TRUST IN HUMAN ACTIVITY RECOGNITION DEEP LEARNING MODELS

TRUST IN HUMAN ACTIVITY RECOGNITION DEEP LEARNING MODELS

BY

AMA SIMONS, B.Eng.

A THESIS

SUBMITTED TO THE SCHOOL OF BIOMEDICAL ENGINEERING

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

Master of Applied Science (2021) McMaster University

(school of biomedical engineering) Hamilton, Ontario, Canada

TITLE:	Trust in Human Activity Recognition Deep Learning Models
AUTHOR:	Ama Simons B.Eng. (Electrical & Biomedical Engineering), McMaster University, Hamilton, Canada
SUPERVISOR:	Dr. Thomas Doyle
NUMBER OF PAGES:	xvi, 113

Lay Abstract

The trustworthiness of artificial intelligence must be explored before society can fully reap its benefits. The element of trust that is explored in this thesis is the robustness of wearable device based artificial intelligence models to changes in data acquisition. The specific changes that are explored are changes in the wearable device used to record the input data as well as input data from different recording sessions. Using human activity recognition models as a vehicle, the results show that performance degradation occurs when the wearable device is changed and when data comes from a different recording session. An out of domain discriminator is developed to alert users when a potential performance degradation can occur.

Abstract

Trust is explored in this thesis through the analysis of the robustness of wearable device based artificial intelligence based models to changes in data acquisition. Specifically changes in wearable device hardware and different recording sessions are explored. Three human activity recognition models are used as a vehicle to explore this: Model A which is trained using accelerometer signals recorded by a wearable sensor referred to as Astroskin, Model H which is trained using accelerometer signals from a wearable sensor referred to as the BioHarness and Model A Type 1 which was trained on Astroskin accelerometer signals that was recorded on the first session of the experimental protocol. On a test set recorded by Astroskin Model A had a 99.07% accuracy. However on a test set recorded by the BioHarness Model A had a 65.74% accuracy. On a test set recorded by BioHarness Model H had a 95.37% accuracy. However on a test set recorded by Astroskin Model H had a 29.63% accuracy. Model A Type 1 an average accuracy of 99.57% on data recorded by the same wearable sensor and same session. An average accuracy of 50.95% was obtained on a test set that was recorded by the same wearable sensor but by a different session. An average accuracy of 41.31% was obtained on data that was recorded by a different wearable sensor and same session. An average accuracy of 19.28% was obtained on data that was recorded by a different wearable sensor and different session. An out of

domain discriminator for Model A Type 1 was also implemented. The out of domain discriminator was able to differentiate between the data that trained Model A Type 1 and other types (data recorded by a different wearable devices/different sessions) with an accuracy of 97.60%.

To my Mom, Dad and Sister

Acknowledgements

I would like to thank my supervisor Dr. Doyle for his continued support and guidance in the writing of this thesis. I would also like to thank the members of my supervisory committee, Dr.Reilly and Dr.Musson, for their valued insights on my research work. In addition, I would like to thank Dr. Heydarian for his suggestions on my research. I would like to thank the entire Biomedic.AI lab for creating a helpful and enjoyable research space.

I would also like to thank my parents, my little sister, my cousins and friends for their continual support throughout all of my academic endeavours.

Lastly, I would like to thank McMaster University, NSERC, and the Canadian Department of National Defence for the funding to complete this research.

Contents

List of Figures

List of Tables

Abbreviations

Abbreviations

AI	Artificial Intelligence
HAR	Human Activity Recognition
ACC	Accelerometer
OOD	Out of Domain
ECG	Electrocardiography
IMU	Inertial Measurement Unit
EEG	Electroencephalography
EMG	Electromyography
SNR	Signal to Noise Ratio

Declaration of Academic Achievement

The research presented in this thesis was conducted by Ama Simons. Ama Simons conducted the literature review, designed and conducted experiments, collected the data for experiments and wrote the manuscript. Dr. Thomas Doyle, Dr. David Musson and Dr. Jim Reilly assisted in this research by providing insights on the research question, methodology and review of the manuscript.

Chapter 1

Introduction

Artificial intelligence (AI)-powered autonomous medical advisory systems are transforming the medical space. An AI algorithm can make decisions from patterns found in various types of medical data, such as imaging, electronic health records, and physiological signals. However, as the use of AI-powered autonomous medical advisory systems becomes more prevalent and model scrutability becomes more opaque, trust in AI must be fostered for a smooth adoption into clinical settings and more generally, into society.

1.1 Problem Statement

Trust in autonomous systems is an ongoing discussion in literature. Lee and Moray (1992) describe trust using the ideas of performance, process, and purpose of the autonomous system:

Process relates to understanding how the autonomous system operates (Lee and Moray, 1992; Lee and See, 2004).

Purpose is defined as the intent of the autonomous system (Lee and Moray, 1992; Lee and See, 2004)

Performance encompasses the reliability, predictability, or robustness an autonomous system (Lee and See, 2004; Lee and Moray, 1992)

When specifically discussing trust in AI-powered autonomous systems this framework continues to provide a structured approach to clearly defining the facets of trust (Siau and Wang, 2018; Hengstler *et al.*, 2016). *Trustworthy* AI has also been discussed by governmental groups. The Ethics for Trustworthy AI (High-Level Expert Group on AI, 2019) details that the seven key requirements of trustworthy AI:

1. Human agency and oversight
2. Technical robustness and safety
3. Privacy and data governance
4. Transparency
5. Diversity, Non-Discrimination and fairness
6. Societal and environmental well-being
7. Accountability

The requirement of technical robustness and safety falls underneath the performance aspect of trust in general autonomous systems. This thesis focuses on the technical robustness and safety element of trust in AI.

Technical robustness and safety includes the reliability or the general safety of the machine learning algorithm (High-Level Expert Group on AI, 2019). One of the

factors may that impact the technical robustness and safety of the AI algorithms is *dataset shift*. Arnold *et al.* (2019) place dataset shift underneath the safety in the elements of trust in the AI services. Generally dataset shift occurs when the conditions in which the system is developed differs from those in which the system is used (Storkey, 2009). There are many forms of dataset shift which are described briefly below (Storkey, 2009):

Simple covariate shift occurs when the feature distributions change between training and testing

Prior Probability shift occurs when the distribution of the labels change between the training set and test set

Sample Selection Bias occurs when there is a change in distrbution as a result of a sample selection procedure

Imbalanced Data occurs when data is balanced in development stages of the model but the deployment distribution data is inherently imbalanced

Domain Shift occurs as a result of a change in measurement system

Source Component Shift occurs because data is made up of different sources and the proportions of the sources can change from development to deployment

In this thesis our specific focus is *domain shift* which "is characterized by the fact that a measurement system, or method of description, can change" (Storkey, 2009, p.19). Domain shift can occur as a result of a change in the chain of data acquisition that provides data for autonomous medial advisory systems. This is not well studied, specifically how modifications or variance in wearable devices will impact the AI.

Storkey (2009) describes an unchanging feature space which can be denoted as x_0. However x_0 can not be observed and what is observed is some mapping of x_0 which is defined as $x = f(x_0)$ and this mapping can change between training and testing distributions (Storkey, 2009). This reasoning can be extended to the measurement of signals by wearable devices. In this scenario x_0 is the signal to be measured and wearable devices provide the mapping f into observable space. However if the wearable devices changes, the mapping f may change as well. This can cause a shift in the data x. This thesis also explores if performance changes can arise when using a deep learning model with data that arose from a different session of recording. Even when using the same wearable device, changes in wearable device placement, or variations in how an activity can be performed from session to session can occur.

1.2 Proposed Research

In this work domain shift is investigated through the lens of human activity recognition (HAR) models. HAR models can use accelerometer (ACC) data to identify if an individual is performing a specific activity such as walking, sitting or standing. The performance of HAR models are evaluated by determining the accuracy on held-out test data set that was recorded by the same wearable device as the training data. However, in deployment it is possible that these models will be used with wearable devices that differ from the wearable devices that were used in development stages. Using different hardware in deployment may introduce domain shift and cause a degradation in model performance. To explore this issue, this thesis:

1. Designs a 2-by-2 experimental set-up in which HAR models were tested on data

that came from a different wearable device that recorded the training data. This set-up will be considered as a type of *cross-domain test* throughout the rest of this thesis

2. Investigates a method to **estimate** the performance degradation (if any) that HAR models experience when deployed on a test set recorded by different wearable device.

Another factor investigated that may cause performance degradation is deploying HAR models on data that arises from different recording sessions. To explore this issue, this thesis:

1. Designs an experimental setup to evaluate if performance degradation occurs when a HAR model is evaluated on data recorded from a different session then the data used to train it

2. Implements a domain discriminator model to alert users of when the data incoming into HAR model is detected as out-of-domain (OOD) meaning that the new incoming data is not the data that was used to train the model

1.3 Thesis Contributions

This thesis has contributed to the field of trust in AI-powered medical advisory systems in the following ways:

1. Performed *cross-domain tests* to determine if performance degradation occurs in HAR models that are used with data that did not come from the same source as the training data

2. Developed a domain discriminator that is able to detect OOD. Out of domain data is when data that the HAR model is deployed on differs from the training data of the HAR model

1.4 Organization of Thesis

This thesis is organized as follows:

1. Chapter 2 presents a rationale for the investigation of domain shift in wearable device based medical advisory systems and a literature review on the current related work in medical advisory systems. The purpose of this chapter is give the reader an overview of how domain shift can present itself in different medical advisory systems and the proposed ways to address it.

2. Chapter 3 describes the data collection process for development of the HAR models. It also describes the methodology for implementing the HAR model.

3. Chapter 4 presents the methodologies for exploring trust through the lens of robustness with the HAR models developed.

4. Chapter 5 presents the results.

5. Chapter 6 discusses the results and its importance in the context of trust.

6. Chapter 7 concludes the thesis and suggests other avenues for future work.

Chapter 2

Literature Review

This literature review addresses in what ways dataset shift, with a particular focus to domain shift, is investigated in medical advisory systems. The methodology of the literature review was inspired by the guidelines for conducting literature reviews in computer science (Kofod-Petersen, 2012).

2.1 Literature Review Methodology

This section covers the search strategy used for the literature review, the primary selection of studies after the search and the inclusion criteria used to screen papers found.

2.1.1 Search Strategy

The overall aim of this literature review is to investigate how existing AI-powered autonomous methods address dataset shift with a particular focus to domain shift in medical advisory systems. The IEEE Xplore database was searched with the

key words: dataset shift, technical robustness, domain shift, machine learning, deep learning, artificial intelligence, health care and medical.

2.1.2 Primary Selection of Studies

Literature dating before than 2018 as well as studies written by the author, literature reviews, comparative studies and surveys were excluded from the literature review.

2.1.3 Assessment of Studies

The full title and abstract of the papers retrieved from the search terms in Section 2.1.1 must have met the following inclusion criteria (IC) to in order to enter the full text screening:

1. IC1 - The paper's main concern is dataset shift/domain shift in medical advisory systems that use AI to make decisions.

2. IC2 - The study is a primary or secondary study that details the results have been derived by empirical methods. In this literature review this means that the study must present results that are drawn from a database. The database used to present results can be an online database or a database that was not created by the authors of the paper

3. IC3 - The study brings awareness to, solves or discusses domain shift in AI-powered medical advisory autonomous systems

If the papers were able to surpass the title and abstract screening the next IC were applied.

1. IC4 - The cause of the dataset/domain shift is a because of a real-world change in data collection due to hardware or recording session changes

The quality criteria of the papers was identified as:

1. QC1 - The paper is written clearly

2.2 Conducting Review

The selected papers are the papers whose abstracts met the IC1, IC2 and IC3 requirements. Table 2.2 shows the search terms that were used, the results from these search terms as well as the papers that were selected.

Table 2.1: Search terms and results

Search	Database	Results	Selected
(dataset shift OR technical robustness OR domain shift) AND (machine learning OR artificial intelligence OR deep learning) AND (health care OR medical)	IEEE Xplore	257	98

After reviewing the abstracts, it was clear that the papers that were reviewed fell into two categories of medical advisory autonomous systems that use different types of data, ones that used medical images and ones that used temporal data. This was summarized in Table 2.2:

Table 2.2: Selected papers by category

Category	Number of Papers
Medical Images	83
Temporal Data	15

The medical imaging category contains papers that discuss dataset/domain shift in the context of autonomous AI systems that use medical images to make decisions. There are 83 papers contained in this category. The temporal data category contains papers that discuss dataset/domain shift in context of autonomous AI systems that use temporal or time series data such as physiological signals to make decisions. There are 15 papers contained in this category. In this initial analysis is it evident that dataset/domain shift has garnered attention in the medical imaging community but less attention in medical data that are time series. Dataset/domain shift in medical imaging applications is outside the scope of this thesis because medical imaging data is a different data type than medical temporal data. Inherently, medical images are typically static data while temporal data is dynamic data. The purpose of including the medical imaging reflection into the study is to demonstrate to the reader that the imaging field receives a lot more attention in dataset/domain shift than in the medical temporal field. IC4 further separated the amount of papers to be included in the data extraction process. Out of the 15 medical temporal data papers that speak about dataset/domain shift, 9 mainly focus on dataset/domain shift that arised due to subject variability and those were excluded as that is out of the scope of this thesis.

2.3 Reporting the Review

The data extracted from the papers contained in this literature review (which is papers that received a quality score of 1) was presented by category of the physiological signals that were investigated. Out of the 15 temporal data papers, 6 papers were reviewed. One paper that was relevant to the topic of interest was included and in total 7 papers were analyzed.

2.3.1 Electrocardiography Signals

AI models that use electrocardiography (ECG) signals can experience domain shift as a result of varying recording protocols that sample ECG signals at different frequencies or apply different gains to the signal (Hasani *et al.*, 2020). Varying recording protocols can be used to record the different online ECG databases that are frequently used to develop AI models for ECG based problems. Each ECG online database can be interpreted as a domain. Hasani *et al.* (2020) used adversarial domain generalization to learn features that are indistinguishable between the different ECG databases. To evaluate their methodology they created 4 domains based on different heart databases. They evaluated their technique using a leave one database out manner. With their adversarial domain generalization they achieved a better score on tests from different databases then when adversarial domain generalization was not employed. The presence of domain shift in ECG signals from different online databases is also present in ECG delination problems (Chen *et al.*, 2020). Chen *et al.* (2020) employed a similar technique as Hasani *et al.* (2020) and learned features that did not change amongst the domains (which were also different databases). Their model had improved performance when compared to a model that did not account of domain shift on inter-dataset experiments.

2.3.2 Inertial Measurement Unit Signals

Domain shift can also be introduced by changes in the placement and orientation of the inertial measurement units (IMUs). This has been investigated in gait monitoring tasks of gait event detection and pathological gait pattern recognition (Mu *et al.*, 2020). Mu *et al.* (2020) state that domain shift can occur due to the changes of the

placement and orientation of IMUs that can occur between trials. In this work an end-to-end position-independent IMU based gait framework that utilized unsupervised domain adaption was successfully able to obtain good results in the presence of domain shift. Their model, like the solutions for domain shift regarding ECG, learns domain invariant features. Pan *et al.* (2021) investigate domain shift when inertial measurement unit are placed on different body parts. The method used to reduce domain shift is joint transfer strategy in which the differences between the domains are reduced but the key structural information is preserved. This method performed well and was able to mitigate the domain shift between sensors placed on different body parts.

2.3.3 Electroencephalography Signals

Accurate classification of electroencephalography (EEG) signals are essential to brain computer interfaces. However EEG signals can be impacted by domain shift because of electrode placement changes as well as changes in the state of a subject (Han and Jeong, 2021). Han and Jeong (2021) use empirical risk minimization in order to improve the inter-session domain shift experienced. Domain shift also presents itself in classifying emotions from EEG signals. Lew *et al.* (2020) aim to use a network that is able to learn emotion representations that are able to correctly classify emotions but are not impacted by different domains. In particular the model that was developed performed better than a base model on an experiment in which the training and testing sets arose from different trials (Lew *et al.*, 2020).

2.3.4 Electromyography Signals

In gesture recognition with electromyography (EMG) signals, domain shift can be caused by intra-session, inter-session and intra-subject factors (Ketykó *et al.*, 2019). Ketykó *et al.* (2019) look specifically at inter-session and inter-subject shift but the inter-session results are summarized here since it covers an the area of study in the thesis. The domain adaption approach used by Ketykó *et al.* (2019) was an addition of a domain adaption layer that performed a linear transformation of the features to the neural network. The domain adaption neural network was able to improve inter-session recognition.

2.4 Literature Review Summary

The survey of literature demonstrates that:

1. Dataset/domain shift in medical temporal data is an understudied field
2. Dataset/domain shift receives more attention in the medical imaging field then it does in the medical temporal data field
3. Papers that focus on medical temporal data mostly focuses on domain shift caused by subject variability
4. Papers that focus on dataset/domain shift because of hardware changes or session changes mostly use pre-existing databases
5. The most common way to reduce domain shift is to learn features that are invariant to domains

This thesis addresses the research gap concerning dataset shift/domain shift in medical temporal data. Although there has been investigation into this work, this area of study is still understudied. In this thesis, we will using data that did not arise from a pre-recorded database. We address domain shift caused by changes in hardware as well as the datast shift due to changes in recording session.

Chapter 3

Human Activity Recognition

This chapter describes the methodology used to develop the HAR models which served as our wearable device based medical advisory system of interest. In Section 3.1 the experimental protocol for recording human activity data is described. The data is divided into Dataset 1 and Dataset 2. The details of both datasets are described in Subsection 3.1.1. The pre-processing steps for Dataset 1 and Dataset 2 are detailed in Section 3.2. In Section 3.3 the development process of the HAR models is described. Figure 3.1 provides a graphical overview of this chapter.

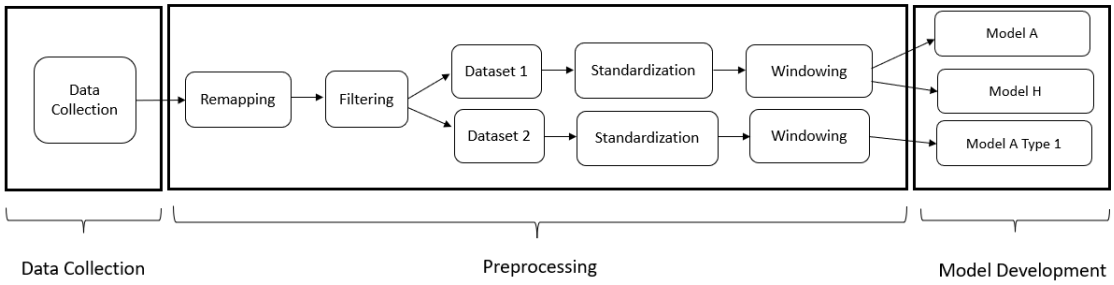


Figure 3.1: Methodology overview

3.1 Data Collection Procedure

The data collection for this thesis was conducted in a home setting with the authour as the sole participant due to Government and University research restrictions around COVID-19. The authour recorded tri-axial acceleration through two wearable devices, the BioHarness and the Astroskin, whilst performing 6 human activities. The 6 human activities performed were:

1. **Walking** in the hallways, living room and kitchen
2. **Sitting** upright at a chair in front of a desk
3. **Standing** in a room in the author's home
4. **Laying** face-up on a mattress
5. **Walking Downstairs**
6. **Walking Upstairs**

A comparison of the Astroskin and the BioHarness wearable devices used in the collection of the data are shown below. The hardware properties such as the range of the accelerometer measurement, the resolution of the analog to converter was taken from the respective datasheets (Ast, 2009; Zep, 2018). The cost are provided by the quotes provided whilst purchasing.

Table 3.1: Properties of Astroskin and BioHarness

Property	Astroskin	Zephyr BioHarness
Range of Accelerometer Measurement	-16 g to 16 g	- 16 g to 16 g
Resolution of Analog to Digital Converter	13 bit	12 bit
Sampling Frequency	50 Hz	100 Hz
Body Worn Position	Right Side of Lower Abdomen	Left Side of Upper Abdomen
Cost	$6500	$901.39

The BioHarness was placed on the left side of the upper abdomen and the Astroskin was placed on the right side of the lower abdomen.

The differences between the BioHarness and the Astroskin devices are:

1. **Resolution of Analog to Digital Converter**: Astroskin has a slightly higher resolution then BioHarness

2. **Sampling Frequency**: BioHarness has a higher sample frequency

3. **Placement of Body Sensor**: As visualized in Figure 3.2 and detailed in Table 3.1 the BioHarness was placed on the left side of the upper abdomen and the Astroskin was placed on the right side of the lower abdomen

Astroskin has a slightly higher resolution then BioHarness meaning it could capture the digital representation of the ACC signal more accurately. Even though the BioHarness has a higher sampling frequency, the sampling rate of the BioHarness was down-sampled to 50 Hz to match the sampling rate of Astroskin. The differences

in the wearable device placement and resolution will translate to a slightly different ACC signal recording.

The Astroskin and the BioHarness also have a stark difference in price. The Astroskin device were used as part of the data collection process because it was a part of a larger project spear-headed by the Canadian National Defence Department. The Zephyr BioHarness was selected as a comparison device because it had a difference in properties compared to the Astroksin and also it was a cheaper alternative.

The activities were not performed in a specific sequence therefore transitions between activities were not recorded. This experimental choice allowed to author to be certain that windows of data taken from specific recordings only contained the desired activity, thus providing certainty for model training ground truth. The potential limitation is that the data does not include the transition between activities.

The BioHarness was placed on the left side of the upper abdomen and the Astroskin was placed on the right side of the lower abdomen. During the data collection process the orientation and placement of both sensors on the authour's body were kept consistent. The BioHarness and the Astroskin wearable devices were synchronized through a light tap at the beginning of each recording session for each activity - this action created a spike in the data that was easily observed. After the synchronization step, the authour performed the designated activity. When activity was completed the wearable sensors were removed from the authors body and the collected data was uploaded to a Mac Mini for further processing.

Table 3.2 describes the amount of data collected for each activity and which recording session of the experimental protocol it was recorded. In each session the activity was performed continually.

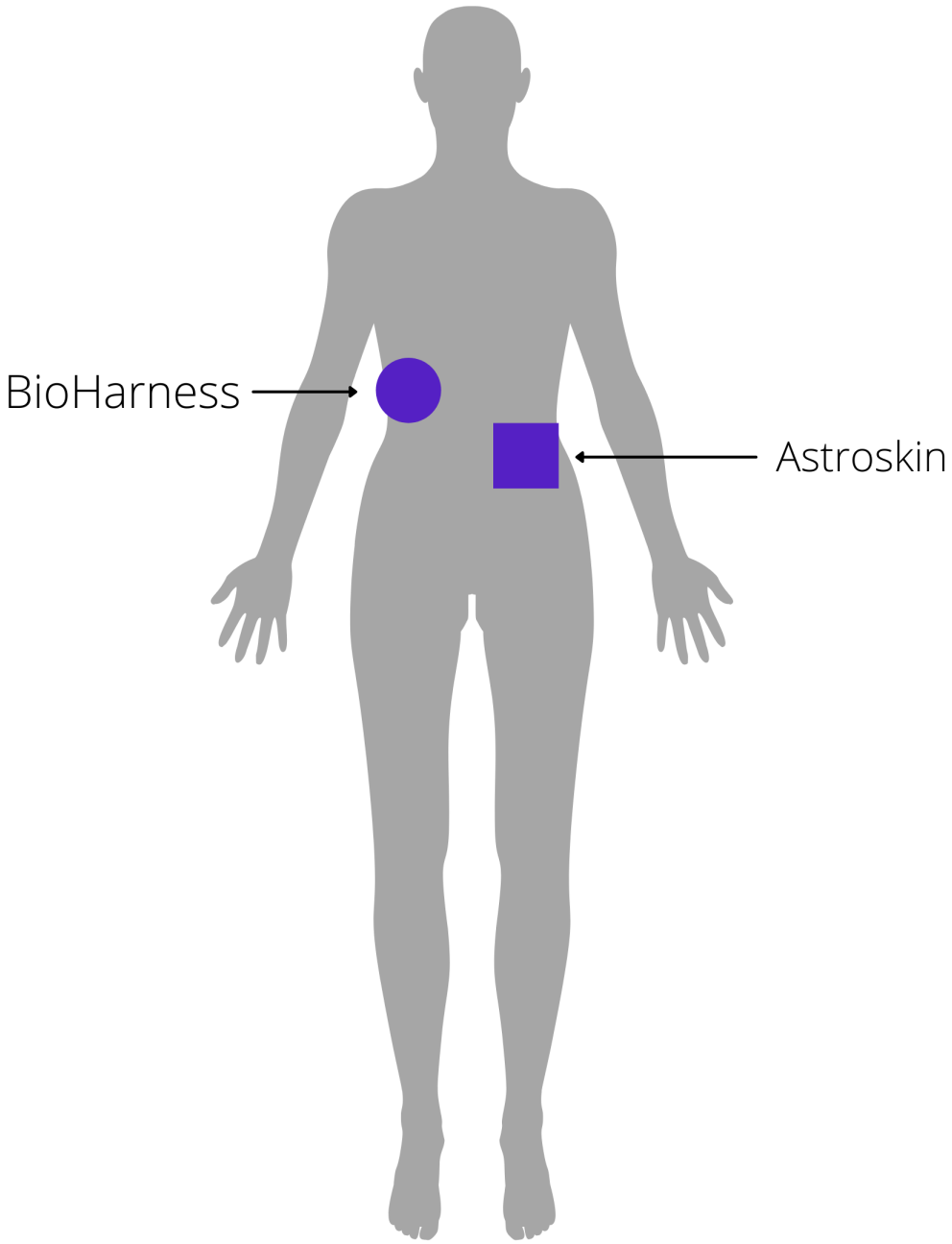


Figure 3.2: Location of sensors

Table 3.2: Amount of data recorded from each session

Activity	Session	Amount of Time (min)
Walking	1	16.3
Walking	2	3.5
Sitting	1	13.3
Sitting	2	3.0
Standing	1	15.0
Standing	2	2.8
Laying	1	12.7
Laying	2	3.0
Walking Upstairs	Multiple	18.5
Walking Downstairs	Multiple	21.2
Total	N/A	109.3

3.1.1 Data Organization

The collected data is organized into two datasets. Dataset 1 includes all the data recorded. Dataset 2 is a subset of Dataset 1 that only includes the activities walking, sitting, standing and laying. Dataset 2 is further organized into types of data based on the session the data was recorded and the wearable device that performed the recording. The different types of data are summarized in Table 3.3. When referring to Dataset 2, this terminology will be used throughout the thesis. The different data types are not to be confused with Type 1 and Type 2 error that are used in statistical hypothesis testing. In this thesis the types of the data refer to different categories of data present in Dataset 2.

Table 3.3: Types of data

	Session	Wearable Sensor
Type 1	1	Astroskin
Type 2	2	Astroskin
Type 3	1	BioHarness
Type 4	2	BioHarness

3.2 Preprocessing

For HAR models that use deep learning methods the preprocessing step is often omitted, especially if the signals are to be used in an end-to-end deep learning fashion. An example of this omitted step in an end-to-end approach was demonstrated by Imran and Latif (2020). However in this thesis some preliminary preprocessing steps are performed. This section describes the preprocessing of the acquired accelerometer (ACC) signals from Astroskin and BioHarness. Initally before the preprocessing steps the ACC signal of the BioHarness device is downsampled by a factor of two to match the sampling frequency of the Astroskin wearable device. The re-mapping and filtering stages of preprocessing is done on Dataset 1. Then the subset of data, Dataset 2 is extracted. Standardization and segmentation were performed on the two data sets separately.

3.2.1 Re-mapping

Astroskin and BioHarness have different frame of references for the x, y and z coordinate system. In order to obtain a common frame of reference for the ACC signal the y-axis and x-axis of the Astroskin ACC signal recordings were exchanged. Figure 3.3 shows frame of reference shared between the BioHarness and Astroskin.

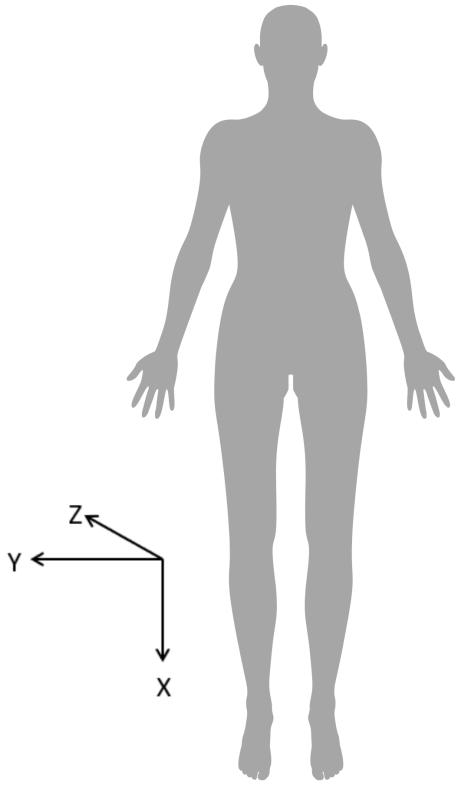


Figure 3.3: Coordinate system

3.2.2 Filtering

To inspect the recorded tri-axial ACC signals for noise (unwanted signals present in the signal of interest), the power spectrum for each recorded activity on each day was computed. The power spectrum was calculated by computing the fourier transform of the ACC signal and then taking the square of the absolute value of the spectrum. There is a constant 1 g force that acts downwards on the body and therefore a large DC component is present in the power spectrum of the data. The power spectrum of the x-axis for walking on Session 1 is shown in Figure 3.4 as an example.

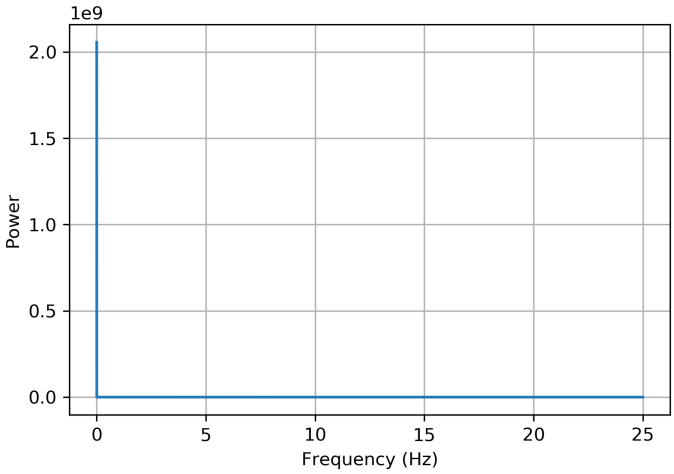
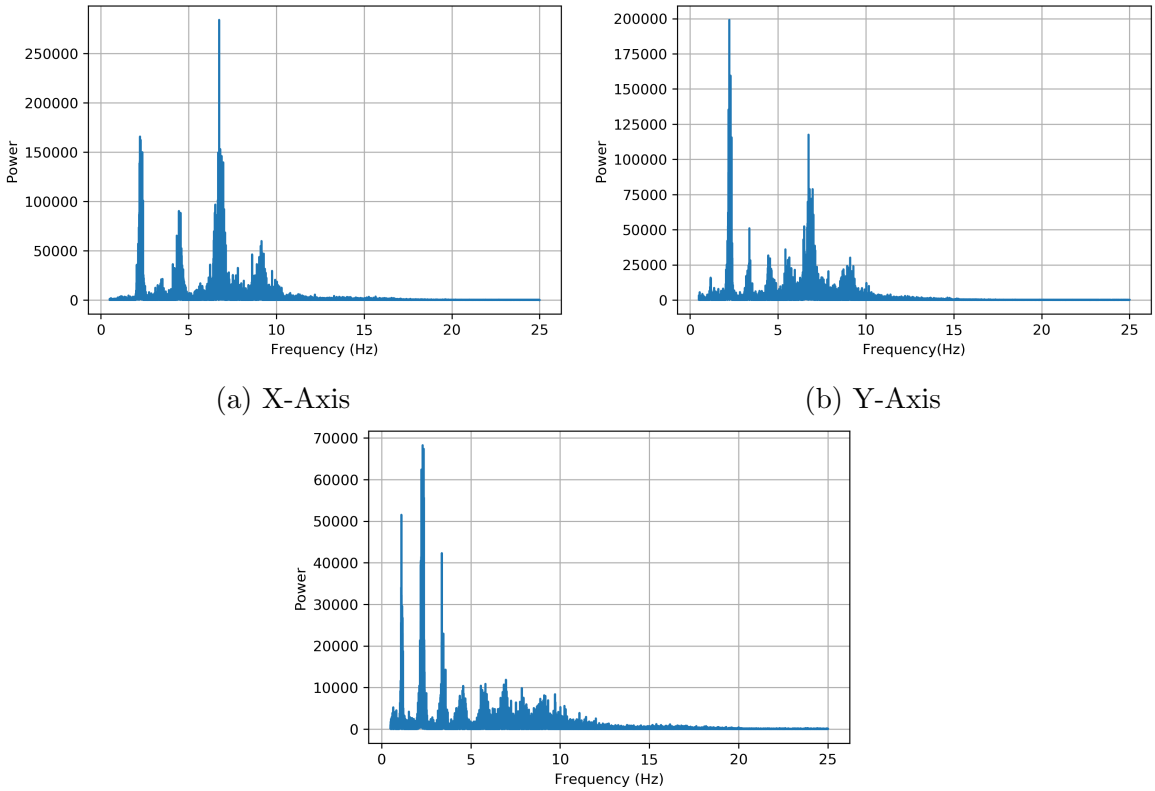


Figure 3.4: Power spectrum of session 1 walking: x-axis

The large DC spike makes it difficult to inspect other frequencies that are present in the signal. Therefore displayed power spectrums start at approximately 0.5 Hz in order to show the frequency content of the signal. For brevity the power spectrums of the walking activity from both sessions and wearable devices are included in this section. The power spectrums of the remaining activities are located in Appendix A. The power spectrum of the walking activity recorded by Astroskin on Session 1

is shown in Figure 3.5. The power spectrum of the walking activity recorded by Astroskin on Session 2 is shown in Figure 3.6. The power spectrum of the Walking activity recorded by the BioHarness on Session 1 is shown in Figure 3.7. The power spectrum of the walking activity recorded by BioHarness on Session 2 is shown in Figure 3.8.

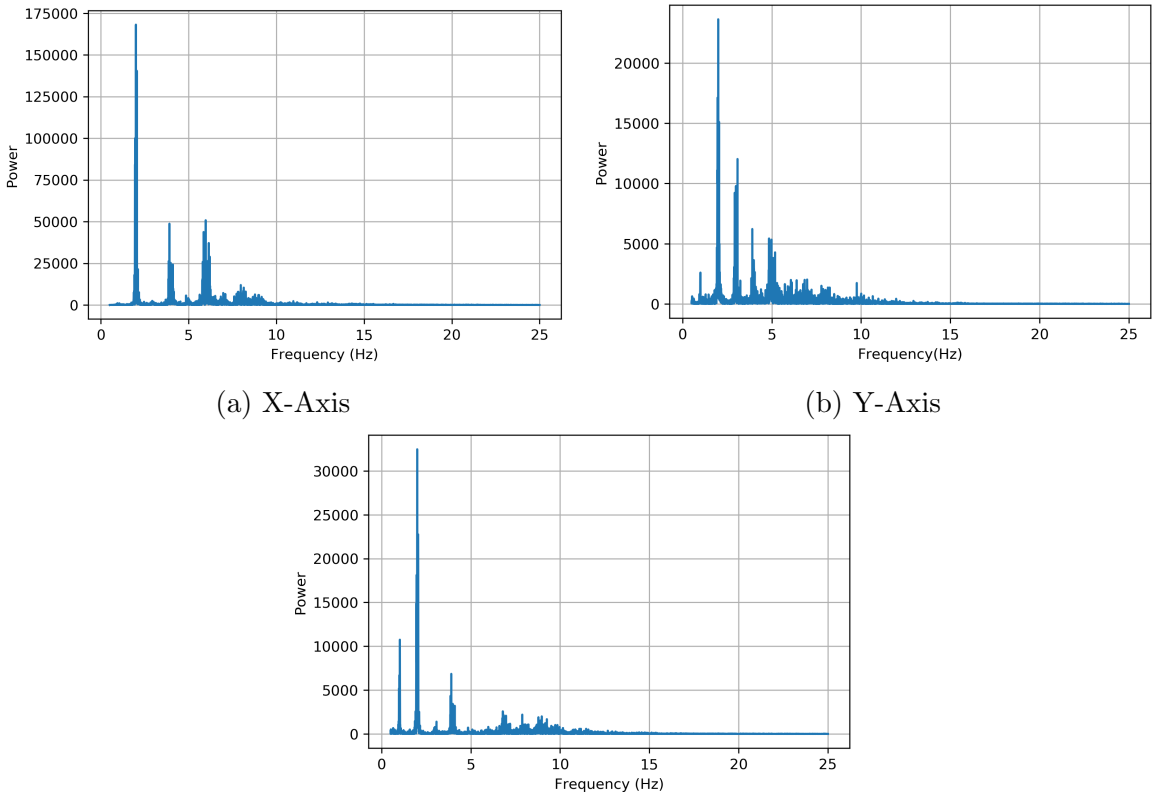


(a) X-Axis

(b) Y-Axis

(c) Z-Axis

Figure 3.5: Walking session 1 recorded by Astroskin

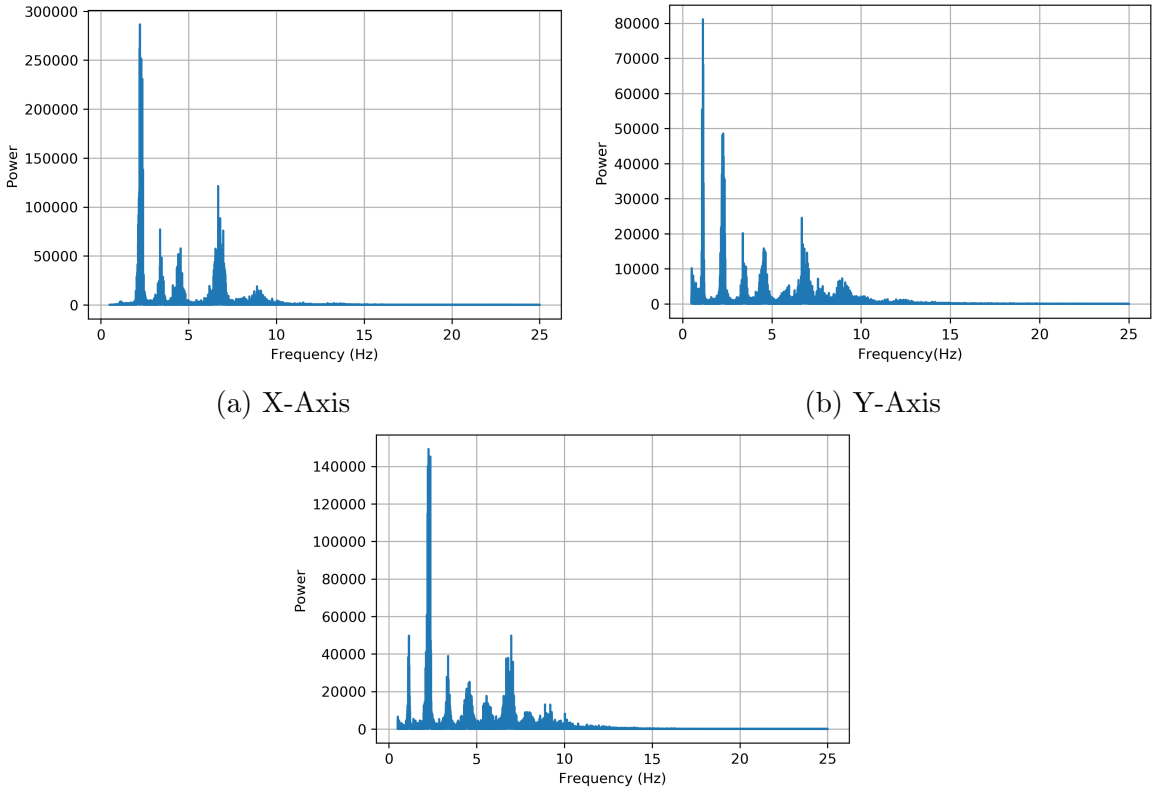


(a) X-Axis

(b) Y-Axis

(c) Z-Axis

Figure 3.6: Walking session 2 recorded by Astroskin

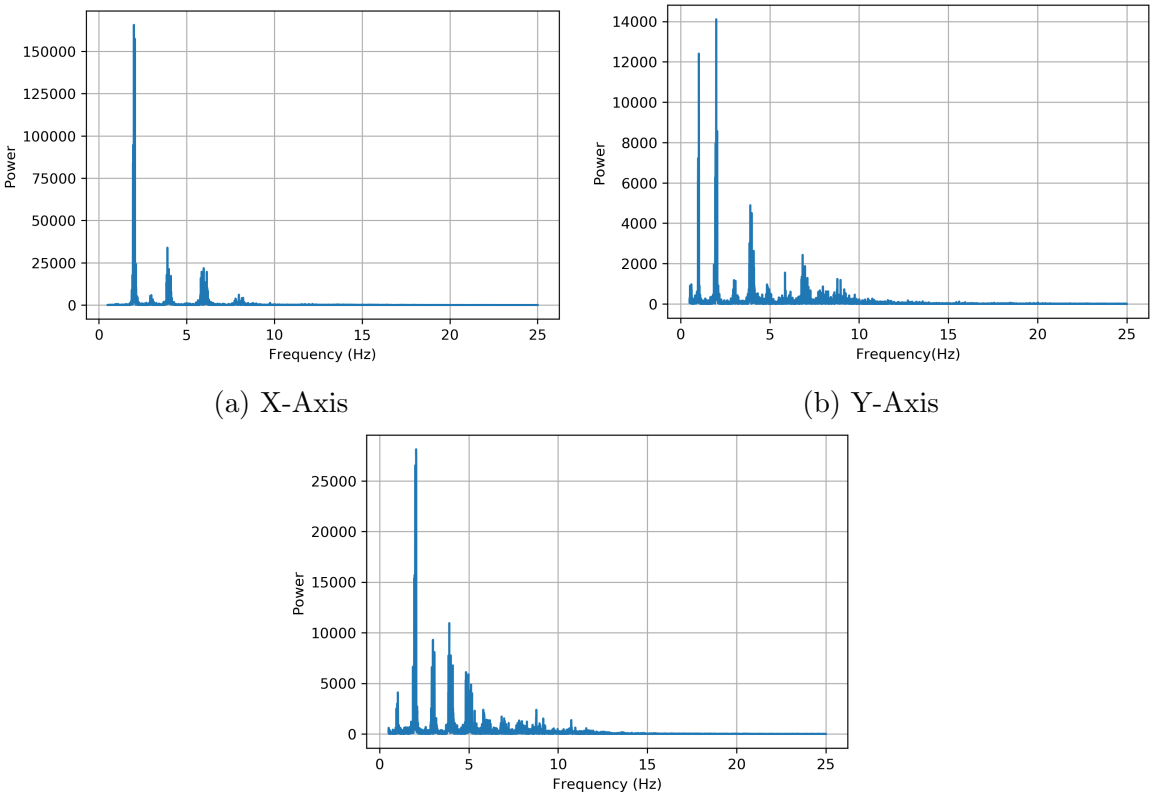


(a) X-Axis

(b) Y-Axis

(c) Z-Axis

Figure 3.7: Walking session 1 recorded by BioHarness



(a) X-Axis

(b) Y-Axis

(c) Z-Axis

Figure 3.8: Walking session 2 recorded by BioHarness

With a quick inspection it can be seen that the power spectrums are all quite different. Furthermore, it can be seen that there are some frequency components beyond 20 Hz. Karantonis *et al.* (2006) states that all measured body movements are contained within frequency components below 20 Hz. Therefore as in Anguita *et al.* (2013) a 3rd-order 20 Hz cutoff Butter-worth filter was implemented remove unwanted frequencies that may contaminate the signal.

3.2.3 Standardization

For a machine learning algorithm, to perform well the features should have similar scales (Géron, 2019). Standardization was used in this work to have similar scales in the features, which in this case was the x, y and z axis of the accelerometer data. "Standardization subtracts the mean value (so standardized values always have a zero-mean) and then divides by the standard deviation so that the resulting distribution has unit variance" (Géron, 2019, p.69). The mean and standard deviation was calculated from what was designated as the training set of the data as advised by Géron (2019). Then the mean and standard deviation was calculated over the training set of data. The standardization was performed following the technique of Zhang *et al.* (2019) were the x, y and z axes were standardized by subtracting the mean and dividing by the standard deviation of each channel of accelerometer data.

3.2.4 Segmentation

The data from each activity was then sectioned into 3 second windows with no overlap. In this thesis the windowing of the data was performed on the signal from each activity individually. The data collection process allowed for this because the activities were not performed sequentially by the participant. This methodology was employed because it would ensure that the windows have ground truth data and are not contaminated with transitional data. The windowing size was selected because it provided reasonable results during model development. Although outside the scope of this thesis, the study of the windowing for human activity recognition has been reviewed in literature and readers may refer to a comprehensive discussion by Banos *et al.* (2014).

3.3 Model Architecture

Firstly, the 1-D convolutional neural network was selected because it is a popular network used in HAR from ACC data because it requires no feature engineering. Kusuma *et al.* (2020) used a 1-D convolutional neural network and raw ACC data to classify walking, upstairs, standing, sitting, jogging and downstairs activities and achieved an accuracy of 95.9%. In terms of the trust aspect, a deep learning model was selected because deep learning models are regarded as opaque. The model architecture of 1-D convolutional neural network used in this thesis is shown in Figure 3.9.

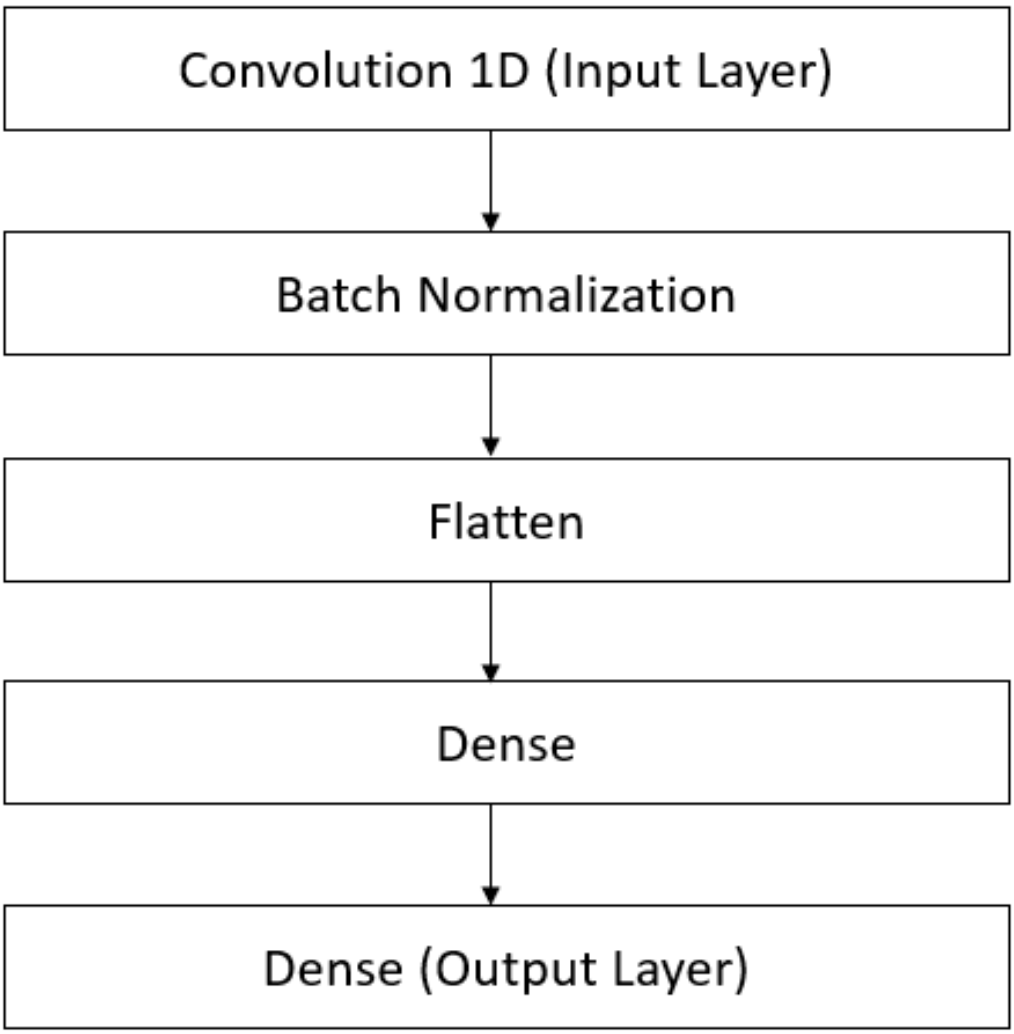


Figure 3.9: Model architecture

3.4 Model Development

This section describes the development of the proposed model architecture given in Section 3.3. In total three models with this architecture were created.

1. Model A

2. Model H

3. Model A Type 1

3.4.1 Model A

Model A was trained with Dataset 1 and therefore has the ability to classify all six activities. However only the data recorded by Astroskin on Session 1 and Astroskin on Session 2 was used to train Model A. The data was partitioned into three sets: the training set which was 80% of the data, the validation set which is 10% of the data and the test set which is 10% of the data. This partition is visualized in Figure 3.10

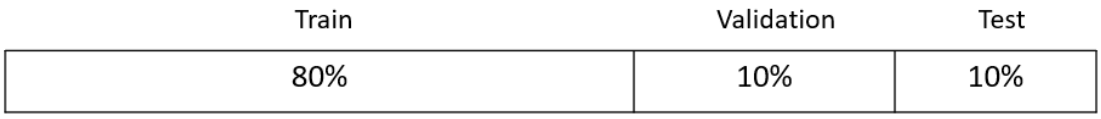
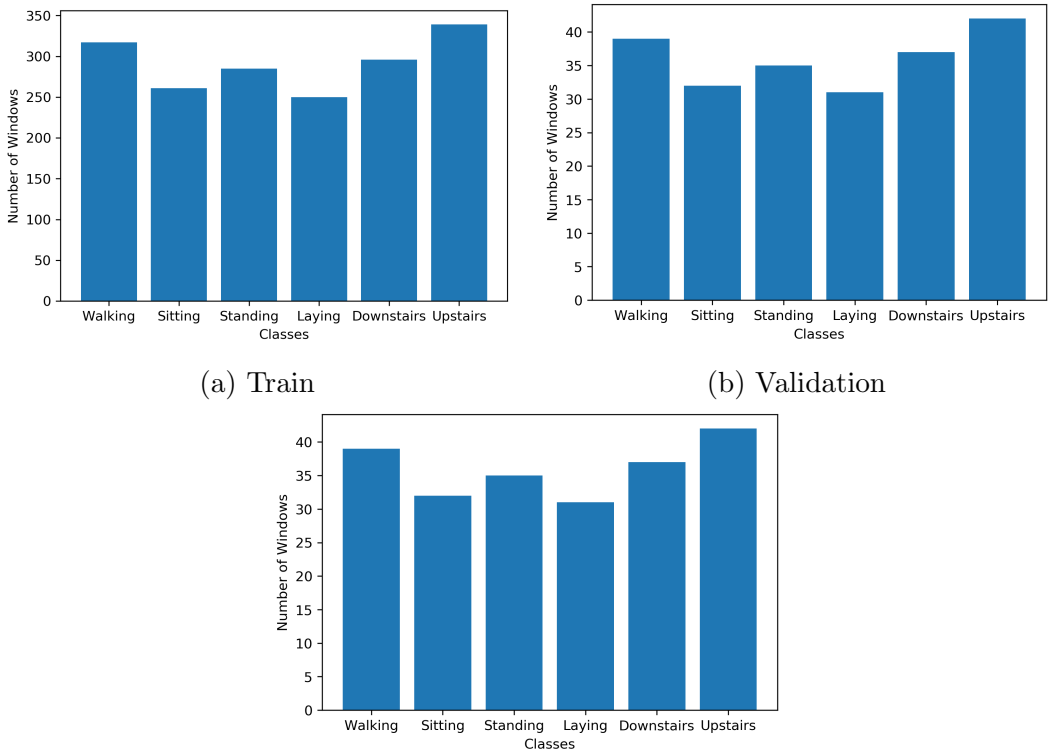


Figure 3.10: Partitioning of the data for evaluation

After this the data was standardized using the method described in Subsection 3.2.3. The mean and the standard deviation was calculated from the training portion of the data. Then the same mean and standard deviation was applied to the validation portion of the data and then the test portion of the data.

In the training set, validation set and the test set an equal amount of walking, sitting, standing and laying recorded on Session 1 and Session 2 was included. This was to ensure that data from Session 1 of the experiment did not end up solely into one set of the data. The upstairs and downstairs data was conducted over a course

of multiple days so there was no attention paid to the distribution over the training, validation and the testing set. After the data was partitioned the windows were generated. The distribution of the windows of the training, validation and test set by class are shown in Figure 3.11.



(a) Train

(b) Validation

(c) Test

Figure 3.11: Window distribution over train validation and test sets

As visualized in Figure 3.11 the data is balanced and activities are distributed relatively equally amongst the training set, the validation set and the test set.

The model was trained for 60 epochs with the Adam optimization algorithm (Kingma and Ba, 2014). A training batch size of 128 samples was used. A learning rate of 0.01 and a learning decay rate of 0.03 was selected. The loss function used during training was categorical cross entropy. The model architecture was determined

by evaluation on the the validation set. After a sufficient model architecture was determined, the validation set was re-added to the training set and the model was retrained on the training set as visualized in Figure 3.12. This technique was used in (Géron, 2019).

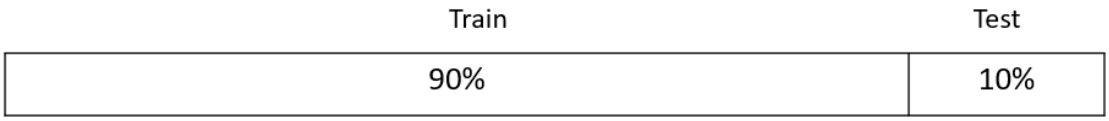


Figure 3.12: Partitioning of the data for evaluation

The model was trained using a GPU cluster. For reproducible results a seed was set.

3.4.2 Model H

Model H was also trained with Dataset 1 and therefore has the ability to classify all six activities. However only the data recorded by the BioHarness on Session 1 and BioHarness on Session 2 was used to train Model H. Like Model A the data was partitioned as visualized in Figure 3.10 and was also standardized in the same manner. The distribution of windows follow the same distribution visualized in Figure 3.11 as the BioHarness and Astroskin Data recorded the same amount of data simultaneously. Model H was trained using the same methodology as Model A. Again like Model H it trained using a GPU cluster. For reproducible results a seed was set.

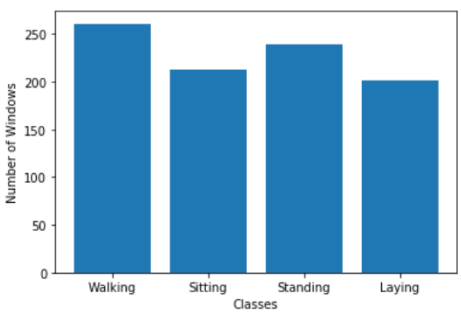
3.4.3 Model A Type 1

Model A Type 1 was trained with Dataset 2 and therefore has the ability to only classify four activities. As mentioned in Chapter 3 Section 3.1.1 Dataset B is further organized into types. The table is included here for convenience:

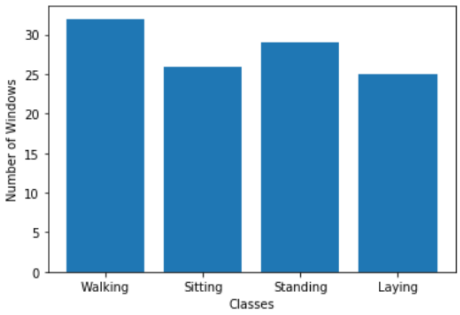
Table 3.4: Different data types

	Session	Wearable Device
Type 1	1	Astroskin
Type 2	2	Astroskin
Type 3	1	BioHarness
Type 4	2	BioHarness

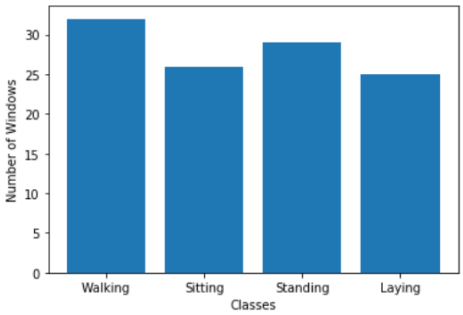
Model A Type 1 is only trained on data from Astroskin that is recorded on Session 1. Type 1 data is partitioned into three sets as visualized in Figure 3.10 and standardized in the same manner as Model A and Model H. The distribution of the windows for Type 1 data across the training, validation are shown in Figure 3.13.



(a) Type 1 train



(b) Type 1 validation



(c) Type 1 test

Figure 3.13: Window distribution over train validation and test sets

The model is trained is a similar way as Model A and Model H. However, setting the seed for this model did not produce reproducible results. Therefore the training of this model was repeated 50 times and the average accuracy as the well as the standard deviation was reported.

Chapter 4

Exploration of Trust Methodology

This chapter outlines the approach to evaluate robustness. Section 4.1 outlines the methodology for the *cross-domain tests*. Section 4.2 outlines the method used to predict performance degradation that may occur when using an alternate wearable sensor for the input ACC data. Section 4.3 discusses the approach of implementing the out of domain discriminator.

4.1 Cross-Domain Tests

The cross-domain test mimics a test that has been used on image-datasets (Torralba and Efros, 2011) in which a classifier is evaluated on a test set that differs from the training set used in development. To perform the cross-domain test for Model A, Model A was evaluated with a test set that was recorded by the BioHarness. Model H was evaluated with a test set that was recorded by the Astroskin. For Model A Type 1 the cross-domain tests were performed with regards to the wearable device as well as the session of recording. First Model A Type 1 was evaluated on Type 2 data

(recorded by the same sensor on a different session), Type 3 data (recorded on the same session simultaneously but by a different sensor) and Type 4 data (recorded on a different session and by a different sensor).

4.2 Performance Prediction

It was then investigated if the performance of Model A on a test set that was recorded by the BioHarness could be predicted. To do this, first different levels of thermal noise were added to the windows of the test set of Model A (which was recorded by Astroskin). Electrical noise caused by the random motion of electrons is called thermal noise (Haykin, 2001). This is visualized in Figure 4.1. The process of thermal noise generation is described in Subsection 4.2.1. The addition of different levels of thermal noise to each of the test set windows serve as *simulations* of different wearable devices. Each level of thermal noise provides a different ACC signal that can be interpreted as coming from a different wearable devices. The distance between the test set with the added thermal noise and the original test set is then calculated. The methodology for this calculation is presented in Subsection 4.2.2. The accuracy of Model A on test sets at different levels of thermal noise is computed. Through fitting a curve, this accuracy and the distance from the different test sets with different levels of added noise and the original test set is related. Then the distance between the test set recorded by the BioHarness and the test set recorded by Astroskin is calculated. The predicted accuracy on the BioHarness test set is obtained by evaluating the equation of the curve. The curve fitting procedure is presented in Subsection 4.2.3.

Figure 4.1: Noise Addition

4.2.1 Generating Thermal Noise

Hammad and El-Sankary (2019) based on work done by Madgwick *et al.* (2013) modelled thermal noise as additive zero-mean Gaussian noise and added it to ACC data. Based on acceptable signal to noise (SNR) levels found in literature Hammad and El-Sankary (2019) generated desired SNR values by adding noise of a specific power to the test set. In this thesis, a similar approach is adopted to add varying levels of noise to the test set. However, only one random simulation of noise is generated. This noise level is made reproducible (pseudo-random) by setting the seed of noise generator in the random Numpy library (Harris *et al.*, 2020) used for noise generation.

SNR is defined as the ratio of the average signal power to the average noise power (Haykin, 2001).

$$SNR = \frac{Power\ of\ Signal}{Power\ of\ Noise} \tag{4.2.1}$$

Often the ratio is expressed in decibels.

$$SNR(\text{dB}) = 10\log(SNR) \tag{4.2.2}$$

To generate a desired SNR, noise of a specific power was added to each of the axes

of the ACC data individually. The power of one axis of ACC is calculated as:

$$Power\ of\ ACC = \frac{1}{N}\sum_{i=1}^{N} X(i)^2 \tag{4.2.3}$$

where N represents the length of ACC signal and X(i) is the value of the ACC signal X at the i^{th} index.

From the given SNR in dB, the ratio is calculated using:

$$SNR = 10^{\frac{SNR(\text{dB})}{10}} \tag{4.2.4}$$

The power of the noise to be added is calculated as:

$$Desired\ Noise\ Power = \frac{Power\ of\ ACC}{SNR} \tag{4.2.5}$$

A zero mean Gaussian random variable with the desired noise power is generated and then added to the signal. This procedure is executed for each axis of the tri-axial ACC data.

4.2.2 Distance Measure

The distance measure between an example window in the original test set with no noise and an example window in the test set with noise uses the distance function described by Bai *et al.* (2012). The distance function (D) is presented in Equation 4.2.6:

$$D[X_1, X_2)] = 1/3\sum_{p=1}^{3}\sqrt{\sum_{b=0}^{m}[X_{1p}(b) - X_{2p}(b)]^2} \tag{4.2.6}$$

In this equation:

1. X_1 represents the original window with no noise

2. X_2 represents the noisy window

3. b is the index of the individual samples in the windows

4. m is the number of samples within the window

To determine the distance measure from the original test set, which contains clean windows and the noisy test set which contains noisy windows the methodology in Figure 4.2 was employed.

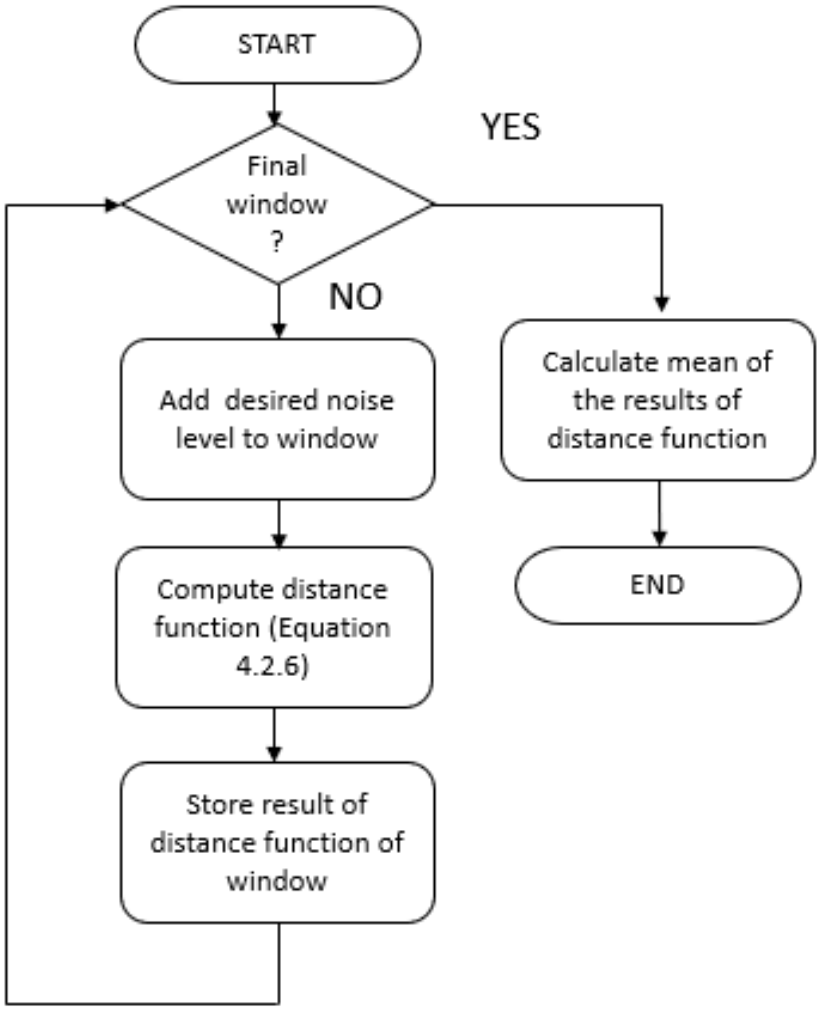


Figure 4.2: Distance Across Test Set

The distance measure between an example window in the original test set with no noise and an example window in the test set that was recorded by the BioHarness is also computed with Equation 4.2.6. The distance measure from the original test set and the test set recorded by the BioHarness is computed following the methodology in Figure 4.2 **excluding** the process of adding noise to the windows.

4.2.3 Curve Fitting

The accuracy of Model A on the test sets with different levels of thermal noise and the distances of these test sets from the original test set were plotted. The type of curve used is the equation below.

$$Accuracy = A + \frac{C}{1 + (\frac{Distance}{B})^3} \tag{4.2.7}$$

where A, B and C are to be determined. The variables A, B and C were determined through using the optimize function provided by the SciPy library (Jones *et al.*, 01). By substituting the distance of the BioHarness test set to the original test set into the equation the expected accuracy Model A on the BioHarness test set is estimated.

4.3 Out of Domain Discriminator

A discriminator is used in generative adversarial networks which was proposed by Goodfellow *et al.* (2014). As a part of a generative adversarial network, the discriminator attempts to distinguish between real and fake generated samples. In this work a domain discriminator is explored with Dataset 2 and Model A Type 1. The domain discriminator is tasked with three different challenges.

1. Challenge 1: Can it differentiate Type 1 and Type 2 data?
2. Challenge 2: Can it differentiate Type 1 data and Type 3 data?
3. Challenge 3: Can it differentiate Type 1 and Type 4 data?

The walking activity was selected to perform this analysis. The challenge of the discriminator can then be more specifically rephrased as:

1. Can the discriminator differentiate between walking recorded by Astroskin on Session 1 and Session 2?

2. Can the discriminator differentiate between walking recorded by Astroskin on Session 1 and walking recorded by the BioHarness on Session 1?

3. Can the discriminator differentiate between walking recorded by the Astroskin on Session 1 and Walking recorded by the BioHarness on Session 2?

The architecture of the domain discriminator is visualized in Figure 4.3.

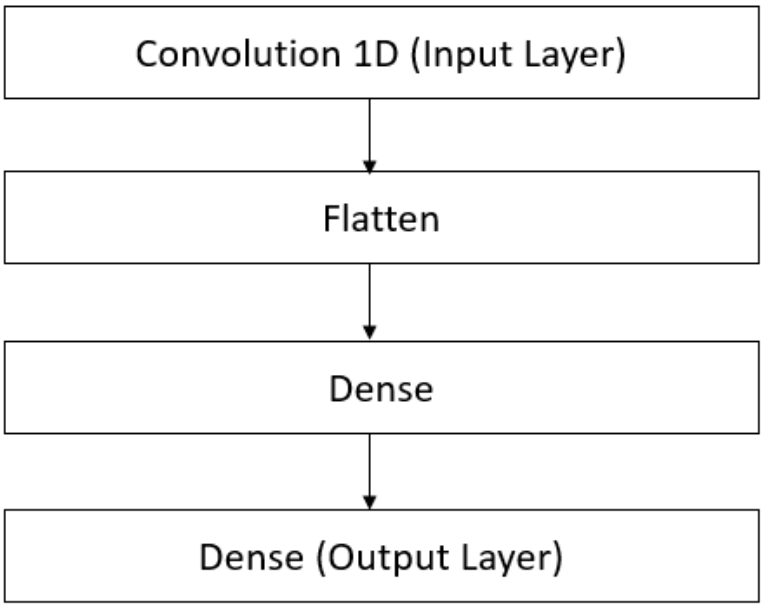


Figure 4.3: Model Architecture Domain Discriminator

It trained using the Adam optimizer(Kingma and Ba, 2014), with the default learning rate and a batch size of 128 training samples for 60 epochs.

4.4 Out of Domain Generalizable Discriminator

The generalizable domain discriminator's task is to be able to alert the user when the data is out of domain. First, the generalizable domain discriminator is trained to differentiate between noisy Type 1 data and non-noisy Type 1 data. The level of thermal noise is treated as a hyperparameter and is tuned until the domain discriminator can achieve perfect classification on the task. Then it is investigated if this same discriminator can differentiate other data types (Type 2, Type 3, Type 4) as out of domain with respect to Type 1.

The walking activity was selected to perform the analysis for this exploratory study. The walking activity was selected because the pattern of walking was is the most variable from the other collected activities in Dataset 2. Sitting, Standing and Laying produce similar accelerometer signatures because there is no defined movement in these activities. The architecture of the generalizable domain discriminator is the same architecture depicted in Figure 4.3. It was trained using the Adam optimizer, with the default learning rate, and a batch size of 128 training samples for 60 epochs.

Chapter 5

Results

This thesis investigated the robustness of Model A, Model H and Model A Type 1. The chapter describes the results of the tests conducted to analyze the robustness of the HAR models. Section 5.1 describes the results obtained in the cross-domain tests. Section 5.2 presents the results obtained for predicting the performance of Model A on test set that was recorded by the BioHarness. Section 5.3 discusses the results obtained from the out of domain discriminator. A summary of the models developed as well as the results for the cross-domain tests is given in Table 5.1 for easy reference.

Table 5.1: Summary of results

Model	Dataset	Data Acquisition Train Set	Data Acquistion Test Set	Accuracy
A	1	Astroskin	Astroskin	99.07%
A	1	Astroskin	BioHarness	65.74%
H	1	BioHarness	BioHarness	95.37%
H	1	BioHarness	Astroskin	29.63%
A Type 1	2	Type 1	Type 1	99.57% +/- 1.76%
A Type 1	2	Type 1	Type 2	50.95% +/- 5.99%
A Type 1	2	Type 1	Type 3	41.31% +/- 13.71%
A Type 1	2	Type 1	Type 4	19.28 +/- 11.83%

5.1 Cross Domain Tests

The cross domain tests evaluated were as follows:

1. The performance of Model A and Model H on test sets that were recorded by a wearable sensor that did not record the training set

2. The performance of Model A Type 1 on the different types of data each of which represented a different domain

5.1.1 Model A

First, Model A is evaluated on the held out test set that was recorded from Astroskin, which is the same sensor that recorded its training set. The accuracy of Model A on this test set was 99.07%. The confusion matrix is presented in Figure 5.1. All confusion matrices values represent the amount of windows that were predicted correctly.

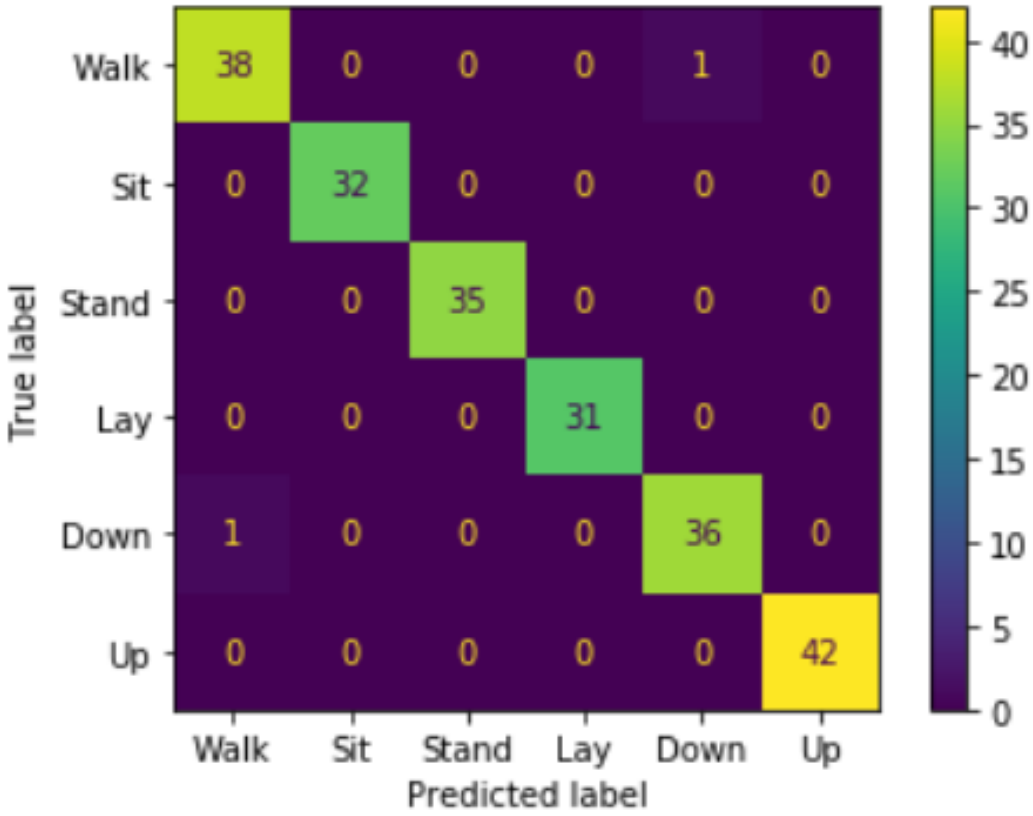


Figure 5.1: Confusion matrix of test set recorded by Astroskin

Model A achieved a good classification on the test set recorded by Astroskin. As seen in Figure 5.1, some walking samples were confused with downstairs samples. This is expected because walking and downstairs ACC signals are similar. However,

the accuracy of Model A on a test set recorded by the BioHarness was 65.74%. The confusion matrix is presented in Figure 5.2.

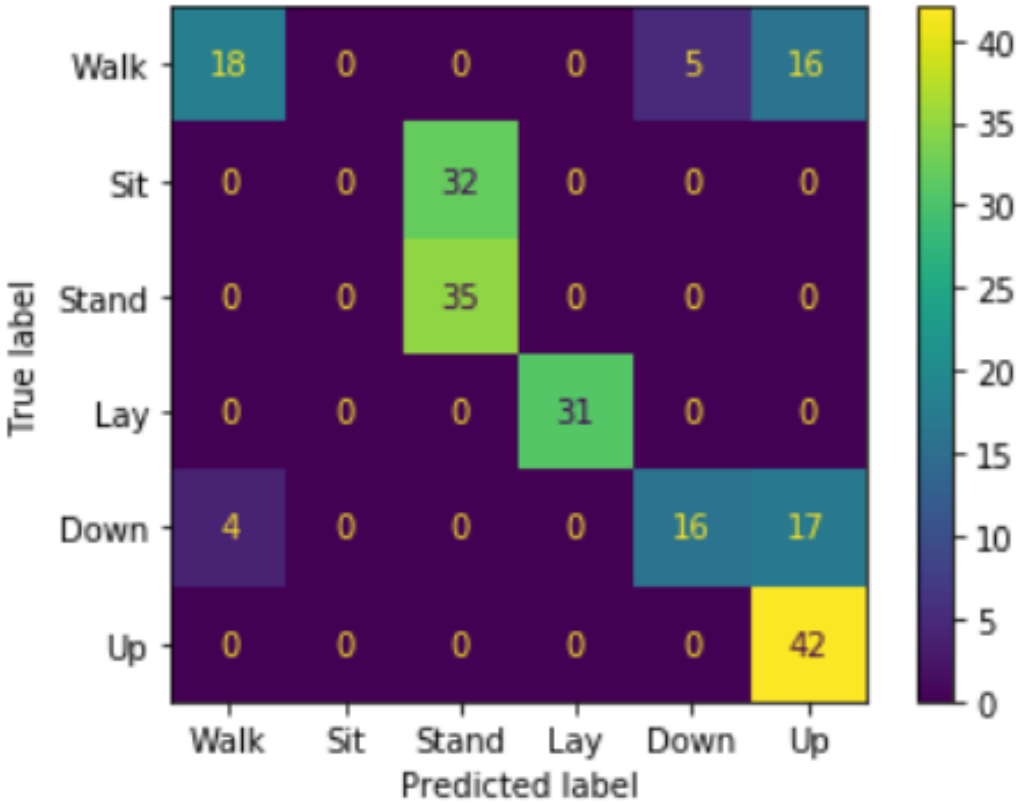


Figure 5.2: Confusion matrix test set recorded by Bioharness

In Figure 5.2, Model A experiences confusion between walking, downstairs and upstairs. Furthermore, sitting which was originally perfectly classified by in the test set recorded by Astroskin is confused with standing. The classes that are confused by Model A are reasonable because the ACC signals of walking, downstairs and upstairs and the ACC signals of sitting and standing are visually similar.

5.1.2 Model H

The accuracy of Model H on a test set recorded by the BioHarness was 95.37% percent. The confusion matrix is presented in Figure 5.3.

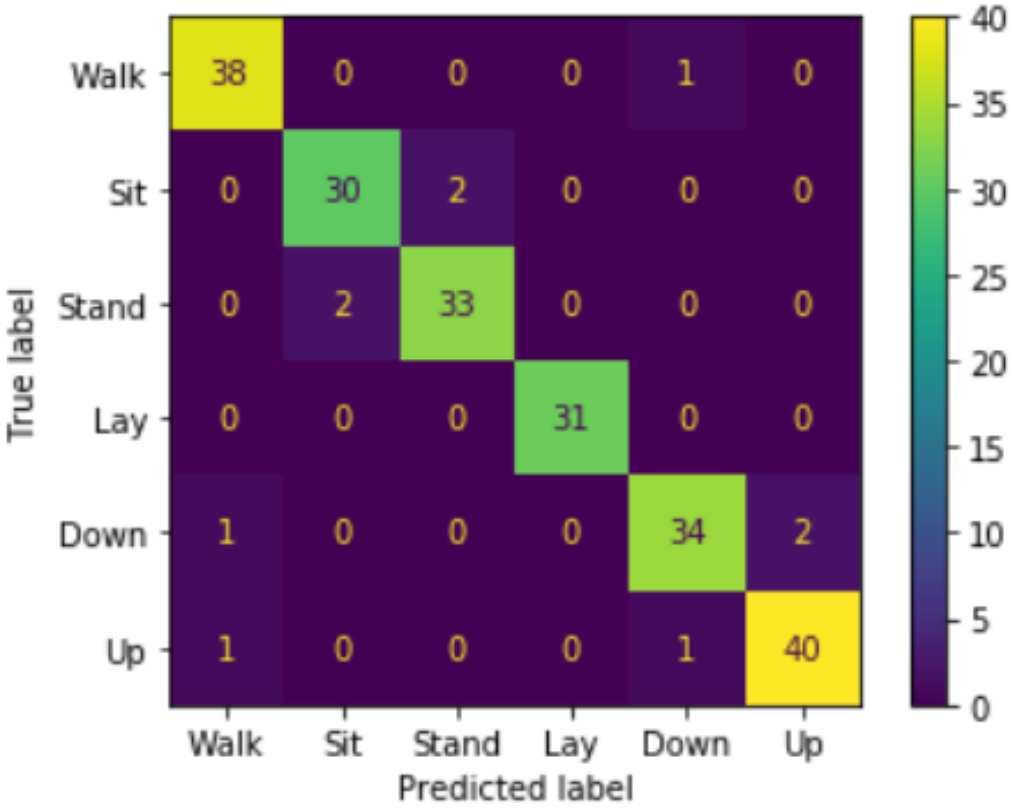


Figure 5.3: Confusion matrix test set recorded by Bioharness

In Figure 5.3 Model H misclassifies 1 walking sample as a downstairs sample and there is some misclassification between the sitting and standing classes. Next Model H was evaluated on the held out test set that was recorded by Astroskin. The accuracy of Model H on this test set was 29.63%. The confusion matrix is presented in Figure 5.4.

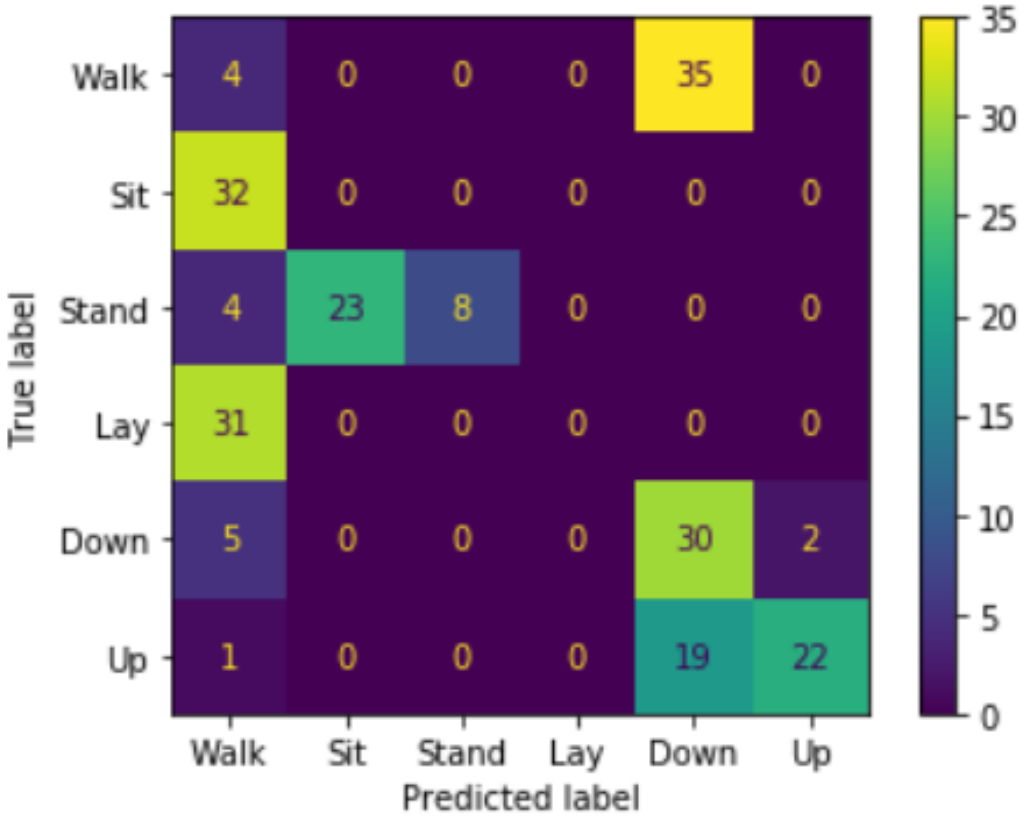


Figure 5.4: Confusion Matrix test set recorded by Astroskin

In Figure 5.4 Model H wrongly predicts the activities of sitting and laying to be walking. It also confuses the downstairs and upstairs activities. Some standing activities are also confused with sitting activities.

5.1.3 Model A Type 1

Firstly, as a benchmark, the performance Model A Type 1 is evaluated on a Type 1 test set. The accuracy of the model was 99.57% +/- 1.76%.

Figure 5.5 presents an example of a confusion matrix on the first one trial run.

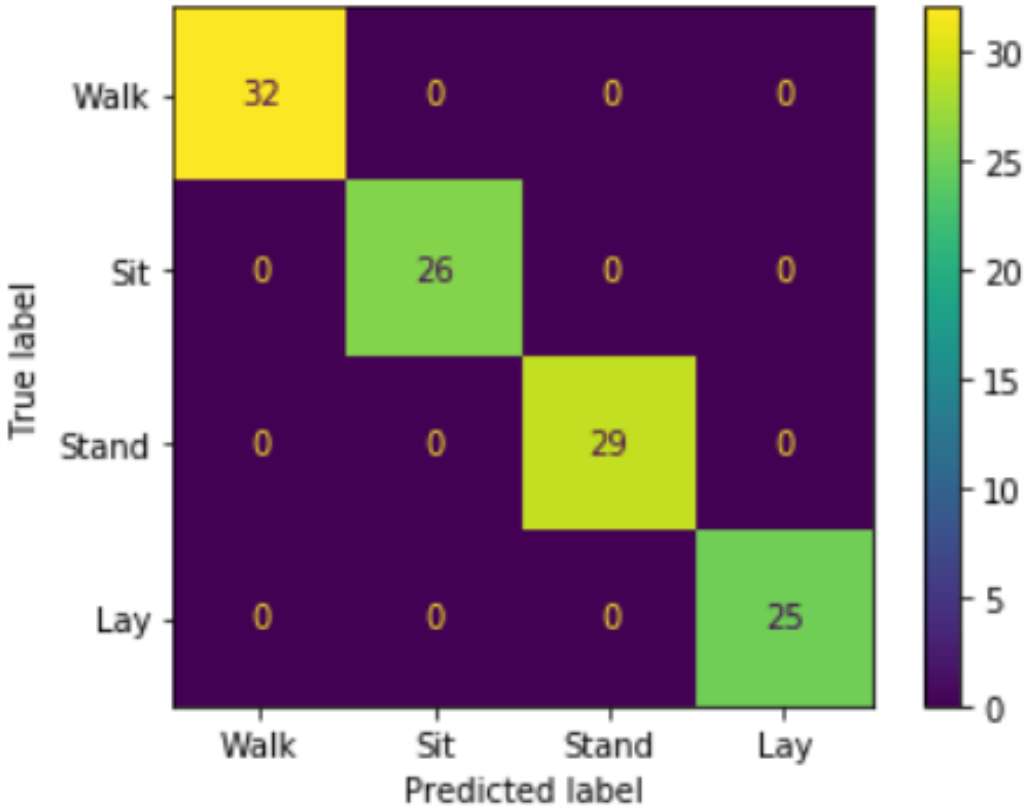


Figure 5.5: Confusion Matrix Type 1 Data

Next Model A Type 1 was evaluated on Type 2 data. The accuracy of this model on this test set was 50.95% +/-5.99%. Figure 5.6 presents an example of a confusion matrix on the first trial run.

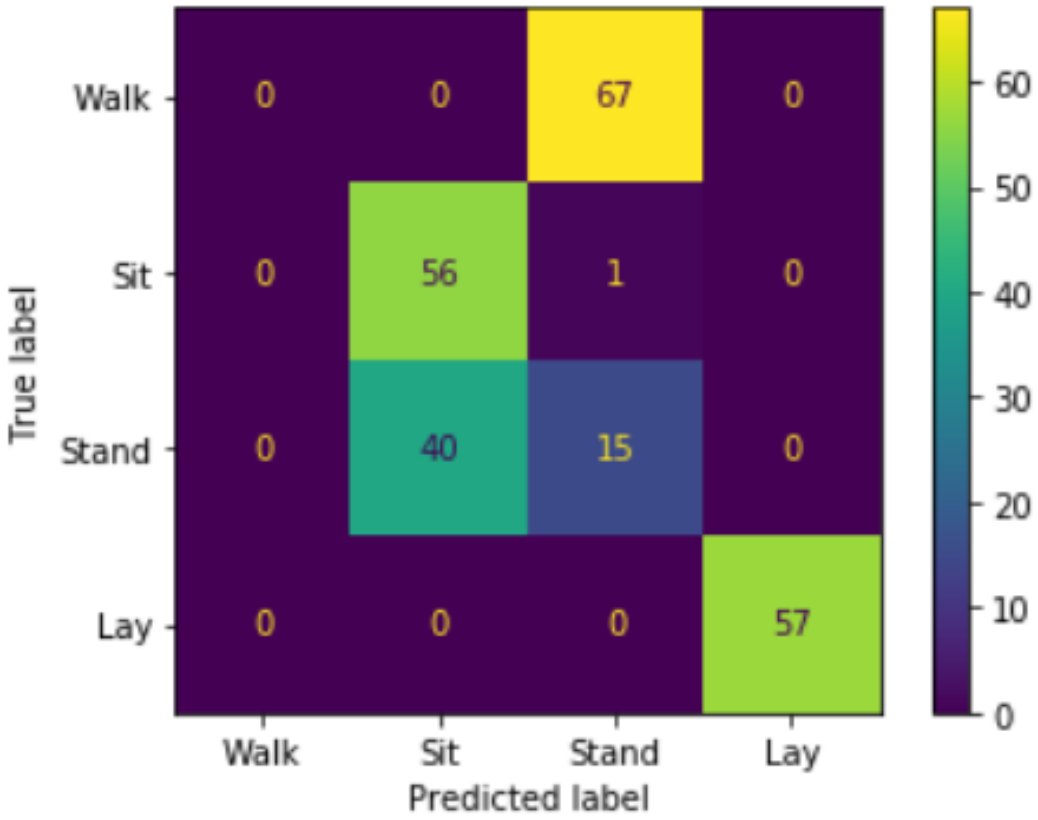


Figure 5.6: Confusion matrix type 2 data

On Type 2 data Model A Type 1 misclassifies the walking activities to be the standing. There is also some misclassification between sitting and standing.

Model A Type 1 is then evaluated on Type 3 data. The achieved accuracy was 41.31% +/-13.71%. Figure 5.7 demonstrates the confusion matrix for the first trial run of the experiment.

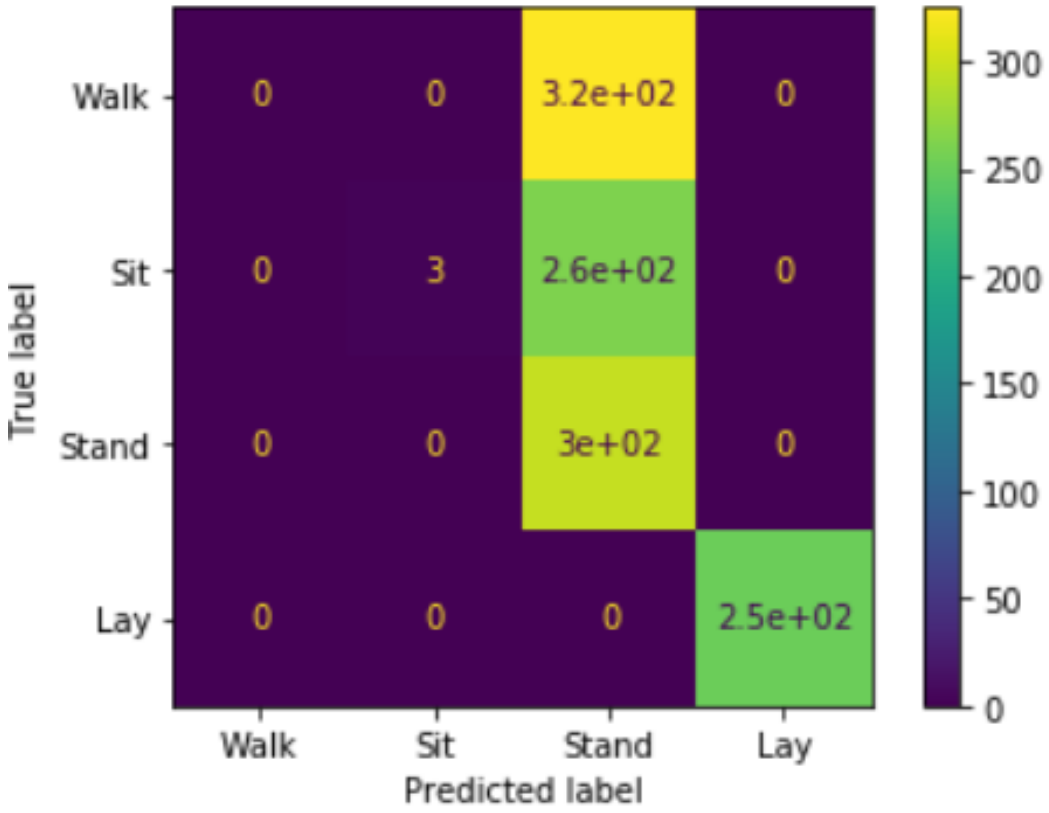


Figure 5.7: Confusion Matrix Type 3 Data

On Type 3 data Model A Type 1 classifies most samples to be standing. Only the laying class correctly classified.

Lastly, the model was then evaluated on Type 4 data. The achieved accuracy is 19.28% +/- 11.83%. Figure 5.8 presents the confusion matrix from one trial run.

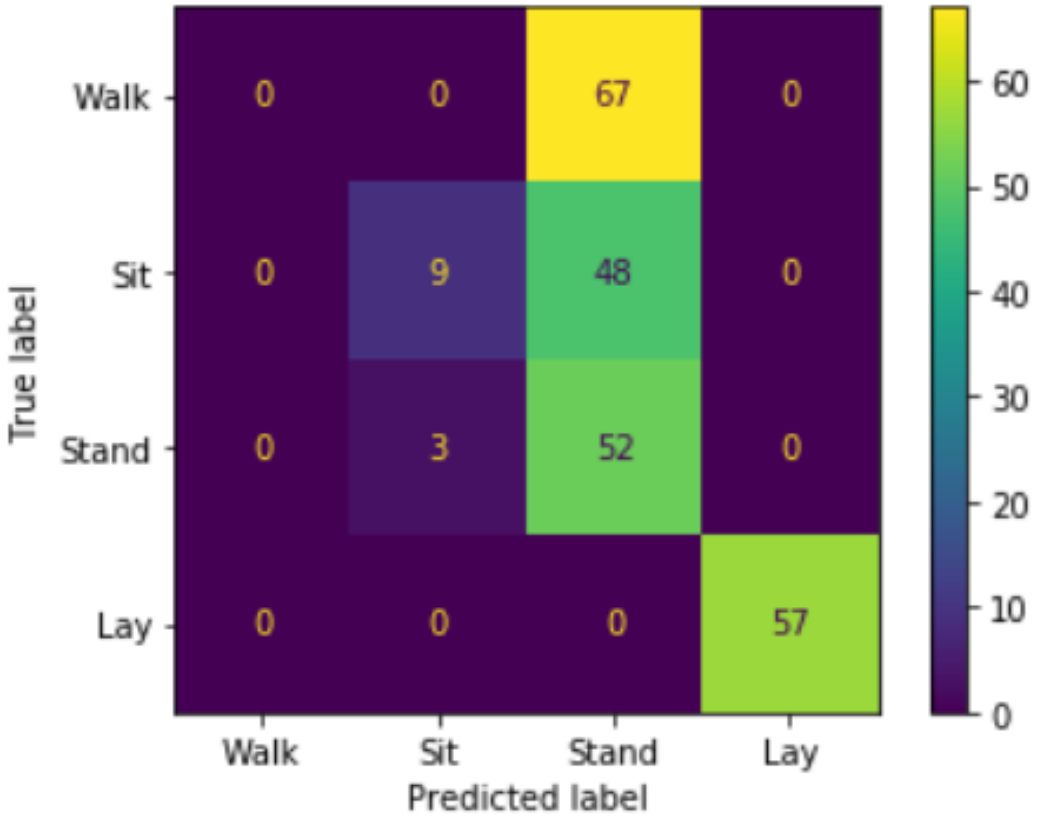


Figure 5.8: Confusion matrix type 4 data

Model A Type 1 confuses the sitting and the standing classes. It also incorrectly classifies the samples from the walking class.

5.2 Performance Prediction

The results for the performance prediction of Model A on the BioHarness test set are provided below. The trend of accuracy with increasing distance from the original test set is shown in Figure 5.9

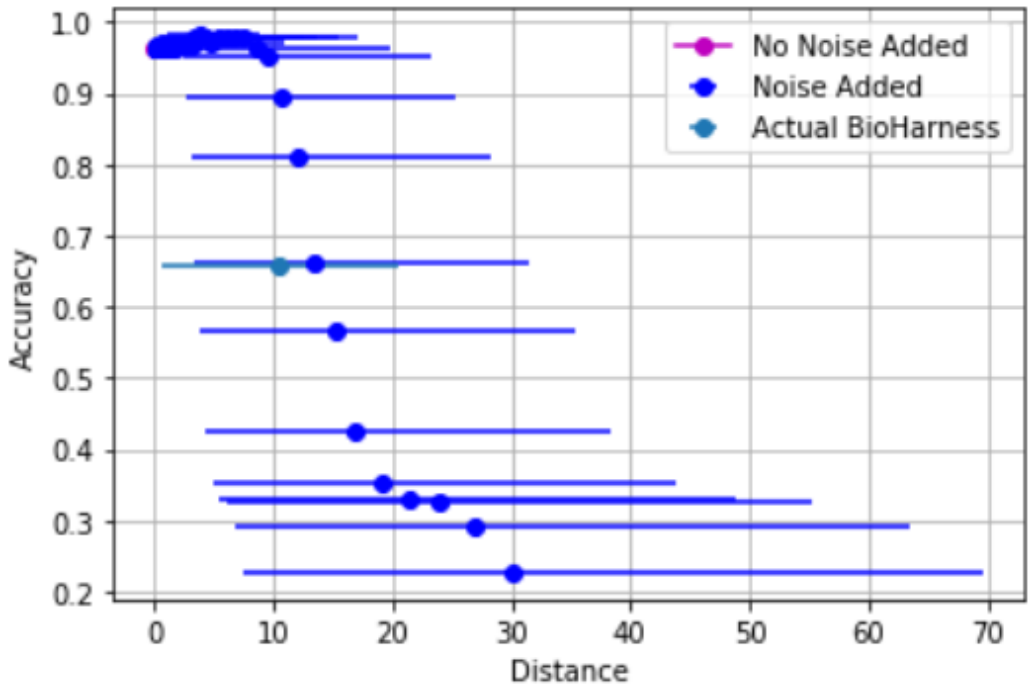


Figure 5.9: Distance vs. Accuracy

The equation of the fitted curve is:

$$Accuracy = 0.034 + \frac{0.941}{1 + (\frac{Distance}{17.601})^3} \tag{5.2.1}$$

The graph of the fitted curve (without the error bars for clarity), the accuracy on the orginial test set, the predicted accuracy on BioHarness test and the actual accuracy on the BioHarness test set of Model A is shown on Figure 5.10.

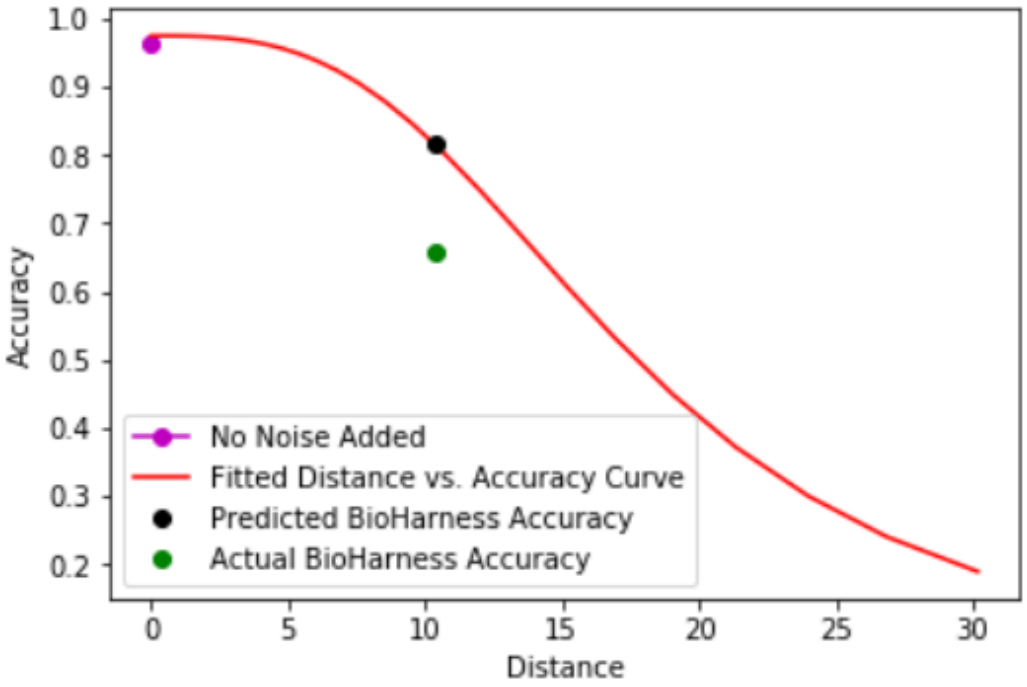
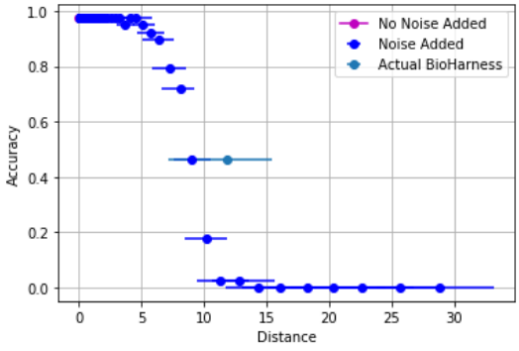
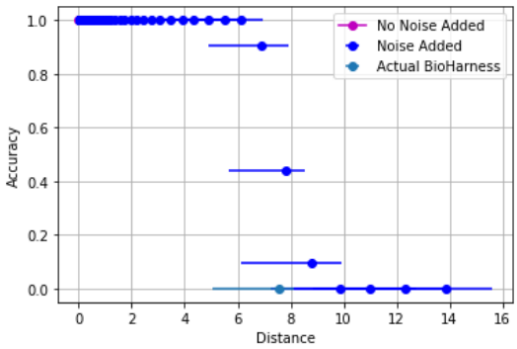


Figure 5.10: Fitted curve

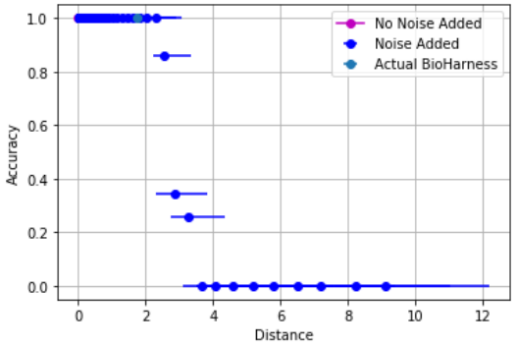
As a further investigation the trends of accuracy with each increasing distance from the original test set in terms of the classes of Walking, Sitting, Standing, Laying, Downstairs and Upstairs and where the actual accuracy of Model A falls on this curve is shown in Figure 5.11. The no noise added data points often overlap with data points around distance 0 and therefore can not always be properly observed in Figure 5.11.



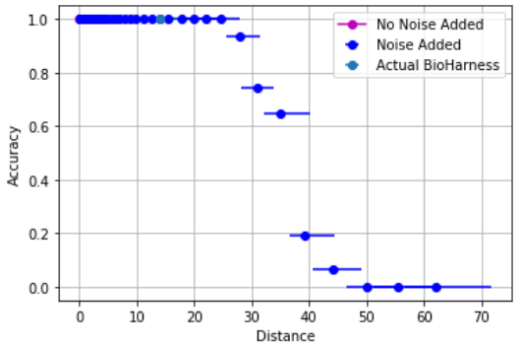
(a) Walking



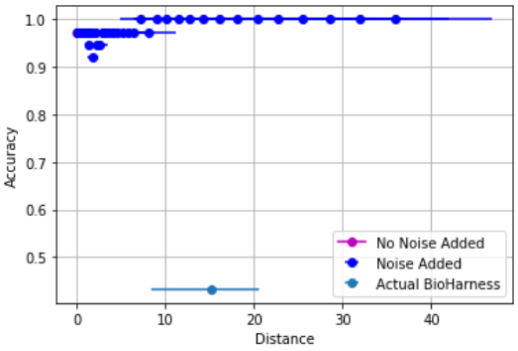
(b) Sitting



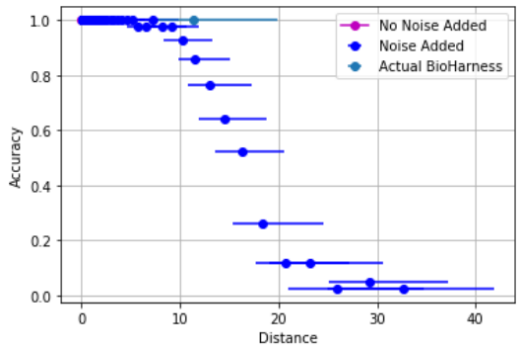
(c) Standing



(d) Laying



(e) Downstairs



(f) Upstairs

Figure 5.11: Distance vs. accuracy - individual activities

5.3 Out of Domain Discriminator

A discriminator was trained in order to differentiate between the different types of data. The results of the 3 different challenges as outlined in Chapter 4 Section 4.3 are presented in the following subsections.

5.3.1 Challenge 1

The first challenge evaluated is if the the discriminator was able to differentiate between Type 1 and Type 2 data. The discriminator differentiated between these two data types with an accuracy of 100%. The confusion matrix is seen below in Figure 5.12.

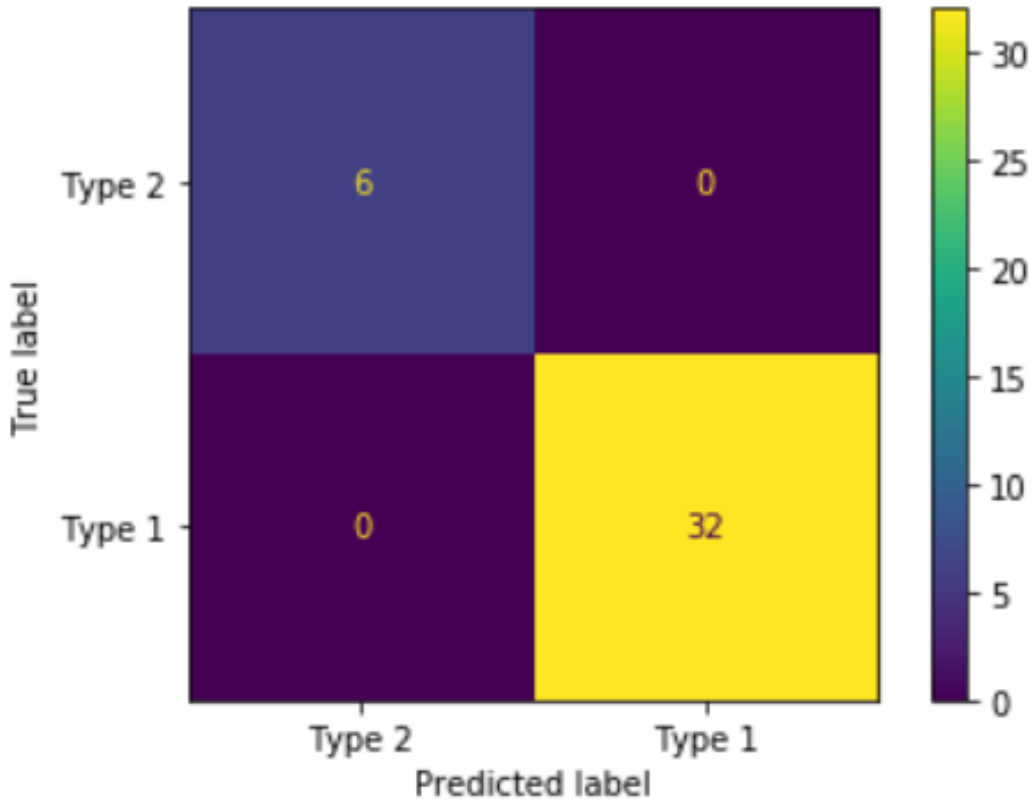


Figure 5.12: Confusion matrix: discrimination between type 1 and type 2 Data

5.3.2 Challenge 2

For the second challenge the discriminator was tasked to differentiate between Type 1 and Type 3 data. The discriminator achieved an accuracy of 100% on this task.

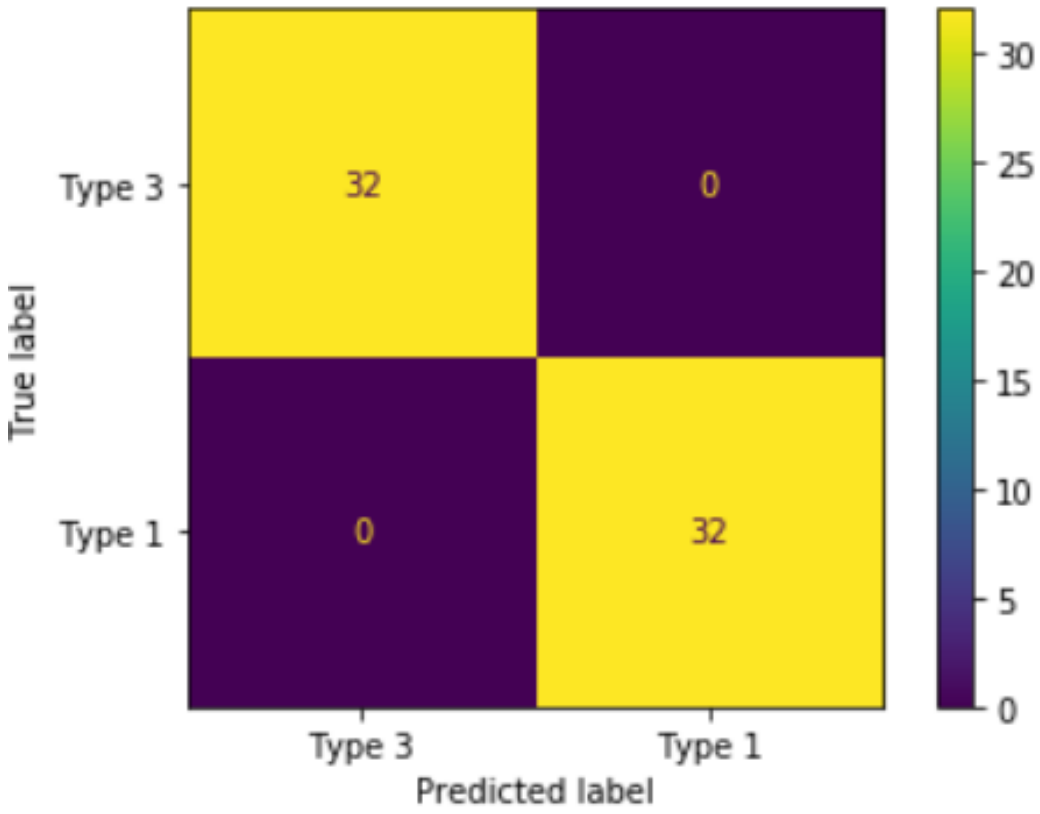


Figure 5.13: Confusion matrix: discrimination Between type 1 and type 3 Data

5.3.3 Challenge 3

For the third challenge the discriminator was tasked with being able to differentiate between between Type 1 and Type 4 data. The discriminator achieved an accuracy of 100% on this task.

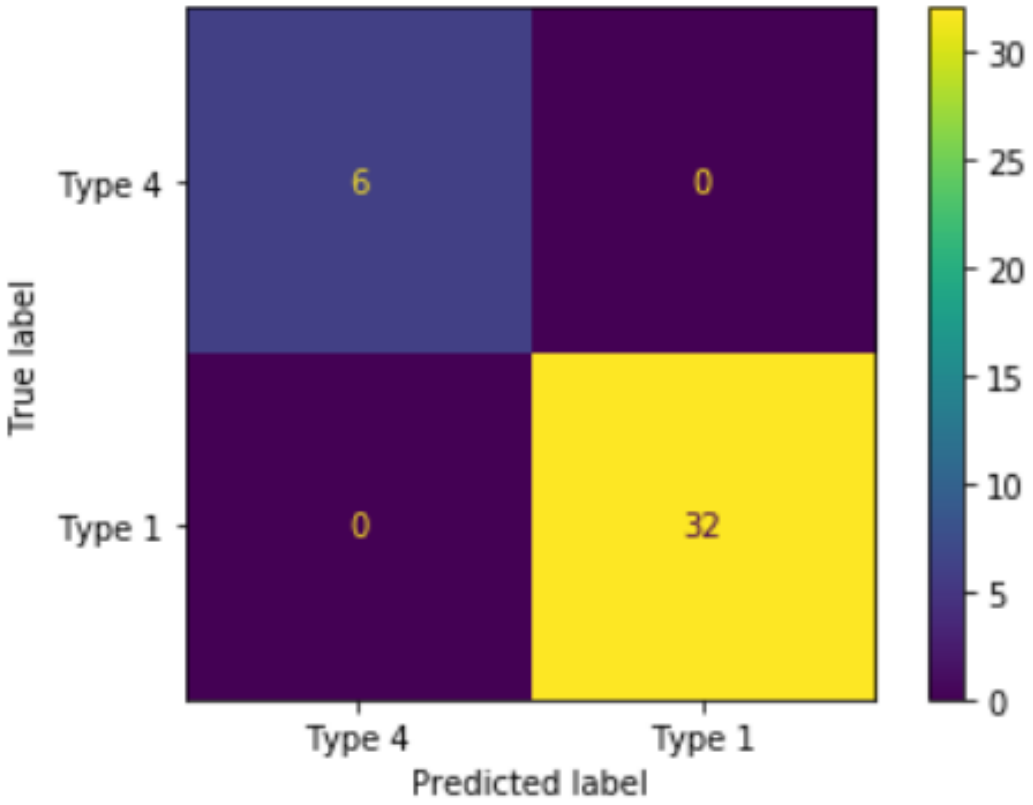


Figure 5.14: Confusion matrix: discrimination between Type 1 and Type 4 Data

5.4 Out of Domain Generalizable Discriminator

The results in Subsection 5.3 demonstrate that the discriminator can differentiate between Type 1 and the other Types of data recorded in this thesis. The following results show a generalizable discriminator trained to discriminate between Type 1 and a noisy version of Type 1 data. This discriminator is then tasked with the challenge of discriminating between Type 1 data and the other types of data present. A discriminator that is able to do this provides a more generalized out of domain detection.

The discriminator achieved an 100% accuracy in the task of discriminating between noisy data with a thermal noise level of 8 dB and non-noisy data. The confusion matrix is presented below.

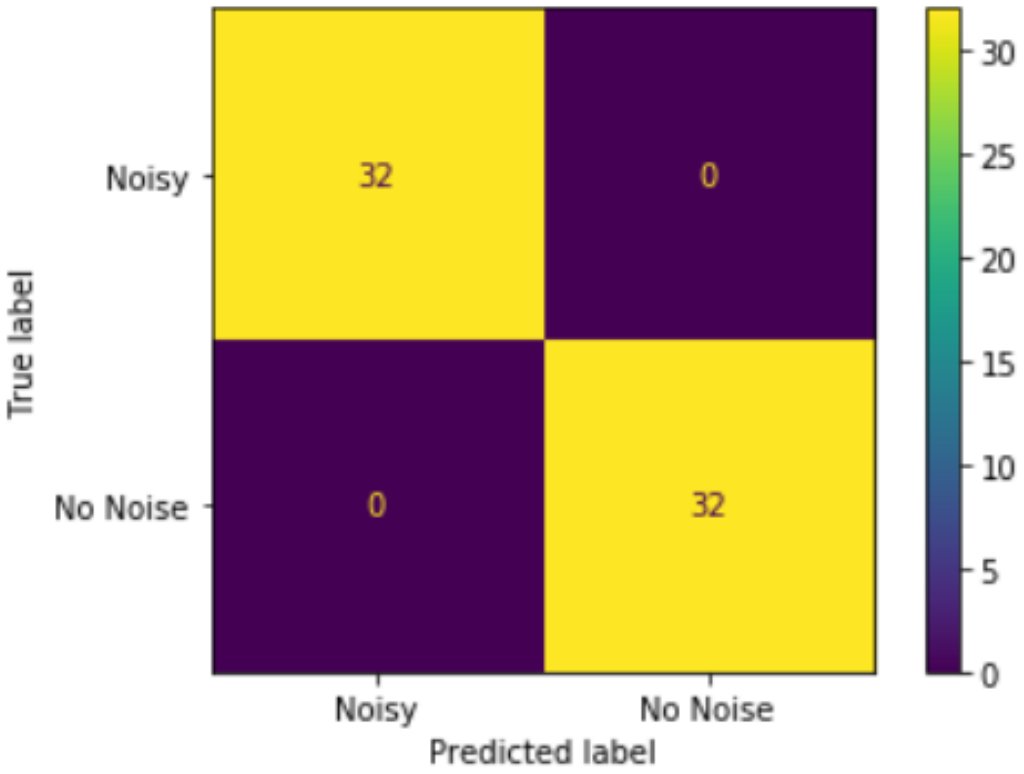


Figure 5.15: Confusion matrix: discrimination between noisy Type 1 and non-Noisy Type 1 data

The next task for this discriminator trained on the noise data is to see if it can discriminate between Type 1 data and the other data, using what it has learned as a different data type from the noisy data. The discriminator achieved an 97.6% accuracy on the task of discriminating between Type 1 data and the data of the other sets. The confusion matrix is shown below.

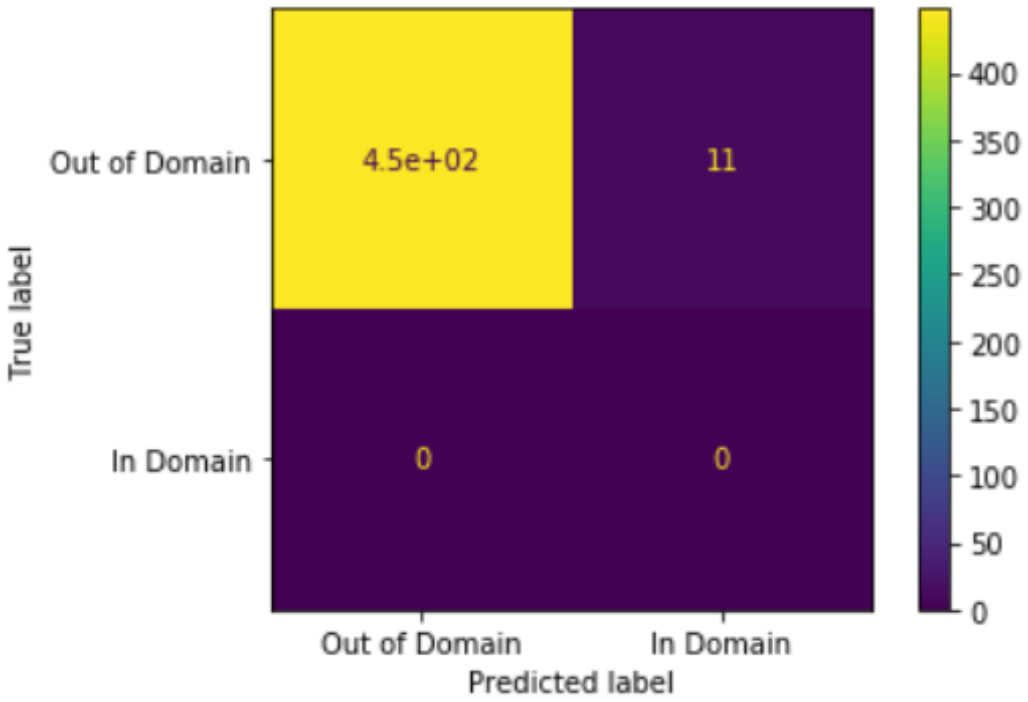


Figure 5.16: Confusion matrix: discrimination between type 1 and other types of data

Chapter 6

Discussion

This chapter discusses the results and the limitations present in this thesis. Section 6.1 discusses the cross-domain test results. Section 6.2 discusses the performance prediction results. Section 6.3 discusses the out of domain discriminator results and Section 6.4 discusses the generalizable out of domain discriminator results. Section 6.5 outlines the limitations of the research work presented in this thesis.

6.1 Cross-Domain Tests

The results of the cross domain tests gives an indication of the robustness of a developed HAR model when deployed in an environment other than the one it was developed in. When Model A was evaluated on a held out test set that came from Astroskin, the wearable devices that also recorded the training set, an accuracy of 99.07% was obtained. However when Model A was deployed on the same test set recorded by the BioHarness it achieved an accuracy of 65.74% (a 33.33% drop). This demonstrates that Model A is not robust to changes in wearable sensors that was

designed to measure the same physical phenomenon (accerolmetery), implying that if the wearable sensor is changed in the data acquisition process, a decrease of model performance and predictive accuracy may be observed.

Likewise when Model H was deployed on a test set recorded by the BioHarness an accuracy of 95.73% was observed. However when Model H was applied to a test set recorded by Astroskin the accuracy was reduced to 29.63% (a 66.10% drop).

This demonstrates that if the wearable device is changed in the data acquisition process, a decrease in may be observed. Model A Type 1, expanded on the initial investigation of Model A and Model H by separating the recordings with respect to the recording session as well as the wearable device that was involved in the recording. The accuracies of Model A Type 1 on data from different types is demonstrated in Table 6.1.

Table 6.1: Model A Type 1 accuracies

Data Type	Accuracy
Type 1	99.57%
Type 2	50.95%
Type 3	41.31%
Type 4	19.28%

The decrease in accuracy on Type 1 data to Type 2 data was 48.62%. Although these two types of data were recorded by the same wearable sensor, the data was from different sessions. This suggests that different sessions could also introduce a decrease in accuracy. However, it is interesting to note that in the example confusion matrix

presented walking was completely misclassified and the rest of the classes had less classification errors. This suggests that walking varies the most in between sessions as compared to the other activities. The decrease in accuracy on Type 3 data was 58.26%, which was a larger decrease than compared to Type 2 data. This suggests that a change in wearable device has more of an impact then a change in session. In Type 3's case almost all the classes were incorrectly classified. For Type 4 data a decrease of 80.29% from Type 1 accuracy for Model A occurred. The large decrease is expected because the both wearable device and the recording session are different from Type 1 data.

6.2 Performance Prediction

The curve that was fitted in order to estimate the accuracy of Model A on a test set that was recorded by the BioHarness was not able to accurately predict the accuracy. The Actual BioHarness accuracy on the test set was 65.74% and the predicted BioHarness accuracy was 81.58%. The curve over-estimated accuracy on the BioHarness test set.

6.3 Out of Domain Discriminator

The preliminary results from the cross-domain test suggest that the reason why the model's performance decreases when evaluated on data other than the data it was trained on (Type 1) is because the data must be in some way different from Type 1 data. The ability of a discriminator to differentiate between the data types with 100% accuracy in all cases, difference between Type 1 and Type 2 data, difference

between Type 1 and Type 3 data, difference between Type 1 and Type 4 data for the walking class supports this assumption.

6.4 Out of Domain Generalizable Discriminator

As the results in Section 6.1 and Section 6.3 suggest that the HAR model would perform poorly on data other than data involved in the training process, a method to account for this is presented in Section 5.4. Although we have determined that the model performed poorly on 3 types of data that differ from the Type 1 data, it is impractical to gather all the different types of data that may exist in the real world. This thesis evaluated two changes, a wearable device change and a session change. The changes that could occur in deployment include this and much more. Although we can not gather all the different types of data that the HAR model may face, we trained a discriminator that is able to differentiate between Type 1 data and Type 1 data with an SNR of 8 dB. This discriminator was successful in discriminating between the Type 1 and Type 1 noisy data. More importantly, it was also successful in determining that the other types of data were "out of domain" or were not alike the Type 1 data that was involved with the training of the model. The model would be useful in deployment because this model can determine if these data differs from the original data. If these data does differ from the model that it was trained with then the user should expect a lower accuracy then the accuracy reported when the model is trained and evaluated on the same data type.

6.5 Limitations

6.5.1 Data Collection

The first limitation encountered in the thesis was the data collection process. The data collection process was performed with only one individual, who was the author. This presents two limitations, the sample size as well as cognitive bias present in the collection of the data process. In terms of sample size, this limitation was introduced because of the COVID-19 pandemic. This resulted in university closures and the prohibition of non-essential human research. This work presented in this thesis was deemed as non-essential because it did not directly respond to the COVID-19 pandemic, did not involve clinical trials and did not have to be continued because of ethical reasons. Therefore, the data size of the experiments had to be reduced to one individual. As only the authour's data trained the HAR models developed in this thesis, these models would not be able to generalize for other subjects. However, it was not the goal of the thesis to create a HAR model that can be used for other subjects, but to create HAR models with the purpose of investigating the aspect of technical robustness and safety. In the experimental process the deep learning models were evaluated on a test set that was never seen before. Bias is also a limiting factor in the data collection process. The author was solely responsible for the generation, collecting as well as the analysis of the data.

6.5.2 Performance Prediction

In order to calculate the distance between windows in the Astroskin test set window and the BioHarness test set windows the Euclidean distances between individual

windows were calculated. However this calculation requires that the data is perfectly synchronized. There are multiple factor that can challenge the attempt for synchronization. Firstly, the synchronization was performed manually with visual inspection. Secondly after synchronization the data from the BioHarness was downsampled. In addition, requiring that accelerometer signals be synchronized in order calculate the Euclidean distances would not work from accelerometer signals from different datasets that can not be synchronized. Therefore the out-of-domain discriminator was implemented to address this shortcoming.

6.5.3 Out of Domain Discriminator and Out of Domain Generalizable Discriminator

The out of domain discriminator as well as the out of domain generalizable discriminator experiments were only conducted on the walking class and therefore conclusions drawn were done all on the walking class. Other classes should be explored in order to determine the applicability of the experiments to other classes of human activity recognition.

Chapter 7

Conclusions and Future Directions

This chapter summarizes the contributions of this thesis as well as directions for future research.

7.1 Summary of Contributions

The goal of this research was to investigate trust in AI-powered autonomous medical advisory systems. Using HAR as a vehicle the author demonstrated that human activity recognition deep learning models are not robust to changes in wearable devices or sessions. The author determined data recorded from a different session and a different wearable device can be differentiated by a discriminator for the walking class. Furthermore the author was able to build a generalizable discriminator that was able to differentiate when data was out of domain for the walking class. This thesis adds to the field of trust in AI in medical advisory systems as it evaluates the robustness of an HAR model in different environments and also investigates a method to inform the user if a model is deployed on data that is out of domain.

7.1.1 Cross Domain Test

The idea of the cross-domain test is not new, however this thesis applies the cross-domain test in a field in which it has gotten less attention, wearable device based AI. Image recognition, which initially inspired the cross-domain test, had seen it fair share of tests that analyze the robustness of models especially when it comes to robustness across hardware. The analysis presented in this thesis, especially the tests in regard to possible performance degradation across wearable devices will alert the AI community to ensure that models developed using wearable devices are robust. This is especially important since new wearable devices is a ever-growing market.

7.1.2 Out of Domain Generalizable Discriminator

In this thesis we harness the power of a generalizeable discriminator to be able to detect when data differs from the training data of the artificial intelligence model. In this regard the generalizable discriminator can alert a user when the model is about to be presented with data that differs from the data used in the development stage, signifying to users the decision made by the model cannot be trusted. This discriminator is generalizable in this experiment meaning that it uses noisy data to learn what an "out of domain" data type is. It then can predict other out of domain data types.

7.2 Future Directions

Firstly, the human activity recognition model was developed using a very small amount of data collected by only one participant. This choice was because of the

University Restrictions concerning research conducted in COVID-19. In future work, these experiments should be tested with larger amount of data tests. This would increase the robustness of the results.

In addition, this thesis concludes that there is a performance degradation when the model is exposed to data that differs from the training data and provides a method to be able to detect when different data is presented. However, the analysis is only conducted using one class, walking. A more thorough investigation of the other HAR classes would provide a more complete view. In addition it would be ideal if the amount of degradation could be estimated.

Appendix A

Appendix

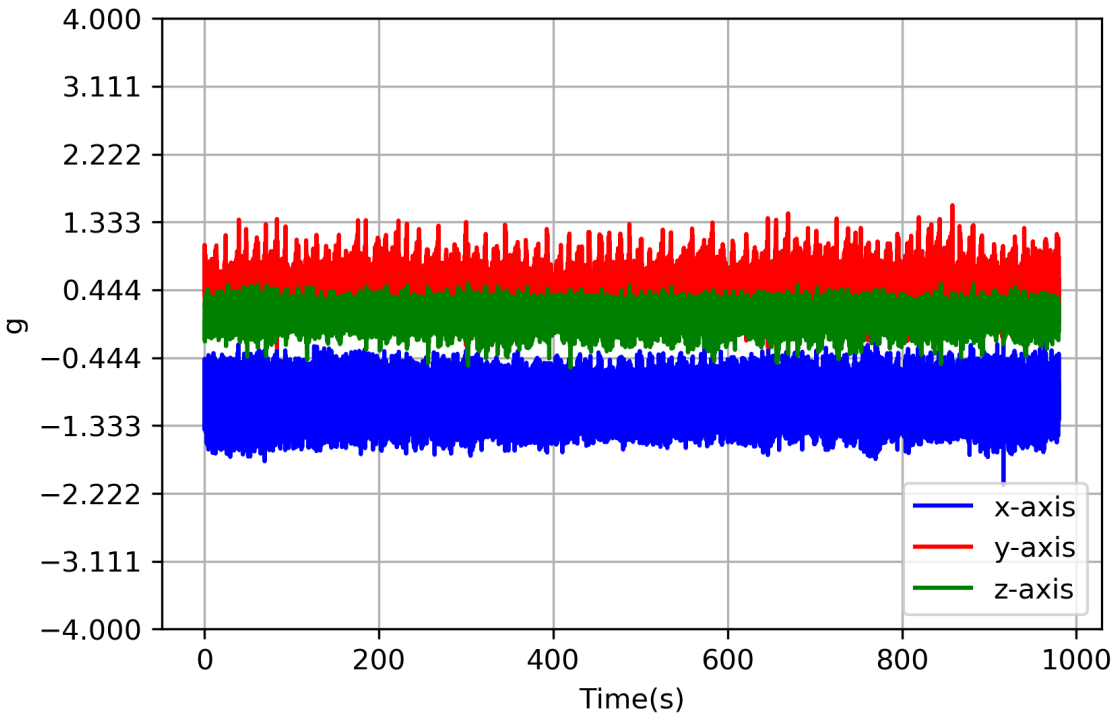


Figure A.1: Astroskin session 1 walking

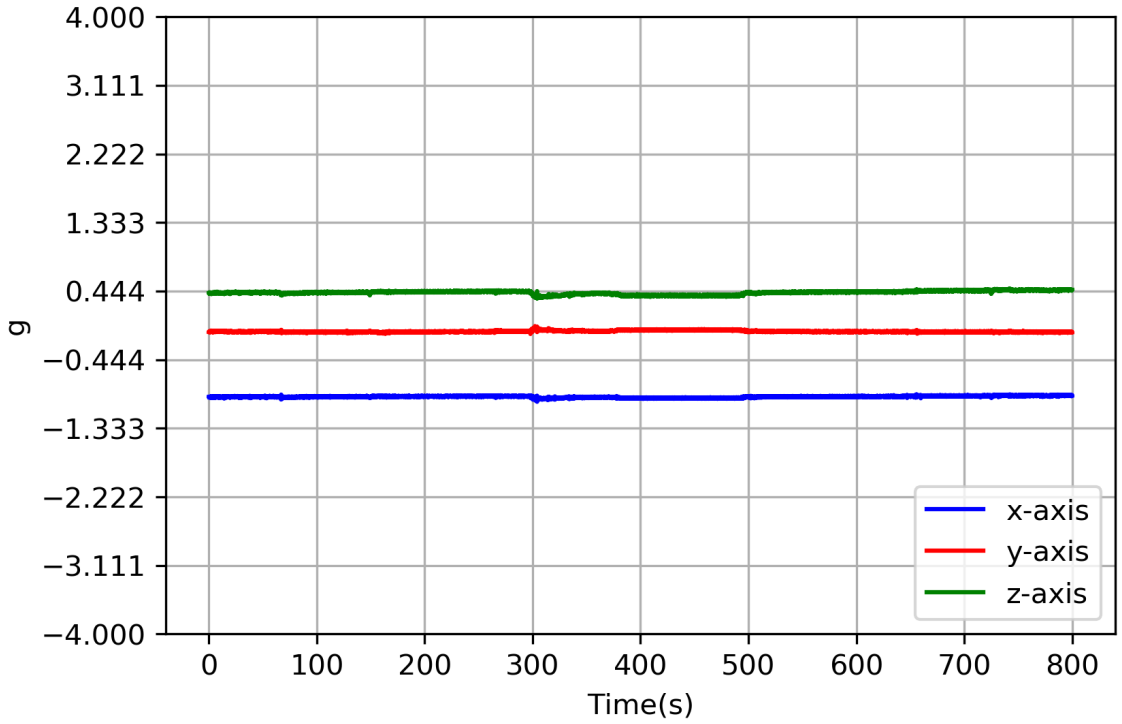


Figure A.2: Astroskin session 1 sitting

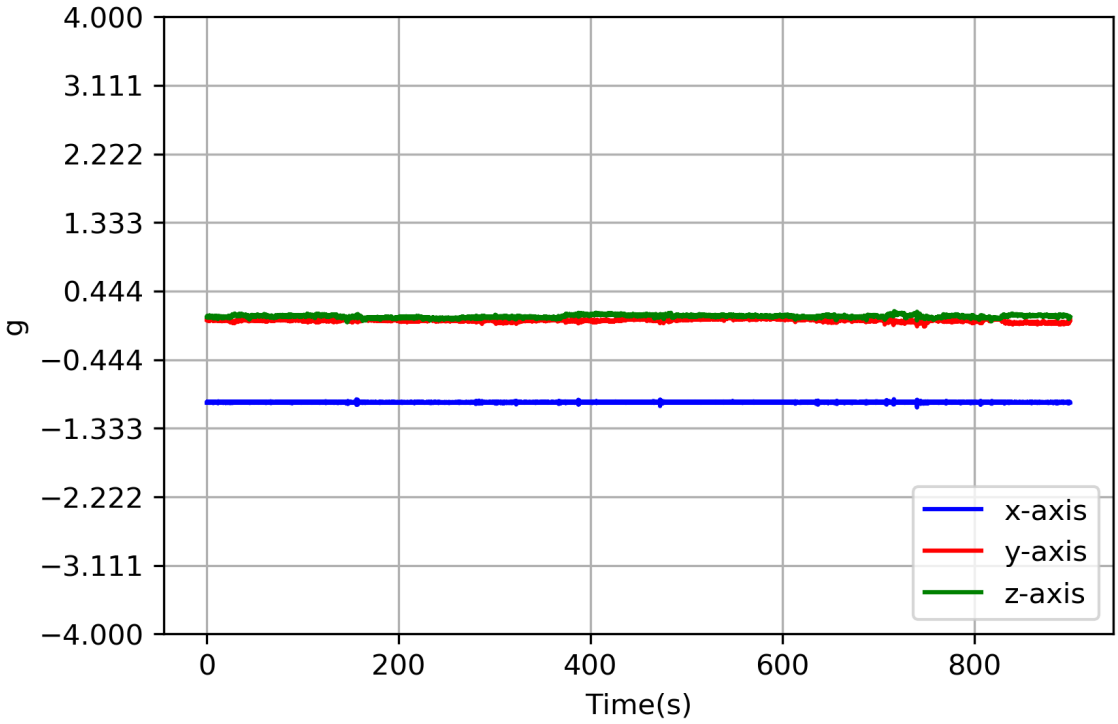


Figure A.3: Astroskin session 1 standing

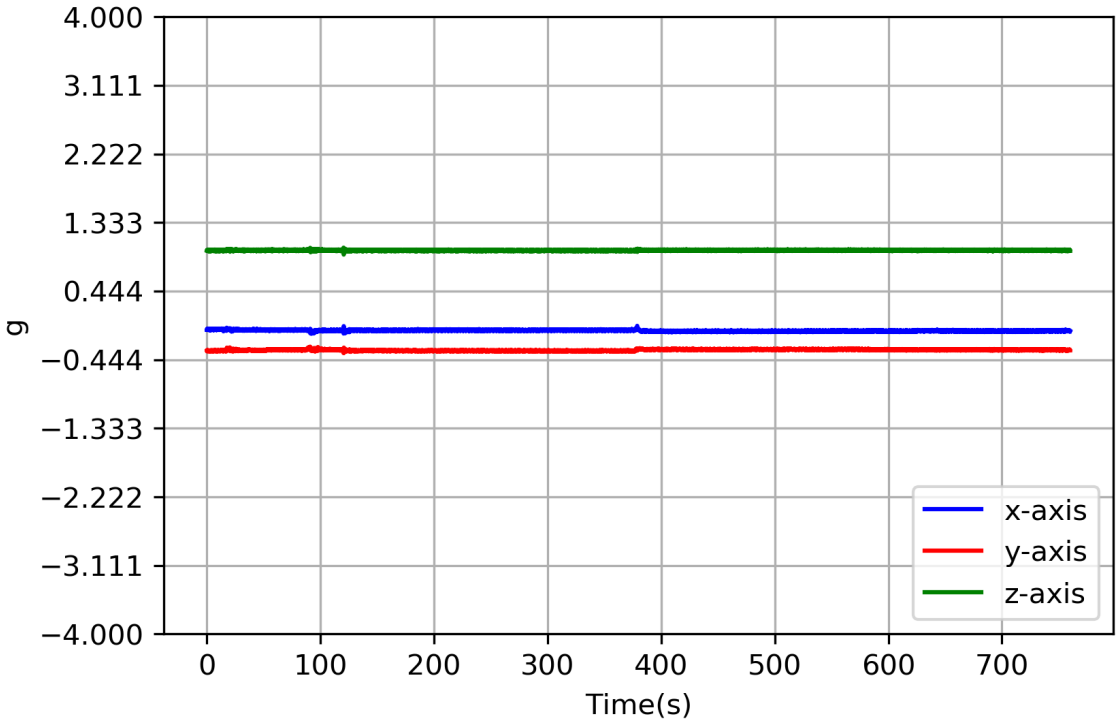


Figure A.4: Astroskin session 1 laying

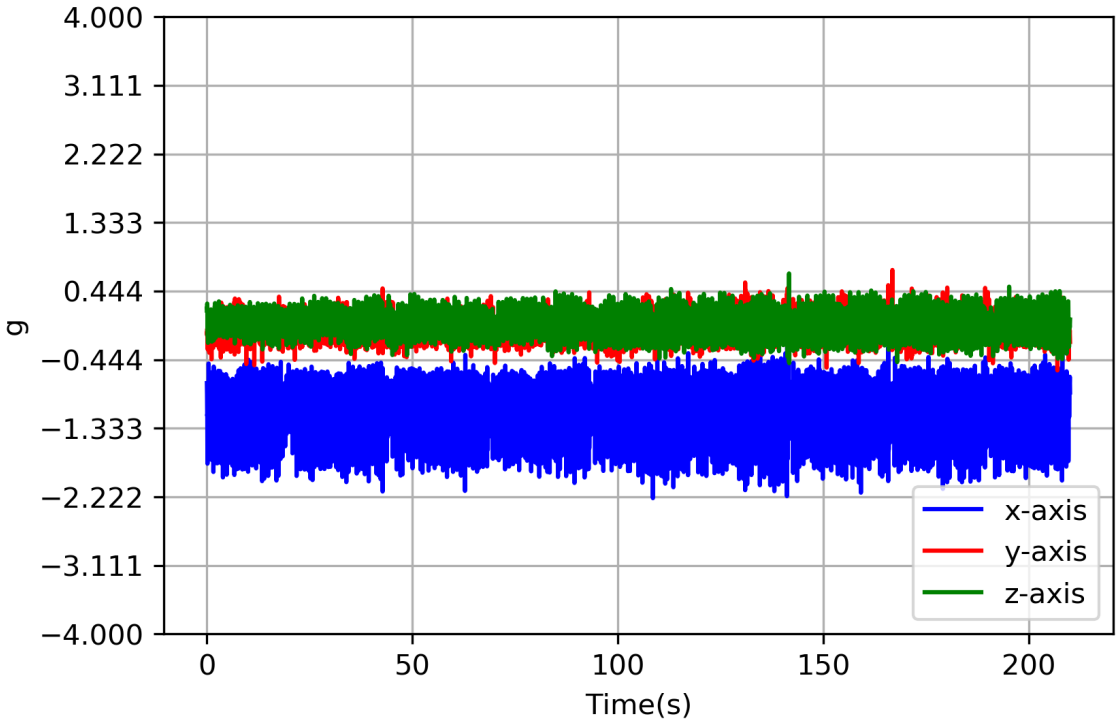


Figure A.5: Astroskin session 2 walking

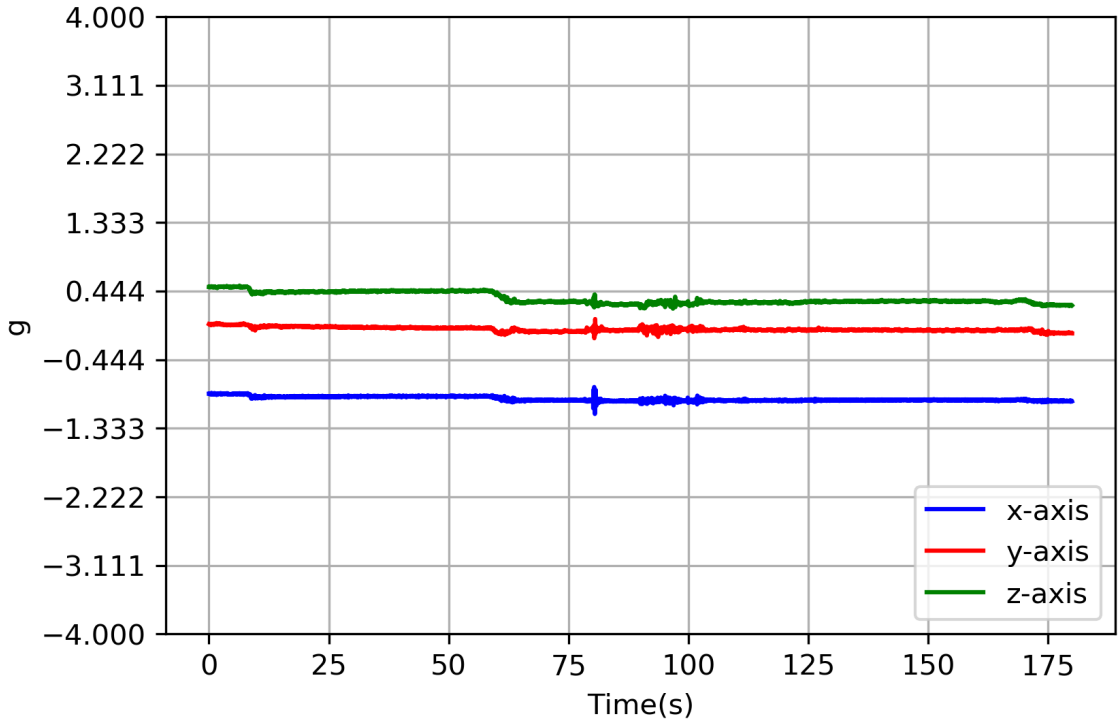


Figure A.6: Astroskin session 2 sitting

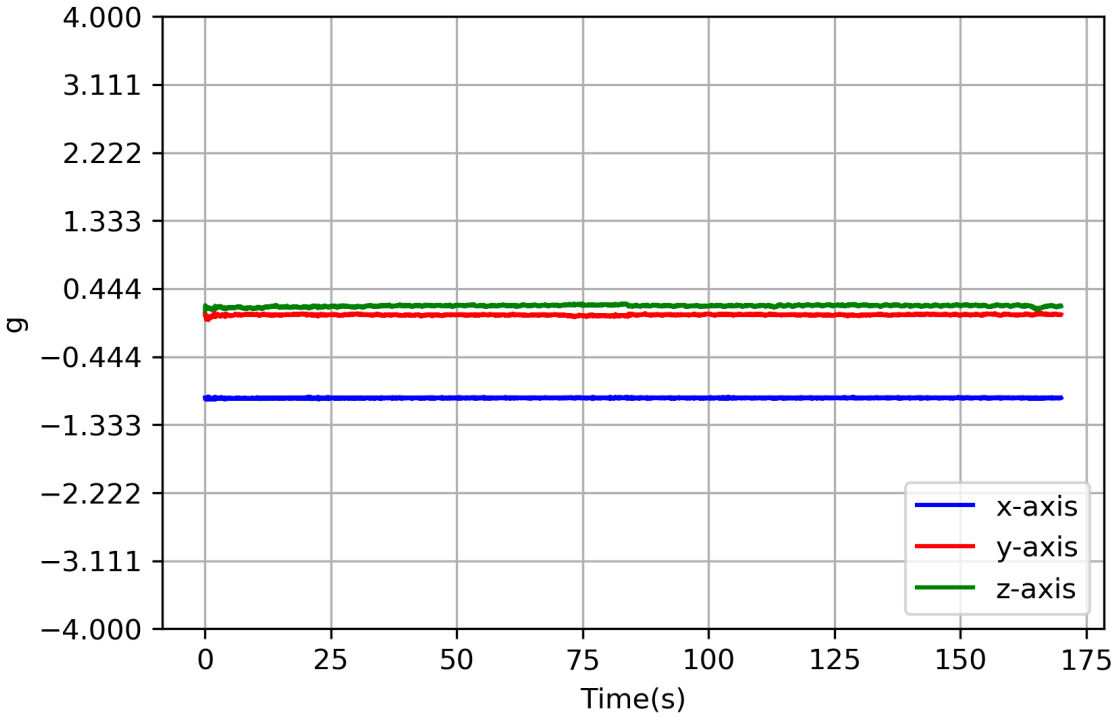


Figure A.7: Astroskin session 2 standing

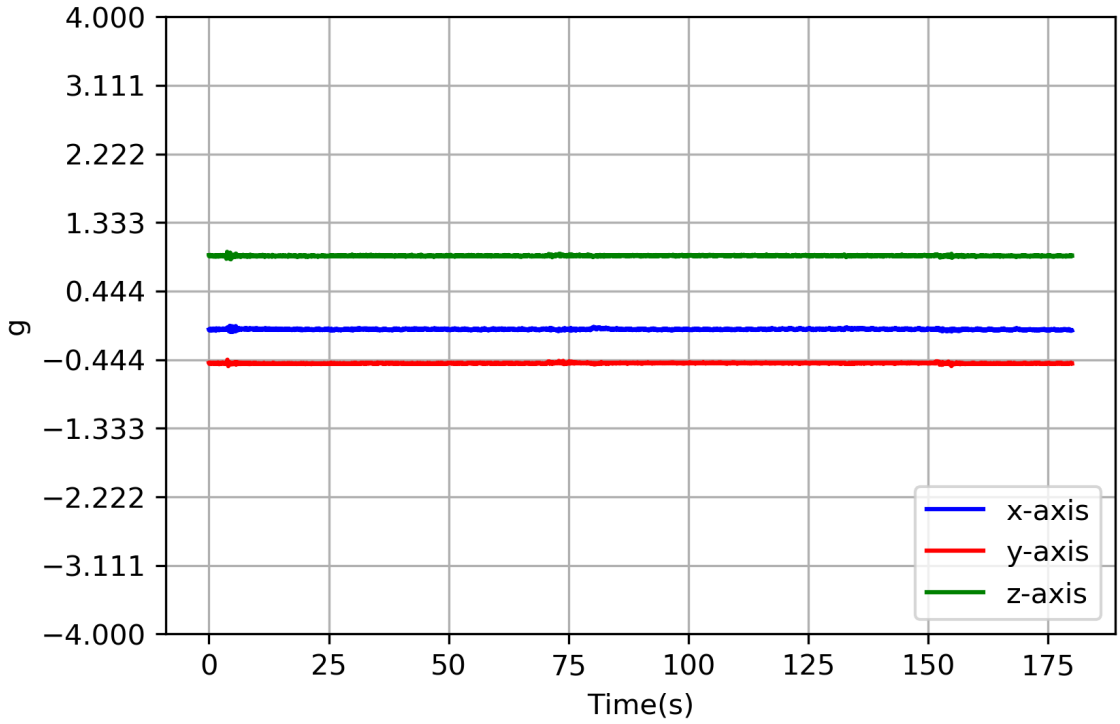


Figure A.8: Astroskin session 2 laying

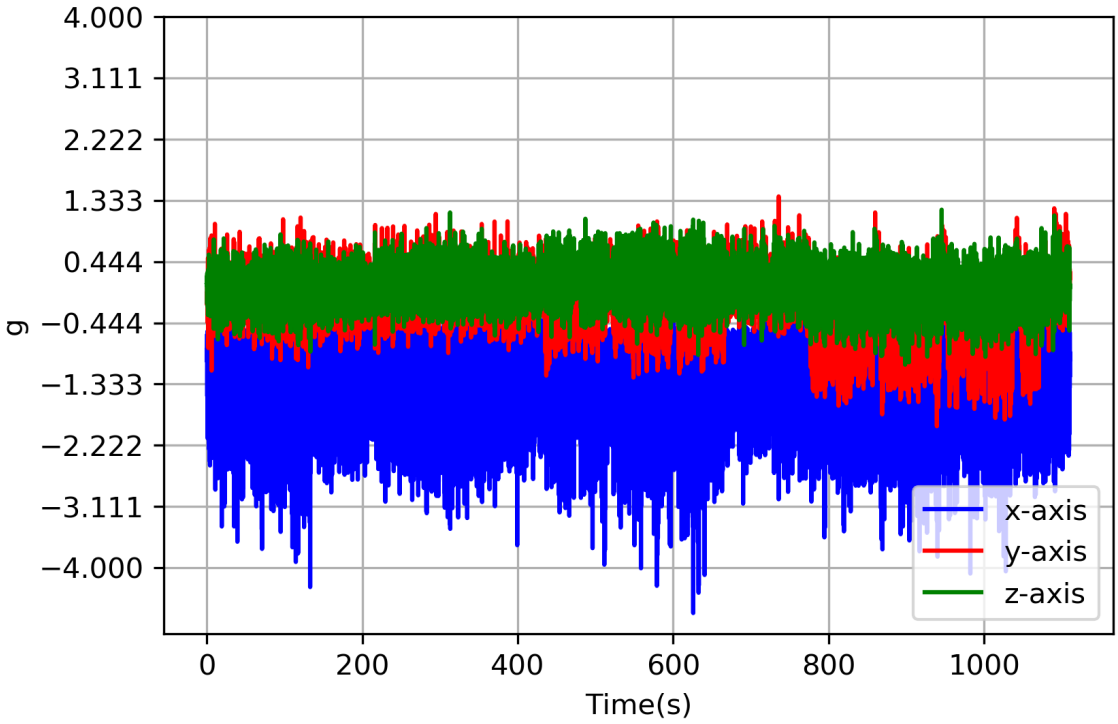


Figure A.9: Astroskin downstairs

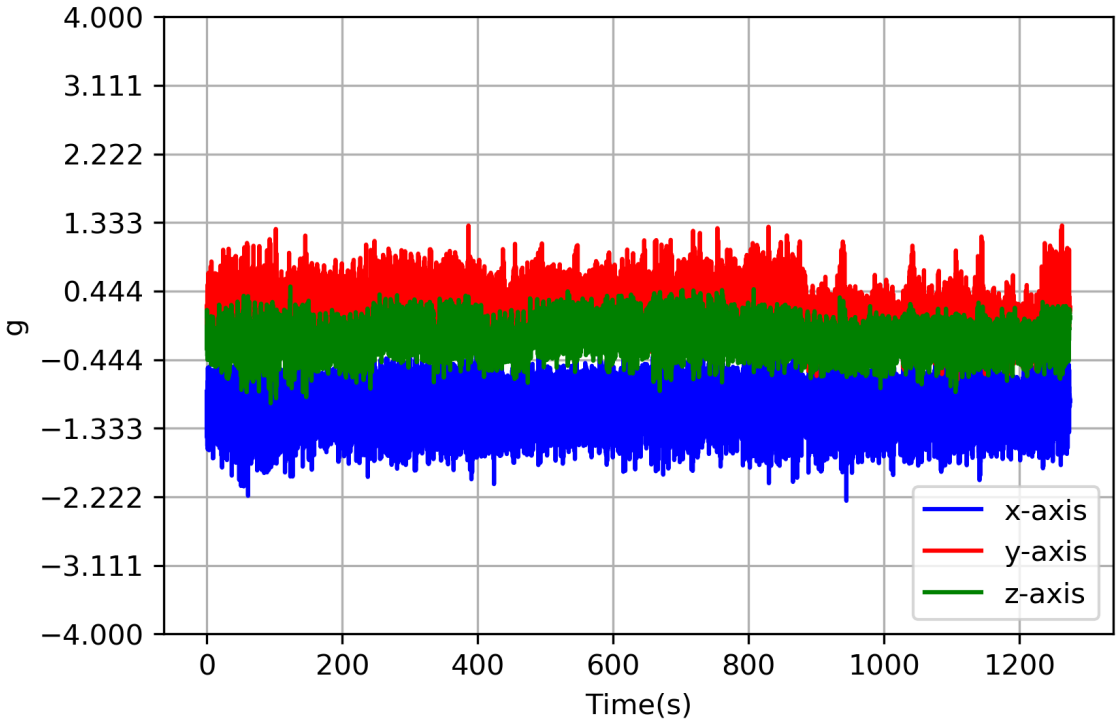


Figure A.10: Astroskin upstairs

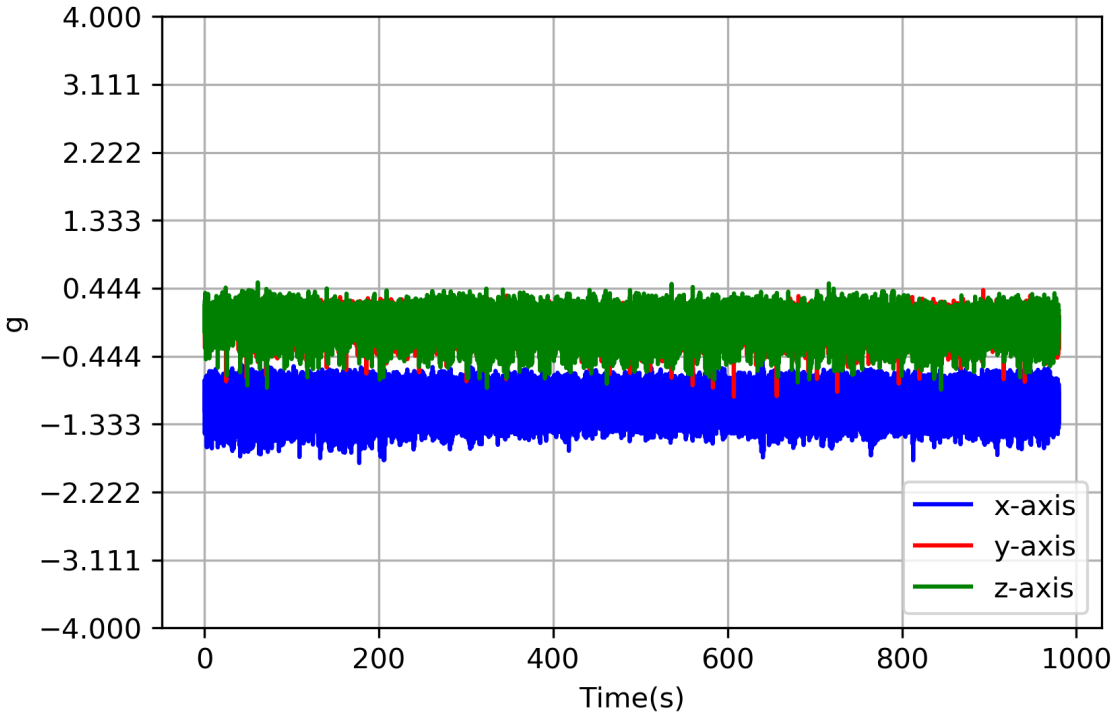


Figure A.11: BioHarness session 1 walking

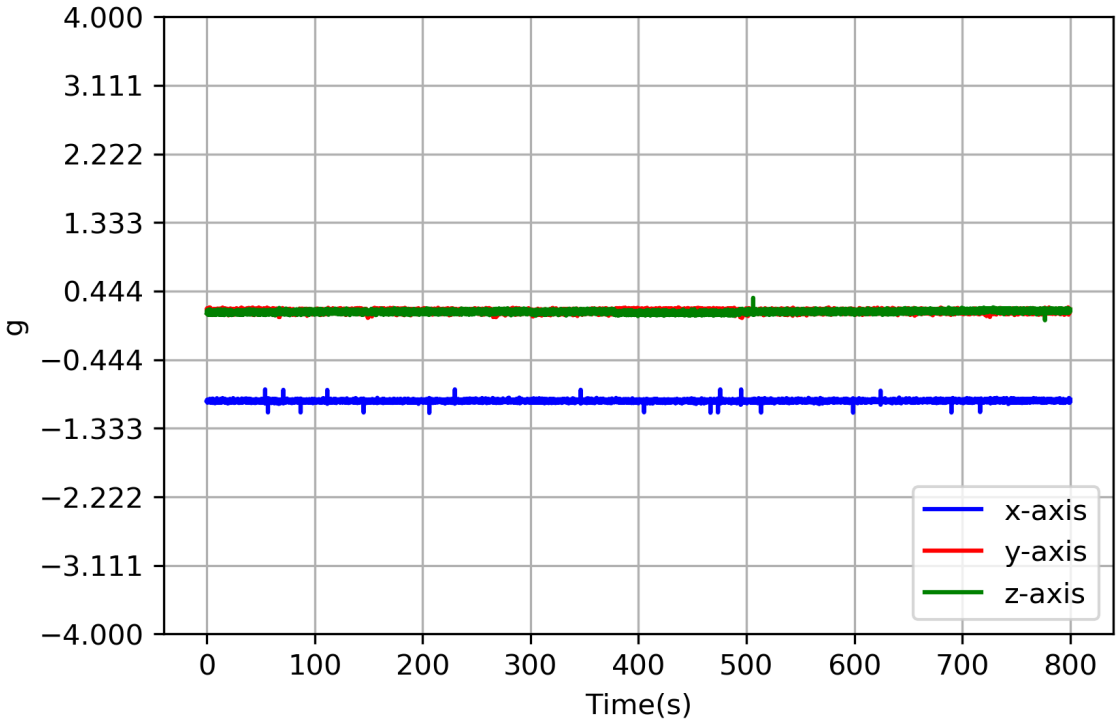


Figure A.12: BioHarness session 1 sitting

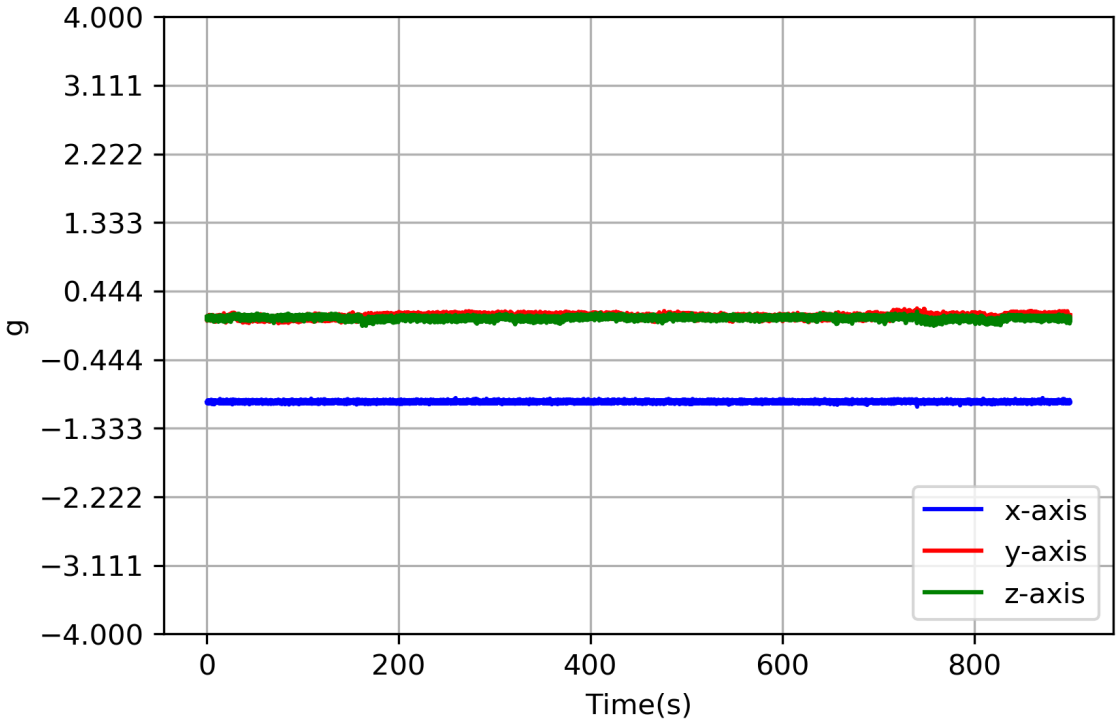


Figure A.13: BioHarness session 1 standing

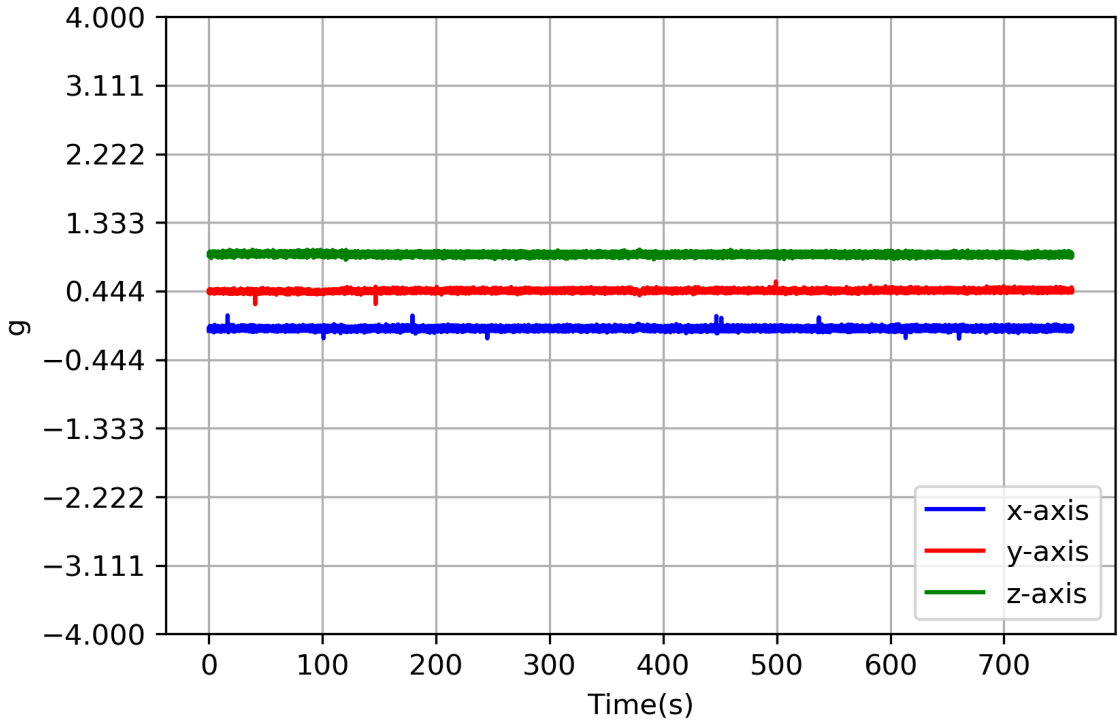


Figure A.14: BioHarness session 1 laying

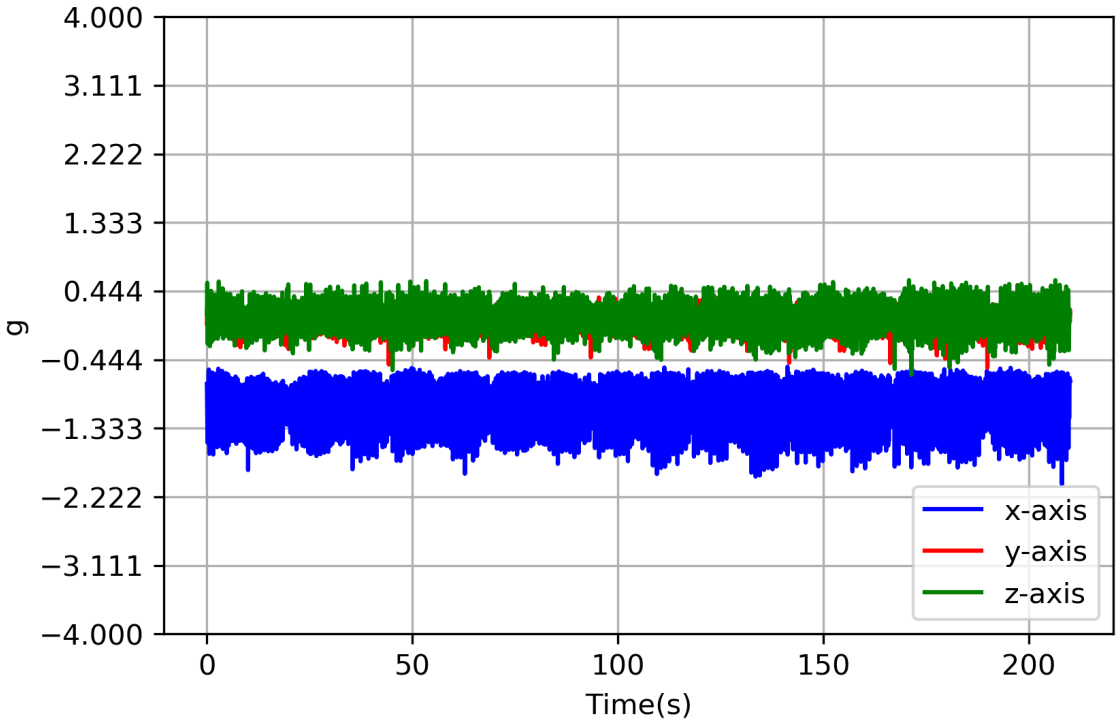


Figure A.15: BioHarness session 2 walking

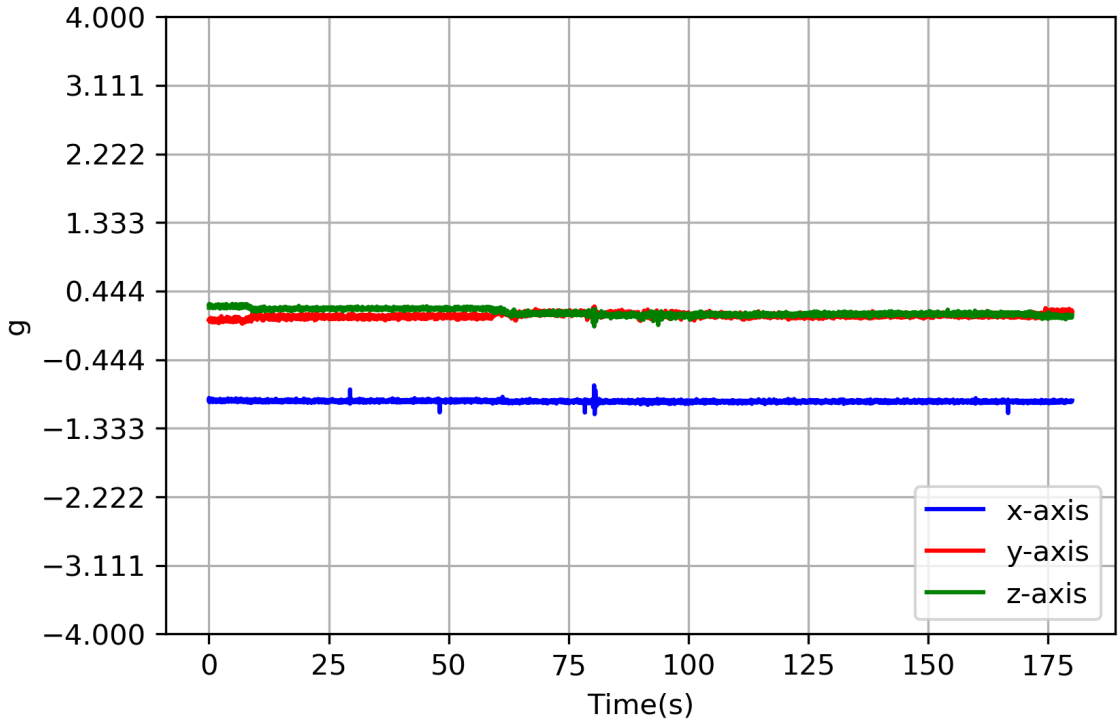


Figure A.16: BioHarness session 2 sitting

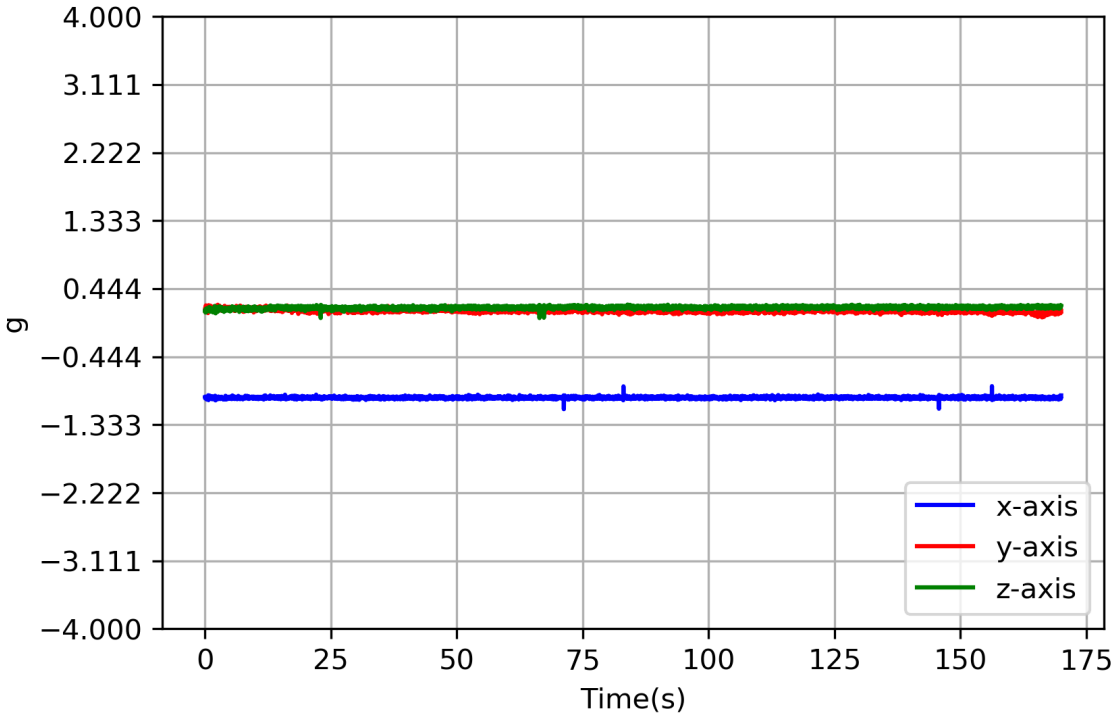


Figure A.17: BioHarness session 2 standing

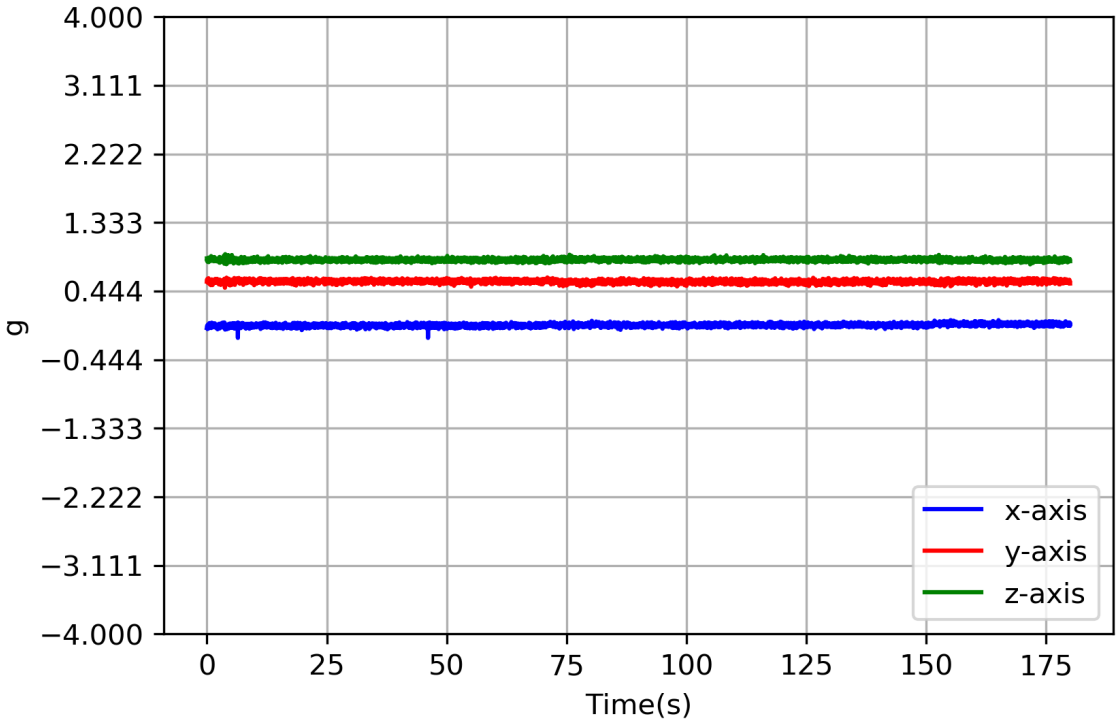


Figure A.18: BioHarness session 2 laying

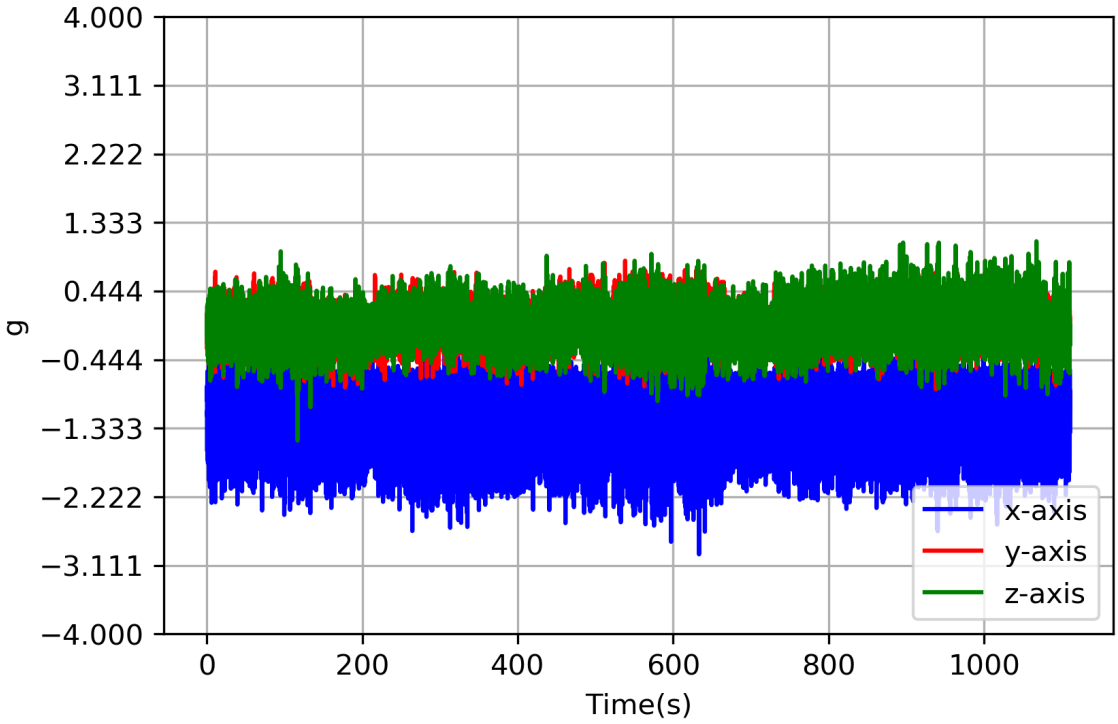


Figure A.19: BioHarness downstairs

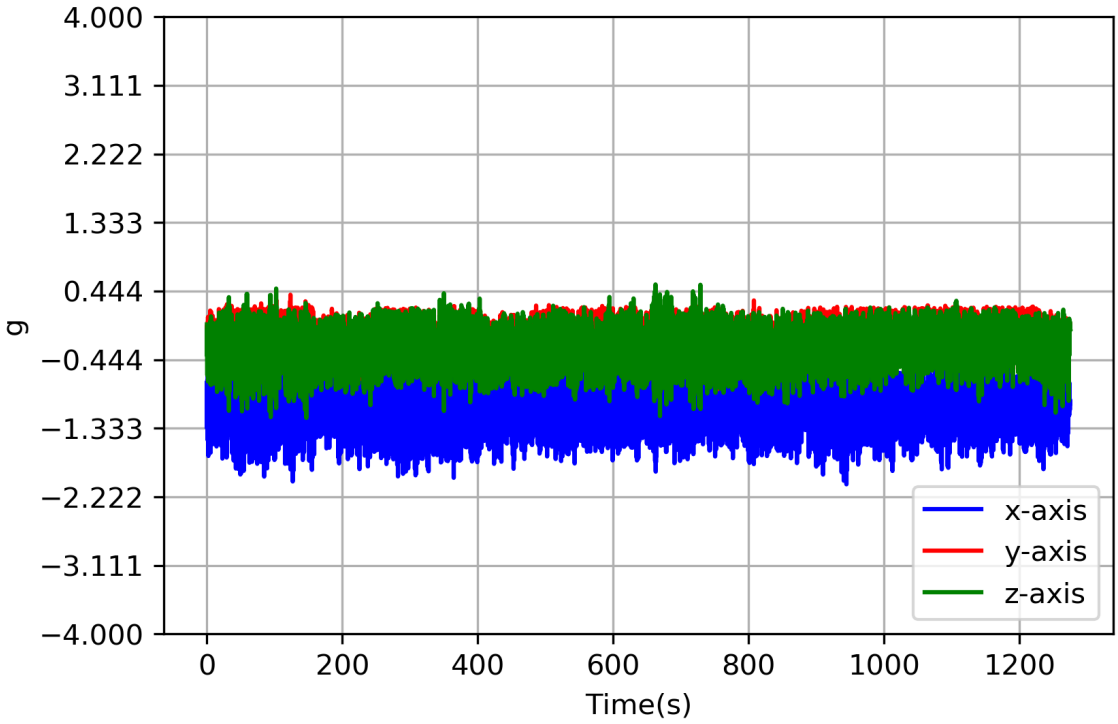
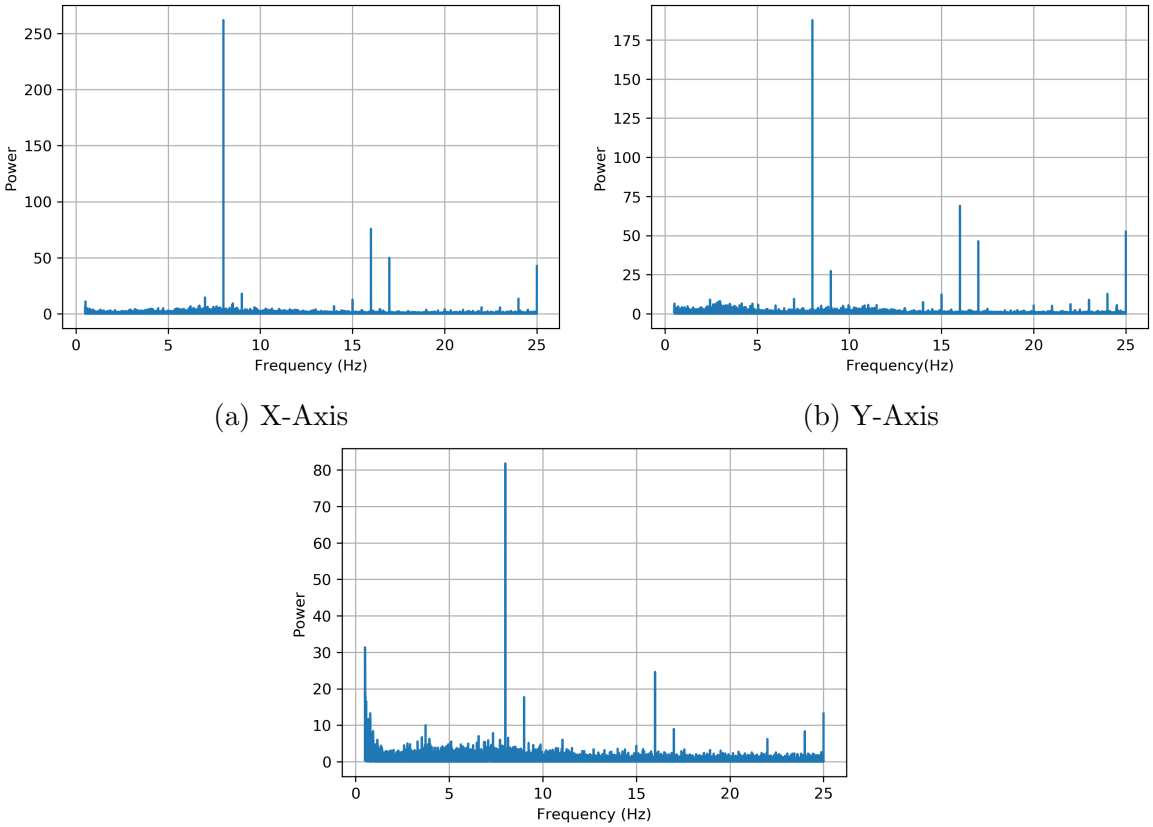


Figure A.20: BioHsrness upstairs

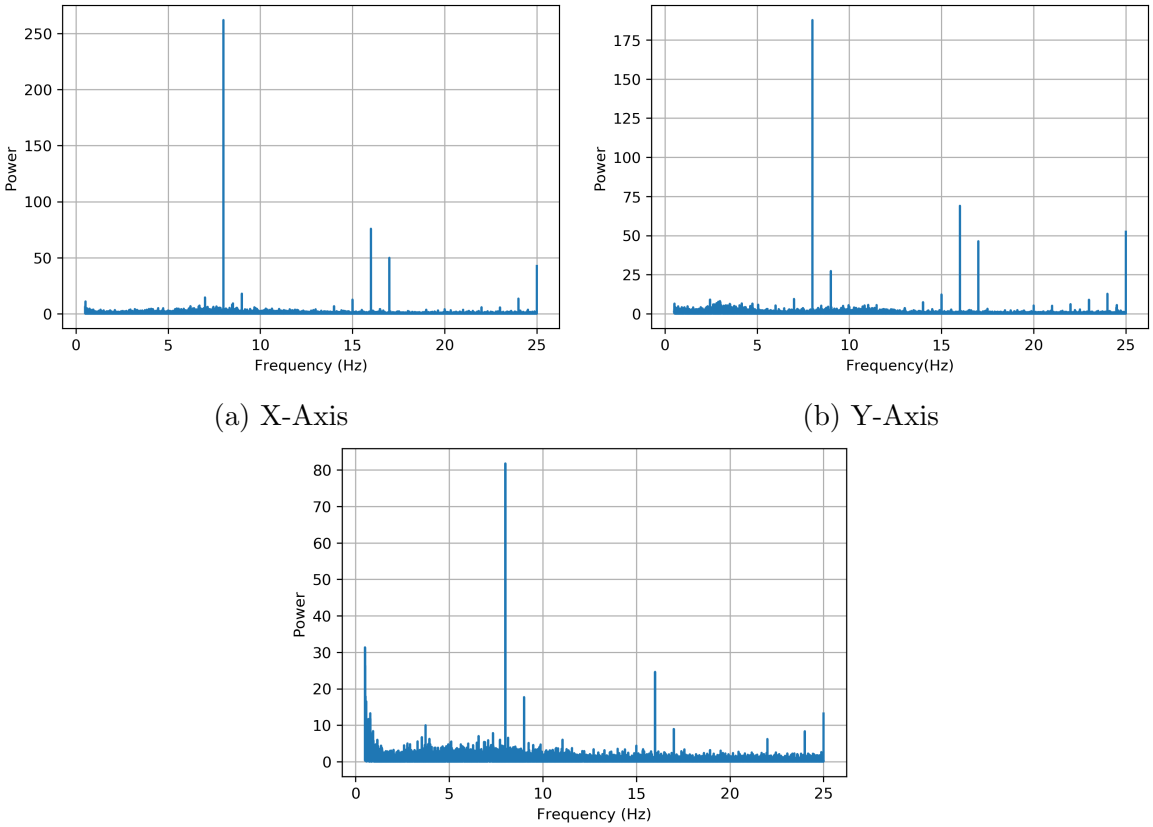


(a) X-Axis

(b) Y-Axis

(c) Z-Axis

Figure A.21: Power spectrum sitting session 1 recorded by Astroskin

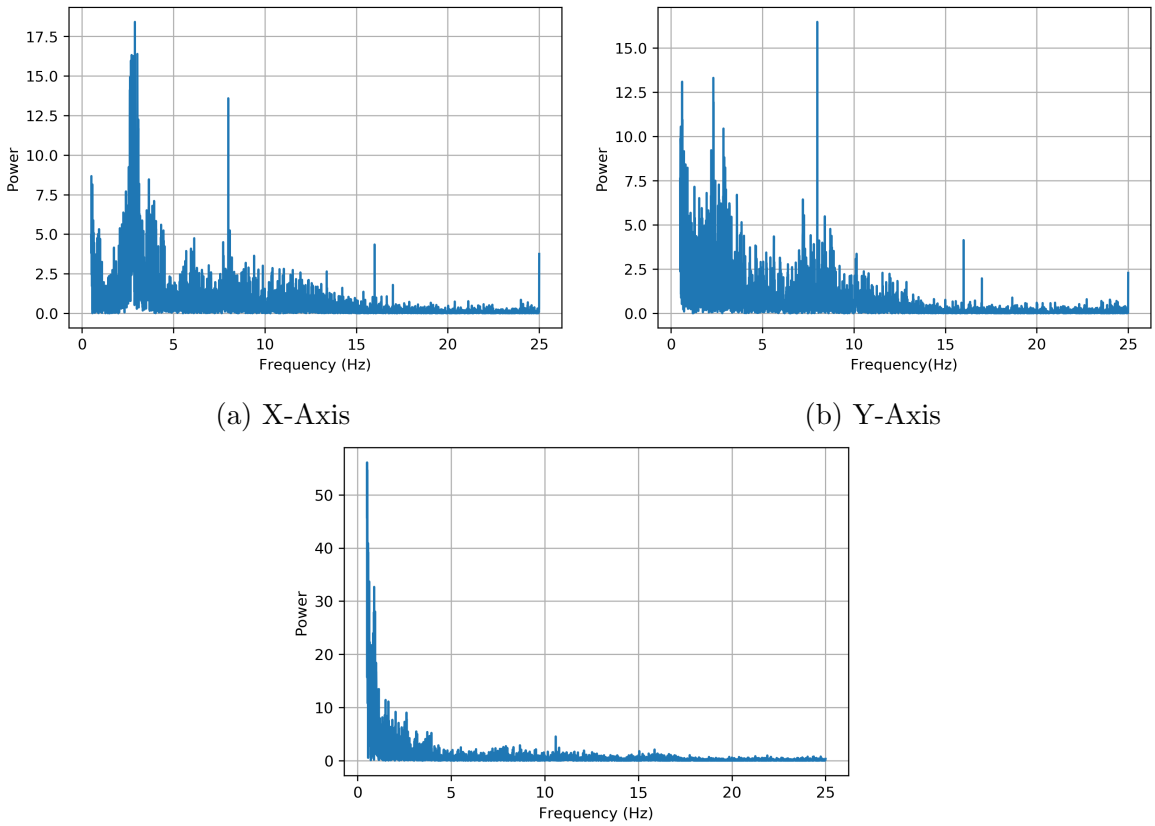


(a) X-Axis

(b) Y-Axis

(c) Z-Axis

Figure A.22: Power spectrum sitting session 1 recorded by Astroskin

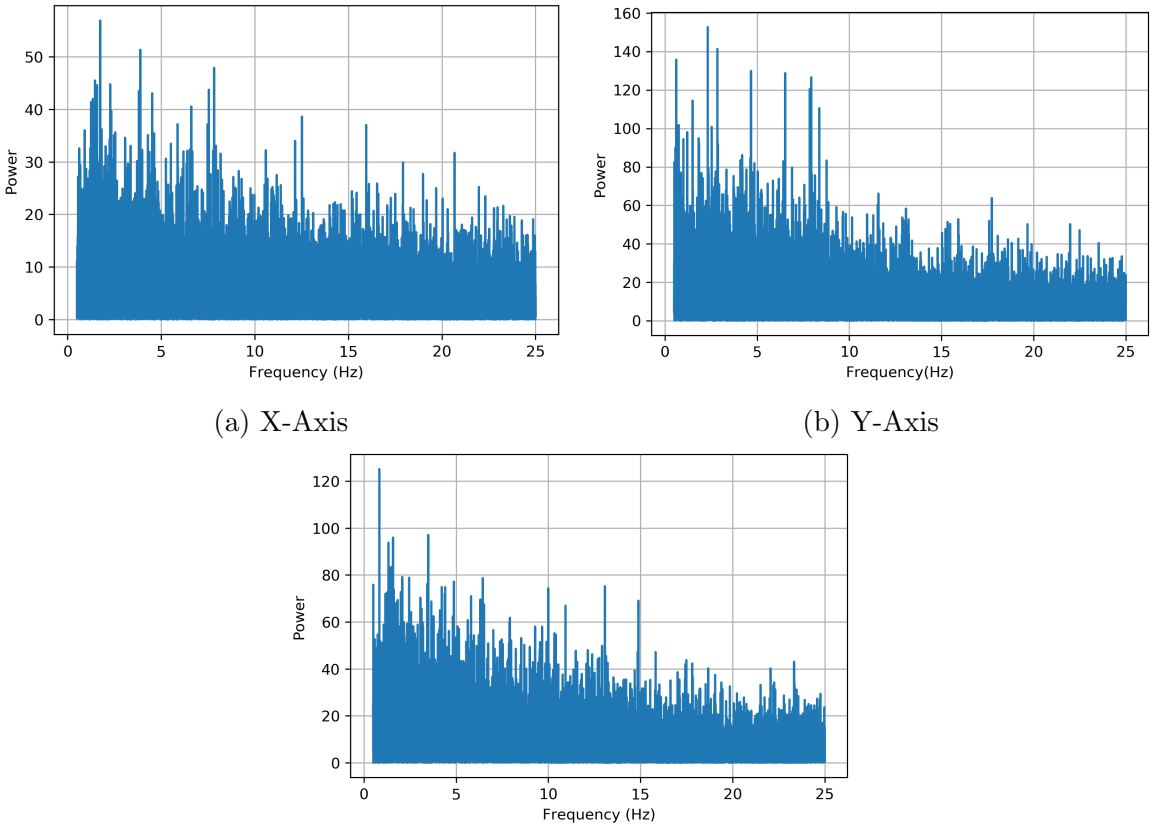


(a) X-Axis

(b) Y-Axis

(c) Z-Axis

Figure A.23: Power spectrum sitting session 2 recorded by Astroskin

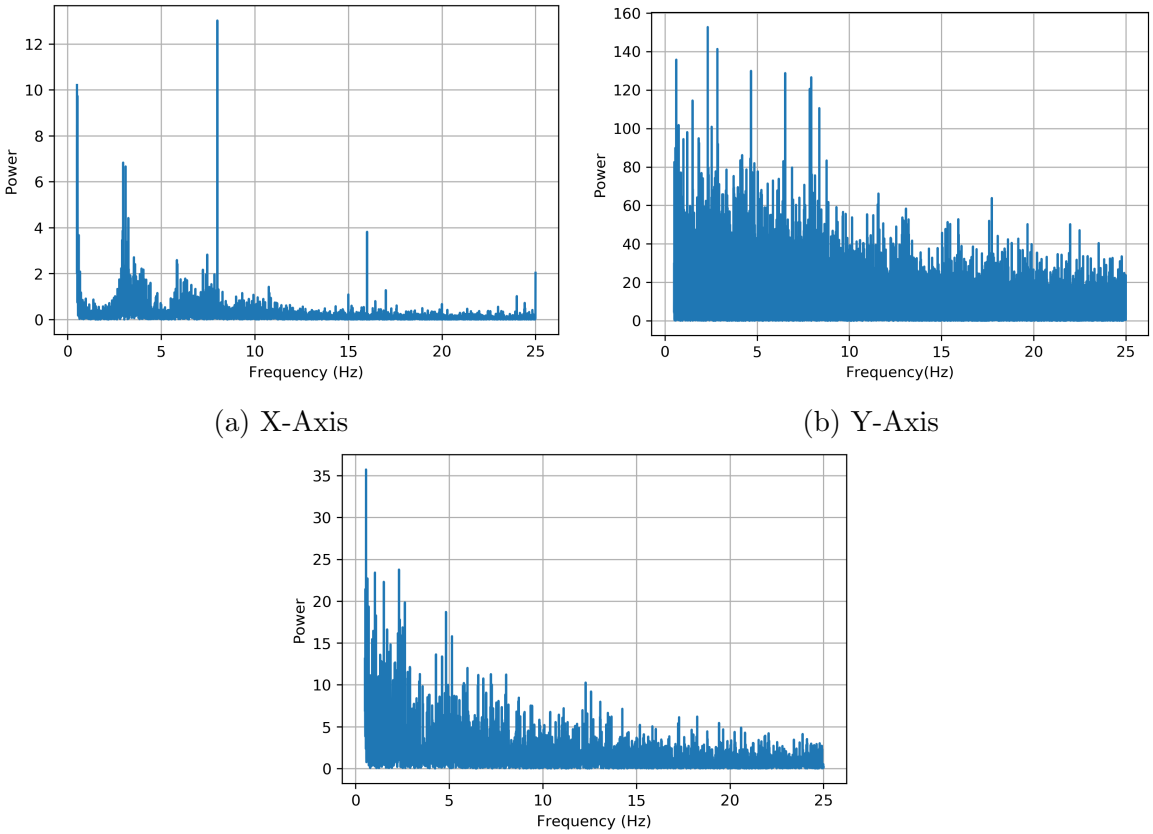


(a) X-Axis

(b) Y-Axis

(c) Z-Axis

Figure A.24: Power spectrum sitting session 1 recorded by BioHarness

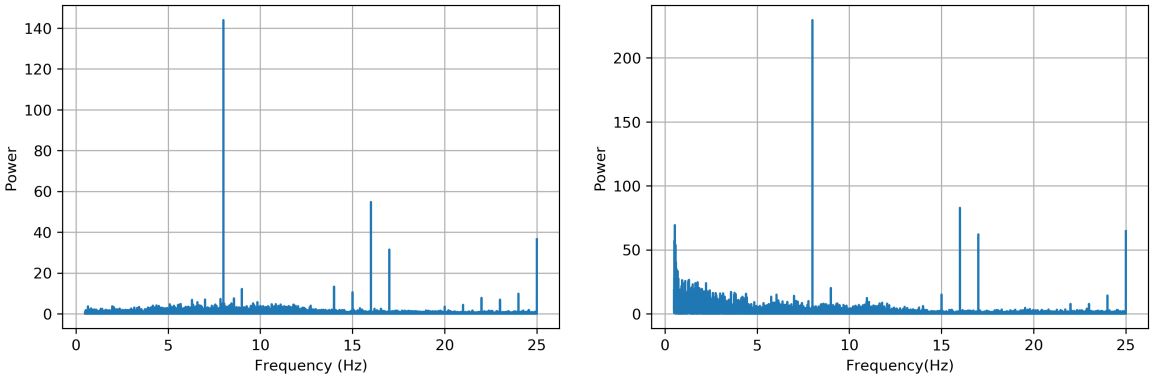


(a) X-Axis

(b) Y-Axis

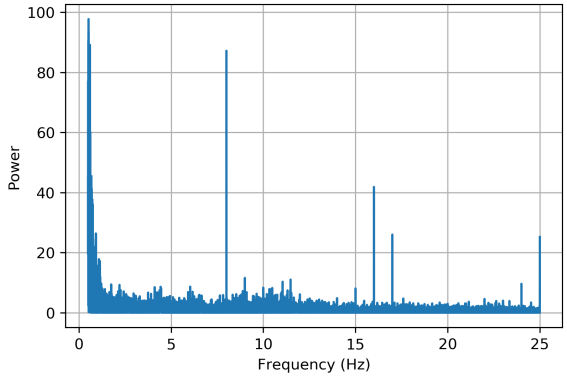
(c) Z-Axis

Figure A.25: Power spectrum sitting session 2 recorded by BioHarness



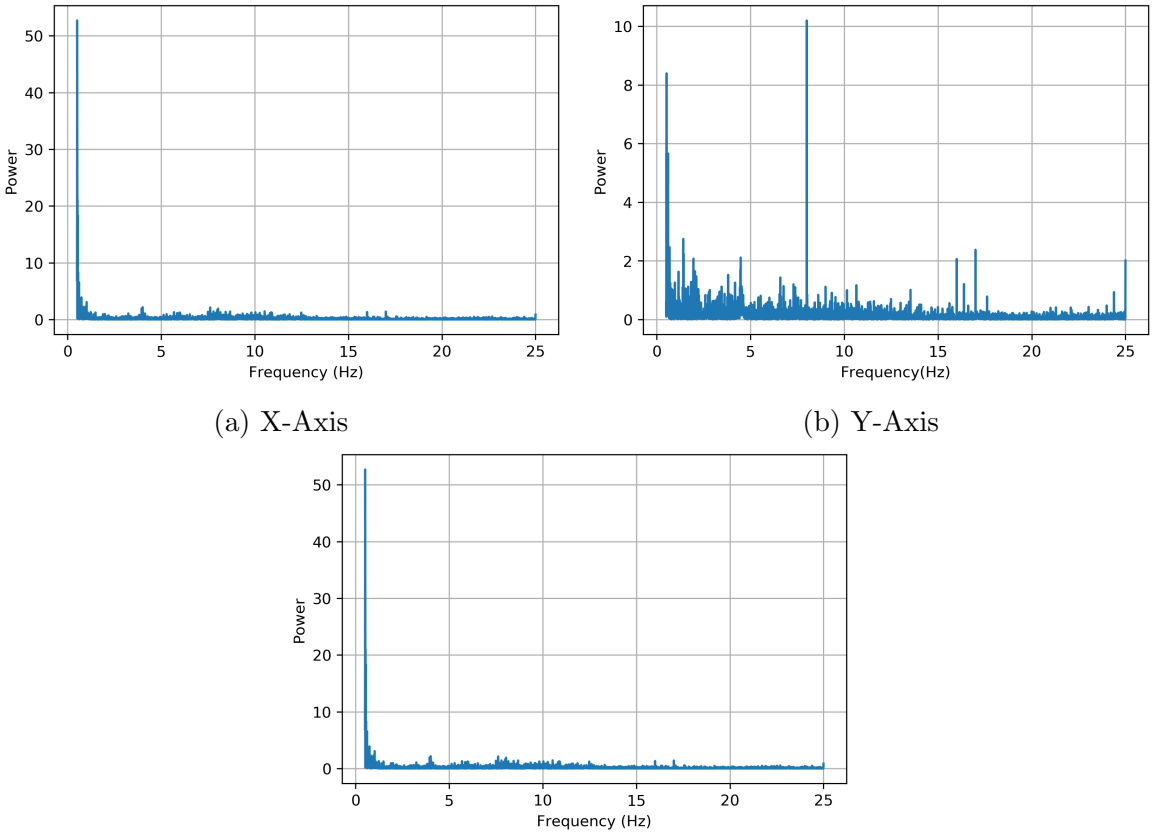
(a) X-Axis

(b) Y-Axis



(c) Z-Axis

Figure A.26: Power spectrum standing session 1 recorded by Astroskin

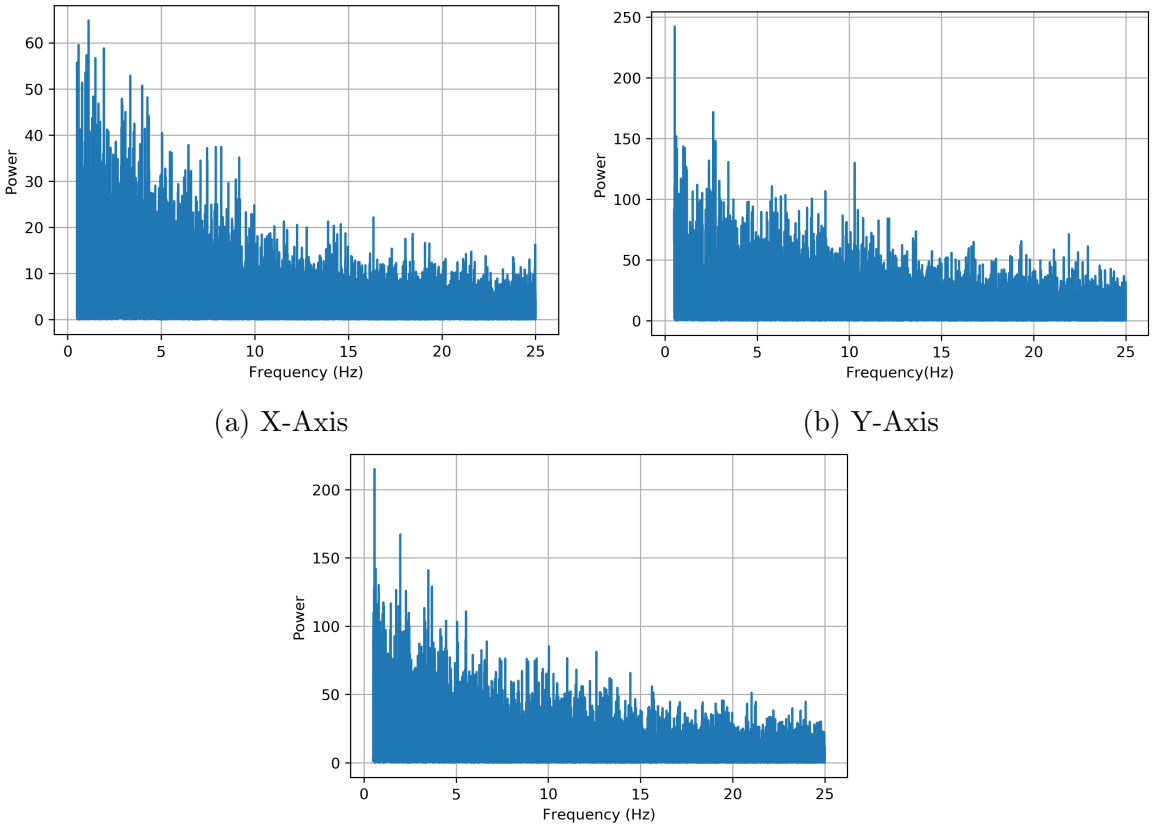


(a) X-Axis

(b) Y-Axis

(c) Z-Axis

Figure A.27: Power spectrum standing session 2 recorded by Astroskin

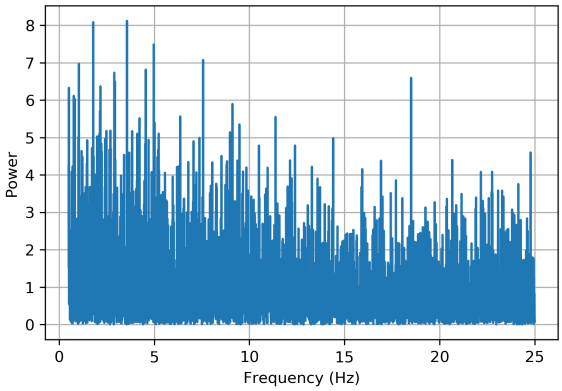


(a) X-Axis

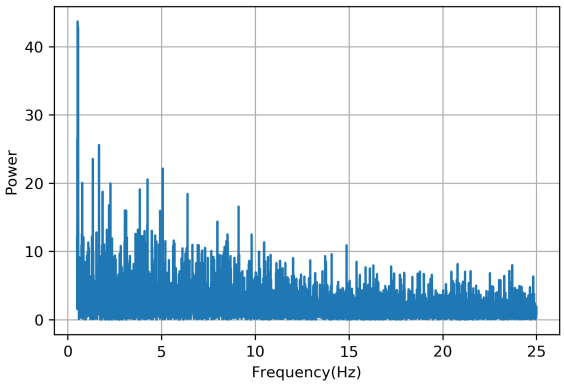
(b) Y-Axis

(c) Z-Axis

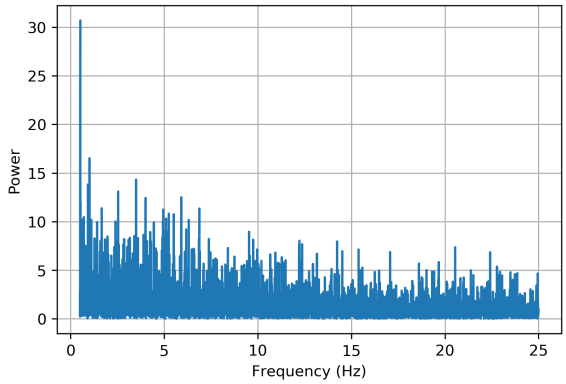
Figure A.28: Power spectrum standing session 1 recorded by BioHarness



(a) X-Axis

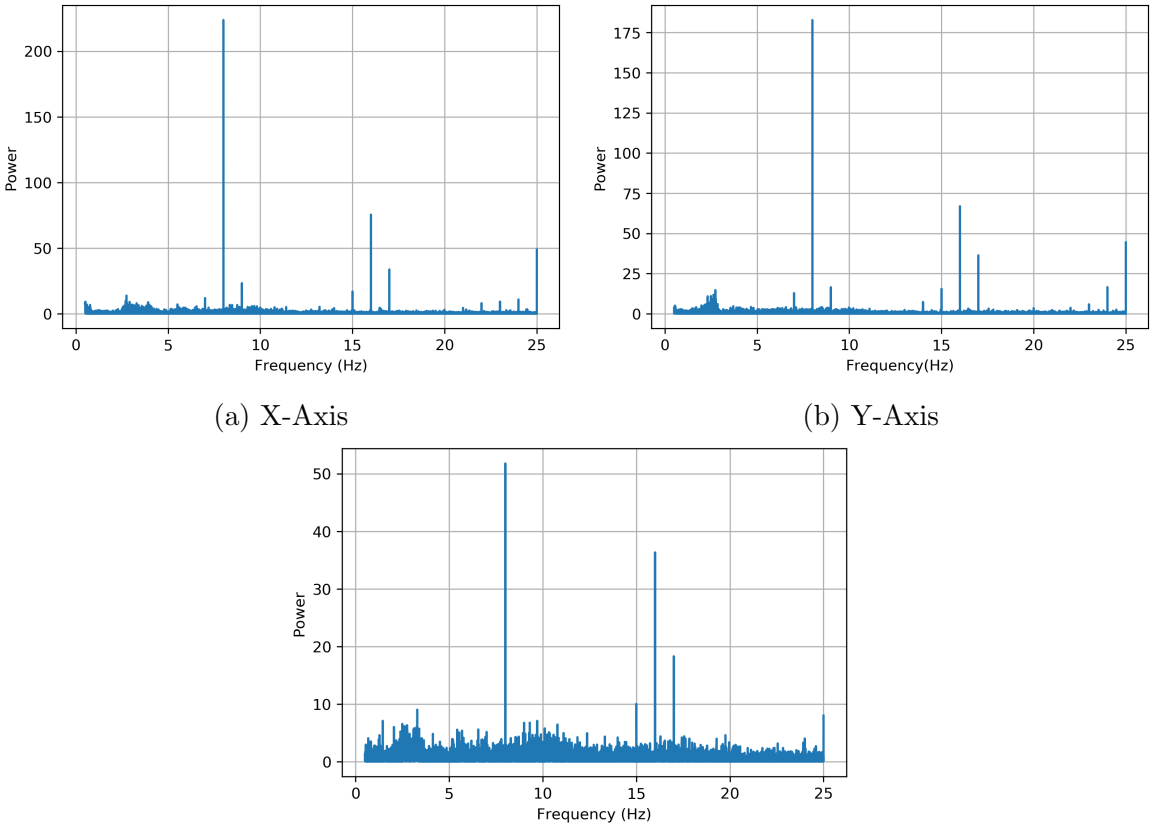


(b) Y-Axis



(c) Z-Axis

Figure A.29: Power spectrum standing session 2 recorded by bioHarness

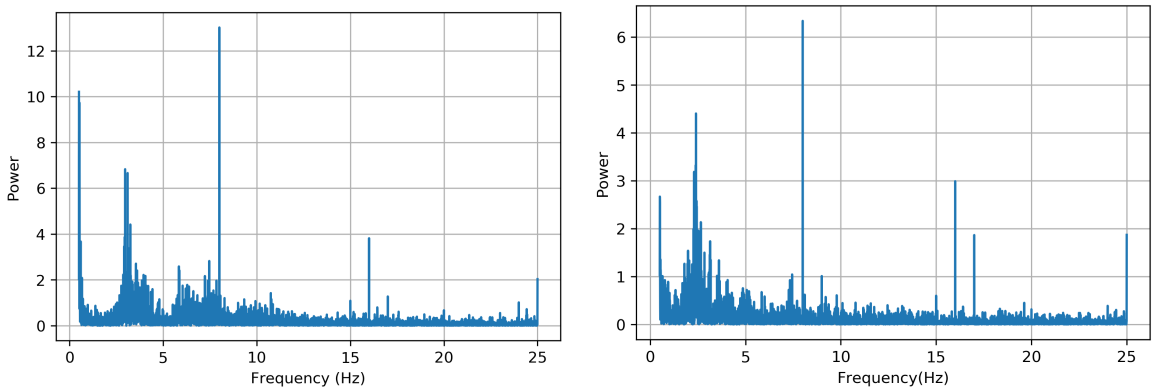


(a) X-Axis

(b) Y-Axis

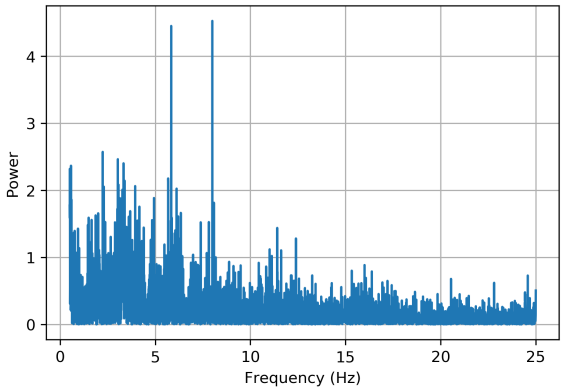
(c) Z-Axis

Figure A.30: Power spectrum laying session 1 recorded by Astroskin



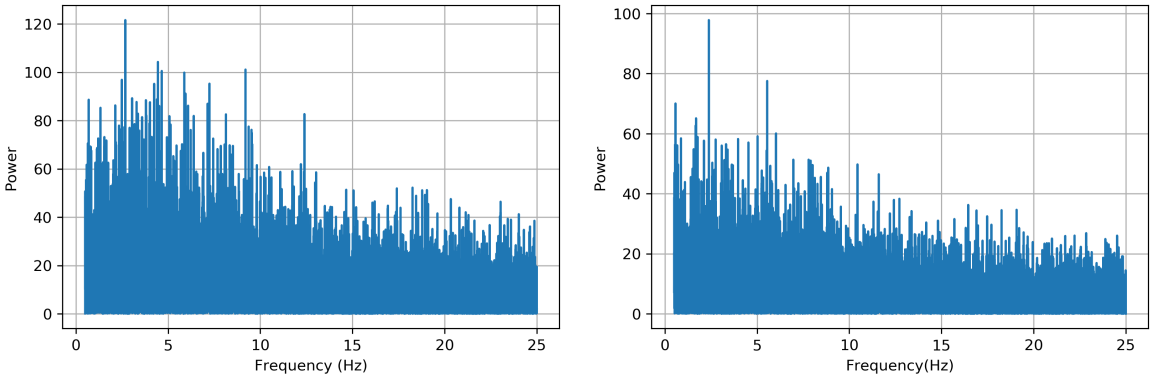
(a) X-Axis

(b) Y-Axis



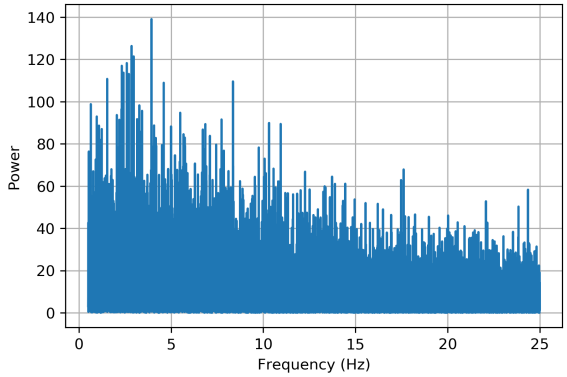
(c) Z-Axis

Figure A.31: Power spectrum laying session 2 recorded by Astroskin



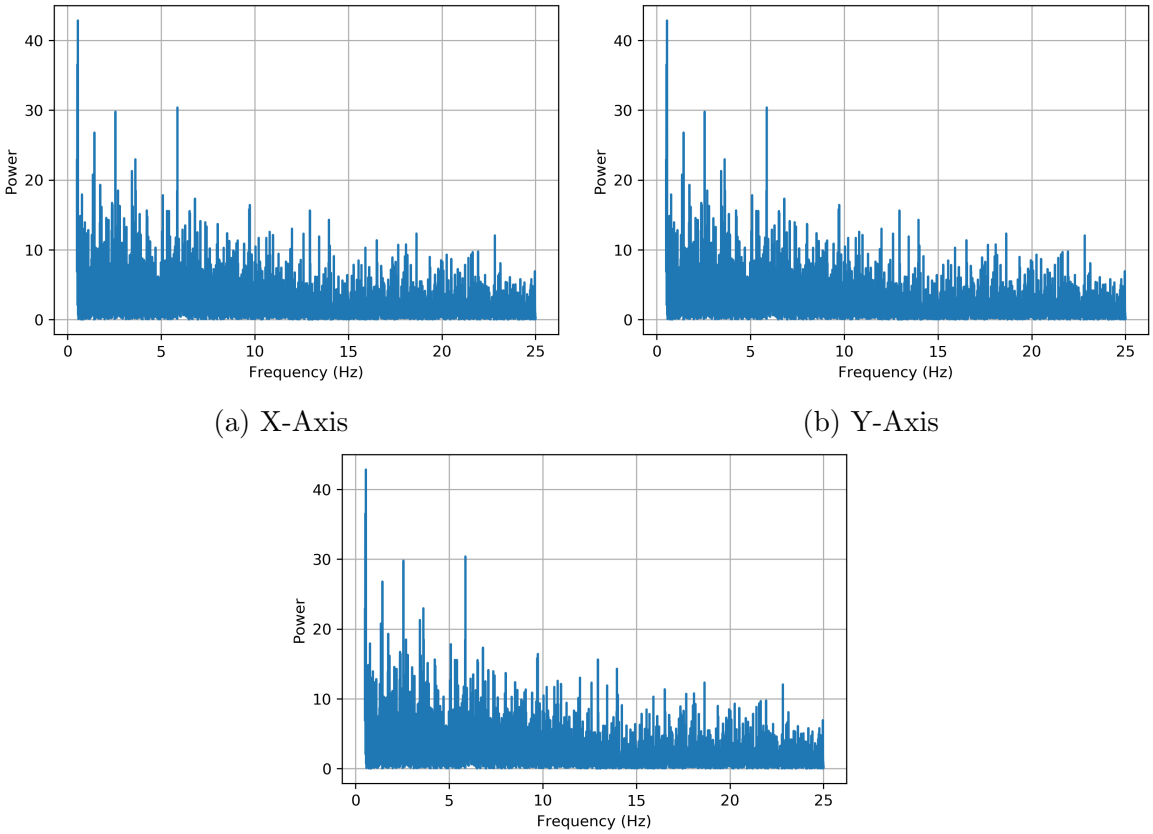
(a) X-Axis

(b) Y-Axis



(c) Z-Axis

Figure A.32: Power spectrum laying session 1 recorded by bioHarness

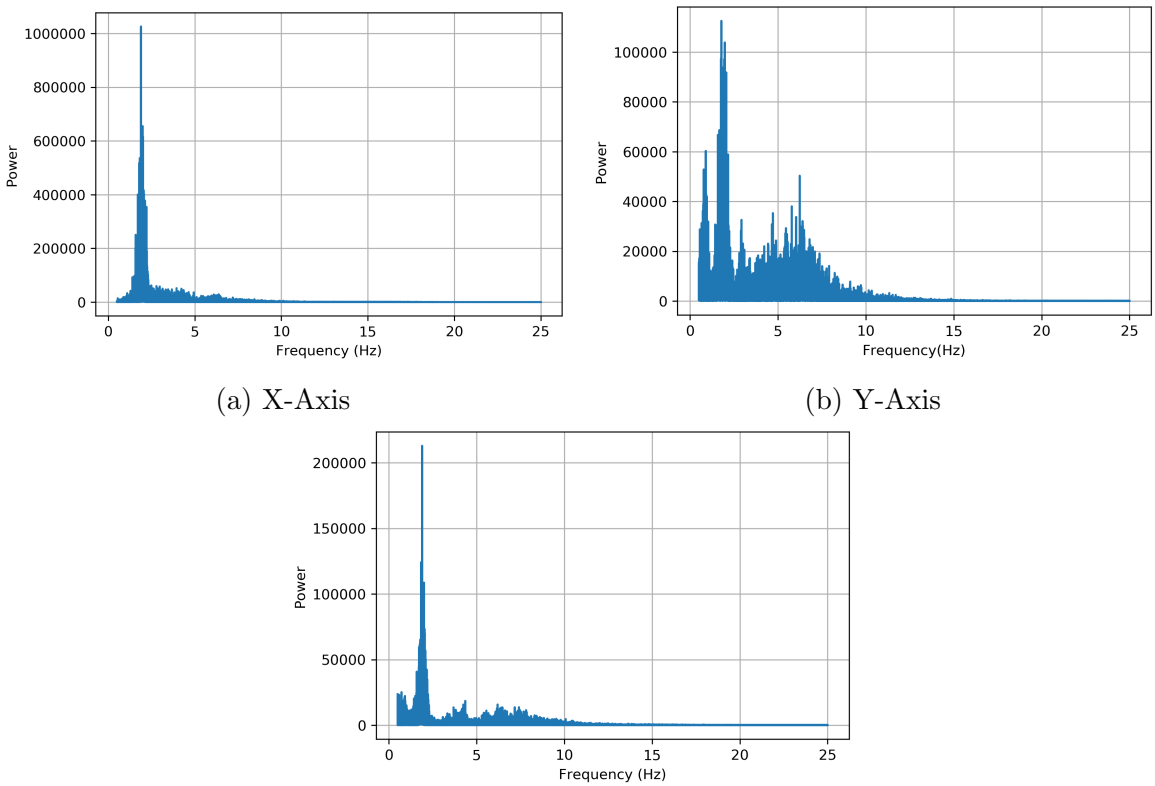


(a) X-Axis

(b) Y-Axis

(c) Z-Axis

Figure A.33: Power spectrum laying session 2 recorded by BioHarness

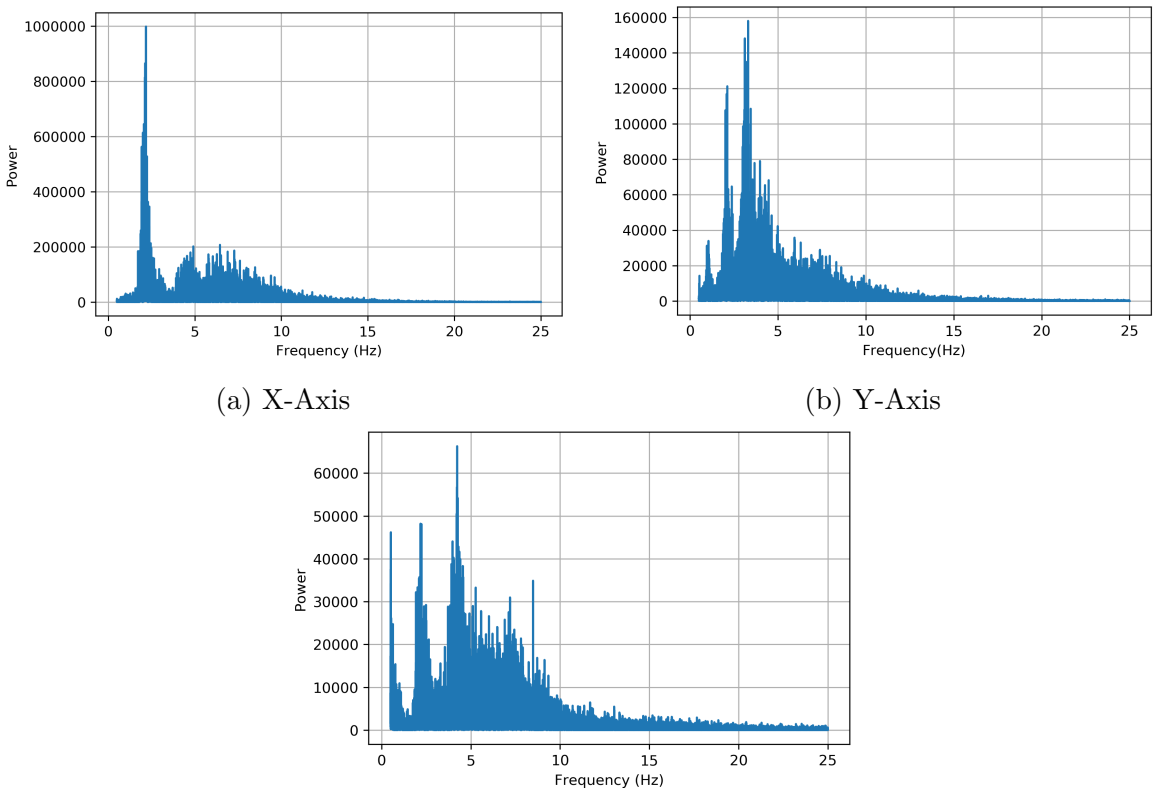


(a) X-Axis

(b) Y-Axis

(c) Z-Axis

Figure A.34: Power spectrum upstairs recorded by Astroskin

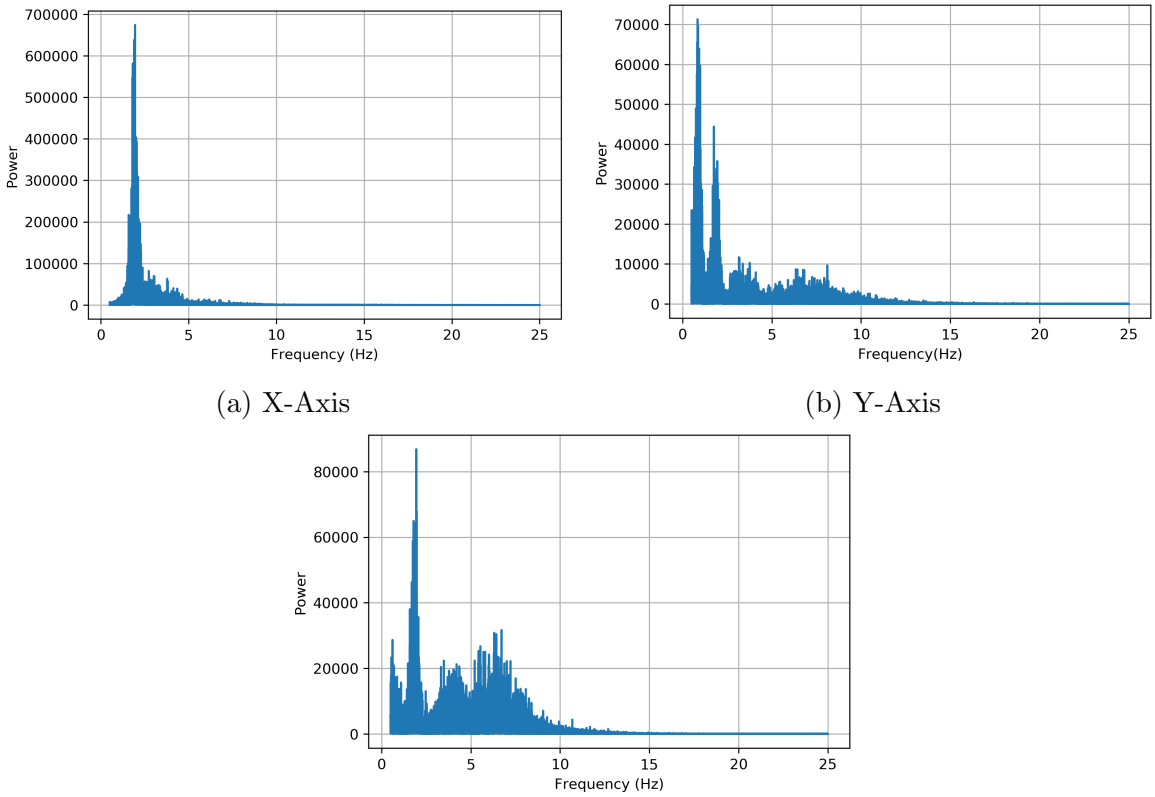


(a) X-Axis

(b) Y-Axis

(c) Z-Axis

Figure A.35: Power spectrum downstairs recorded by Astroskin

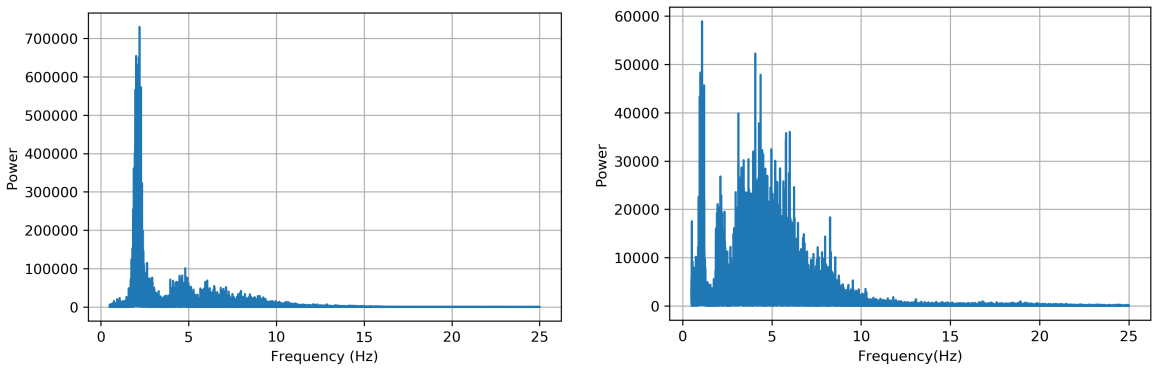


(a) X-Axis

(b) Y-Axis

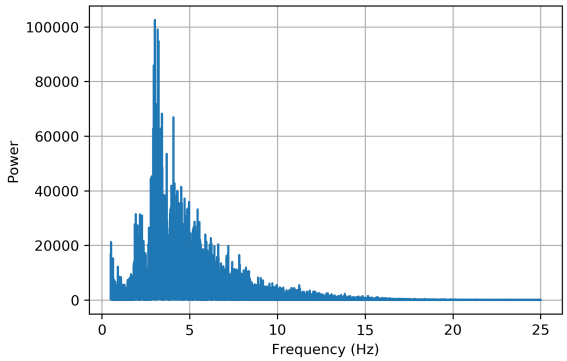
(c) Z-Axis

Figure A.36: Power spectrum upstairs recorded by BioHarness



(a) X-Axis

(b) Y-Axis



(c) Z-Axis

Figure A.37: Power spectrum downstairs recorded by Bioharness

Bibliography

(2009). *Digital Accelerometer*. Analog Devices. Rev. E.

(2018). *Zephyr BioModule 3*. Medtronic. Rev. E.

Anguita, D., Ghio, A., Oneto, L., Parra, X., Reyes-Ortiz, J. L., *et al.* (2013). A public domain dataset for human activity recognition using smartphones. In *Esann*, volume 3, page 3.

Arnold, M., Bellamy, R. K. E., Hind, M., Houde, S., Mehta, S., Mojsilović, A., Nair, R., Ramamurthy, K. N., Olteanu, A., Piorkowski, D., Reimer, D., Richards, J., Tsay, J., and Varshney, K. R. (2019). Factsheets: Increasing trust in ai services through supplier's declarations of conformity. *IBM Journal of Research and Development*, **63**(4/5), 6:1–6:13.

Bai, J., Goldsmith, J., Caffo, B., Glass, T. A., and Crainiceanu, C. M. (2012). Movelets: A dictionary of movement. *Electronic journal of statistics*, **6**, 559.

Banos, O., Galvez, J.-M., Damas, M., Pomares, H., and Rojas, I. (2014). Window size impact in human activity recognition. *Sensors*, **14**(4), 6474–6499.

Chen, M., Wang, G., Chen, H., and Ding, Z. (2020). Adaptive region aggregation

network: Unsupervised domain adaptation with adversarial training for ecg delineation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1274–1278. IEEE.

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Hammad, I. and El-Sankary, K. (2019). Practical considerations for accuracy evaluation in sensor-based machine learning and deep learning. *Sensors*, **19**(16), 3491.

Han, D.-K. and Jeong, J.-H. (2021). Domain generalization for session-independent brain-computer interface. In *2021 9th International Winter Conference on Brain-Computer Interface (BCI)*, pages 1–5. IEEE.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, **585**(7825), 357–362.

Hasani, H., Bitarafan, A., and Baghshah, M. S. (2020). Classification of 12-lead ecg

signals with adversarial multi-source domain generalization. In *2020 Computing in Cardiology*, pages 1–4. IEEE.

Haykin, S. (2001). *Communication systems*. John Wiley Sons, 4th edition.

Hengstler, M., Enkel, E., and Duelli, S. (2016). Applied artificial intelligence and trust—the case of autonomous vehicles and medical assistance devices. *Technological Forecasting and Social Change*, **105**, 105–120.

High-Level Expert Group on AI (2019). Ethics guidelines for trustworthy ai. Report, European Commission, Brussels.

Imran, H. A. and Latif, U. (2020). Hharnet: Taking inspiration from inception and dense networks for human activity recognition using inertial sensors. In *2020 IEEE 17th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET)*, pages 24–27. IEEE.

Jones, E., Oliphant, T., Peterson, P., *et al.* (2001–). SciPy: Open source scientific tools for Python.

Karantonis, D. M., Narayanan, M. R., Mathie, M., Lovell, N. H., and Celler, B. G. (2006). Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring. *IEEE transactions on information technology in biomedicine*, **10**(1), 156–167.

Ketykó, I., Kovács, F., and Varga, K. Z. (2019). Domain adaptation for semg-based gesture recognition with recurrent neural networks. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kofod-Petersen, A. (2012). How to do a structured literature review in computer science. *Ver. 0.1. October*, **1**.

Kusuma, W. A., Minarno, A. E., and Wibowo, M. S. (2020). Triaxial accelerometer-based human activity recognition using 1d convolution neural network. In *2020 International Workshop on Big Data and Information Security (IWBIS)*, pages 53–58. IEEE.

Lee, J. and Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, **35**(10), 1243–1270.

Lee, J. D. and See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, **46**(1), 50–80.

Lew, W.-C. L., Wang, D., Shylouskaya, K., Zhang, Z., Lim, J.-H., Ang, K. K., and Tan, A.-H. (2020). Eeg-based emotion recognition using spatial-temporal representation via bi-gru. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 116–119. IEEE.

Madgwick, S. O., Harrison, A. J., Sharkey, P. M., Vaidyanathan, R., and Harwin, W. S. (2013). Measuring motion with kinematically redundant accelerometer arrays: Theory, simulation and implementation. *Mechatronics*, **23**(5), 518–529.

Mu, F., Gu, X., Guo, Y., and Lo, B. (2020). Unsupervised domain adaptation for position-independent imu based gait analysis. In *2020 IEEE Sensors*, pages 1–4. IEEE.

Pan, T., Huang, Z., Ye, Y., Cheng, Y., He, W., and Wang, C. (2021). Joint transfer strategy for cross-domain human activity recognition. In *2021 IEEE 4th International Conference on Electronics Technology (ICET)*, pages 1261–1265. IEEE.

Siau, K. and Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter business technology journal*, **31**(2), 47–53.

Storkey, A. (2009). When training and test sets are different: Characterizing learning transfer. In J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, editors, *Dataset Shift in Machine Learning*, chapter 1, pages 1–28. The MIT Press, Cambridge.

Torralba, A. and Efros, A. A. (2011). Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE.

Zhang, Y., Zhang, Z., Zhang, Y., Bao, J., Zhang, Y., and Deng, H. (2019). Human activity recognition based on motion sensor using u-net. *IEEE Access*, **7**, 75213–75226.