3 Analysis of Results

The interviews were audio recorded for future reference, together with notes taken during the interview. Due to technical issues, some of these recordings were lost. However, the existing recording would be transcribed. As proposed by Miles, et al. (2014) coding is a way to analyze information gathered during fieldwork, and to categorize pieces of data to synthesize their message. Therefore, the transcripts, together with the notes from those interviews missing recordings would be coded to capture their essence. However, since the interviews had fairly different focuses depending on which of the presented role was interviewed, it was difficult to generate a descriptive set of codes that would be common for all interviewees. Instead, the general framework for the report was used, with categories "FEASIBLE", "DESIRABLE", and "VIABLE", together with subcategories "POSITIVE", "NEGATIVE" and "NEUTRAL" for each. These codings were later summarized as presented in *4 Empirical findings.*

3.1.3 **Observations**

Observations can be defined as the systematic description of events, behaviors, and artifacts in the social setting chosen for a study (Marshall & Rossman, 2014). With the basis in this definition and as a complement to the conducted interviews and part of the qualitative data collection, observations were made on a construction site, supervised by some of the construction workers that had previously been interviewed. The observations were made to get a better understanding of the work environment that construction workers act in, providing an opportunity to understand what type of movements construction workers often perform, and which of those that may cause occupational injuries. Furthermore, it was found rewarding to involve with the construction workers in their normal environment, where they might express thoughts and opinions which they would not be expressed in a formal interview setup. The results from the observations were mainly used to motivate the choice of activities performed in the data collection, further explained in *3.2.1.3 Activity selection and performance.*

3.2 Body Movement Analysis Approach

As suggested by Khanna & Awad (2015) the process of developing a machine learning algorithm can be decomposed into seven steps: (1) collect the data that is to be analyzed, (2) preprocess the data through formatting, cleaning and sampling, (3) transform the data, (4) train the algorithm, (5) test the algorithm, (6) apply reinforcement learning, and (7) execute.

Inspired by this approach, but with the prerequisites of this thesis in consideration, a five-step approach was used. First, the *data collection* involved selecting the subset of all available data attributes generated from the smart helmet, and then gathering the data through the smart helmet. Second, *choosing a relevant machine learning algorithm* for the problem was done through evaluation of results from previous works. Third, *preprocessing of the data* was conducted on the collected raw data stream, to be understood by the chosen algorithm. Fourth, a *feature optimization* was performed, as this according to Yi, et al. (2015) can reduce the dimensions of the features, contributing to a more efficient training and better performance of the algorithm. Fifth, *training and testing of the algorithm* was conducted, including analyzing the prediction results using various evaluation metrics which as suggested by Khanna & Awad (2015) can determine whether the model needs improvements or is sufficient.

As a final part of this section, limitations that occurred due to technical issues during the data collection, and how the effect of those were mitigated will be presented.

3.2.1 **Data Collection**

In the following subsection the smart helmet will be introduced with its hardware, embedded software and backend server components, through which the data would be collected. Thereafter, a description of the data collection will be explained including research subjects, activity selection, and performance.

3.2.1.1 Description of the Smart Helmet

The smart helmet prototype used for the data collection was retrieved from the IT-consultancy firm Cybercom. The smart helmet was equipped with a 3-axis accelerometer and a 3-axis gyroscope connected to an Arduino board, which would generate a raw data stream during the data collection. Furthermore, a buzzer was attached to the Arduino board to increase the communication between the hardware and the user environment. The buzzer would beep to inform the beginning and end of each recording sequence of the embedded software. Lastly, a battery was used as the system's power supply. Embedded software code was produced for the helmet to communicate with an external server. However, this code cannot be shared because of confidentiality policy at Cybercom.

A full description of the hardware is described in Table 3.2. The full arrangement of hardware (later referred to as the IoT-device) was located at the back of the helmet, as seen in Figure 3.1.

Table 3.2 Description of the hardware components used in the IoT-device of the smart helmet.

Hardware component	***Brand***	***Model***
Arduino board	Arduino MKR WiFi 1010	Arduino
3-axis accelerometer and 3-axis gyroscope	MPU6050 accelerometer and gyroscope 3-axis UEXT	Olimex
Buzzer	Buzzer 3.8 kHz	-
Battery	Battery LiPo 3.7V 1500mAh	-

Figure 3.1. The arrangement of hardware mounted on the smart helmet.

3.2.1.2 Activity Set Selection

The selected activity set for the data collection consisted of five activities, which could be divided into two main groups: lifting and walking. The lifting activities would be differentiated as light lifting and heavy lifting. Likewise, the walking activities would be differentiated as walking, walking while carrying something heavy, and walking while looking upwards. All activities are further described in *Table 3.3*.

The activities were chosen with regards to the ease with which they could be understood and performed by the participants, while also being related to the observed movements carried out by construction workers during the study visit. Furthermore, it was found interesting to evaluate rather similar activities within each

of the main groups, to see if it was possible to distinguish minor differences in movement patterns which could be a valuable result for future developments.

Table 3.3. Detailed descriptions of the activity set performed by the research subjects.

Activity	*Description*
Light lifting	Lifting a 2 kg object from the ground onto a table, approximately a lift of 0.75 m.
Heavy lifting	Lifting a 10 kg object from the ground onto a table, approximately a lift of 0.75 m.
Walking	Walking at a regular pace, approximately 1.4 m/s
Walking while carrying something heavy	Walking at a regular pace, approximately 1.4 m/s, while carrying a 10 kg object in each hand
Walking while looking upwards	Walking at a regular pace, approximately 1.4 m/s, while bending the head backwards to look at the ceiling.

3.2.1.3 Research Subjects

It is problematic to quantify the size of a dataset that will turn the trained model from good into great in advance. A rule of thumb is however that more complex problems and models need more data points. Simultaneously, the goal when training an algorithm is to build a model which will understand the relationship and the patterns of the data. Therefore, the quality of the data is an equally important factor to consider as the quantity, since the limit of the used data will be the limit of the trained model. From this, it is therefore a good approach to simply begin with a general estimation on the dataset required, to be able to work with the model. With time and as results appear to evaluate, it will become more obvious if more data is needed (Lionbridge, 2019).

A representative set of research subjects consists of selective samples from the original target population, which efficiently capture significant information with low redundancy (Pan, et al., 2005). Therefore, a group of eleven research subjects of various age, height and weight were selected when collecting data to generate a representative dataset, which was intended to simulate the spread that are occurring at a construction site. However, due to the outbreak of Covid-19 the spread in gender was highly unequal with only two male subjects out of eleven in total. Among the accepted subjects, the ages ranged from 18 to 60 years, with a median of 25 years. The heights ranged from 162 to 195 cm, with a median of 170 cm. Similarly, weights ranged from 55 to 90 kg, with a median of 66 kg. Details of each subject's characteristics are illustrated in *Appendix B.2.1 Table B.2.*

3.2.1.4 Setting the Sampling Frequency

The sampling frequency is the frequency with which each data sample is collected, in this study from recordings of performed activities during a timespan. Previous

research from Anguita, et al. (2013) argues for a sampling frequency of 50 Hz being sufficient for HAR projects. This statement is further affirmed by the studies of Maurer, et al. (2006), concluding that a sampling frequency higher than 20 Hz provides no increased gain in precision for ambulation activities, which represents three out of five of the chosen activities.

With the aim to distinguish rather similar movements within the two main groups of activities, i.e. walking and lifting, it was considered adequate to use a higher sampling frequency such as the one Anguita, et al. (2013) proposes. Furthermore, as mentioned by Lara & Labrador (2013), several of the activities could be considered overlapping, e.g. walking while looking upwards, which still is subject for research. This made it difficult to argue for the sampling frequency of 50 Hz being high enough to make these kinds of distinctions. However, as argued by Anguita, et al. (2013), a too high sampling frequency may create a superfluous quantity of data or result in significant performance loss of the hardware components, why 50 Hz was settled with as an outset.

A fixed sampling frequency was never set on the helmet. However, it was during initial test recordings discovered that the raw data stream would be sampled with a frequency around 50 Hz, or 1500 samples over 30 seconds. The achieved frequency would vary depending on the physical environment in which the movements were recorded, as well as the prevailing communication setup (e.g. the Wi-Fi capacity). To achieve a frequency level as even as possible over all sampling sequences, an accepted sample size of 1500±150 samples over 30 seconds was set, corresponding to a frequency of 50±5 Hz.

3.2.1.5 Communication and Server Setup

The raw data stream generated from the smart helmet was transferred to a backend server at a frequency of 5.5 Hz using a wireless connection. Data was formatted to JavaScript Object Notation (JSON), which is a file and data interchange format used to transfer data between servers and web applications. Each transfer to the server consisted of nine JSON-objects, as this would fit the static random-access memory (SRAM) of the Arduino board, which otherwise would be overloaded and block the sending (Arduino, 2020b; Arduino, 2020a).

Each JSON-object would include the current research subject coded with letters, the sample number to distinguish between different objects, a Unix timestamp, and the x-, y- and z-values for acceleration and angular velocity respectively. The full structure of the JSON-object is presented in *Appendix B.3.1 Table B.7*.

Furthermore, an employee at Cybercom established a server in Azure, which made it possible to extract the three-axial accelerometer and gyroscope values respectively from each JSON-object. These values were then put into separate CSV-files, which could be downloaded locally to be further processed, as described in *3.2.3 Data Preprocessing*.

3.2.1.6 Activity Performance

At the beginning of each test session, each participant was given information regarding the study, and was told the terms and conditions of the participation. They were also asked to share their information regarding gender, age, height, and weight for statistics.

Each research subject was instructed to follow a protocol of the chosen activities while wearing the smart helmet. Each activity was performed during a sequence of 30 seconds, with a longer break between each sequence to prepare the smart helmet for the next recording. For the lifting movements, there was a three second break between each lift during the 30 second sequence, to allow the experiment supervisor to replace the weight on the ground again. The samples collected during these breaks would later on be excluded from the dataset, to simulate a repetitive lifting movement. To receive a balanced dataset, i.e. equal number of samples from each movement, twice as many sequences of lifting was recorded compared to walking sequences to compensate for the excluded breaks. In total, walking activities were recorded three times, while lifting activities were recorded six times. A summarized description of the protocol can be found in *Appendix B.2.1 Table B.3*.

3.2.2 **Choice of Machine Learning Algorithm**

As motivated in several sections of *2 Background*, the Random Forest machine learning algorithm was chosen for evaluation. This, due to the results it had shown in classification problems similar to this project, and comparison to other algorithms.

Further, due to the time limitation and the scope of the project, it was decided to retrieve the Random Forest algorithm from the open source Python library called Scikit-Learn with inbuilt machine learning libraries. The algorithm can be found by browsing `sklearn.ensemble.RandomForestClassifier`.

3.2.3 **Data Preprocessing**

The algorithm retrieved from Scikit-Learn cannot handle features made up of strings, and more critically a feature cannot consist of time series of data (Scikit-Learn, 2020). Therefore, the categorical activities were encoded from string values to integer values. Moreover, since the algorithm is evaluating the input data on a per-event basis, i.e. row by row, it not possible to use time series of data as input values. This, as time series of data capture snapshots of the movement, instead of capturing a complete movement.

Instead, inspiration was taken from the signal processing of a similar research project by Anguita, et al. (2013), further described in *2.3.4 Research Topics and Future Developments*. Like this thesis, the authors used time series of data from a

three-axial accelerometer and a three-axial gyroscope for HAR, why the method seamed relevant. Each step of the preprocessing will be further described in the subsections below.

3.2.3.1 Noise Reduction

As a first step in the signal processing, and inspired by the methodology of Anguita, et al. (2013), a noise reduction was performed on the raw accelerometer and gyroscope signals. This means eliminating frequencies which are not caused by actual movements but caused by oversensitivity to vibrations etc. in the sensors.

For this, a median filter and a 3rd order low-pass Butterworth filter with a 20 Hz cutoff frequency was used. According to Anguita, et al. it is sufficient to use the 20 Hz as a threshold to capture human body motion, since 99 percent of its energy is contained below 15 Hz. The model used for the median filter was `scipy.signal.medfilt`, while `scipy.signal.butter` was used for the low-pass Butterworth filter.

3.2.3.2 Separation of Gravitational Force from Accelerometer Values

The original data collected from the accelerometer would include both the acceleration caused by the actual movements, as well as the force of gravity which is always naturally present. Because of this it would not be possible to recognize any movements from the unprocessed signals. Therefore, the second step in the signal processing would be to separate the acceleration signal into body acceleration and gravity components.

Again, the signals were filtered with a 3rd order low-pass Butterworth filter. The gravitational force only consists of low frequency components (van Hees, et al., 2013). Therefore, a cutoff frequency of 0.3 Hz for the Butterworth filter was used for the constant gravity signal as performed by Anguita, et al. The same Butterworth filter model was used as presented for noise reduction.

After the gravity signal had been separated, the values were subtracted from the original acceleration values to obtain the body acceleration values.

3.2.3.3 Calculation of Euclidean Magnitude and Time Derivative

To extend the possible number of features as input variables for the algorithm, additional signals were calculated for the time domain, influenced by Anguita, et al. These were derived from the body acceleration and angular velocity values, and are here mentioned as the Euclidean magnitude, jerk, and angular acceleration.

The Euclidean magnitude was calculated for both body acceleration and angular velocity values. It can be described as the magnitude, i.e. the length, of the vector obtained from each combination of three-axial values that were sampled. For this, the model `numpy.linalg.norm` was used. Mathematically, it can be formulated as:

$$\|a\| = \sqrt{x^2 + y^2 + z^2} \quad (3.1)$$

The jerk, $j(t)$ is the time derivative of the acceleration, $a(t)$. Similarly, the angular acceleration, $\alpha(t)$, is the time derivative of the angular velocity, $\omega(t)$. For each of the $i = 1,2,3 \ldots n$ sampled values of each recorded sequence these values were calculated as:

$$j(t) = \frac{da(t)}{dt} = \frac{a(t_{n+1}) - a(t_n)}{t_{n+1} - t_n} \quad (3.2)$$

$$\alpha(t) = \frac{d\omega(t)}{dt} = \frac{\omega(t_{n+1}) - \omega(t_n)}{t_{n+1} - t_n} \quad (3.3)$$

3.2.3.4 Separation Into Time Windows

As a next step, the collected time signals were separated into equally sized time windows. This approach has been described in *3.2.4 The HAR Problem Definition and* is as proposed by Lara & Labrador (2013) a method to make the HAR problem deterministically solvable.

To further minimize the transition errors that may occur, Lara & Labrador suggest using overlapping windows. The overlap was set to 50 percent, meaning that the second half of each window would contain the same values as the first half of the following window, as proposed by Anguita, et al. (2013).

Furthermore, each window would consist of 128 values. The chosen number of values can be argued for. Firstly, it is a power of 2, which has been shown to result in a more efficient Fast Fourier Transform (FFT) (The SciPy Community, 2020). The FFT is a part of the data processing and will be further described in the following subsection. Secondly, if movements are recorded with a sampling frequency of 50 Hz, each window will fit samples from a period of 2.56 seconds. As proposed by Anguita et al. (2013) at least a full walking cycle of two steps is preferred to fit each window. They also propose that the average step rate when walking is 90 to 130 steps per minute, meaning that each window would fit 1.9 to 2.8 walking cycles.

The time windows were produced with the Python package `window_slider 0.8` and model `window_slider.Slider`.

3.2.3.5 Frequency Domain Mapping

A signal mapped in the time domain provides information regarding how the signal changes over time. A representation of a signal in the frequency domain enables the observation of other signal characteristics, which may otherwise be difficult to notice. For example, when mapped in the frequency domain, it is possible to observe

how the signal's energy is distributed over a range of frequencies (MathWorks, 2020).

Inspired by Anguita, et al. (2013) FFT was applied to the obtained sets of time domain signals to map them in the frequency domain. The model used for FFT was `numpy.fft.rfft`. The full list of time and frequency domain signals can be seen in *Appendix B.3.2 Table B.8 and Table B.9.*

3.2.3.6 Measurements

As a last step in the preprocessing of data, calculations of a set of features for the time and frequency signals were performed. These features would be the final input variables for the algorithm to process.

As proposed by Lara & Labrador (2013) acceleration signals are highly fluctuating, leading to difficulties in observing underlying patterns from the raw data only. It is therefore proposed to use various feature extraction methods to better describe patterns. Anguita, et al. (2013) added several new sets of features to the standard measurement previously used in HAR projects to improve the learning performance of the algorithm. Influenced by Anguita, et al. a set of 14 statistical measurements was used to describe the obtained time and frequency domain signals, with a full description in *Appendix B.3.2 Table B.10*. Not all measurements were calculated for the time and frequency domain respectively, due to their signal characteristics.

By performing these calculations, the final number of features would end up at 348. Hence, from each set of time windows consisting of manipulated attribute values. 348 features would be extracted and sorted into a new time window. These could then be fed as input values to the algorithm. A full list of their names and heritage is presented in *Appendix B.3.2 Table B.11.*

3.2.3.7 Handling Categorical Activities

As mentioned the `sklearn.ensemble.RandomForestClassifier` cannot handle features made up of strings, why the categorical activities were manually encoded using integers. These encodings are presented in *Appendix B.3.3 Table B.11.*

3.2.4 **Feature Optimization**

To maximize the performance of the algorithm and contributing to more efficient training and as suggested by Yi, et al. (2015), an available feature optimization technique from Scikit Learn called `feature_importances` was used. The model builds on setting a higher value to indicate greater importance of a feature. Using the model would enable validating each feature against its value of contribution in a prediction through recursion, to determine if the dimension of features had to be reduced.

3.2.5 **Train and Test the Algorithm**

To ease the performance assessment, the preprocessed data was randomly partitioned into two independent sets, influenced by the approach of Anguita, et al. (2013). From the total, 70 percent of the data was used to train the algorithm and the remaining 30 percent was selected for testing.

Depending on what problem the machine learning algorithm is trying to solve, different evaluation metrics are often used. However, the confusion matrix often stands as a base through the evaluation of most supervised algorithms, as explained in *2.5 Evaluation of Machine Learning Algorithms*. Therefore, the confusion matrix was found to be a sufficient evaluation metric to start with. For example, it answered questions such as how many of the activities were correctly classified and at what cost in terms of false positives.

Caruana & Niculescu-Mizil (2006) highlight that an algorithm may perform well on one metric while it performs badly on others. Therefore, to make better use of the data from the confusion matrix, it was chosen to use the spread of evaluation metrics presented in *2.5 Evaluation of Machine Learning Algorithms* to ensure no misleading results.

3.2.6 **Adjustments Due to Technical Issues**

After the data collection several technical issues were discovered, which would have great impact on the further work and results. This subsection aims to describe these issues as they were discovered, and which adjustments to the presented method were made to reach a result as similar to the initial hypothesis as possible.

3.2.6.1 False Impressions Due to a Faulty Sample Counter

Before the data collection, a sample counter was set up with the purpose to state how many samples were collected during each sequence of 30 seconds. This was considered a safeguard to know when to accept or reject a sequence depending on the obtained sampling frequency.

However, after all test sessions had been conducted it was discovered that this counter was defect, and not consistent with the number of samples that could be downloaded from the server. Instead, the number of samples that had reached the server would be highly erratic. For each subject, the first couple of recorded sequences would result in an approved sampling frequency, i.e. within the interval of 50±5 Hz. However, the number of samples would decrease for each recorded sequence, reaching a frequency as low as 10 Hz. This was later understood to be caused by a server capacity limitation due to the free trial used.

The sample counter was later corrected to give the true results, which would be valuable for future tries. The server was also upgraded to a paid version to increase

the capacity and receive data with the desired frequency. Both applications were tested before starting a new data collection.

3.2.6.2 Data Preprocessing with Regards to the New Conditions

Had the activities been performed in the same order for all subjects, without any retakes, the achieved frequency for each activity would be more similar since it would decrease with a rather consistent rate. However, since the activity protocol was conducted in the order preferred by the subject, with retakes due to the faulty sample counter signaling too low sample counts, the frequency of different activities would vary a lot. Furthermore, within the CSV-files of samples from each recorded sequence, there would be slack during periods where no samples were received.

Thus, each file had to be further edited, and solid recordings separated into new files, which would be labeled with their final achieved frequency. Due to the sorting of data, the final dataset would not be balanced, as the total solid recordings would differ between the activities performed.

Those files representing the same activities and with rather similar frequencies would be merged and further preprocessed as explained in *3.2.3 Data Preprocessing*. An exception to this occurred when processing the signals of those data sequences with a lower sampling frequency. It was discovered that the filtering was not as effective as for data sequences of higher sampling frequencies. This was assumed to be due to the cutoff frequency being too close to the sampling frequency, disabling filtration.

Nonetheless, the proposed methodology for preprocessing the data was still considered valuable, since it would calculate measurements based on the general characteristics of the solid value gatherings. As acceleration metrics are naturally fluctuating as earlier mentioned and further described by Lara & Labrador (2013), the result did not necessarily have to be completely distorted. However, it was assumed that it would not either give as accurate results as it would with the initial methodology.

3.2.6.3 Using an External Dataset

To make up for the encountered issues and still being able to evaluate the methodology from the initial prerequisites, an additional external dataset was evaluated. As proposed by Lara & Labrador (2013) each dataset has distinct characteristics which may be either beneficial or unfavorable for a specific algorithm. Therefore, using an additional dataset was also seen as a possibility to explore what results the algorithm would show for a different activity set.

The dataset was collected for a similar research study at Cybercom by Kock & Sarwari (2020). Their study aimed to identify and classify three different activities (jumping, squatting and stomping) as real or fake for a mobile game application. The activities would be defined as fake when the research subject manipulated the

phone in an attempt to simulate the actual activity. The activities are further described in *Appendix B.2.1 Table B.4*.

The set of experiments were carried out by a group of 12 volunteers with ages ranging from 18 to 38. Each person was instructed to perform the three real and fake activities respectively. Worth noticing is that three of the volunteers participated twice. While the research subjects were performing the activities, they were holding an iPhone 6 in landscape mode at shoulder height extended from the body with slightly bent arms.

The data was collected using the smartphone's acceleration and gyroscope at a sampling rate at 50 Hz. Moreover, the attributes from the raw data stream collected from the smartphone can be seen together with a description of the total dataset in *Appendix B.3.1 Table B.6*. Apart from the three-axial acceleration and angular velocity values, and due to the use of an iPhone, the authors did automatically obtain three additional attributes called roll, pitch and yaw. These attributes represent the relative rotation in x-, y-, and z-axis respectively.

We further processed the additional dataset according to *3.2.3 Data Preprocessing*. However, the noise reduction and separation of gravity components were not performed, as this was done automatically through the iPhone. Furthermore, the values for roll, pitch and yaw were not used since two datasets as similar as possible were desired. However, the feature importance of these attributes was evaluated separately as described in *3.2.4 Feature Optimization*.

3.2.6.4 Notations for the Different Datasets

In the following chapters a consistent set of prefixes will be held for the different datasets used, presented in *Table 3.4*.

Table 3.4 The set of prefixes used to describe the different datasets used in this thesis.

Prefix	***Description***
Smart helmet dataset	The dataset collected with the smart helmet as part of this thesis methodology.
External dataset	The dataset collected by Kock & Sarwari (2020) and used for validation in this thesis.
Raw time-series	The original time-series of data of each dataset, before being preprocessed.
Preprocessed	The preprocessed data of each dataset, consisting of the 348 features presented in the methodology.

4 Interview Results

The following chapter aims to describe the empirical findings and results made from interviews and observations in the pre-study with the five different stakeholders which were presented in the methodology chapter: the employees, the employers, the trade union, the legal system, and the technical expert. Naturally, these will contribute differently to the evaluation of the innovation criteria depending on their role and authority.

4.1 Feasibility

From the perspective of the legal system, and consistent with theory, the data processing approach with regards to the smart helmet must follow one of the legal bases of GDPR to be lawful. It is normally not possible to use the basis of consent in power relationships. Hence, this basis cannot be used to motivate the implementation of the smart helmet in construction companies, as employees will be subordinate to employers. However, the processing of personal data could be considered a trade-off between the employee's vital interests and the interests of the employer and any third parties. Furthermore, the relevance and the benefits brought by the smart helmet through higher safety and decreased injury rates may carry more weight than only protecting the employee's personal integrity. (LEG)

Considering the risks of implementing the smart helmet, data may be used by the employer to identify the worker's activity, bringing disadvantages to workers appearing to be less active (LEG). It is therefore important to ensure that the data is only accessible by a few selected people within the organization, or even by an external party such as an occupational health service that the employer may be hiring. Using the latter would make the collected data confidential and not accessible by the employer (TECH).

This was also considered a worthwhile solution by the employer, who expressed an understanding of the complexity of the problem while showing little to no interest in taking part in the data themselves. The construction company is today connected to an occupational health service that performs various mobility and ergonomic check-ups in their routine, which would align well with the introduction of a smart helmet to prevent related injuries. Actions are today only taken on an individual

level if injuries are discovered, but the employer also receives a yearly report on a group level, stating statistics and trends of their employee health (NCC1).

Many employees own a smartphone today, and are therefore used to wearing GPS applications, making them less hesitant about wearing these in their daily work (NCC1). Hence, according to both employers and employees, the sensors of the smart helmet would most likely not be perceived as a threat towards the personal integrity rights (NCC2). Again, the employer states that they want to avoid all situations where the efficiency and performance of employees can be measured, or where this type of data can be stored and analyzed (NCC1).

The employer also expressed that although employees may worry about being supervised while using these types of technological applications, a bigger obstacle would still be the technology resistance among users. The technology resistance is both based on lack of knowledge, but also habitual with a determination of continuing to do things the way they have always been done. The technological adoption of the smart helmet might vary between employees depending on age and level of digitalization, and this might have to be considered if implementing a final solution (NCC1).

From the perspective of the trade union it is also observed that the construction industry today is overall rather analog, especially so on the construction site compared to the planning levels of the construction industry (TU). This confirms the background found on the topic, stating that the construction industry is not as digitalized as other industries are as of today. Many workers have never used digital tools in their daily work, and if the smart helmet would be introduced it is important to simplify any user application that may come with it, to minimize resistance towards it (NCC1). From the technical expert's perspective the outlook is positive, as technology has ramped up in later years. Earlier, these devices needed several sensors and cords connected to them, while it today is easier to gather movement data from smartphones or other unobtrusive devices (TECH).

Furthermore, it is said that several initiatives have been taken towards digitalization which have not always been successful. One example is the ID06 card which is an electronic register to make sure only registered construction workers are at the site, preventing illegal workforce. However, this system has been bypassed with time, which may create a disbelief for new digitalization initiatives to success among employers and employees in the construction industry (TU).

4.2 Desirability

The liability of the trade union includes acting for their members' interests in all areas where the organization can make an impact. This may for example include

negotiations on retiring age or lawful employment contracts for construction workers (TU).

It is reported by the trade union that approximately 80 percent of Byggnads' members belong to small construction companies, where Byggnads is an external actor to which workers can turn for support. The remaining part of the members work at larger construction companies such as NCC, Peab, and Skanska, where Byggnads have internal organizations consisting of elected representatives (TU).

Furthermore, it is noticed by the trade union representative that the hazards which construction workers are exposed to in their daily work is not given a lot of attention, and that few people are aware of the injury and death tolls, compared to those in other industries. This creates a big interest within the trade union and their initiatives for the construction workers wellbeing. Furthermore, it is said that many initiatives are taken towards the trade union members in order to change the industry culture and way of working, e.g. by encouraging workers to de-emphasize the macho culture in the industry (TU).

The overall attitude towards the smart helmet among construction workers was also positive. The interviewees stated that there are many activities in their daily work causing occupational hazards and long-term injuries, where a preventative approach would be greatly appreciated by many. Examples of demonstrated hazards could be heavy lifts and carrying of construction materials, as well as being situated in unergonomic positions for a long period of time (NCC2). The technical expert agreed that there is an overall need for systems and devices which can prevent these types of injuries connected to ergonomics and overload factors (TECH).

When ordering a new helmet, a critical attribute to consider both among construction workers and the purchasing unit is the weight of the helmet. As the work demands a considerate amount of movement, and often so angular movements of the head and neck, the helmet weight is more apparent and may also amplify the damage made to these body parts. More specifically, when construction workers have a liability to always observe the crane when located at a construction site to make sure not to walk under its load. This leads to construction workers spending a lot of time walking while looking upwards, making the helmet weight obvious. Therefore, it would be important to consider this aspect, and to minimize the weight of the IoT-device alone, in case of an implementation of the smart helmet (NCC1; NCC2).

From a legal perspective, the desirability aspect was vaguely discussed. However, it was mentioned that these types of questions are encouraged, as they may set new examples for practice, in a digital environment with rather weak guidelines at this point in time (LEG).

4.3 Viability

As of today, NCC has a central purchasing unit which is assists the divisions with a selected range of workwear and equipment, in which several construction helmets are included. This range is continuously reevaluated to adjust to the market offerings as well as the construction workers' demand. NCC also has a representative group consisting of construction workers, which from time to time try out new clothes and equipment that may be added to the assortment. Today, there are about six helmets included in this assortment, which the construction workers are free to choose among (NCC2).

If the smart helmet would fulfill the requirements that NCC set for their equipment, such as safety certifications and weight, it should follow the same business model as other helmets. Although it is difficult to state any exact numbers, it was expressed that there would be a willingness to pay a higher price than for regular construction helmets if it would increase the safety of workers. If the set up would be developed and sold as an additional application to be mounted onto any existing helmet, this would be considered as a new type of product for which it is more difficult to estimate the willingness to pay among construction companies. However, as this solution would allow using an already approved construction helmet, the purchasing and implementation process was thought to be dramatically shortened. (NCC2)

4.4 Summary

To provide an overview of the interview findings, the opinions considered most apparent among respondents was concluded in *Table 4.1*. Vertically the four interviewed stakeholder groups are listed, and horizontally the feasibility and desirability criteria are listed along with potential risks or limitations. The viability criterion is precluded from the table, as it was considered difficult to draw any conclusions from the data collected in this area, while the smart helmet was assumed to follow the same business model as existing construction helmets.

Table 4.1. A summary of the different stakeholder attitudes with regards to the feasibility and desirability criteria, as well as proposed risks.

	Feasibility	***Desirability***	***Risks***
Employee	Are used to wearing smartphones, making them less hesitant towards sharing data.	Long-term injuries are common and an approach to prevent these is welcomed.	The purpose of the helmet may be misunderstood due to technology resistance.
Employer	Promotes a third party to handle data, avoiding violation of personal integrity rights.	Supportive to equipment improving occupational safety if consistent with certifications and requirements.	The additional helmet weight caused by the IoT-device may exceed what is acceptable for safety reasons.
Trade union	Sensors are accepted if they do not provide video recordings or GPS.	Positive to any initiatives in improving occupational safety.	Digitalization initiatives have sometimes failed, risking a disbelief in new solutions.
Legal system	The interest of the individual must overweigh that of the responsible companies.	It provides a definition of praxis in similar cases where regulations may be ambiguous.	An appropriate responsible unit must be chosen to handle the data.
Technological expert	Proposes an occupational health service to manage the data.	Sees a big need for further development in the science of HAR to lower the injury rate.	

5 Body Movement Analysis Results

In this chapter the results from the movement data analysis and the performance of the Random Forest classification will be presented. First, the classification results and statistics for the smart helmet dataset will be described. Second, the results from a comparative evaluation of the raw time-series external dataset will be presented, followed by the corresponding evaluation of the preprocessed external dataset. Last, a summary of the results and an objective comparison between the two datasets will be presented to reinforce the discussion and analysis following in the next chapter.

5.1 Analysis of the Preprocessed Smart Helmet Dataset

The results obtained from evaluating the smart helmet dataset will be presented in this section. As mentioned in the methodology several issues were encountered during the data collection and processing of data. Therefore, and due to the complexity of this topic the results may not be as intuitive to the reader as they otherwise would be, why they will be followed by a more thorough explanation.

5.1.1 **Acceleration Signal Analysis**

In Figure 5.1 and Figure 5.2 the body acceleration data collected during a walking and lifting movement respectively is illustrated. The illustrated data was obtained from movements within the highest frequency interval reached, i.e. similar to the desired frequency. It is possible to see some repetitive patterns in the acceleration data for both activities, however, it is somewhat irregular which may be due to the variation of frequencies.

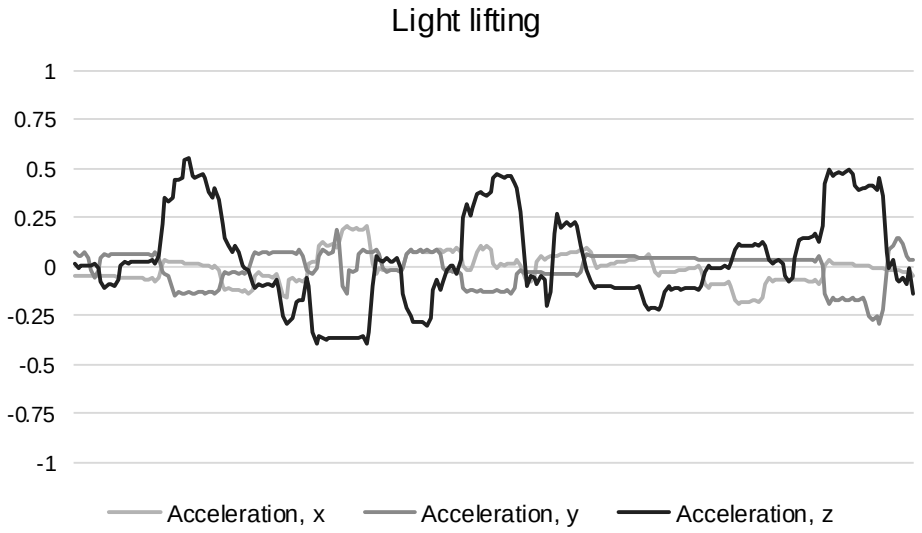

Figure 5.1 An illustration of the separated body acceleration data in three axes, when performing the light lifting activity.

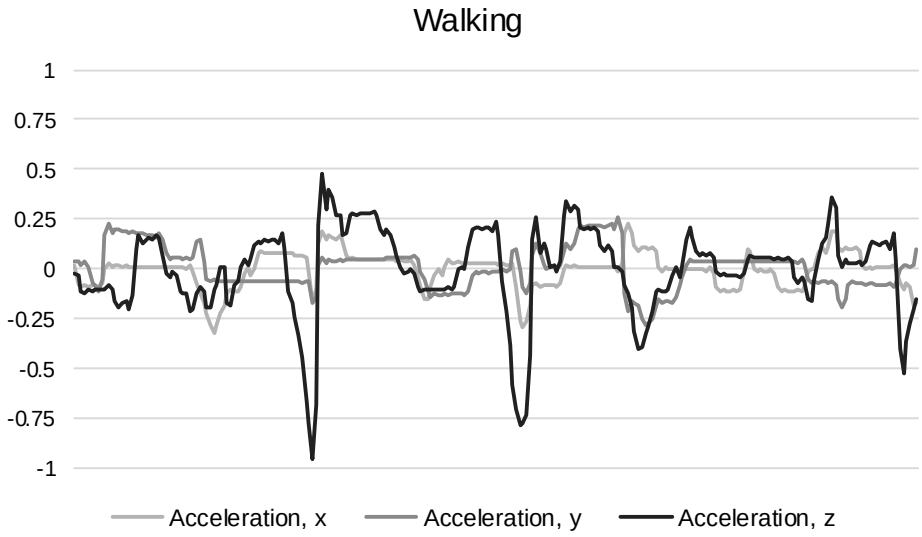

Figure 5.2 An illustration of the separated body acceleration data in three axes, when performing the walking activity.

5.1.2 **Initial Confusion Matrix and Evaluation Metrics**

Furthermore, a confusion matrix was generated with regards to the collected and preprocessed movement data, presented in *Table 5.1*. As explained in the background chapter, the optimal result is reached when the matrix shows as high values as possible along the diagonal from the top-left corner to the bottom-right corner. It is possible to see that the classification results for the smart helmet dataset are relatively high.

The minimum precision of 0.8276 is for *walking while carrying something heavy*, and the maximum is for *walking while looking upwards* with no confusion at all. The remaining activities range from 0.8686 to 0.8832, showing a rather low variance. Furthermore, it can be seen that *walking while carrying something heavy* is most commonly mistaken for *walking*, and vice versa. A rather high confusion can be seen between the two lifting movements as well. Likewise, the recall ranges from 0.8182 to 0.8936 for all activities but *walking while looking upwards*, which scores 0.9881.

Table 5.1 The confusion matrix obtained from the initial classification results on the preprocessed smart helmet dataset.

	Light lifting	***Heavy lifting***	***Walking***	***Walking, carrying heavy***	***Walking, looking upwards***	***Recall***
Light lifting	**119**	20	1	0	0	0.8500
Heavy lifting	18	**168**	2	0	0	0.8936
Walking	0	1	**121**	14	0	0.8897
Walking, carrying heavy	0	3	13	**72**	0	0.8182
Walking, looking upwards	0	0	0	1	**83**	0.9881
Precision	0.8686	0.8750	0.8832	0.8276	1.0000	

The additional evaluation metrics derived from the confusion matrix results are presented in *Table 5.2*. It can be seen that the accuracy, precision, recall, and F1-score all range from 0.8850 to 0.8851. Meanwhile, the MCC shows a bit lower score at 0.8528, which may be explained by the adjustment it does to the imbalanced dataset. However, all metrics are relatively high, especially so with regards to the limitations during the data collection.

Table 5.2 The overall evaluation metrics when calculated for the initial classification performance of the preprocessed smart helmet dataset.

Accuracy	*Precision (Weighted)*	*Recall (Weighted)*	*F1-score (Weighted)*	*MCC*
0.8850	0.8851	0.8850	0.8850	0.8528

5.1.3 **Feature Optimization Confusion Matrix and Evaluation Metrics**

Lastly, the feature importance was evaluated. Because of the high number of features each one will not be presented. However, a list of the five highest importance scores are presented together with those that scored zero in *Table 5.3*. In total, ten features were eliminated, resulting in 338 features. It can be seen that several of the gravitational force vectors had high importance, and especially so in the x-axis. By looking at the raw time-series signals, it is apparent that the movements in the x-axis are the most significant, why it may also create the most significant results. Among those features that scored zero these are exclusively entropy and maximum index vectors for various attributes.

Table 5.3 The most and least important features when performing a feature optimization on the preprocessed smart helmet dataset. A more detailed description of each feature can be found in *Appendix B.3.2 Table 8-11*.

Most important features	*Score*	*Least important features*	*Score*
tGravityAcc-mean()-X	0.040385	tBodyAccJerk-entropy()-X	0.00000
tGravityAcc-max()-X	0.033889	fBodyAcc-entropy()-X	0.00000
tGravityAcc-sma()	0.026421	fBodyAcc-entropy()-Y	0.00000
tGravityAcc-min()-X	0.021605	fBodyAcc-entropy()-Z	0.00000
fBodyGyro-energy()-Y	0.020515	fBodyAccJerk-entropy()-Y	0.00000
		fBodyAccJerk-entropy()-Z	0.00000
		fBodyAccMag-maxInds()	0.00000
		fBodyBodyAccJerkMag-maxInds()	0.00000
		fBodyBodyGyroMag-maxInds()	0.00000
		fBodyBodyGyroJerkMag-maxInds()	0.00000

After evaluating the dataset again, but excluding unimportant features, the confusion matrix presented in *Table 5.4* was obtained. The precision is increased for light lifting and walking, equal for walking while looking upwards, while decreased for heavy lifting and walking while carrying something heavy.

The recall for *light lifting* and *walking while carrying something heavy* is almost precisely the same as the result obtained before the feature optimization. Furthermore, it is rather increased for *heavy lifting*, slightly increased for walking while looking upwards, and slightly decreased for *walking*. The sum of the predictions for each activity differs from the original classification presented in *Table 5.1*, as the training and testing data was randomly partitioned between the two classifications.

Table 5.4 The confusion matrix obtained from the classification results of the preprocessed smart helmet dataset, after a feature optimization resulting in 338 features.

	Light lifting	***Heavy lifting***	***Walking***	***Walking, carrying heavy***	***Walking, looking upwards***	***Recall***
Light lifting	**136**	22	0	2	0	0.8500
Heavy lifting	7	**166**	1	0	0	0.9540
Walking	2	0	**131**	18	0	0.8733
Walking, carrying heavy	0	6	8	**63**	0	0.8181
Walking, looking upwards	0	0	0	0	**73**	1.0000
Precision	0.9379	0.8558	0.9357	0.7590	1.0000	

The overall evaluation metrics do increase slightly after the feature optimization, as can be seen in *Table 5.5*.

Table 5.5 The overall evaluation metrics after feature optimization when calculated for the classification results of the preprocessed smart helmet dataset, together with their absolute increase from the previous results in Table 5.2.

Accuracy	***Precision (Weighted)***	***Recall (Weighted)***	***F1-score (Weighted)***	***MCC***
0.8961 (+0.0111)	0.9003 (+0.0152)	0.8960 (+0.0110)	0.8965 (+0.0115)	0.8674 (+0.0146)

5.2 Analysis of the Raw Time-Series External Dataset

The results obtained from the evaluation of the raw time-series external dataset will be presented in the following subsections. This dataset will exclude the roll, pitch, and yaw attributes to simulate the smart data set as much as possible. First, the general movement data characteristics will be presented below as was done in *5.1 Analysis of the Preprocessed Smart Helmet Dataset*. In the following two

subsections the results from the evaluation of the raw time-series external dataset with and without the roll, pitch, and yaw attributes included will be presented, to evaluate the importance of these attributes which were only included in the external dataset.

5.2.1 **Acceleration Signal Analysis**

In Figure 5.3, Figure 5.4 and Figure 5.5 the body acceleration data collected during jumping, squatting, and stomping respectively is illustrated. It is for all tree activities possible to distinguish repetitive movement patterns, where the body acceleration in x- and z-axis is especially apparent.

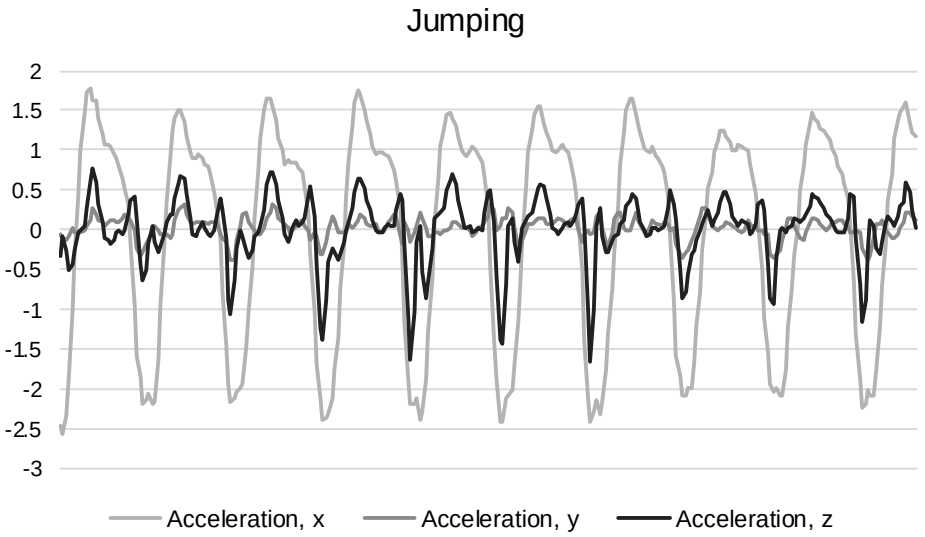

Figure 5.3 An illustration of the separated body acceleration data in three axes, when performing the true jumping activity.

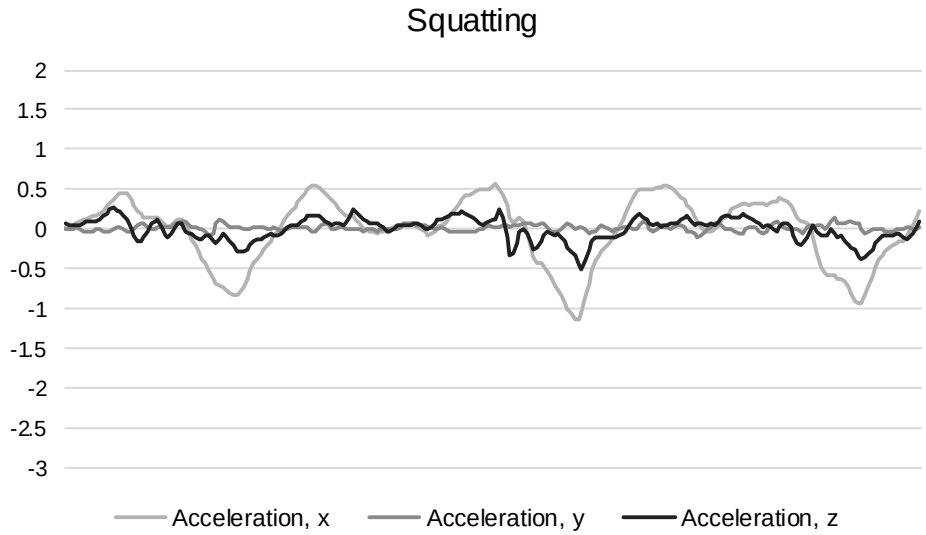

Figure 5.4 An illustration of the separated body acceleration data in three axes, when performing the true squatting activity.

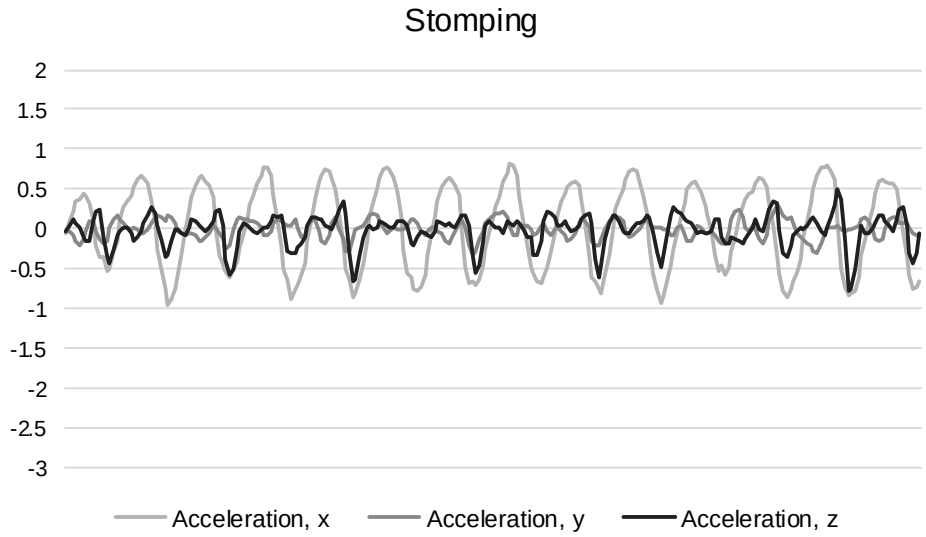

Figure 5.5 An illustration of the separated body acceleration data in three axes, when performing the true stomping activity.

5.2.2 **Initial Confusion Matrix and Evaluation Metrics**

As the external raw time-series of data also contains the attributes for roll, pitch, and jaw, each evaluation step of this subsection will be presented in two parts. The first will exclude the additional attributes, to simulate the smart helmet dataset. The

second part will include the additional attributes to illustrate the difference they make.

5.2.2.1 Excluding Roll, Pitch and Yaw Attributes

The confusion matrix with regards to the raw time-series external dataset is presented in *Table 5.3*. As can be observed, the precision ranges from 0.7321 to 0.8977 for all activities but jumping, which only reaches 0.5317. The precision is specifically high for the activities labeled as fake, ranging from 0.7955 to 0.8977. Likewise, the recall ranges from 0.7068 to 0.8482 for all activities but *fake stomping*, which only reaches 0.6010. Regarding the recall it is not possible to see any significant difference between true and false movements.

Table 5.3 Confusion matrix of the classification results on the raw time-series external dataset, with the roll, pitch and yaw attributes excluded.

	Fake jumping	***Squatting***	***Fake stomping***	***Jumping***	***Fake squatting***	***Stomping***	***Recall***
Fake jumping	**3796**	119	150	152	197	142	0.8332
Squatting	102	**3928**	14	155	208	224	0.8482
Fake stomping	187	34	**4038**	2153	107	101	0.6010
Jumping	159	278	108	**3305**	176	650	0.7068
Fake squatting	330	361	117	160	**3353**	254	0.7329
Stomping	130	249	71	291	174	**3747**	0.8037
Precision	0.8070	0.7905	0.8977	0.5317	0.7955	0.7321	

The additional evaluation metrics derived from the confusion matrix results are presented in *Table 5.5*. The accuracy, precision, recall, and F1-score all range from 0.7993 to 0.8010 when evaluating the data with roll, pitch, and yaw attributes excluded, showing significantly similar values in comparison to each other. The MCC is 0.7600, hence lower than the other measurements which may be due to the dataset not being perfectly balanced.

Table 5.5 The overall evaluation metrics when calculated for the resulting classification performance of the raw time-series external dataset, with the roll, pitch and yaw attributes excluded.

Accuracy	*Precision (Weighted)*	*Recall (Weighted)*	*F1-score (Weighted)*	*MCC*
0.7997	0.8010	0.7998	0.7993	0.7600

5.2.2.2 Including Roll, Pitch and Yaw Attributes

The confusion matrix with regards to the raw time-series external dataset is presented in *Table 5.6*. It can be observed that the precision is ranging from 0.8225 to 0.9587. As for the classification with roll, pitch, and yaw excluded, the precision is specifically high for the activities labeled as fake, ranging from 0.9241 to 0.9587. The precision for all activities is noticeably higher when including roll, pitch, and yaw if compared to *Table 5.3*. Recall ranges from 0.8359 to 0.9534, which again is higher than the respective values in *Table 5.3*.

Table 5.6 Confusion matrix of the classification results on the external raw time-series external dataset, with the roll, pitch and yaw attributes included.

	Fake jumping	*Squatting*	*Fake stomping*	*Jumping*	*Fake squatting*	*Stomping*	*Recall*
Fake jumping	**4143**	77	41	116	102	73	0.9101
Squatting	46	**4345**	2	26	46	142	0.9431
Fake stomping	48	2	**4439**	70	45	52	0.9534
Jumping	82	104	67	**3809**	79	416	0.8359
Fake squatting	87	127	48	104	**4129**	199	0.8794
Stomping	64	212	33	191	67	**4087**	0.8782
Precision	0.9268	0.8927	0.9587	0.8825	0.9241	0.8225	

The additional evaluation metrics derived from the confusion matrix results are presented in *Table 5.7*. When including roll, pitch, and yaw attributes, the accuracy, precision, recall, and F1-score range from 0.9001 to 0.9013. Furthermore, the MCC is 0.8804 which again is lower than the other measurements due to the unbalanced dataset. Similar to the confusion matrix results, the scores are overall higher for the dataset including roll, pitch, and yaw attributes.

Table 5.7 The overall evaluation metrics when calculated for the resulting classification performance of the raw time-series external dataset, with the roll, pitch and yaw attributes included. Together with their absolute increase from the previous results in Table 5.5.

Accuracy	*Precision (Weighted)*	*Recall (Weighted)*	*F1-score (Weighted)*	*MCC*
0.9001 (+0.1004)	0.9013 (+0.1003)	0.9001 (+0.1003)	0.9002 (+0.1009)	0.8804 (+0.1204)

5.2.3 **Feature Optimization**

Lastly, the feature importance is also presented in two parts, excluding and including the roll, pitch, and yaw attributes.

5.2.3.1 Excluding Roll, Pitch and Yaw Attributes

The feature importance with the roll, pitch, and jaw attributes excluded is presented in Figure 5.6, in which the most significant feature is body acceleration in the x-axis, scoring 0.149679. The remaining features would range from 0.082299 for gravitational force in the x-axis, to 0.129519 for angular velocity in the y-axis. However, since none of the features were considered unimportant, the already presented result maintains.

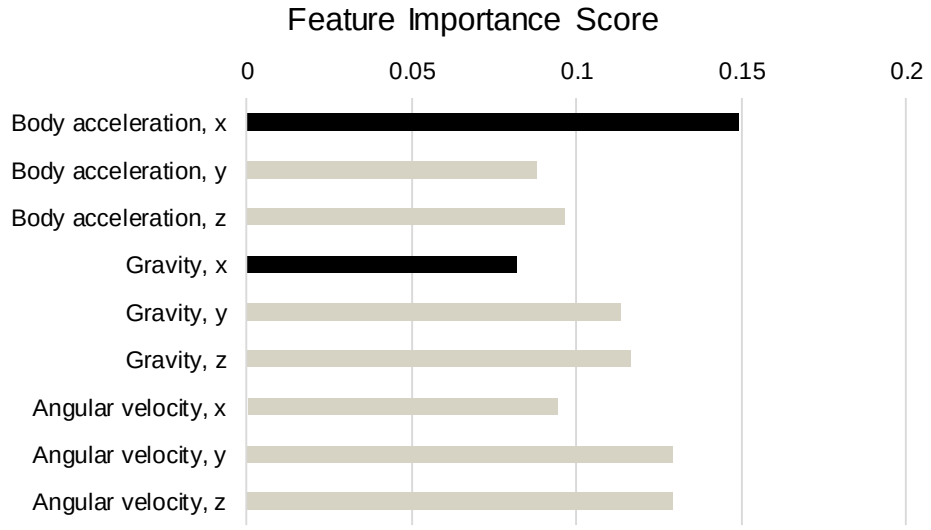

Figure 5.6 The feature importance when evaluating the raw time-series external dataset, with roll, pitch and yaw attributes excluded. The highest and lowest values are marked in black.

5.2.3.2 Including Roll, Pitch and Yaw Attributes

When including the roll, pitch, and jaw attributes, as seen in Figure 5.7 the scoring is observed to be clearly different. The yaw attribute shows significant importance, scoring 0.176285 followed by body acceleration in the x-axis, scoring 0.120188. The lowest score was seen for gravitational force in the x-axis, with a score of 0.052661. Similar to above, none of the features were considered unimportant, hence the already presented result maintains.

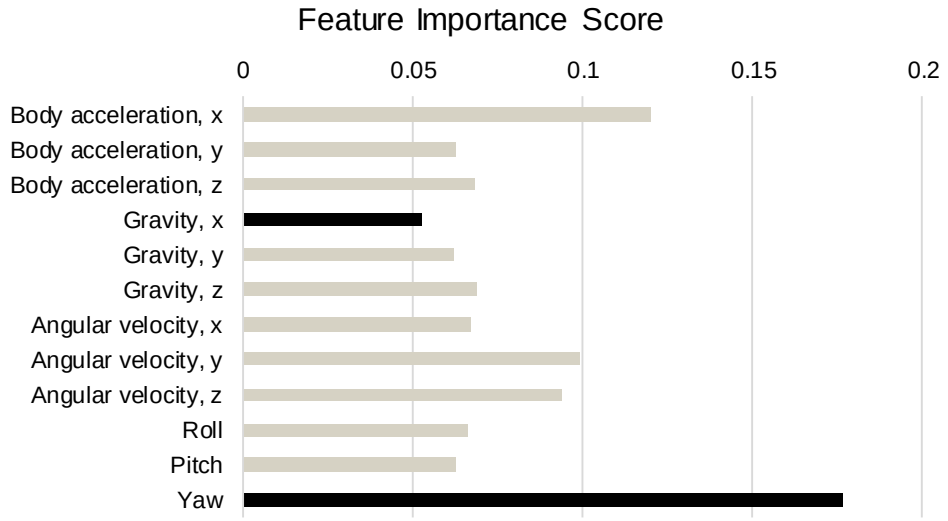

Figure 5.7 The feature importance when evaluating the raw time-series external dataset, with roll, pitch and yaw attributes included. The highest and lowest values are marked in black.

5.3 Analysis of the Preprocessed External Dataset

The results obtained from the evaluation of the preprocessed external dataset will be presented below. This dataset was as earlier mentioned used to validate the feature extraction methodology. First, the confusion matrix and evaluation metrics will be presented. Second, the feature importance will be evaluated for the preprocessed dataset.

5.3.1 **Initial Confusion Matrix and Evaluation Metrics**

The confusion matrix with regards to the preprocessed data of 348 features is presented in *Table 5.8.* As can be observed, the precision reaches the maximum of 1.0000 for *squatting, fake stomping,* and *jumping*. Furthermore, it ranges from

0.9474 to 0.9759 for the remaining three activities. Any difference between true and fake activities cannot be seen. The recall does too reach the maximum of 1.000 for *squatting*, *fake stomping*, and *stomping*. For the remaining three activities it ranges from 0.9467 to 0.9740.

Table 5.8 Confusion matrix of the classification results on the preprocessed external dataset.

	Fake jumping	***Squatting***	***Fake stomping***	***Jumping***	***Fake squatting***	***Stomping***	***Recall***
Fake jumping	**75**	0	0	0	2	0	0.9740
Squatting	0	**65**	0	0	0	0	1.0000
Fake stomping	0	0	**59**	0	0	0	1.0000
Jumping	0	0	0	**71**	0	4	0.9467
Fake squatting	3	0	0	0	**81**	0	0.9643
Stomping	0	0	0	0	0	**72**	1.0000
Precision	0.9615	1.0000	1.0000	1.0000	0.9759	0.9474	

The additional evaluation metrics derived from the confusion matrix results are presented in *Table 5.9*. It is observed that all scores range from 0.9792 to 0.9797, where the MCC also is similar to the other evaluation metrics.

Table 5.9 The overall evaluation metrics when calculated for the resulting classification performance of the preprocessed external dataset.

Accuracy	***Precision (Weighted)***	***Recall (Weighted)***	***F1-score (Weighted)***	***MCC***
0.9797	0.9792	0.9797	0.9792	0.9792

5.3.2 **Feature Optimization Confusion Matrix and Evaluation Metrics**

Lastly, the feature importance was evaluated. Again, due to the high number of features, a list of the five highest scores are presented in *Table 5.10*, while 38 features scored zero and will not be presented. Among these, 28 features were of the entropy attribute, and four of the max index attribute.

Table 5.10 The most important features when performing the feature optimization on the preprocessed external dataset. A more detailed description of each feature can be found in *Appendix B.3.2 Table 8-11*.

Most important features	***Score***
tGravityAcc-std()-Z	0.027169
tGravityAcc-mad()-Z	0.025138
tBodyGyroMag-sma()	0.024024
tGravityAcc-iqr()-Z	0.022622
fBodyBodyGyroJerkMag-meanFreq()	0.019296

After evaluating the dataset again, but excluding unimportant features, the confusion matrix presented in *Table 5.11* was obtained. The precision has increased for fake squatting, while it has decreased slightly for fake jumping and squatting. The recall for fake jumping has increased, and the result for fake squatting has an increased marginally. Furthermore, the recall has slightly decreased for jumping.

Table 5.11 The overall evaluation metrics when calculated for the resulting classification performance of the preprocessed external dataset, after feature optimization resulting in 310 features.

	Fake jumping	***Squatting***	***Fake stomping***	***Jumping***	***Fake squatting***	***Stomping***	***Recall***
Fake jumping	**80**	0	0	0	0	0	1.0000
Squatting	0	**73**	0	0	0	0	1.0000
Fake stomping	0	0	**64**	0	0	0	1.0000
Jumping	0	0	0	**66**	0	4	0.9429
Fake squatting	4	0	0	0	**72**	0	0.9474
Stomping	0	0	0	0	0	**69**	1.0000
Precision	0.9523	1.0000	1.0000	1.0000	1.0000	0.9452	

The overall evaluation metrics do increase slightly except for MMC after the feature optimization, as can be seen in *Table 5.12*.

Table 5.12 The overall evaluation metrics when calculated for the resulting classification performance of the external preprocessed dataset, after feature optimization resulting in 310 features. Together with their absolute increase from the previous results in Table 5.9.

Accuracy	*Precision (Weighted)*	*Recall (Weighted)*	*F1-score (Weighted)*	*MCC*
0.9815 (+0.0018)	0.9824 (+0.0032)	0.9815 (+0.0018)	0.9815 (+0.0023)	0.9780 (-0.0012)

5.4 Summary

From the presented results above it can be seen that the evaluation metrics from the preprocessed external dataset is significantly better than the ones obtained when using the raw time-series of data, both when including and excluding the roll, pitch and, yaw attributes. The results from using the preprocessed smart helmet dataset outperform the results of the raw time-series external dataset when excluding the roll, pitch, and yaw attributed, but do not reach the performance of the preprocessed external dataset. Neither do they reach the performance of the raw time-series external dataset with roll, pitch and yaw included.

Furthermore, from the evaluation of the preprocessed smart helmet dataset some confusion can be seen within the two main groups of activities, i.e. lifting and walking, apart from for the activity *walking while looking upwards* which shows high precision and recall. From the raw time-series external dataset it is possible to see that the fake movements are better distinguished than the true movements. In cases of confusion, the fake activities are randomly misclassified, and not specifically with their respective true activities.

Finally, the feature importance shows that the yaw attribute is of high importance for the raw time-series external dataset. When comparing the most important features of the two preprocessed datasets, the results will vary. However, the least important features will mainly be represented by entropy and max index attributes.

6 Discussion

In this thesis, the construction industry environmental barriers in which a smart helmet would be implemented is presented and evaluated, with regards to five stakeholder perspectives. Furthermore, the performance of the Random Forest algorithm is evaluated, when analyzing human activity data collected from sensors in the smart helmet. This chapter will provide a discussion based on the results achieved from the research questions of this thesis. First, the results of each question will be separately discussed to give the reader a deeper understanding of their significance. Further, these will be summarized to conclude the discussion. Lastly some future research areas that are interesting for further investigation will be presented.

6.1 The Environmental Barriers for the Smart Helmet

This section will present a discussion related to the research question asking what the environmental barriers in the Swedish construction industry are for the smart helmet. The discussion is based on the conducted literature review and the interview results, to provide the reader with new perspectives on the obtained results.

6.1.1 **Initiatives against Occupational Injuries**

As disclosed by the background chapter of this thesis, the construction industry is facing a major challenge in decreasing the number of occupational injuries. From the empirical results, it is also clear that occupational injuries are an obvious problem in the daily life of construction workers. This awareness can be acknowledged by the clear communication both employers and employees had regarding the helmet weight, one main attribute to consider when choosing between construction helmets. As a heavier helmet could cause overload injuries to the neck, choosing a light-weight helmet is an active choice to avoid injuries.

However, there is little engagement expressed to change the overall situation in the industry from the individual construction workers' perspective, although external initiatives are welcomed. This may be due to three major factors discovered.

Firstly, construction work has historically used rather analog techniques, demanding heavy physical labor, and hence causing injuries. Going so far back in time, this may have become an accepted condition by workers, and therefore little is done to prevent it.

Secondly, from a cultural perspective, this issue could be amplified by the construction industry being heavily dominated by men. This creates a macho culture, where it may be less accepted among construction workers to observe and take action towards perceived physical issues. Furthermore, it may generate a stubbornness to manage tasks oneself, which in turn can lead to a higher injury rate.

Thirdly, from a knowledge perspective, construction workers may miss vital knowledge of various prevention methods to avoid injuries, but also of the technological possibilities that enable prevention, resulting in a passive attitude towards injury prevention. However, as the attitude towards external initiatives for injury prevention is positive, the imposition of preventative methods such as the smart helmet should be possible as long as it does not demand too much of the construction worker.

6.1.2 **Technology Resistance**

While digitalization is a strong driver in the development of other industries, the same cannot be said for the construction industry. The backlog of technological initiatives, and even a pronounced technology resistance was apparent from the empirical results. As previously mentioned, when discussing the lack of engagement among construction workers towards preventative methods, the issue of technology resistance may also be due to the analog history of the construction industry in combination with the lack of knowledge regarding technological applications. These results may potentially cause hindrance in the implementation of the smart helmet. However, similar to the contention above, it could be bridged by not demanding too much of the construction workers with regard to physical interaction in the use of the smart helmet. Furthermore, communicating the underlying purpose of the helmet and why it should be adopted towards construction companies and their employees would be crucial for maximizing the retention rate.

Although it is said that the technology resistance is high within the construction industry, the employers also state that most employees own a smartphone and are therefore used to that certain level of technology. However, it is understandable that it might be different to use new technologies in private than to start using it in work tasks that are otherwise not as digitalized.

Nonetheless, the implications of a technology resistance in the construction industry as a whole can be further discussed. It may set an even higher barrier to the implementation than the individual construction worker’s attitude will. This, as the construction industry as a whole sets the framework for the internally recommended

regulations and principles. This will obviously also affect which equipment and technological systems will be introduced in construction companies, such as when the assortment of construction helmets is evaluated.

6.1.3 **Privacy Issues**

It is stated by the employer that they have no interest in handling the data collected from the helmet. This may be interpreted as them only being interested in obtaining the true purpose of the smart helmet, i.e. preventing injuries among their employees. If this is the case, it implies that the risk of employers misusing data, e.g. to monitor the efficiency of employees, is eliminated. However, it could also be interpreted as the employer not having the resources or knowledge to handle that type of data, resulting in a lack of interest to engage in it.

Regardless of the employer's intention, if the principles of privacy by design proposed by Cavoukian (2010) are followed when implementing the helmet, the privacy issue should be eliminated. It would ensure the setup of a comprehensive structure for handling the data between convenient responsible, without sharing the information with potentially harmful stakeholders. Furthermore, as discussed by Johansson Stålnacke & Pettersson (2016) it is important to understand the difference in attitudes between the sender and receiver, regarding what type of data is more sensitive to share.

Interview results did also show that one important factor to consider is how the collection of data is interpreted among employees. It was emphasized that the employee should not feel directly monitored, e.g. through GPS or video recordings where data is more descriptive. This was also an argument by Lara & Labrador (2013) for not using external sensors in HAR systems. As the movement (i.e. acceleration and angular velocity) data from the smart helmet can be considered more difficult for a layman to draw any conclusions from, the collection of movement data through the sensors on the smart helmet could be more acceptable.

Considering the empirical results presented from a legal perspective, this approach should also support the legal basis saying that the interest of and value for the employee should weigh heavier than that of the organization or system processing and storing the data.

The organization proposed by interviewees to be responsible for this is the potential occupational health service that employers may hire. Since this organization has no power relationship with the employee, its interest in the data can be neglected. Furthermore, information processed by health services is confidential, meaning that it cannot be shared with the employer. However, this idea would demand from construction companies that they are hiring an occupational health service to enable the use of the helmet, which may not be the case for all companies.

Nonetheless, introducing a smart helmet in the construction industry could make a great example for the further digitalization of the industry, and for new practices related to privacy by design if done right. Furthermore, as mentioned in the interview results, it provides a definition of praxis in similar cases where regulations may be ambiguous, which could be valuable and encouraging future developments.

6.2 Body Movement Analysis Performance

In this section the performance of the body movement analysis with regards to the preprocessing method and the Random Forest classifier algorithm will be discussed. The results of both datasets used will be analyzed and compared to each other, with a discussion on why the differences would occur.

6.2.1 **Activity Set and the Complexity of the Activities**

When comparing the two activity sets with regards to their recognition performance, several distinctive results can be seen. As the overall recognition performance of the preprocessed external dataset was higher than that for the smart helmet dataset, a contributing factor to this could be the complexity of activities in the latter. The activities of Kock & Sarwari are rather distinguishable with regards to their frequency and amplitude, while those in the smart helmet dataset are more overlapping and/or complex to distinguish within the two main groups of activities.

Furthermore, when performing the feature optimization it was apparent that the most important features would differ between the two datasets. This argues for different characteristics in the activity set, why the recognition performance also may differ between them.

Since the misclassification of fake activities in the raw time-series dataset of Kock & Sarwari was rather randomly distributed while the classification overall was experiencing high precision, it could be argued that the true and corresponding fake activities all could be perceived as different activities with regards to their acceleration and angular velocity data. This would also contribute to their higher overall evaluation metrics compared to those of the smart helmet dataset.

From the dataset collected with the smart helmet, the activity with the highest precision and recall was *walking while looking upwards*. It was initially hypothesized that this activity would be confused with the other two walking activities, which was rejected by the results. This can be explained by the tilt of the neck when looking upwards, causing the axes in which the acceleration and angular velocity are measured to shift. From this, it can be concluded that the algorithm should be able to distinguish between similar movements performed with the head in different positions. Recognizing differences as these could be argued for being

an advantage in the recognition of movement patterns with the head in a harmful position, such as when monitoring the crane at a construction site which was presented as a cause of injuries in the interview results.

6.2.2 **Choice of Sensor Attributes and Their Positioning**

From the interviews it was suggested that the sensors on the smart helmet should not cause any trouble as long as they do not provide video recordings or GPS. Although this was expressed as a condition for protecting the privacy of construction workers, it is also supported from a technological perspective, as the additional sensors would cause a substantial increase of the power and computational expenditures without necessarily creating any extra value.

However, it can be discussed whether the chosen sensor positioning is optimal to recognize more advanced motions. The experiment might generate more consistent results from the sensor being positioned elsewhere, or by using additional sensors. For example, when analyzing the lifting activities, a wrist positioned sensor might give more accurate results, as proposed by Lara & Labrador (2013). Adding this sensor would also give information about the arm movement and could provide possibilities to observe how the arm strength reacts to the lifting activities depending on different weights. Nonetheless, this would also be a trade-off considering the obtrusiveness of the system, as it would demand users to wear not only a helmet but also some sort of wrist-mounted device.

Furthermore, the roll, pitch, and yaw attributes included in the initial raw times-series dataset of Kock & Sarwari can be discussed to be valuable for this type of movement analysis. Partly to get a more precise result, but also because these rotations may be valuable in the process of understanding movements. This, since they can detect e.g. harmful rotations of the head or neck in the three axes, which may be essential in predicting and preventing injuries. After analyzing the external dataset it was especially apparent that yaw, i.e. the relative angle rotation around the x-axis, is a significant feature for the concerned activities. Furthermore, the improvement of the results when including these attributes in the raw time-series data analysis was apparent. However, roll and pitch scored lower in feature importance and may therefore not have an as big impact on the algorithm result.

Despite this, the additional value added by using a magnetometer could be discussed in future developments of the setup, to evaluate if it worth the extra financial and energy expenditure as well as increasing the physical size of the device. Furthermore, the importance could not be validated by also analyzing the smart helmet dataset and may therefore not apply to all types of activities.

6.2.3 **The Training Data Quality and Quantity**

The overall quality of the two datasets would differ for two reasons: the initial purpose of their use, and the limitations that occurred during the data collection of the smart helmet dataset. This would in turn affect both the quality and quantity of the training data.

For the smart helmet dataset, the sampling frequency would vary between approximately 10 to 50 Hz, affecting the data quality, and implying that the time windows of 128 samples would contain a varying number of movement cycles depending on the average frequency of one window. This might have contributed to some misclassifications in the recognition performance. Furthermore, as the sampling frequency sometimes only reaches approximately 10 Hz at some points, it can be questioned whether this was enough to capture small differences in movements.

However, from the results it is visible that the recognition performance was in first hand high between the main groups of activities, but also rather high within the groups. This argues for a good recognition performance also for small differences between movements, which could most likely be improved if using a higher and fixed sampling frequency. However, it is difficult to say whether the sensors could catch up even smaller differences than those in the activities presented, for example when distinguishing between lifting movements of several different weights.

As already discussed, more sensors would possibly distinguish the differences, but a higher sampling frequency could possibly do so too. This, without risking obtrusiveness with regards to the size and weight of the IoT-device, which was discussed to be an important factor when purchasing a construction helmet.

Especially when considering the different types of walking, a higher sampling frequency might be needed to for example distinguish whether a person is carrying weights could be distinguished through a more apparent wobbling from side to side. However, these movement characteristics could vary in their distinctiveness among individuals, and thereby still be mistaken for each other.

The quality of the external dataset will not be commented on, as the sampling was performed by Kock & Sarwari, and it is assumed that the iPhone used for data collection was well functioning. Furthermore, the sampling showed accurate results, with easily identifiable movements from the signal analysis of the body acceleration data. Although the signals from the smart helmet dataset showed repetitive movement patterns, they were not as easily distinguishable as those of the external dataset. This could be explained by at least three reasons: (1) the frequency was not constant for the illustrated signal as several samples of an average frequency were merged, causing variations in the cycle time, (2) the accelerometer and gyroscope may not have been positioned correctly or not set in a steady position during the data collection, possibly causing distortion to the data collected for each axis, either from angular misplacement or vibrations, and (3) the noise filtering and separation

of the body acceleration may not have been correctly conducted, possibly causing data to be mistakenly eliminated.

Lastly, the quantity of data would differ for the two datasets, also affecting the final results. For both lifting activities and walking, the data quantity would be larger than for remaining activities in both datasets with a factor of two to three. Although the smart helmet dataset was overall larger than the external dataset for all activities, the higher imbalance between activities may have caused confusion which resulted in lower classification results. Considering the high classification results of the external dataset, the quantity of data can be considered sufficient for the corresponding activity set. However, it is not possible to tell without further experiments whether the quantity is high enough also for the smart helmet dataset.

6.2.4 **The Preprocessing Method**

It is possible to argue that the data preprocessing method used makes significant improvements to the algorithm's performance.

Firstly, the classification results of the external dataset were improved when extracting new features from the raw time-series data when roll, pitch, and yaw attributes are excluded. The performance of the algorithm improved even further when the features of the least importance for the prediction were eliminated.

Secondly, the classification results of the smart helmet dataset were satisfying with regards to the encountered limitations during the data collection, and were similarly improved when eliminating the features of least importance. However, as discussed in previous sections the results of the classification for the smart helmet dataset could most likely be enhanced if the data quality was higher.

Among the features eliminated due to low importance, a prevailing majority were of entropy and max index attributes, disclosing that these may not be relevant to consider in the first place. However, it could be further evaluated whether the unimportance is general, if it depends on the characteristics of the datasets, and/or if the programmed formula for calculating these is faulty.

The method of analyzing the raw time-series dataset of Kock & Sarwari (2020) showed good results specifically in recognizing the activities labeled as fake. This distinction could not be seen for the preprocessed dataset. However, this may be due to the evaluation metrics for the preprocessed dataset being so high that almost all activities were correctly classified.

The overall evaluation metrics did not show any significant variations to be commented on for any of the datasets. It was shown that the MCC was higher and more similar to the other metrics when preprocessing the external dataset, which would advocate for using the proposed method also when working with slightly unbalanced datasets such as that of Kock & Sarwari. The improvement could not be

validated with the results of the smart helmet dataset. However, the MCC did not either differ noticeably from the other evaluation metrics of the smart helmet dataset which was even more unbalanced. Hence, the method could be considered suitable also for somewhat unbalanced datasets.

6.3 Summary

Although occupational injuries are a problem in the daily life of construction workers, there is barely any engagement expressed to change the situation from an individual perspective. However, the attitude towards external initiatives for injury prevention is positive. In the case of the smart helmet, it is crucial that the purpose of the smart helmet, i.e. prevention of harmful movement patterns, is communicated clearly both to the construction companies and their employees, since most construction helmets today are chosen based on its weight. It is therefore important to minimize the added weight of the IoT-device for it to be overlooked for its additional value of safety provided by the smart helmet.

Further, it can be argued if adding a magnetometer to the smart helmet to obtain the yaw values only would increase the additional weight without offering more value to the Random Forest classifier. The same argument can be held if other sensors were to be added to collect more data about the user. Further, it is also a question regarding how much information should be collected with the smart helmet since the employees do not want to feel directly monitored. However, since the current setup of the smart helmet is more accepted by the employees than data obtained from GPS and/or video recording there should not be a problem regarding privacy. Moreover, if the principle of privacy by design is applied, and by giving the responsibility of the data to the occupational health service, the issue of feeling monitored can be further eliminated.

The performance of the classification of the Random Forest is satisfactory. The classification result can however be further improved since the data obtained from the smart helmet were insufficient. However, the algorithm was effective to distinguish between movements with greater difference (walking vs lifting) and it could differentiate more similar movements (light lifting vs heavy lifting) too. Further, the Random Forest algorithm showed even more precise results when predicting movements involving a change in the position of the head and/or rotation of the neck (walking, while looking upwards). Lastly, the method used to preprocess the data made significant improvements to the algorithm’s performance.

6.4 Future Research

For future research, several topics were during the process of this report found interesting for further work. Below, a number of ideas are presented that could act subject to developments of this project, related to the technological aspects.

To reach the future purpose of the smart helmet, i.e. being able to prevent long-term and irreversible injuries from overload factors among construction workers, two criteria must be reached. First, it must be ensured that the helmet can also tell the difference between movements in a natural environment with different individual characteristics included in the data collection. Furthermore, to enable the identification of harmful movement patterns, this work would have to be combined with the science of ergonomics to develop methods for distinguishing characteristics of harmful activities.

It is discussed whether more sensor attributes would reach a higher recognition performance, or if they would have to be positioned on additional body parts to do so. Therefore, in future works it would be interesting to evaluate the limit where no additional value is generated from extra sensor attributes to the helmet.

It has not yet been discovered whether the algorithm could identify more complex movements or overlapping movements. Neither has it been evaluated how well it would work over a floating time period, where movement data would be cut into time windows according to their natural happening, causing transition errors to some of the windows. This could do as a first step towards adjusting the setup for a more natural environment. An even further step would be to enable the analysis of real-time data collected from the construction site, and simultaneously communicate warnings of harmful movement patterns to the construction worker. However, these kinds of techniques would also demand an extremely large and historically collected training dataset combined with knowledge in ergonomics as previously mentioned.

An alternative way of processing the data which could be subject to further research is to use the integration device, i.e. the built-in processor, to store and process the data instead of sending it to a server. This would reduce the energy expenditure from the transmission, which could be an advantage when implementing the helmet in a large-scale environment.

The existing recognition model is general for all users, but an alternative implementation to investigate is using an individual model. This could possibly be more aware of the user’s individual movement patterns and hence be able to recognize changes in those that may not apply for other users.

7 Conclusion

It is shown that the stakeholder attitudes towards the smart helmet are overall positive. While privacy was hypothesized as the biggest barrier for the implementation of the helmet, it was later found out that the technology resistance of the construction industry together with the additional weight to the IoT-device were considered larger barriers. Nonetheless, it is emphasized that monitoring and storage of personal movement data should follow the privacy by design principles to avoid intruding on the privacy of construction workers.

From the results achieved through this study it can be said that the Random Forest machine learning algorithm together with the presented methods of data preprocessing is suitable for classifying a movement dataset collected through the smart helmet. It performs well also in recognizing rather detailed differences between similar activities, which could be valuable if the helmet was to be implemented in the construction industry to recognize and prevent harmful movement patterns. However, the technological architecture of the smart helmet should be considered in future developments to guarantee the best prerequisites for data collection and processing.

8 References

AFA Försäkring, 2017. *Arbetsolyckor och sjukskrivningar i byggbranschen,* Stockholm: s.n.

Anguita, D. et al., 2013. *A Public Domain Dataset for Human Activity Recognition Using Smartphones.* Bruges, Belgium, ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning.

Arbetsmiljöverket, 2015. *Byggbranschen: Korta arbetsskadefakta Nr 3/2015,* Stockholm: Arbetsmiljöverket.

Arbetsmiljöverket, 2018. *Fysisk belastning i arbete: Korta arbetsskadefakta Nr 1/2018,* Stockholm: Arbetsmiljöverket.

Arbetsmiljöverket, 2019. *Ansvar vid byggnads- och anläggningsarbete.* [Online]
Available at: https://www.av.se/produktion-industri-och-logistik/bygg/ansvar-vid-byggnads--och-anlaggningsarbete/
[Accessed 15 May 2020].

Arduino, 2020a. *Arduino MKR WiFi 1010.* [Online]
Available at: https://store.arduino.cc/arduino-mkr-wifi-1010
[Accessed 15 May 2020].

Arduino, 2020b. *Memory.* [Online]
Available at: https://www.arduino.cc/en/tutorial/memory
[Accessed 15 May 2020].

Bao, L. & Intille, S. S., 2004. Activity recognition from user-annotated acceleration data. In: *Pervasive.* s.l.:Springer-Verlag Berlin Heidelberg, pp. 1-17.

Barriball, L. K. & While, A., 1994. Collecting data using a semi-structured interview: a discussion paper. *Journal of Advanced Nursing,* Volume 19, pp. 328-335.

Bernfort, L. et al., 2013. *Från RÖD till GRÖN - En intervention inom byggbranschen för bättre levnadsvanor och ett hållbart arbetsliv,* s.l.: s.n.

BhanuJyothi, K., Himabindu, K. & Suryanarayana, D., 2017. *A Comparative Study of Random Forest & K – Nearest Neighbors on HAR dataset Using Caret,* s.l.: s.n.

Boughorbel, S., Jarray, F. & El-Anbari, M., 2017. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS ONE,* 12(6).

Brill, J., 2014. The Internet of Things: Building Trust and Maximizing Benefits Through Consumer Control. *Fordham Law Review,* 83(1), pp. 205-217.

Brown, S., 2020. *The Innovation Ultimatum: How six strategic technologies will reshape evety business in the 2020s.* 1st ed. New York: John Wiley & Sons Inc.

Bryman, A., 2012. *Social Research Methods.* 4th ed. New York, NY: Oxford University Press.

Byggföretagen, 2020a. *30 största byggföretagen efter omsättning i Sverige.* [Online]
Available at: https://byggforetagen.se/app/uploads/2020/01/30-St%C3%B6rsta-2018.pdf
[Accessed 15 May 2020].

Byggföretagen, 2020b. *Branschens struktur.* [Online]
Available at: https://byggforetagen.se/statistik/branschens-struktur/
[Accessed 15 May 2020].

Byggnads, 2019. *Svenska Byggnadsarbetareförbundet.* [Online]
Available at: https://www.byggnads.se/om-oss/om-oss/
[Accessed 15 May 2020].

Byggvärlden, 2015. *Undersöker förebyggande hälsa inom bygg.* [Online]
Available at: https://www.byggvarlden.se/undersoker-forebyggande-halsa-inom-bygg-89996/nyhet.html
[Accessed 15 May 2020].

Caruana, R. & Niculescu-Mizil, A., 2006. *An Empirical Comparison of Supervised Learning Algorithms.* Pittsburgh, PA, Association for Computing Machinery.

Cavoukian, A., 2010. *The 7 Foundational Principles: Implementation and Mapping of Fair Information Practices.* [Online]
Available at: https://www.ipc.on.ca/wp-content/uploads/Resources/pbd-implement-7found-principles.pdf
[Accessed 15 May 2020].

Chicco, D. & Jurman, G., 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics,* 21(6).

Datainspektionen, 2020. *Dataskyddsförordningens syfte och tillämpningsområde.* [Online]
Available at: https://www.datainspektionen.se/lagar--regler/dataskyddsforordningen/dataskyddsforordningens-syfte-och-tillampningsomrade/
[Accessed 15 May 2020].

Deng, X., Liu, Q., Dengad, Y. & Mahadevan, S., 2016. An improved method to construct basic probability assignment based on the confusion matrix for classification problem. *Information Sciences,* Volume 340-341, pp. 250-261.

Dewi, C. & Chen, R.-C., 2019. *Human Activity Recognition Based on Evolution of Features Selection and Random Forest.* Bari, Italy, IEEE.

Dubois, A. & Gadde, L.-E., 2002. Systematic combining: an abductive approach to case research. *Journal of Business Research,* Volume 55, pp. 553-560.

Feelgood, 2008. *Hälsoföretaget Feelgood och NCC har tecknat förnyat avtal om företagshälsovård för cirka 11 000 anställda.* [Online]
Available at: https://mb.cision.com/Main/4242/9331207/63705.pdf
[Accessed 15 May 2020].

Feelgood, 2019. *Peab blir ny stor kund till Feelgood.* [Online]
Available at: https://news.cision.com/se/feelgood-svenska-ab/r/peab-blir-ny-stor-kund-till-

feelgood,c2814377
[Accessed 15 May 2020].

Geurts, P., 2002. *Contributions to Decision Tree Induction: Bias/Variance Tradeoff and Time Series Classification*, s.l.: Université de Liège.

Hastie, T., Tibshirani, R. & Friedman, J., 2009. *The Elements of Statistical Learning.* 2nd ed. New York: Springer.

Höst, M., Regnell, B. & Runeson, P., 2006. *Att genomföra ett examensarbete.* 1st ed. s.l.:Studentlitteratur.

IDEO, 2020. *What is Design Thinking?*. [Online]
Available at: https://www.ideou.com/blogs/inspiration/what-is-design-thinking
[Accessed 4 April 2020].

IDG, 2019. *Del 5: Så skiljer sig gdpr från pul.* [Online]
Available at: https://cio.idg.se/2.1782/1.674864/gdpr/sida/5/del-5-sa-skiljer-sig-gdpr-fran-pul
[Accessed 15 May 2020].

James, G. M., 2003. Variance and Bias for General Loss Functions. In: R. E. Schapire, ed. *Machine Learning.* s.l.:Kluwer Academic Publishers, pp. 115-135.

Jiang, C. et al., 2017. Machine Learning Paradigms for Next-Generation Wireless Networks. *IEEE Wireless Communications,* 24(2), pp. 98-105.

Jia, Y., 2009. *Dietetic and Exercise Therapy Against Diabetes Mellitus.* Tianjin, China, IEEE.

Johansson Stålnacke, L. & Pettersson, F., 2016. *Sakernas internet: Personalisering inom den digitala sfären,* Umeå: Umeå Universitet.

Johansson, M., 2016. *En företagshälsovård för byggbranschen - En kartläggning och analys av behov och förutsättningar,* Luleå: Luleå Tekniska Universitet.

Jordao, A., Borges Torres, L. A. & Robson Schwartz, W., 2018. Novel approaches to human activity recognition based on accelerometer data. *Signal, Image and Video Processing,* Volume 12, pp. 1387-1394.

Khanna, R. & Awad, M., 2015. *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers.* 1st ed. Berkeley, CA: Apress.

Kines, P. et al., 2011. Nordic Safety Climate Questionnaire (NOSACQ-50): A new tool for diagnosing. *International Journal of Industrial Ergonomics,* Issue 41, pp. 634-646.

Kock, E. & Sarwari, Y., 2020. *Dataset.* Malmö, Sweden: s.n.

Lara, Ó. D. & Labrador, M., 2013. A Survey on Human Activity Recognition using Wearable Sensors. *IEEE Communications Surveys & Tutorials,* 15(3), pp. 1192-1209.

Lavanya, B. & Gayathri, G. S., 2017. *Exploration and Deduction of Sensor-Based Human Activity Recognition System of Smart-Phone Data.* Coimbatore, India, IEEE.

Lekvall, P. & Wahlbin, C., 2001. *Information för marknadsföringsbeslut.* 4th ed. Gothenburg, Sweden: IHM Publishing.

Lionbridge, 2019. *How Much AI Training Data Do You Need?.* [Online]
Available at: https://lionbridge.ai/articles/how-much-ai-training-data-do-you-need/
[Accessed 15 May 2020].

Marshall, C. & Rossman, G. B., 2014. *Designing Qualitative Research.* 6th ed. Newbury Park, CA: SAGE Publications.

MathWorks, 2020. *Practical Introduction to Frequency-Domain Analysis.* [Online]
Available at: https://se.mathworks.com/help/signal/examples/practical-introduction-to-frequency-domain-analysis.html
[Accessed 15 May 2020].

Maurer, U., Smailagic, A., Siewiorek, D. P. & Deisher, M., 2006. *Activity Recognition and Monitoring Using Multiple Sensors on Different Body Positions.* Washington, DC, IEEE Computer Society.

McKinsey & Company, 2016. *Imagining construction's digital future.* [Online]
Available at: https://www.mckinsey.com/industries/capital-projects-and-infrastructure/our-insights/imagining-constructions-digital-future#
[Accessed 15 May 2020].

Medium, 2017. *Random Forest Simple Explanation.* [Online]
Available at: https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d
[Accessed 15 May 2020].

Medium, 2018. *Regression Versus Classification Machine Learning: What's the Difference?.* [Online]
Available at: https://medium.com/quick-code/regression-versus-classification-machine-learning-whats-the-difference-345c56dd15f7
[Accessed 16 May 2020].

Miles, M. B., Huberman, A. M. & Saldana, J., 2014. *Qualitative Data Analysis: A Methods Sourcebook.* 3rd ed. s.l.:SAGE Publications.

Mitchell, R. K., Agle, B. R. & Wood, D. J., 1997a. Toward a Theory of Stakeholder Identification and Salience: Defining the Principle of Who and What Really Counts. *The Academy of Management Review,* 22(4), pp. 853-886.

Mitchell, T. M., 1997b. *Machine Learning.* 1st ed. s.l.:McGraw-Hill Education.

Mohammadi, M. o.a., 2015. Data mining EEG signals in depression for their diagnostic value. *BMC Medical Informatics and Decision Making,* 15(1).

Noy, C., 2008. Sampling Knowledge: The Hermeneutics of Snowball Sampling in Qualitative Research. *International Journal of Social Research Methodology,* 11(4), pp. 327-344.

Oliver Wyman, 2018. *Digitalization of the construction industry: The revolution is underway,* s.l.: Oliver Wyman.

Oshiro, T. M., Perez, P. S. & Baranauskas, J. A., 2012. *How Many Trees in a Random Forest?.* s.l., Springer-Verlag Berlin Heidelberg.

Pan, F., Wang, W., Tung, A. K. H. & Yang, J., 2005. *Finding Representative Set from Massive Data.* Houston, TX, IEEE.

Parmar, H. S., n.d. *Human Activity Recognition using Machine,* s.l.: s.n.

Poh, C. Q. X., Ubeynarayana, C. U. & Goh, Y. M., 2018. Safety leading indicators for construction sites: A machine learning approach. *Automation in Construction,* Volume 93, pp. 375-386.

Prescient & Strategic Intelligence, 2020. *Automated Machine Learning Market Research Report: By Offering, Deployment Type, Enterprise Size, Application, Industry - Industry Size, Share, Development and Demand Forecast to 2030,* s.l.: s.n.

Ravi, N., Dandekar, N., Mysore, P. & Littman, M. L., 2005. *Activity Recognition from Accelerometer Data.* s.l., AAAI Press.

Samuelson, B., 2018. *Arbetsskador inom byggindustrin 2018,* Stockholm: Byggindustrins Centrala Arbetsmiljöråd.

Scikit-Learn, 2020. *3.2.4.3.1. sklearn.ensemble.RandomForestClassifier.* [Online]
Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
[Accessed 15 May 2020].

Seref, B. & Bostanci, E., 2018. *Sentiment Analysis using Naive Bayes and Complement Naive Bayes Classifier Algorithms on Hadoop Framework.* s.l., IEEE.

Statens Medicin-Etiska Råd, 2020. *Integritet.* [Online]
Available at: http://www.smer.se/etik/integritet/
[Accessed 15 May 2020].

Statistiska Centralbyrån, 2020. *Arbetskraften ökade.* [Online]
Available at: https://www.scb.se/hitta-statistik/statistik-efter-amne/arbetsmarknad/arbetskraftsundersokningar/arbetskraftsundersokningarna-aku/pong/statistiknyhet/arbetskraftsundersokningarna-aku-4e-kvartalet-2019/
[Accessed 15 May 2020].

Stergiou-Kita, M. et al., 2015. Danger zone: Men, masculinity and occupational health and safety in high risk occupations. *Safety Science,* Issue 80, pp. 213-220.

Sutton, R. S. & Barto, A. G., 2017. *Reinforcement Learning: An Introduction.* 2nd ed. Cambridge, MA: The MIT Press.

Sveriges Byggindustrier, 2017. *Fakta om kvinnor och män i byggbranschen,* s.l.: s.n.

Sveriges Riksdag, 2016. *The Constitution of Sweden: The Fundamental Laws and the Riksdag Act.* [Online]
Available at: https://www.riksdagen.se/globalassets/07.-dokument--lagar/the-constitution-of-sweden-160628.pdf
[Accessed 15 May 2020].

The SciPy Community, 2020. *numpy.fft.fft.* [Online]
Available at: https://numpy.org/doc/1.18/reference/generated/numpy.fft.fft.html
[Accessed 17 May 2020].

Tillväxtverket, 2018. *Digitalisering i svenska företag,* Stockholm: Ruth.

Towards Data Science, 2018. *Understanding the Bias-Variance Tradeoff.* [Online]
Available at: https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229
[Accessed 16 May 2020].

Unionen, 2020. *This is how the Swedish labour market works.* [Online]
Available at: https://www.unionen.se/in-english/how-swedish-labour-market-works
[Accessed 15 May 2020].

van Hees, V. T. et al., 2013. Separating Movement and Gravity Components in an Acceleration Signal and Implications for the Assessment of Human Daily Physical Activity. *PLOS One,* 8(4).

Weinberg, B. D., Milne, G. R., Andonova, Y. G. & Hajjat, F. M., 2015. Internet of Things: Convenience vs. privacy and secrecy. *Business Horizons,* 58(6), pp. 615-624.

Xu, J., Zhang, Y. & Miao, D., 2020. Three-way Confusion Matrix for Classification: A Measure Driven View. *Information Sciences,* Volume 507, pp. 772-794.

Yang, L. et al., 2018. Towards Smart Work Clothing for Automatic Risk Assessment of Physical Workload. *IEEE Access,* Volume 6, pp. 40059-40072.

Yi, B.-J., Lee, D.-G. & Rim, H.-C., 2015. The Effects of Feature Optimization on High-Dimensional Essay Data. *Mathematical Problems in Engineering,* Volume 2015.

Yin, J., Yang, Q. & Pan Junfeng, J., 2008. Sensor-Based Abnormal Human-Activity Detection. *IEEE Transactions on Knowledge and Data Engineering,* 20(8), pp. 1082-1090.

Åström Paulsson, S. et al., 2014. *Stimulerar avtal mellan arbetsgivare och företagshälsovård till samarbete för hälsosamma arbetsplatser? – En genomlysning av avtal och avtalsprocess,* Uppsala: Arbets- och miljömedicin.

Appendix A Interview Guides (in Swedish)

A.1 Trade Union (TU)

A.1.1 Fackets roll

Nr	*Fråga*	*Kommentar*
1	Vilket ansvar och roll har facket gentemot byggarbetare?	Politiskt (arbetsrätt), samarbete med arbetsmiljöverket, etc.
2	Vilken makt har facket att besluta om vilka produkter, tjänster, och förändringar som får genomföras inom byggbranschen?	Har någon av kategorierna extra fokus eller inflytande?
3	Ett exempel på en nyare digital lösning i byggbranschen är införandet av ID06. Hur har detta mottagits av facket, byggföretagen och byggarbetarna?	
4	Har synen på ID06 förändrats sedan införandet för några år sedan?	
5	Finns det andra liknande exempel?	Hur har de mottagits?
6	Har det förekommit fall där produkter, tjänster och förändringar inte tillåtits?	Vad har det berott på, och vilken typ av produkt tjänst eller förändring har det handlat om?

A.1.2 Den tekniska lösningen

Nr	*Fråga*	*Kommentar*
1	Vår tanke är att behandling och analys av indata för att identifiera ett felaktigt rörelsemönster ska innebära ökad säkerhet för byggarbetare, men kan samtidigt uppfattas som övervakande och integritetskränkande. Ser du några andra eventuella fördelar och brister med produkten ur ett fackligt och/eller användarvänligt perspektiv?	Exempel på erfarenheter kring liknande produkter som införts och som skapat värde eller orsakat komplexitet som inte förutsågs?

2	Var går gränsen mellan ökad säkerhet och den integritetskränkning eller övervakning som byggarbetare utsätts för genom den här produkten?	
3	Kan man överväga produkten tack vare det värde den skapar och på så sätt få den att mottas bättre?	Hos facket/användarna?
4	Vilka är de faktorer ni kollar på för att avgöra om integritet hos en byggarbetare överskrids av dess arbetsgivare?	Hur måste vi tänka kring produkten och insamling av information för att minimera att den personliga integriteten hos byggarbetare kränks.
5	Får och kan en arbetsgivare enlig er ställa krav på vilken data vill samla in och analysera från anställdas arbete och beteende?	Exempelvis genom kontrakt eller specifika krav på ett projekt.
6	Hur ställer ni er till balansen mellan säkerhet och integritet i en produkt?	Vilket väger tyngre, och hur definieras vinningen av vardera sida?

A.2 Employer (NCC1)

Nr	***Fråga***	***Kommentar***
1	Hur ser processen ut när ni köper in ny skyddsutrustning, som hjälmar?	Påverkas det av användarna eller er som arbetsgivare?
2	Har ni intresse att skydda era medarbetare på ett sätt som "överträffar" standarden?	
3	Jobbar ni någonsin proaktivt med säkerhet på arbetsplatser och i så fall hur?	
4	Vår tanke är att behandling och analys av indata för att identifiera ett felaktigt rörelsemönster ska innebära ökad säkerhet för byggarbetare, men kan samtidigt uppfattas som övervakande och integritetskränkande. Ser du några andra eventuella fördelar och brister med produkten ur ett arbetsgivar- och/eller användarvänligt perspektiv?	Exempel på erfarenheter kring liknande produkter som införts och som skapat värde eller orsakat komplexitet som inte förutsågs?
5	Hur ställer ni er till balansen mellan säkerhet och integritet i en produkt?	Vilket väger tyngre, och hur definieras vinningen av vardera sida?
6	Om arbetarna vill ha en produkt för deras ökade säkerhet, vilken makt har arbetsgivaren i frågan?	
7	Hur arbetar ni med personlig integritet på arbetsplatsen?	Ser ni det ur ett grupperspektiv eller på individnivå?
8	Hur ser ni på personlig integritet kopplat till inspelning, tid, position i hjälmens fall?	Har ni några erfarenheter av sådana fall?

9	Ponera att ni skulle använda er av den här hjälmen. Hur hanterar ni data i verksamheten?	
10	Från intervju med Jonas och Jörgen har vi plockat upp att man t.ex. skulle kunna hantera data genom företagshälsovård för att skydda den med sekretess. Fungerar det i den här typen av verksamhet?	

A.3 Employer & Employee (NCC2)

Nr	***Fråga***	***Kommentar***
1	Hur ofta upplever du att det introduceras nya skyddsutrustningar som exempelvis hjälmar i ditt arbete?	
2	Är man generellt van vid att introducera nya produkter, t.ex. skyddsutrustningar i arbetet?	
3	Hur ställer du dig till nya produkter? Är mottagandet olika beroende på om produkten är teknologisk eller ej?	
4	Om en produkt skulle introduceras med syfte att öka säkerheten på arbetsplatsen, skulle den mottas mer positivt än om påverkan var neutral?	
5	Förhåller man sig olika beroende på om det är individen eller gruppen som skyddas?	
6	Vår tanke är att behandling och analys av indata för att identifiera ett felaktigt rörelsemönster ska innebära ökad säkerhet för byggarbetare, men kan samtidigt uppfattas som övervakande och integritetskränkande. Hur ser du på detta utifrån ditt arbete, din vardag, och den information du idag delar med dig till arbetsgivare?	
7	Var anser du att gränsen går för vad som är integritetskränkande och vad som inte är det?	T.ex. tid, position, filmning. Vilket väger tyngre, och hur definieras vinningen av vardera sida?
8	Kan du tänka dig ett exempel på någon teknologi eller liknande som skulle överskrida den gränsen?	
9	Om hjälmen skulle användas av er byggarbetare, hur skulle du önska att data behandlas och hanteras?	
10	Motsätter du dig att arbetsgivaren skulle få hantera den insamlade datan? Skulle exempelvis företagshälsovården vara en mer legitimerad enhet att göra det?	
11	Hur upplever du idag att personlig integritet hanteras på arbetsplatsen?	Ser ni det ur ett grupperspektiv eller på individnivå?

A.4 Legal services (LEG)

A.4.1 Rättsväsendets roll

Nr	*Fråga*	*Kommentar*
1	Utifrån det vi läst om bl.a. Datainspektionen, GDPR, Arbetsmiljöverket och Arbetsdomstolen, finns det flera olika organ med såväl lägre som högre makt att ge riktlinjer och besluta kring hur produkter som hjälmen bör brukas i arbetslivet. Hur skulle du säga att dessa samverkar, och finns det några andra organ som är viktiga att nämna i sammanhanget?	
2	Vår tanke är att behandling och analys av indata för att identifiera ett felaktigt rörelsemönster ska innebära ökad säkerhet för byggarbetare, men kan samtidigt uppfattas som övervakande och integritetskränkande. Ser du några andra eventuella fördelar och brister med produkten ur ett antingen rättsligt, fackligt eller användarvänligt perspektiv?	
3	Har det förekommit fall i historien där produkter, tjänster och förändringar inte tillåtits? Vad har detta berott på och vilken typ av produkt, tjänst eller förändring är det?	
4	Har det varit produkter liknande den smarta bygghjälmen riktad mot byggarbetarna eller andra typer av produkter?	
5	Finns det någon typ av pågående förarbete för att förtydliga den här typen av frågor i arbetslivet? Om ja, vilken riktning ser det ut att ta?	

A.4.2 Övervakning och integritet

Nr	*Fråga*	*Kommentar*
1	Kan du ge några exempel på fall av övervakning i arbetslivet som varit tillåtna samt några som inte tillåtits?	
2	Hur skulle ett fall av övervakning där de anställda inte anser att situationen är integritetskränkande dömas, om den trots allt skulle bryta mot en lag?	
3	Kan ett fall frias för att inblandade inte har några motsättningar? Bygger domen på individen eller rättsliga riktlinjer?	
4	Vilka rättsliga rättigheter har arbetsgivaren att övervaka arbetstagaren med hjälp av olika teknologiska hjälpmedel såsom positioneringsteknik, e-post och annan data kring beteende som kan samlas in och lagras av arbetsgivaren?	

5	Eftersom vår produkt inte är uppbyggd av någon GPS-sändare, utan endast av en accelerometer, skulle du säga att det finns färre juridiska motsättningar till varför produkten skulle få användas?	

A.4.3 Säkerhet och integritet

Nr	*Fråga*	*Kommentar*
1	Var går gränsen mellan ökad säkerhet och den integritetskränkning eller övervakning som byggarbetare utsätts för genom en produkt lik den smarta bygghjälmen?	
2	Kan man överväga produkten tack vare den ökade säkerhet den bidrar med och på så sätt få den att mottas bättre av byggbranschen? Finns det t.ex. lagar för integritetskränkande och för arbetsmiljö som kan tala för eller emot varandra?	
3	Är det möjligt att i ett anställningsavtal eller i ett projekt kräva av en anställd eller projektarbetare att delge den här typen av information då det finns krav på att bära hjälm på arbetsplatsen?	

A.5 Technological Expert (TECH)

Nr	*Fråga*	*Kommentar*
1	Vad är din bakgrund rent studie- och yrkesmässigt? Hur har det lett dig in på spåret med smarta byggkläder och hur man mäter arbetsmiljö?	
2	Vi fick ju tag på dig genom Viktor (BuildSafe). Hur har ditt arbete med dem sett ut, vilka kopplingar kan man dra mellan era olika projekt?	
	Vad var startpunkten till arbetet, vilken idé började ni från och hur har den utvecklats?	Vilka anledningar har funnits till detta - svårigheter, “enkelheter”, osv?
	Hur har ditt engagemang med smarta byggkläder mottagits av byggbranschen?	
3	Kan du ge några exempel på vart i ditt arbete som du har stött på mest problem och hur du löst dessa?	Vi har t.ex. haft intervjuer för att få bättre förståelse för facket och den rättsliga sidan - är det någon av dessa som har begränsat? Om, hur?
	De sensorer vi kommer använda på hjälmen, är en accelerometer med inbyggd gyroskop 3-axel? Tror du det är möjligt att samla in tillräckligt med information genom denna setup eller måste fler sensorer adderas?	

Har du koll på några liknande projekt med sensorer och rörelseanalys?	Gärna som inkluderar machine learning, men inte nödvändigtvis.
Avslutningsvis, har du några andra tips, erfarenheter eller områden som du vill dela med dig som du tror kan vara värdefullt för oss?	

Appendix B Explanatory Tables

This appendix will present a number of tables which were found overly detailed to present in the report, risking to confuse the reader. Therefore, the information in these tables is directed to those readers with a high technological interest, or those encouraged to read the details of the methodology.

B.1 Supervised Machine Learning Algorithm Characteristics

In this section, a table of selected supervised machine learning algorithms' key characteristics will be described.

Table B.1 A selection of supervised learning algorithms (Jiang, et al., 2017; BhanuJyothi, et al., 2017; Mohammadi, et al., 2015).

Learning techniques	***Key characteristics***
Linear regression models	Estimate the variables' linear relationships.
K-nearest neighbor (KNN)	Classify a new data point through a majority vote of K-nearest neighbors' classifications.
Support vector machines (SVM)	Prediction model is developed by separating the dataset into two classes with the use of a hyperplane. Non-linear mapping to high dimension.
Naive Bayesian learning	A statistical classification technique based on the Bayes Theorem.
Linear Discriminant Analysis (LDA)	Prediction model is developed by mapping the dataset into a new feature space which are more linearly discriminant compared with the original features of the dataset.
Parallel Random Forest	Parallel implementation (instead of sequential implementation) of the Random Forest classification algorithm.

B.2 Activities for Data Collection

In this section tables containing details from the physically conducted data collection be presented.

B.2.1 Research Subject Details and Activity Protocol

Table B.2 The distribution of characteristics among the research subjects.

Person	*Gender*	*Age*	*Height*	*Weight*
P1	Male	58 year	195 cm	90 kg
P2	Male	18 year	190 cm	87 kg
P3	Female	60 year	168 cm	70 kg
P4	Female	53 year	169 cm	66 kg
P5	Female	29 year	170 cm	69 kg
P6	Female	26 year	162 cm	55 kg
P7	Female	25 year	173 cm	64 kg
P8	Female	24 year	177 cm	66 kg
P9	Female	24 year	170 cm	64 kg
P10	Female	24 year	162 cm	55 kg
P11	Female	23 year	168 cm	58 kg

Table B.3 Experiment protocol of activities for the initial data collection.

Activity	*Recording time*	*Hold time*	*Repetitions*
Light lifting	30 seconds	3 seconds	6
Heavy lifting	30 seconds	3 seconds	6
Walking	30 seconds	Continuous	3
Walking while looking upwards	30 seconds	Continuous	3
Walking while carrying something heavy	30 seconds	Continuous	3

Table 3.4. Detailed descriptions of the activity set performed by the research subjects.

Activity	*Description*
Jumping	Jumping on the spot for 30 seconds, while holding the smartphone in landscape mode at shoulder height extended from the body with slightly bent arms.
Fake Jumping	Manipulating the phone in an attempt to simulate actual jumping for 30 seconds, while holding the smartphone in landscape mode at shoulder height extended from the body with slightly bent arms.
Squatting	Squatting on the spot for 30 seconds, while holding the smartphone in landscape mode at shoulder height extended from the body with slightly bent arms.
Fake squatting	Manipulating the phone in an attempt to simulate actual squatting for 30 seconds, while holding the smartphone in landscape mode at shoulder height extended from the body with slightly bent arms.
Stomping	Squatting on the spot for 30 seconds, while holding the smartphone in landscape mode at shoulder height extended from the body with slightly bent arms.
Fake stomping	Manipulating the phone in an attempt to simulate actual stomping for 30 seconds, while holding the smartphone in landscape mode at shoulder height extended from the body with slightly bent arms.

B.3 Data Preprocessing

In this section, tables of the raw time-series of data will firstly be presented. Thereafter, tables for the measurements used along with the feature vectors in both the time domain and in the frequency domain. Lastly, the tables for the encoded categorical activities and a full descriptions of the datasets used will be presented.

B.3.1 Features of the raw time-series of data and JSON-object

Table B.5 Description of the raw time-series features of the smart helmet dataset.

Feature	*Description*
Acc_X	Acceleration value in x-axis.
Acc_Y	Acceleration value in y-axis.
Acc_Z	Acceleration value in z-axis.
Rotation_X	Angular velocity in x-axis.
Rotation_Y	Angular velocity in y-axis.
Rotation_Z	Angular velocity in z-axis.

Table B.6 Description of the raw time-series features of the external dataset.

Feature	*Description*
Acc_X	Body acceleration value in x-axis.
Acc_Y	Body acceleration value in y-axis.
Acc_Z	Body acceleration value in z-axis.
Grav_X	Gravitational acceleration value in x-axis.
Grav_Y	Gravitational acceleration value in y-axis.
Grav_Z	Gravitational acceleration value in z-axis.
Rotation_X	Angular velocity in x-axis.
Rotation_Y	Angular velocity in y-axis.
Rotation_Z	Angular velocity in z-axis.
Roll	Relative rotation around the x-axis.
Pitch	Relative rotation around the y-axis.
Yaw	Relative rotation around the z-axis.
Action	The activity performed by the research subject.

Table B.7 Description of the parameters in a JSON-object.

JSON-parameters	*Description*
Person	Identifier to distinguish between different research subjects.
Sample	Identifier to distinguish between different JSON-objects.
Unix-time	Unix time stamp.
Param {	
- Acc_X	Acceleration value in x-axis.
- Acc_Y	Acceleration value in y-axis.
- Acc_Z	Acceleration value in z-axis.
- Gyro_X	Angular velocity in x-axis.
- Gyro_Y	Angular velocity in x-axis.
- Gyro_Z}	Angular velocity in x-axis.

B.3.2 Measurements and vectors in time domain & frequency domain

Table B.8 An overview of the final vectors produced in the time domain

Notation	*Description*
tBodyAcc_XYZ	Body acceleration for x-, y-, and z-axis.
tBodyAccMag	Body acceleration Euclidean magnitude.
tGravityAcc-XYZ	Gravity for x-, y- and z-axis.
tGravityAccMag	Gravity Euclidean magnitude.
tBodyAccJerk-XYZ	Jerk for x-, y- and z-axis.
tBodyAccJerkMag	Jerk Euclidean magnitude.
tBodyGyro-XYZ	Angular velocity for x-, y- and z-axis.
tBodyGyroMag	Angular velocity Euclidean magnitude.
tBodyGyroAngularAcc-XYZ	Angular Acceleration for x-, y- and z-axis.
tBodyGyroAngularAccMag	Angular Acceleration Euclidean magnitude.

Table B.9 An overview of the final vectors produced in the frequency domain

Notation	*Description*
tBodyAcc-XYZ	Body acceleration for x-, y- and z-axis.
tBodyAccMag	Body acceleration Euclidean magnitude.
tBodyAccJerk-XYZ	Jerk for x-, y- and z-axis.
tBodyAccJerkMag	Jerk Euclidean magnitude.
tBodyGyro-XYZ	Angular velocity for x-, y- and z-axis.
tBodyGyroMag	Angular velocity Euclidean magnitude.
tBodyGyroAngularAccMag	Angular acceleration Euclidean magnitude.

Table B.10 List of the measurements for computing the feature vectors.

Measurements	*Description*
Mean (mean)	The mean of all values in one window.
Standard deviation (std)	The standard deviation of the values in one window.
MAD (mad)	The median average deviation for one window.
Maximum (max)	The maximum value in one window.
Minimum (min)	The minimum value in one window.
Signal magnitude area (sma)	The integral of the x-, y-, and z-values from their corresponding windows, from time zero to the end time of the window.
Energy (energy)	The sum of the square of values in one window, divided with the number of values.
Interquartile range (iqr)	The middle 50 percent of values in one window when ordered from lowest to highest, calculating the difference between the median values of the upper and lower half.
Signal entropy (entropy)	The level of uncertainty in the values of a window, based on their probability.
Correlation coefficient (correlation)	The correlation between the xyz-signals: the correlation (x,y), (x,z), and (y,z). Only for time domain values.
Maximum index (maxInds)	The largest frequency component in one window. Only for frequency domain values.
Mean frequency (meanFreq)	Frequency signal weighted average in one window.
Skewness (skewness)	The skewness of the frequency signal, describing the asymmetry of the normal distribution. Only for frequency domain values.
Kurtosis (kurtosis)	The kurtosis of the frequency signal, describing the tail of the normal distribution. Only for frequency domain values.

Table B.11 A full list of the 348 features obtained after preprocessing the data, for time domain and frequency domain respectively.

Time domain	*Frequency domain*
tBodyAcc-mean()-X	fBodyAcc-mean()-X
tBodyAcc-mean()-Y	fBodyAcc-mean()-Y
tBodyAcc-mean()-Z	fBodyAcc-mean()-Z
tBodyAcc-std()-X	fBodyAcc-std()-X
tBodyAcc-std()-Y	fBodyAcc-std()-Y
tBodyAcc-std()-Z	fBodyAcc-std()-Z
tBodyAcc-mad()-X	fBodyAcc-mad()-X
tBodyAcc-mad()-Y	fBodyAcc-mad()-Y
tBodyAcc-mad()-Z	fBodyAcc-mad()-Z
tBodyAcc-max()-X	fBodyAcc-max()-X
tBodyAcc-max()-Y	fBodyAcc-max()-Y
tBodyAcc-max()-Z	fBodyAcc-max()-Z
tBodyAcc-min()-X	fBodyAcc-min()-X
tBodyAcc-min()-Y	fBodyAcc-min()-Y
tBodyAcc-min()-Z	fBodyAcc-min()-Z
tBodyAcc-sma()	fBodyAcc-sma()
tBodyAcc-energy()-X	fBodyAcc-energy()-X
tBodyAcc-energy()-Y	fBodyAcc-energy()-Y
tBodyAcc-energy()-Z	fBodyAcc-energy()-Z
tBodyAcc-iqr()-X	fBodyAcc-iqr()-X
tBodyAcc-iqr()-Y	fBodyAcc-iqr()-Y
tBodyAcc-iqr()-Z	fBodyAcc-iqr()-Z
tBodyAcc-entropy()-X	fBodyAcc-entropy()-X
tBodyAcc-entropy()-Y	fBodyAcc-entropy()-Y
tBodyAcc-entropy()-Z	fBodyAcc-entropy()-Z
	fBodyAcc--maxInds-X
	fBodyAcc--maxInds-Y
	fBodyAcc--maxInds-Z
	fBodyAcc--meanFreq-X
	fBodyAcc--meanFreq-Y
	fBodyAcc--meanFreq-Z
	fBodyAcc--skewness()-X
	fBodyAcc--kurtosis()-X

	fBodyAcc--skewness()-Y
	fBodyAcc--kurtosis()-Y
	fBodyAcc--skewness()-Z
	fBodyAcc--kurtosis()-Z
tBodyAcc-correlation()-X,Y	
tBodyAcc-correlation()-X,Z	
tBodyAcc-correlation()-Y,Z	
tGravityAcc-mean()-X	
tGravityAcc-mean()-Y	
tGravityAcc-mean()-Z	
tGravityAcc-std()-X	
tGravityAcc-std()-Y	
tGravityAcc-std()-Z	
tGravityAcc-mad()-X	
tGravityAcc-mad()-Y	
tGravityAcc-mad()-Z	
tGravityAcc-max()-X	
tGravityAcc-max()-Y	
tGravityAcc-max()-Z	
tGravityAcc-min()-X	
tGravityAcc-min()-Y	
tGravityAcc-min()-Z	
tGravityAcc-sma()	
tGravityAcc-energy()-X	
tGravityAcc-energy()-Y	
tGravityAcc-energy()-Z	
tGravityAcc-iqr()-X	
tGravityAcc-iqr()-Y	
tGravityAcc-iqr()-Z	
tGravityAcc-entropy()-X	
tGravityAcc-entropy()-Y	
tGravityAcc-entropy()-Z	
tGravityAcc-correlation()-X,Y	
tGravityAcc-correlation()-X,Z	
tGravityAcc-correlation()-Y,Z	
tBodyAccJerk-mean()-X	fBodyAccJerk-mean()-X
tBodyAccJerk-mean()-Y	fBodyAccJerk-mean()-Y

tBodyAccJerk-mean()-Z	fBodyAccJerk-mean()-Z
tBodyAccJerk-std()-X	fBodyAccJerk-std()-X
tBodyAccJerk-std()-Y	fBodyAccJerk-std()-Y
tBodyAccJerk-std()-Z	fBodyAccJerk-std()-Z
tBodyAccJerk-mad()-X	fBodyAccJerk-mad()-X
tBodyAccJerk-mad()-Y	fBodyAccJerk-mad()-Y
tBodyAccJerk-mad()-Z	fBodyAccJerk-mad()-Z
tBodyAccJerk-max()-X	fBodyAccJerk-max()-X
tBodyAccJerk-max()-Y	fBodyAccJerk-max()-Y
tBodyAccJerk-max()-Z	fBodyAccJerk-max()-Z
tBodyAccJerk-min()-X	fBodyAccJerk-min()-X
tBodyAccJerk-min()-Y	fBodyAccJerk-min()-Y
tBodyAccJerk-min()-Z	fBodyAccJerk-min()-Z
tBodyAccJerk-sma()	fBodyAccJerk-sma()
tBodyAccJerk-energy()-X	fBodyAccJerk-energy()-X
tBodyAccJerk-energy()-Y	fBodyAccJerk-energy()-Y
tBodyAccJerk-energy()-Y	fBodyAccJerk-energy()-Z
tBodyAccJerk-iqr()-X	fBodyAccJerk-iqr()-X
tBodyAccJerk-iqr()-Y	fBodyAccJerk-iqr()-Y
tBodyAccJerk-iqr()-Z	fBodyAccJerk-iqr()-Z
tBodyAccJerk-entropy()-X	fBodyAccJerk-entropy()-X
tBodyAccJerk-entropy()-Y	fBodyAccJerk-entropy()-Y
tBodyAccJerk-entropy()-Z	fBodyAccJerk-entropy()-Z
	fBodyAccJerk-maxInds-X
	fBodyAccJerk--maxInds-Y
	fBodyAccJerk--maxInds-Z
	fBodyAccJerk--meanFreq-X
	fBodyAccJerk--meanFreq-Y
	fBodyAccJerk--meanFreq-Z
	fBodyAccJerk--skewness()-X
	fBodyAccJerk--kurtosis()-X
	fBodyAccJerk--skewness()-Y
	fBodyAccJerk--kurtosis()-Y
	fBodyAccJerk--skewness()-Z
	fBodyAccJerk--kurtosis()-Z
tBodyAccJerk-correlation()-X,Y	
tBodyAccJerk-correlation()-X,Z	

tBodyAccJerk-correlation()-Y,Z
tBodyGyro-mean()-X
tBodyGyro-mean()-Y
tBodyGyro-mean()-Z
tBodyGyro-std()-X
tBodyGyro-std()-Y
tBodyGyro-std()-Z
tBodyGyro-mad()-X
tBodyGyro-mad()-Y
tBodyGyro-mad()-Z
tBodyGyro-max()-X
tBodyGyro-max()-Y
tBodyGyro-max()-Z
tBodyGyro-min()-X
tBodyGyro-min()-Y
tBodyGyro-min()-Z
tBodyGyro-sma()
tBodyGyro-energy()-X
tBodyGyro-energy()-Y
tBodyGyro-energy()-Z
tBodyGyro-iqr()-X
tBodyGyro-iqr()-Y
tBodyGyro-iqr()-Z
tBodyGyro-entropy()-X
tBodyGyro-entropy()-Y
tBodyGyro-entropy()-Z
fBodyGyro-mean()-X
fBodyGyro-mean()-Y
fBodyGyro-mean()-Z
fBodyGyro-std()-X
fBodyGyro-std()-Y
fBodyGyro-std()-Z
fBodyGyro-mad()-X
fBodyGyro-mad()-Y
fBodyGyro-mad()-Z
fBodyGyro-max()-X
fBodyGyro-max()-Y
fBodyGyro-max()-Z
fBodyGyro-min()-X
fBodyGyro-min()-Y
fBodyGyro-min()-Z
fBodyGyro-sma()
fBodyGyro-energy()-X
fBodyGyro-energy()-Y
fBodyGyro-energy()-Z
fBodyGyro-iqr()-X
fBodyGyro-iqr()-Y
fBodyGyro-iqr()-Z
fBodyGyro-entropy()-X
fBodyGyro-entropy()-Y
fBodyGyro-entropy()-Z
fBodyGyro--maxInds-X
fBodyGyro--maxInds-Y
fBodyGyro--maxInds-Z
fBodyGyro--meanFreq-X
fBodyGyro--meanFreq-Y
fBodyGyro--meanFreq-Z
fBodyGyro-skewness()-X
fBodyGyro--kurtosis()-X
fBodyGyro--skewness()-Y
fBodyGyro--kurtosis()-Y
fBodyGyro--skewness()-Z

	fBodyGyro--kurtosis()-Z
tBodyGyro-correlation()-X,Y	
tBodyGyro-correlation()-X,Z	
tBodyGyro-correlation()-Y,Z	
tBodyGyroJerk-mean()-X	
tBodyGyroJerk-mean()-Y	
tBodyGyroJerk-mean()-Z	
tBodyGyroJerk-std()-X	
tBodyGyroJerk-std()-Y	
tBodyGyroJerk-std()-Z	
tBodyGyroJerk-mad()-X	
tBodyGyroJerk-mad()-Y	
tBodyGyroJerk-mad()-Z	
tBodyGyroJerk-max()-X	
tBodyGyroJerk-max()-Y	
tBodyGyroJerk-max()-Z	
tBodyGyroJerk-min()-X	
tBodyGyroJerk-min()-Y	
tBodyGyroJerk-min()-Z	
tBodyGyroJerk-sma()	
tBodyGyroJerk-energy()-X	
tBodyGyroJerk-energy()-Y	
tBodyGyroJerk-energy()-Y	
tBodyGyroJerk-iqr()-X	
tBodyGyroJerk-iqr()-Y	
tBodyGyroJerk-iqr()-Z	
tBodyGyroJerk-entropy()-X	
tBodyGyroJerk-entropy()-Y	
tBodyGyroJerk-entropy()-Z	
tBodyGyroJerk-correlation()-X,Y	
tBodyGyroJerk-correlation()-X,Z	
tBodyGyroJerk-correlation()-Y,Z	
tBodyAccMag-mean()	fBodyAccMag-mean()
tBodyAccMag-std()	fBodyAccMag-std()
tBodyAccMag-mad()	fBodyAccMag-mad()
tBodyAccMag-max()	fBodyAccMag-max()
tBodyAccMag-min()	fBodyAccMag-min()

tBodyAccMag-sma()	fBodyAccMag-sma()
tBodyAccMag-energy()	fBodyAccMag-energy()
tBodyAccMag-iqr()	fBodyAccMag-iqr()
tBodyAccMag-entropy()	fBodyAccMag-entropy()
tGravityAccMag-mean()	
tGravityAccMag-std()	
tGravityAccMag-mad()	
tGravityAccMag-max()	
tGravityAccMag-min()	
tGravityAccMag-sma()	
tGravityAccMag-energy()	
tGravityAccMag-iqr()	
tGravityAccMag-entropy()	
tBodyAccJerkMag-mean()	fBodyBodyAccJerkMag-mean()
tBodyAccJerkMag-std()	fBodyBodyAccJerkMag-std()
tBodyAccJerkMag-mad()	fBodyBodyAccJerkMag-mad()
tBodyAccJerkMag-max()	fBodBodyAccJerkMag-max()
tBodyAccJerkMag-min()	fBodyBodyAccJerkMag-min()
tBodyAccJerkMag-sma()	fBodyBodyAccJerkMag-sma()
tBodyAccJerkMag-energy()	fBodyBodyAccJerkMag-energy()
tBodyAccJerkMag-iqr()	fBodyBodyAccJerkMag-iqr()
tBodyAccJerkMag-entropy()	fBodyBodyAccJerkMag-entropy()
tBodyGyroMag-mean()	fBodyBodyGyroMag-mean()
tBodyGyroMag-std()	fBodyBodyGyroMag-std()
tBodyGyroMag-mad()	fBodyBodyGyroMag-mad()
tBodyGyroMag-max()	fBodBodyGyroMag-max()
tBodyGyroMag-min()	fBodyBodyGyroMag-min()
tBodyGyroMag-sma()	fBodyBodyGyroMag-sma()
tBodyGyroMag-energy()	fBodyBodyGyroMag-energy()
tBodyGyroMag-iqr()	fBodyBodyGyroMag-iqr()
tBodyGyroMag-entropy()	fBodyBodyGyroMag-entropy()
	fBodyBodyAccJerkMag-maxInds()
	fBodyBodyGyroMag-meanFreq()
	fBodyBodyGyroMag-skewness()
	fBodyBodyGyroMag-kurtosis()
tBodyGyroJerkMag-mean()	fBodyBodyGyroJerkMag-mean()
tBodyGyroJerkMag-std()	fBodyBodyGyroJerkMag-std()

tBodyGyroJerkMag-mad()	fBodyBodyGyroJerkMag-mad()
tBodyGyroJerkMag-max()	fBodBodyGyroJerkMag-max()
tBodyGyroJerkMag-min()	fBodyBodyGyroJerkMag-min()
tBodyGyroJerkMag-sma()	fBodyBodyGyroJerkMag-sma()
tBodyGyroJerkMag-energy()	fBodyBodyGyroJerkMag-energy()
ttBodyGyroJerkMag-iqr()	fBodyBodyGyroJerkMag-iqr()
tBodyGyroJerkMag-entropy()	fBodyBodyGyroJerkMag-entropy()
	fBodyBodyGyroJerkMag-maxInds()
	fBodyBodyGyroJerkMag-meanFreq()
	fBodyBodyGyroJerkMag-skewness()
	fBodyBodyGyroJerkMag-kurtosis()

B.3.3 Categorical Activities

Table B.11 Encoded categorical activities of the smart helmet dataset.

Activity	***Integer***
Light lifting	1
Heavy lifting	2
Walking	3
Walking while carrying something heavy	4
Walking while looking upwards	5

Table B.12 Encoded categorical activities of the external dataset.

Activity	***Integer***
Fake jumping	1
Squatting	2
Fake stomping	3
Jumping	4
Fake squatting	5
Stomping	6

B.3.4 Description of Datasets

Table B.13 Total description of the preprocessed smart helmet dataset.

Number of instances	*Classes, i.e. activities*	*Samples per class*	*Features*
92 400	5	549 (Light lifting)	348 before feature optimization
		624 (Heavy lifting)	338 after feature optimization
		411 (Walking)	
		265 (Walking, carrying heavy)	
		267 (Walking, looking upwards)	

Table B.14 Total description of the preprocessed external dataset.

Number of instances	*Classes, i.e. activities*	*Samples per class*	*Features*
1440	6	218 (Squatting)	348 before feature optimization
		262 (Fake Squatting)	310 after feature optimization
		240 (Others)	

Table B.18 Total description of the raw time-series external dataset.

Number of instances	*Classes, i.e. activities*	*Samples per class*	*Features*
92 400	6	14 000 (Squatting)	9 without roll, pitch and yaw
		16 800 (Fake Squatting)	12 with roll, pitch, and yaw
		15 400 (Others)	

Appendix C Implementation Code

All the implemented code for data preprocessing and running the Random Forest algorithm is publicly available and can be found at:

https://github.com/sarahjohannesson/RandomForestClassification