Using Wearable Inertial Measurement Units and Deep Learning to Recognise Strength Training Exercises

by

Alex Bakker-Noguer

A Masters Dissertation

Submitted in partial fulfilment
of the requirements for the award of

Master of Science
in
Data Science
of
Loughborough University

Supervised by Dr Mohamad Saada

17th September 2023

Acknowledgements

I would firstly like to thank my supervisor Dr Mohamad Saada for his invaluable advice and support for the duration of this project. I would also like to thank my close friends and family (you know who you are) who have played an integral part in the successful completion of this thesis. I would lastly like to thank my lecturers and my course mates for making the last year not just a great educational experience but also a very enjoyable last year at Loughborough University.

Abstract

Advances in technology have led to the extensive availability and use of inertial measurement units (IMUs) in the field of human activity recognition (HAR). The widespread integration of IMUs into smart devices has led to the saturation of IMU-based HAR research in the application of aerobic exercise. This posed a unique opportunity to target the anaerobic/strength training industry that has garnered a lot of attention in recent years. Accordingly, this thesis used labelled IMU data to 1) investigate the effect that different IMU positions had on the performance of a Long Short-Term Memory (LSTM) model and 2) to assess the performance of an LSTM model between aerobic and strength training exercises. To address these objectives, four different LSTM models were trained and tested. Three LSTMs were built using data from different individual IMU positions (wrist, pocket and leg) and one LSTM was built using data from all three IMU positions combined. The results revealed that the data from the wrist and leg IMUs resulted in the highest performing LSTM models (91% and 87% accuracy respectively). In addition, the LSTM models were consistently better at recognising aerobic exercises compared to strength training exercises. Future studies should concentrate their attention on the success of the wrist IMU position and on the possibility of an IMU positioned on the foot/shoe. Moreover, it is recommended that future studies focus on determining the limitations of strength exercise recognition.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Human activity recognition (HAR) is a prominent field of research that uses contemporary sensing technologies and machine learning (ML) algorithms to automatically recognise human actions when they perform them [4]. The growing interest in this field is evident by the increasing number of research studies published in recent years [5]. This upward trend is likely due to the well-established notion that technology is constantly evolving and this has therefore made HAR systems ever more viable, accurate and reliable. Of all the different types of sensing technologies available, one type in particular has been used extensively and shown a great deal of promise in HAR systems. This type of sensing technology is the use of wearable inertial measurement units (IMUs) [6, 7].

The reason wearable IMUs are so popular in HAR is because they are commercially available at a low cost, they are already embedded into our contemporary smart devices (e.g., smartphones and smartwatches) and they have the ability to be directly attached to limbs. In addition, the wearable technology market is on a clear upward trajectory. For instance, in 2021, global shipments of wearable sensors reached 533.6 million units, an increase of 20% compared to 2020 [8]. The market is mostly shared by the large companies Apple, Xiaomi, Samsung and Huawei (they shared a total of 57.5% of the market in 2021) [8], who are constantly pushing new products into this rapidly growing market. Moreover, the wearable market is forecasted to continue to grow as another report from 2016 [9] predicted that the total number of wearable units sold would be just under 5 billion by 2027 compared to just over 3 billion in 2015. It is worth noting that the 2016 report [9] included many more wearable product categories than the 2022 report [8] (including smartwatches, hearables, medical devices, etc). Furthermore, wear-

able technology was the number 1 ranked fitness trend in the Worldwide Survey of Fitness Trends for 2023 [10]. These statistics demonstrate the lucrative potential of positive research findings in this application of HAR.

In the literature, seminal work in wearable IMU based HAR can be traced back to the late 1990's [11] and early 2000's [12]. For example, in 2000 [12], Van Laerhoven and Cakmakci conducted a study where they installed two sensors in the trousers and used ML techniques to recognise seven different activities (sitting, standing, walking, running, climbing stairs, descending stairs and riding a bicycle). Since then, wearable IMU based HAR has been widely applied to a variety of different activity types including activities of daily living (ADL), kitchen activities, transitional activities, etc [13]. Additionally, HAR as a whole field has been extensively and effectively applied to a variety of real-world applications. These include detecting abnormal behaviour using video surveillance [14], assisting the elderly and reducing the risk of disease in healthcare [15], facilitating peoples' lives in smart homes [16], etc. However, an application of particular significance, due to the advances in wearable technology, is the utilization of HAR for sports and exercise-related endeavours [17].

It is common knowledge that engaging in regular exercise induces a plethora of health benefits. This includes an improvement in physical health, such as cardiovascular fitness [18], reducing the risk of chronic diseases, such as type 2 diabetes [19], and even improving social, cognitive and emotional health [20]. Additionally, not engaging in exercise (i.e., being physically inactive) is problematic and it has been found on multiple occasions to increase the risk of mortality [21]. Therefore, performing regular exercise is not only very important for individual humans but also for reducing the worldwide healthcare burden.

Presently, the World Health Organisation (WHO) recommends that adults between the ages of 18 and 64 should engage in 150 to 300 minutes of aerobic exercise and a minimum of two sessions of muscle strengthening activities (on two separate days) per week [22]. However, a comprehensive study across 32 countries found that only around 17% of adults aged 18 and above met the guidelines of aerobic exercise and muscle strengthening activities [23]. Therefore, there is a need to create interventions to increase worldwide exercise levels. This is where HAR comes in. Research publications have shown that the automated tracking and documentation of exercise can substantially enhance individuals' motivation to engage in regular exercise [24, 25]. It is worth noting that these studies are based on pedometry and the increased motivation to engage in walking. However,

because this is a field that has been comprehensively studied these results likely carry over to other forms of activity recognition.

At present, HAR has been extensively applied to aerobic exercise and ADL [26, 27, 28, 29, 30, 31]. This is reflected in the wide array of commercially available wearable products [32], as they are capable of monitoring and recording various aerobic exercises, including walking, running, cycling and swimming. Furthermore, emerging devices (e.g., Fitbit Sense series, Apple Watch with OS 5 or later and WHOOP 4.0) exhibit promising capabilities in automatically detecting aerobic activities without requiring explicit user input. However, it should be noted that some of these devices necessitate a minimum activity duration of 15 minutes before successful detection can occur.

Conversely, there is a noticeable absence of commercial wearables that can effectively track and record strength training exercises. This is replicated in the research with a limited number of studies contributing to the recognition of strength training exercises [33]. This is slightly alarming for a few reasons. Firstly, as mentioned previously, the WHO recommends that adults engage in weekly strength training exercise [22]. Secondly, strength training is trending highly at the moment as it was ranked 2nd in the Worldwide Survey of Fitness Trends 2023 [10]. Thirdly, strength training is a crucial part of an athletes training programme as it can help to prevent injury and improve performance [34]. Lastly, strength training is associated with a whole host of health benefits [35]. Given these reasons, the limited capabilities of commercial wearables and the limited research studies, HAR for strength training exercise presents an opportunity to tap into a flourishing commercial market, to improve public health and to support the development of athletes worldwide.

1.2 Aims and Objectives

Accordingly, this thesis aims to recognise and classify strength training exercises using labelled data from wearable IMU sensors and ML. In order to achieve this, the thesis implements various Long Short-Term Memory (LSTM) neural networks. As a result of the literature review, the aim is broken down into two key objectives:

1. To compare the effects that different IMU sensor positions have on the classification accuracy of an LSTM model.

2. To compare the classification accuracy of an LSTM model between aerobic vs strength training exercises.

1.3 Thesis Structure

To successfully achieve these objectives the thesis was structured as follows. Chapter 2 contains the literature review which was conducted to gain a deeper understanding of the field of HAR and to identify gaps in the research that enabled the aims and objectives to be formed. In parallel to this, a dataset, containing IMU data, was searched and selected to enable the aims and objectives to be addressed. Chapter 3 details a description of the methodology implemented, including the exploratory data analysis, any data preprocessing, and the model design. Chapter 4 displays and analyses the results of the experiments and discusses the meaning of these results against the relevant literature. Finally, Chapter 5 concludes the thesis and deduces what the findings imply for future research.

Chapter 2

Literature Review

Following on from the introduction, which provided an overview of the rationale and aims of this thesis, this chapter provides a more detailed discussion of the current literature. Conducting a literature review is crucial to understanding a subject area. It involves the investigation and critique of similar work in the area, which results in the identification of gaps in the research. This literature review contains three key subsections detailing 1) the different sensing modalities used in HAR, 2) a few of the most common deep learning (DL) techniques used for HAR and 3) HAR within the sport and exercise domain. A summary is also provided which highlights the key findings in the current literature and relates them back to the aims and objectives from Chapter 1.

2.1 Human Activity Recognition (HAR) Sensing Types

It is widely acknowledged within the scientific community that technology is continuously progressing. This advancement in technology has presented new avenues for the field of HAR to develop sophisticated systems capable of addressing real-world challenges. The hardware employed in HAR has witnessed significant advancements in several aspects, including computational power, data quality, set-up requirements, size, and cost [36]. Consequently, these advancements have made HAR more accessible and attainable. The practice of HAR incorporates a diverse range of sensing technologies which can be broadly classified into two primary categories: wearable sensors and external sensors [4, 7, 13]. This subsection of the literature review provides an overview of each type, their strengths and limitations, as well as giving examples of each in the literature.

2.1.1 External Sensors

External sensors are devices that remain stationary and measure the surrounding environment. They are often deployed in smart homes or in other indoor environments. Examples of these sensors include radio frequency identification (RFID), WiFi, and video cameras [37]. RFID identifies and tracks tags by detecting the electromagnetic pulse from a nearby reader. The strength of the electromagnetic pulse is altered by human movement, and this is what enables the movement to be recognised. Fan et al [38] used deep neural networks and RFID data to build HAR systems. As WiFi is built into many indoor environments and urban buildings it has also been widely used for activity recognition. Most of the HAR literature in WiFi has used WiFi channel state information [39].

Although video cameras make up only a subsection of all external sensors used in HAR they are one of the most prominent sensors used in the literature [6]. Typically, RGB cameras are used in HAR as they can recreate what the human eyes see. A notable example is the study conducted by Ji et al [40] who used RGB videos from surveillance footage in an airport to recognise human actions. Regular RGB video cameras and specialised cameras have also been used to create different data modalities that can provide alternative information. For example, RGB videos can be used to create superimposed human skeletons [41]. From these, human movements can be detected. However, skeleton video-based HAR is sensitive to pose variations which can affect the accuracy of activity recognition, especially for complex movements. Furthermore, depth cameras (such as Microsoft Kinect and Time-of-Flight Cameras) are another type of video-based HAR [42] which use infrared light to measure the distance between the camera and the objects in the environment (field of view).

The main reason video cameras have become so popular in HAR is because video data can supply a vast amount of spatial and temporal information. This encourages the building of deep neural networks which can achieve a higher accuracy for recognising human activities than classical ML algorithms. Video data also offers a realistic representation of the activity and can be applied to a wide range of scenarios. However, due to these large sequences of image data, video-based HAR requires a very high amount of computational power compared to other sensing types. This in turn increases the processing time and the cost of processing the video data. Other limitations of video-based HAR include the field of view, the viewpoint, the environmental conditions and privacy (as not many people like to be constantly recorded by cameras).

2.1.2 Wearable Sensors

Wearable sensors, as the name suggests, encompass sensors that can be worn or integrated into various forms of attire. Examples of such sensors include those embedded in clothing, specialized medical equipment and personal devices such as smartphones or smartwatches [7]. Wearable sensors offer many practical advantages over external sensors. With the advancement of technology, wearable sensors have become smaller in size and generally more cost-effective. As previously stated, this enables them to be integrated into everyday clothing or accessories. They are also not affected by occlusion, background and privacy issues associated with external sensors.

Wearable sensors encompass various types of built-in sensors, such as biosensors and IMUs. Biosensors are devices that measure the activity of human organs and biological systems like blood pressure, heart rate and respiratory information. Whilst IMUs are the most popular wearable sensor used in HAR systems [43] and are also one of the most popular sensors used for HAR overall [6, 7]. IMUs contain 3 separate sensors (tri-axial accelerometers, gyroscopes and magnetometers), that are either used separately [44] or in combination [45]. The two most commonly used are accelerometers and gyroscopes [46]. Accelerometers measure the acceleration in metres per second squared, whilst gyroscopes measure the angular acceleration (usually in radians per second).

One key challenge of IMU-based HAR is the placement of the sensor. IMUs have been attached to many parts of the body (including, but not limited to, the wrist, chest, waist and leg) to gauge which position gives the highest accuracy for detecting activities [47, 48]. In these studies, it is often observed that more than one IMU gives significantly higher accuracy. This is supported by the review paper conducted by Chaurasia and Reddy [4]. However, studies using separate or single IMUs have also resulted in a high detection accuracy for human movement [44, 49]. Furthermore, many of the positions used are not practical for everyday life and can be uncomfortable for the user. There is therefore a need for investigating and comparing IMU locations that reflect the position of everyday smart devices and clothing. This would ensure that the IMU-based HAR system can be practically implemented into day-to-day life.

2.1.3 Hybrid Sensors

In HAR, different sensing modalities have often not been used in isolation. Many have been combined in order to achieve higher accuracy, to counteract their

separate disadvantages and to combine their separate advantages. For example, Hayashi et al [50] found that the accuracy of their HAR system was improved when they combined IMU and acoustic data. In addition, Bharti et al [51] combined different sensing types (including IMU, temperature and location from Bluetooth) and achieved a model accuracy of 95%. Furthermore, Zou et al [52] investigated gait recognition by combining data from IMU sensors, RGB video cameras and depth video cameras. They found their model achieved higher recognition accuracy when compared to other gait recognition models using a single type of sensor.

It is clear that merging different sensing modalities can improve model reliability, robustness and accuracy. However, the use of more than one technology incurs a higher cost and also involves more complexity which usually demands more processing capabilities. For these reasons using hybrid sensing modalities was outside of the scope of this thesis.

2.2 Machine Learning (ML) Methods

In the field of HAR, researchers employ ML and DL algorithms to recognize and classify various activities. ML algorithms adhere to the conventional pipeline, encompassing steps such as raw data collection, preprocessing, feature extraction, ML model training, and classification. Examples of commonly utilized ML algorithms include Decision Tree, Support Vector Machine, and K-nearest neighbour [7]. In recent years, DL techniques have gained significant traction in HAR and other domains [37] owing to their remarkable performance [53]. DL methods typically adopt a modified pipeline wherein feature extraction is automated during the model training process. This subsection of the literature review provides examples of DL algorithms used in the field of HAR, including the DL algorithm chosen for this thesis.

2.2.1 Artificial Neural Networks (ANN)

Artificial Neural Networks (ANNs), also known as Neural Networks (NNs), are a group of ML algorithms that contain artificial neurons (or nodes). They are called NNs as they are said to resemble the human brain [54]. They are known for their layered structure which increases their complexity and lends to them being referred to as DL algorithms. The concept of NNs has been around for 80 years as the theory behind them was first coined by McCulloch and Pitts [55]. The next turning point was in 1986 when Rumelhart and his colleagues [56] introduced the idea of backpropagation. However, it was not until 2012 where Krizhevsky, and

his two colleagues, [57] developed the NN famously known as AlexNet. Their research demonstrated constructing very deep NNs was viable with the advancement in computing power and the amount of data. In addition, this deeper NN delivered much greater performance.

In general, two of the most common types of NNs used in computing research are Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) [58]. This is also true in the field of HAR, as the two prominent DL algorithms are CNNs and LSTM networks (a type of RNN) [7]. This section provides an overview of key research contributions and advancements in these two algorithms, as well as examples of their use in the field of HAR.

Convolutional Neural Network (CNN)

CNNs are a type of NN that are particularly suited for image or video classification tasks as they have been widely applied to the fields of computer vision [59] and face recognition [60]. They use convolutional layers and kernels to extract pertinent features from the data and they use pooling layers to reduce the complexity of the model and control for overfitting. The seminal research in CNNs was conducted by LeCun when he constructed the famous LeNet [61]. This was later improved in 1998 with LeNet version 5 [62].

In the field of HAR, two different CNN frameworks have been applied: two-stream 2D CNNs and 3D CNNs. Two of the early two stream methods were conducted by Karpathy et al [63] and Simonyan and Zisserman [64]. Simonyan and Zisserman [64] constructed a model that comprised of two different networks, temporal and spatial. Each network learnt the relevant appearance and action features separately. As is typical for two-stream 2D CNNs [65], the final classification scores were constructed by the fusion of the classification scores from each stream. A classical 3D CNN approach was conducted by [40] to recognize human actions in video sequences. This single model extracted both temporal and spatial features and achieved state-of-the-art results compared to baseline methods. It is worth noting that these early models were mainly trained on short video clips. Later models focused on improving the long-range spatio-temporal dependencies [66].

Even though CNNs can combine spatial and temporal features to produce a high level of accuracy and have been improved to longer sequences of time, they still only use relatively short time sequences. NNs that focus on temporal sequence

modelling (e.g., LSTM) are a high performing alternative that are powerful enough to overcome longer temporal information.

Recurrent Neural Network (RNN)

RNNs are a type of ANN that emerged to model and recognise patterns from sequential data [67]. Their structure enables them to have a memory of past states which they can use to predict the sequence in the following timestep. The concept of RNNs was introduced by Elman [68] and since then they have been applied to a variety of domains such as speech recognition [69] and time-series analysis [70]. However, basic RNNs suffer from the vanishing/exploding gradient problem in backpropagation and they also only have a short-term memory. Due to this, improved versions of RNNs have been created, such as the LSTM.

Long Short Term Memory (LSTM)

LSTMs are a type of RNN that were designed to address the limitations of vanilla RNNs [1]. They overcome the vanishing/exploding gradient problem and have an increased memory capacity for learning long-term dependencies. LSTMs use different gated units to regulate the flow of information, enabling them to remember information from long sequences. These include the forget, input and output gates (Figure 2.1). LSTMs also use two separate paths to make predictions. One is known as the "cell state", this acts as the main pathway that stores long-term memory. The other is known as the "hidden state", this is the pathway for short-term memory. The flow of information within the cell and hidden states is controlled by the gated units. They essentially assist the LSTM in determining whether specific information should be remembered or forgotten. Contrary to vanilla RNNs, LSTMs transmit only the most valuable information by making small information adjustments.

The LSTM network was first proposed by Hochreiter and Schimdhuber in 1997 [71]. Since then, it has become a state-of-the-art model being used by Google for speech recognition [72]. As alluded to earlier, LSTMs have been used extensively in the field of HAR [7]. Examples of LSTM-based HAR models are the papers by Mekruksavanich and Jitpattanakul [73], Uddin and Soylu [74] and Pienaar and Malekian [75]. Mekruksavanich and Jitpattanakul [73] conducted a comprehensive study comparing LSTMs with different numbers of layers and interestingly found similar performance for each LSTM network. Uddin and Soylu [74] used data from multiple different IMU measures (accelerometer, gyroscope and magnetometer) to recognize different activities such as running, walking, and cycling.

The results from this study suggest that the integration of these different IMU measures (accelerometer, gyroscope and magnetometer) allows for a more comprehensive representation of human activities, leading to enhanced recognition accuracy. Pienaar and Malekia [75] proposed an LSTM model using the WISDM dataset [76] and achieved an activity recognition accuracy of 94%.

LSTMs have emerged as a widely preferred type of ANN for tackling sequential problems. This preference stems from their capability to retain information over long sequences, leading to remarkably high levels of accuracy. Given its popularity and its advantageous characteristics, the LSTM network was selected as the model for this thesis.

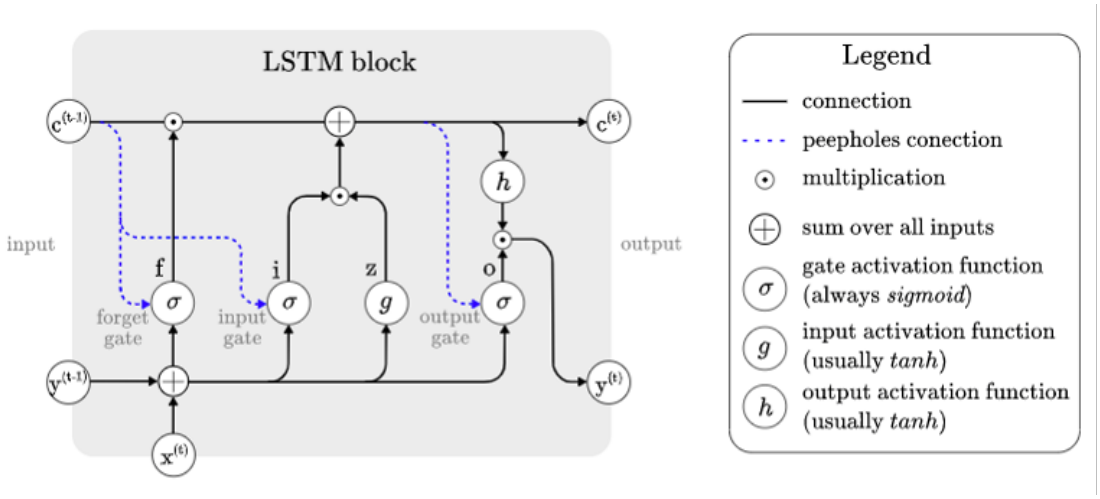


Figure 2.1: Architecture of a typical LSTM block. Source: [1]

Hybrid Methods

Hybrid models use a combination of DL models in the aim of achieving a better performing model by leveraging their respective strengths and overcoming their respective limitations. A promising hybrid model involves integrating CNN and RNN architectures. Notable studies in the field of HAR by Ordóñez and Roggen [30] and Singh et al [77] exemplify effective approaches to combining CNNs and RNNs. In the paper by Ordóñez and Roggen [30] the CNN and RNN dense layers outperformed the regular CNN with just dense layers. This improved performance was likely due to the CNN capturing spatial relationships in the data whilst the RNN captured temporal relationships. Singh et al [77] came to a comparable conclusion. In addition, CNNs have been used in combination with restricted boltzmann machines [78] and stacked autoencoders [79]. In these studies, CNNs are utilized for feature extraction, while the generative models aid in accelerating the training process.

Despite the advantages hybrid models bring, similarly to combining different sensing types, they are associated with an increase in complexity and computational requirements. In addition, integrating different models requires an even more careful hyperparameter selection and tuning process. This adds further complexity and therefore hybrid models were also outside the scope of this thesis.

2.3 Human Activity Recognition (HAR) for Sport and Exercise

HAR has been successfully employed in a diverse array of real-world applications, encompassing areas such as security and surveillance [14], healthcare [15] and smart homes [16]. As alluded to in Chapter 1, due to the advances in wearable technology, HAR has sparked a significant interest in the area of sports and exercise. This is because exercise is associated with many positive effects for one's physical [18] and mental [20] health. In addition, HAR can help to increase one's motivation to engage in physical activity, thus reducing one's chronic disease risk. Exercise can be broadly split into two categories: aerobic and anaerobic. This section of the literature review covers each type in the context of HAR.

2.3.1 Aerobic Exercise

Aerobic exercise, also known as endurance or cardiovascular exercise, is a form of physical activity (PA) that uses oxygen to meet the energy demands of that activity. It can be any form of PA (low to high intensity) that raises heart rate and promotes the circulation of oxygenated blood around the body and towards the working muscles. Aerobic exercise involves exercises that are highly repetitive and that last for extended periods of time. Countless research papers have evidenced the numerous benefits associated with performing regular aerobic exercise. For these reasons HAR has been extensively applied to sporting exercises that are aerobic in nature.

For example, an early paper by Ermes et al [26] used tri-axial accelerometers on the hip and wrist to classify 10 different activities. They implemented a hybrid classifier and achieved a maximum accuracy of 90%. The activities included aerobic exercises such as running, cycling on a regular bike and rowing on a rowing machine. However, they also included activities of daily living (ADL), such as lying down, standing and sitting. These are not necessarily aerobic exercises, but they are one of the most popular types of activity HAR has been applied to and

are commonly combined with aerobic exercises in research studies [27, 28]. This is due to many of the benchmark HAR datasets containing just ADL (OPPORTUNITY, SKODA) [80, 81] or a combination of both ADL and aerobic exercise (WISDM, UCI) [76, 82]. For this reason, HAR applied to ADL was also reviewed in this section of the literature review.

A few examples of studies investigating HAR of aerobic exercise and ADL include [27, 28, 29, 30, 31]. Bayat et al [27] and Ayu et al [29] conducted similar studies in that they both collected their own data from a total of 4 participants (2 males and 2 females). Ayu et al [29] used 2 IMU sensors (one held in the hand and the other in the trouser pocket) and found that the more training data they used the higher the accuracy of their model. Bayat et al [27] compared the position of an IMU in the hand vs in the pocket. Interestingly, they were able to achieve a moderately high accuracy of 91.15% with 1 IMU sensor. This is likely due to the fusion method they applied to enhance the classification accuracy. These studies suggest that a small number of sensors may still be able to provide high classification accuracy, however they both use very small numbers of participants so these results may not generalise very well to the wider population.

The other three studies [28, 30, 31] created models using benchmark datasets. Alsheikh et al [28] used the WISDM [76], SKODA [81] and DAPHNET [83] datasets, Ordóñez and Roggen [30] used the OPPORTUNITY (ADL) [80] and SKODA (ADL) [81] datasets and Ignatov [31] used the WISDM (aerobic + ADL) [76] and UCI (aerobic + ADL) [82] datasets. All three employed DL models and achieved state-of-the-art performance. These studies further demonstrate the superior performance of DL algorithms and in turn support the use of DL in this thesis.

It is evident that there is an extensive amount of literature examining HAR in aerobic exercises and ADL. This has saturated the possible research avenues available in this application of HAR. The application of HAR to anaerobic exercise also exists and merits discussion and critique.

2.3.2 Anaerobic Exercise

Anaerobic exercise, also known as strength training, is a form of PA that does not use oxygen to meets its energy demands. It usually involves exercise that is of much higher intensity and is of much shorter duration than aerobic exercise. Anaerobic exercise also raises the heart rate however, it is less about encouraging blood circulation and more about maximising strength and power. Anaerobic ex-

ercises are also usually broken down into repetitions and sets (e.g., 3 sets and 10 repetitions in each set). Over the years, several studies have highlighted the numerous benefits of engaging in anaerobic exercise [35]. These include increasing muscular strength and power, attenuating muscle wasting and interestingly improving cardiovascular health. Although these benefits exist, and (as alluded to in Chapter 1) the WHO recommends adults to complete anaerobic exercise as well as aerobic exercise, there is limited research in this application of HAR. A reason for this could be due to the higher complexity of the movements and the large variety of different strength training exercises that exist.

One of the earliest studies was conducted by Chang et al [84] who classified 9 free weight exercises using 2 IMUs (one on a glove and one clipped to the front of the waist). 10 participants were recruited for this study and the model achieved an accuracy of 90%. This was pioneering work, however the IMUs were placed in positions that are not practical when performing weight training exercises. The only other notable free weight exercise study was conducted by Depari et al [85]. They achieved a classification accuracy of 93% when exploring 10 free weight exercises using data from 1 IMU (on the wrist) and 7 participants. This study suggests that using only 1 IMU may have the capability to provide adequate activity recognition.

Two other examples of HAR for anaerobic exercises are studies conducted by Koskimaki and Siirtola [86] and Morris et al [87]. Koskimaki and Siirtola [86] investigated 30 different upper body exercises using 2 IMU's (one on the left wrist and one on the chest). They achieved good results, however they only collected data from 1 participant, so results have a low validity, and they also likely have low generalisation to the general population. Morris et al [87] investigated 13 different gym exercises using 1 IMU (on the forearm) for 114 participants. They achieved outstanding classification accuracy (the lowest was 96%). However, a complex hybrid model was used and the IMU was placed in a position that is not typical when performing strength exercises.

Although there have been fewer studies in strength exercise based HAR, they have shown great promise. A relatively low number of participants (7 - 10) and even 1 IMU was shown to produce high recognition accuracy. However, a key limitation to most of these studies was the position of the IMU(s) used. There are a lack of studies that look at strength exercise based HAR for IMU's in multiple practical locations. This presents a gap in the literature that this thesis aims to investigate. In addition, there are a lack of studies that directly investigate whether strength training exercises (potentially due to their higher complexity) are more

difficult for HAR systems to detect compared to aerobic exercises. Understanding this is another key aim of this thesis and has implications for the viability of future work in strength exercise HAR using IMUs.

2.4 Summary

In summary, HAR is a considerably large research field and has many different branches within it. The review of the three subtopics (HAR sensing modalities, HAR ML methods and HAR for sport and exercise) highlighted the rationale behind the aims of this thesis.

Firstly, the sensing modality chosen was wearable IMUs due to their many practical advantages, especially when compared to video cameras. These include, their cost, portability and how they are integrated into smart devices that people already carry around and wear. This means that HAR systems built with data collected from these sensors can be more easily implemented to the masses and therefore create more impact.

Secondly, the model chosen was an LSTM model. It is the obvious model of choice due to the high accuracy it can achieve (as it is a type of DL model) and its popularity in the field of HAR. In addition, it is specialised for sequential data (which is the data collected by IMUs).

Thirdly, this thesis is targeting the classification of anaerobic/strength/gym exercises. This HAR application has not received nearly as much attention as the application of aerobic exercise and ADL. This may be due to the complexity of the movements and the plethora of different gym exercises that people perform. This relates directly to Objective 2 of this thesis (a comparison between the classification of aerobic and anaerobic exercises), which aims to investigate whether the movements of strength training exercises are in fact too complex for HAR. In addition, there is a lack of research exploring multiple practical IMU positions for strength exercise based HAR. This relates directly to Objective 1 of this thesis (a comparison between different practical IMU locations, in isolation and in combination), and aims to address this gap within the literature. The overall findings of this thesis have implications for whether strength exercise based HAR systems have the potential to be seamlessly embedded into our daily lives.

Chapter 3

Methodology

This chapter follows on from Chapter 2 and describes the methodology used to address the aims and objectives outlined in Chapter 1. The chapter starts with a description of the tools used throughout the processes and a description of the dataset used. This is followed by the exploratory data analysis (EDA), where the dataset was analysed to gain an understanding of what it contained and to assess its quality before employing preprocessing techniques. The rest of this chapter details the data preprocessing steps performed and the design of the model, including the hyperparameter tuning, evaluation methods used, and the different experiments conducted.

3.1 Tools

The methods described in this chapter were all undertaken using Python. Python is a high-level and general-purpose programming language that is known for its readability and simplicity. The Python code was written using the Jupyter Notebook integrated development environment (IDE), as it provides a great interface for structuring files due to the ability to create separate code and markdown cells. Python contains many libraries that enable complex tasks to be completed easily using specific methods within these libraries. The libraries Pandas and NumPy were used for general data analysis and transformation processes. For data visualisation the libraries Matplotlib and Seaborn were used. For model building and specific preprocessing tasks the libraries TensorFlow and Sci-kit Learn were used. TensorFlow is a popular library used for building deep learning models, whilst Sci-kit Learn is a popular library for many classical machine learning tasks such as data preprocessing, model building and model evaluation.

3.2 The Dataset

Figure 3.1: Example of each workout being performed. Source: [2]

The dataset used was a labelled dataset obtained from the following website [88]. The dataset has been previously used in the following three studies [2, 89, 90]. This dataset was chosen as it was suitable to answer the research aims and objectives detailed in Chapter 1. This is because it contained both anaerobic and aerobic gym workouts and because it contained data from IMU sensors positioned in 3 different locations. These were the wrist, the pocket and the leg (more specifically the lower calf). The dataset contained data from 10 participants (5 males and 5 females) who performed 11 different popular gym workouts. The workouts include 5 aerobic (Riding, Ropeskipping, Running, Stairsclimber and Walking) and 6 anaerobic (Adductor, Armcurl, Benchpress, Legcurl, Legpress and Squat) type exercises. An example of each of the exercises being performed is displayed in Figure 3.1. This enabled a comparison between the aerobic and anaerobic type

exercises. This is a total of 11 workout classes, however there was also a labelled Null class where the participants were not performing the exercises or resting between the exercises.

Each participant completed all the gym exercises consecutively in one testing battery. For the anaerobic exercises this consisted of 3 sets of 10 repetitions (except for Squats), whilst the aerobic exercises were performed for around 2 minutes each. Each participant completed the exercise battery 5 times (5 consecutive days). This addressed the limitation of a low participant number and increased the amount of data available as many standard datasets in HAR contain data from more than 10 participants: 30 participants (UCI-HAR) [82], 51 participants (WISDM) [76] and 90 participants (KU-HAR) [91]. Further details about the experimental protocol conducted can be found in the study conducted by Bian et al [2].

3.3 Exploratory Data Analysis (EDA)

EDA is the process of analysing and understanding the characteristics of a dataset before implementing a machine learning algorithm. It is a crucial step that enables data scientists to identify patterns and data quality issues that will inform the most suitable preprocessing steps required before model implementation. Similarly, to model building, it is usually an iterative process as initial model performance and insights from EDA can be used to further manipulate the dataset to enhance model performance. The EDA section will cover the following subsections: data description, category distribution, correlations analysis and time series analysis.

3.3.1 Data Description

Initially the dataset was loaded into a Jupyter Notebook file and the .info() method was executed to provide a general overview of the contents of the raw dataset (Figure 3.2a). The raw dataset contained 4703200 instances and 11 columns/attributes: object (participant number), day, workout (class), sensor position, A_x (acceleration in x direction), A_y (acceleration in y direction), A_z (acceleration in z direction), G_x (angular acceleration in x direction), G_y (angular acceleration in y direction), G_z (angular acceleration in z direction) and human body capacitance (HBC). For the purposes of this thesis the HBC column was removed. In addition, the types of the columns object, day, workout and sensor position were converted to categorical to make data analysis easier. The tri-axial acceleration and angular acceleration were the 6 features used for model training

and the workout column contained the labelled classes. The categorical columns contained important information that enabled the data to be filtered for the EDA and for training different models.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4703200 entries, 0 to 4703199
Data columns (total 11 columns):
 #   Column            Non-Null Count    Dtype
---  ------            --------------    -----
 0   Object            4703200 non-null  int64
 1   Day               4703200 non-null  int64
 2   Workout           4703200 non-null  object
 3   Sensor_Position   4703200 non-null  object
 4   A_x               4703200 non-null  float64
 5   A_y               4703200 non-null  float64
 6   A_z               4703200 non-null  float64
 7   G_x               4703200 non-null  float64
 8   G_y               4703200 non-null  float64
 9   G_z               4703200 non-null  float64
 10  Body_Capacitance  4703200 non-null  float64
dtypes: float64(7), int64(2), object(2)
memory usage: 394.7+ MB
```

(a) Before field type change.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4703200 entries, 0 to 4703199
Data columns (total 10 columns):
 #   Column           Non-Null Count    Dtype
---  ------           --------------    -----
 0   Object           4703200 non-null  category
 1   Day              4703200 non-null  category
 2   Workout          4703200 non-null  category
 3   Sensor_Position  4703200 non-null  category
 4   A_x              4703200 non-null  float64
 5   A_y              4703200 non-null  float64
 6   A_z              4703200 non-null  float64
 7   G_x              4703200 non-null  float64
 8   G_y              4703200 non-null  float64
 9   G_z              4703200 non-null  float64
dtypes: category(4), float64(6)
memory usage: 233.2 MB
```

(b) After field type change.

Figure 3.2: Information about the dataset.

The non-null count for each column indicates that the dataset had no missing values (Figure 3.2a). Checking for missing values is another crucial step before implementing ML models. This is because missing values pose several threats to ML models including: compromised data integrity, poorer model performance and biased or inaccurate predictions. Missing data is often handled in two ways either 1) the rows corresponding to the missing value is removed or 2) the missing value is filled using an imputation technique. There are many different imputation methods, and the choice of method depends on the missing data problem you are trying to solve. Figure 3.2a displays the type of the acceleration measures as float type. This means they are continuous numbers that contain decimal points. Figure 3.2b displays the general information of the dataset after the initial cleaning steps.

The .describe() method was then used to obtain descriptive statistics of the dataset (Figure 3.3). It can be seen that the minimum value is 0 and the maximum value is 1 for all of the acceleration and angular acceleration features. This indicated that the dataset had already been normalised between 0 and 1. This is a standard step in data preprocessing when building DL models as they tend to work best when the input values are small (i.e., between 0 and 1). When the input values are too large, the computational requirements of the model are significantly increased as the model hast to learn larger weights. Therefore, it is important to normalise the input values to reduce the computational cost of model training. Normalisation is also important for achieving better model performance, as it converts all the features to a similar scale and effectively handles outliers. In

addition, in DL algorithms, normalising the features enables gradient descent to converge faster as it avoids oscillations in different directions caused by different scales. The normalised input values are mathematically calculated as follows:

$$x' = \frac{x - min(x)}{max(x) - min(x)} \tag{3.1}$$

Where x indicates an input vector of sensor readings, and x' indicates the resulting normalised vector ranging from 0 and 1. Where min(x) and max(x) indicate the minimum and maximum values within that input vector (x).

	A_x	A_y	A_z	G_x	G_y	G_z
count	4.703200e+06	4.703200e+06	4.703200e+06	4.703200e+06	4.703200e+06	4.703200e+06
mean	5.035507e-01	4.985133e-01	4.977778e-01	4.997786e-01	4.995448e-01	5.002019e-01
std	4.246152e-02	4.969804e-02	3.060151e-02	4.460951e-02	3.632589e-02	4.586413e-02
min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
25%	4.965000e-01	4.931250e-01	4.945000e-01	4.931250e-01	4.945312e-01	4.929688e-01
50%	5.000000e-01	4.997500e-01	4.998750e-01	4.999625e-01	5.000000e-01	4.999625e-01
75%	5.056250e-01	5.047500e-01	5.032500e-01	5.064437e-01	5.051937e-01	5.054312e-01
max	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00

Figure 3.3: Descriptive statistics of the dataset.

In order to visualise the spread of the data, histograms were plotted of each feature (Figure 3.4). It is clear that each feature observed a normal distribution and there were no extreme outliers present in the data. For LSTM models normal distribution is not a key assumption however, visualising data in this way is typical in EDA and facilitates a deeper understanding of the dataset. Whilst for some classical ML algorithms, such as Naïve Bayes and Logistic Regression, normal distribution is a key assumption.

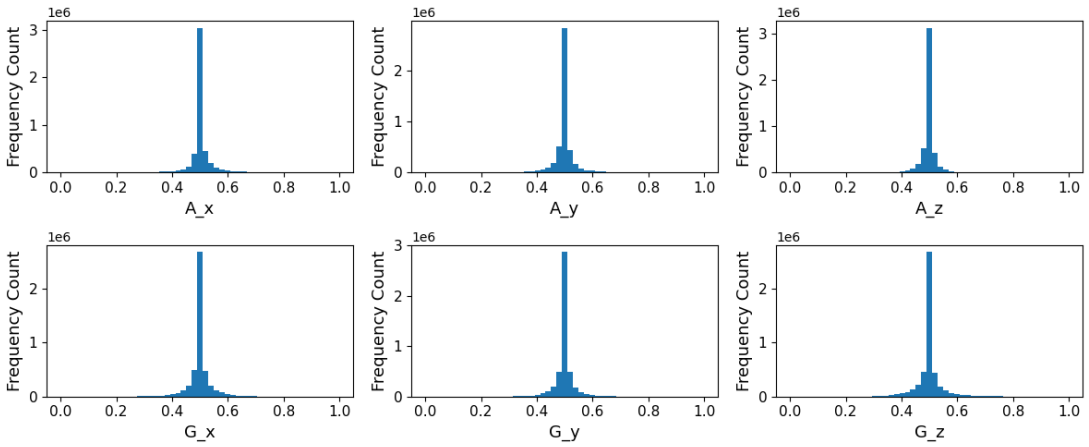


Figure 3.4: Histograms of each feature.

3.3.2 Category Distribution

The next step in the EDA was investigating how the data was distributed in each categorical column. This is a crucial step before implementing classification machine learning algorithms as it reveals whether the class distribution is balanced or not. Figure 3.5 displays the counts in each category for each categorical column. Bar graphs enable the distribution to be visualised clearly. The most important graph is the bottom left graph, where the counts of each of the classes (workouts) is displayed. It can be seen that there was a significant class imbalance, as a large majority of the class labels were Null. In addition, there were much less training examples for RopeSkipping compared to Walking.

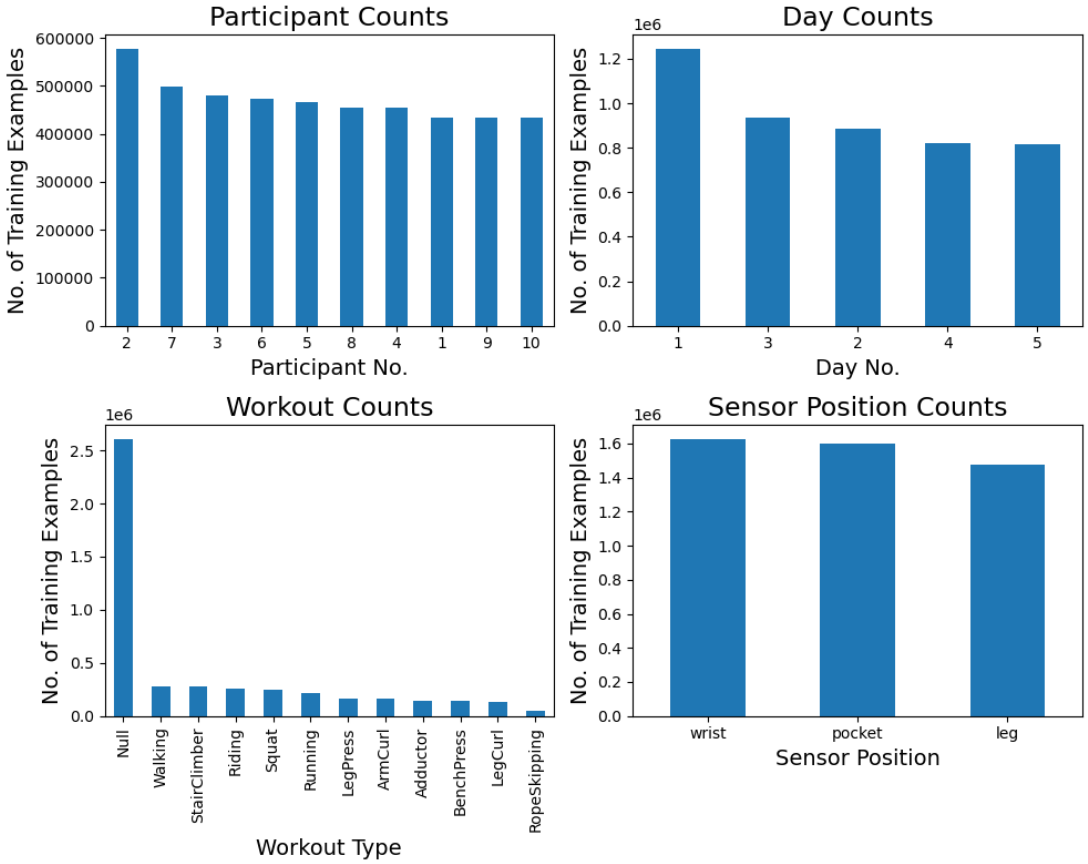


Figure 3.5: Bar graphs displaying the distribution of different categories in the dataset.

Class imbalance can be problematic when training models as it can lead to biased training towards the majority class. In addition, evaluation metrics that are commonly used for classification tasks, such as accuracy, can be misleading. For example, if 80% of the instances belong to class X and 20% belong to the other classes (Y and Z), a model that predicts class X 100% can achieve a high overall accuracy of 80%. However, it will then fail to predict the other minority classes.

For these reasons class imbalance was addressed as part of the preprocessing step before training the model. The class imbalancing techniques were carefully selected as the data was sequential and some of these techniques can damage the sequential nature and render the data meaningless.

Imbalance in the other categories did not affect model training as they were not the class labels. However, they did provide some valuable insights. Firstly, it was good to know that the total number of instances for each sensor position was similar. This means the sensors could be compared fairly. In addition, the number of instances for each participant was similar. So, if the data was split by participants, it would provide enough representative data for training and testing.

3.3.3 Correlations Analysis

Exploring the correlations between features is important as multicollinearity can lead to unstable and unreliable predictions. In addition, it can provide insights into patterns and trends that facilitate better interpretation of the results of the final model. Figure 3.6 displays the overall correlations of each feature. From this it was deduced that there was no multicollinearity between the features and that they would each provide strong individual effects that would contribute towards the model performance. However, Figure 3.7 told a slightly different story.

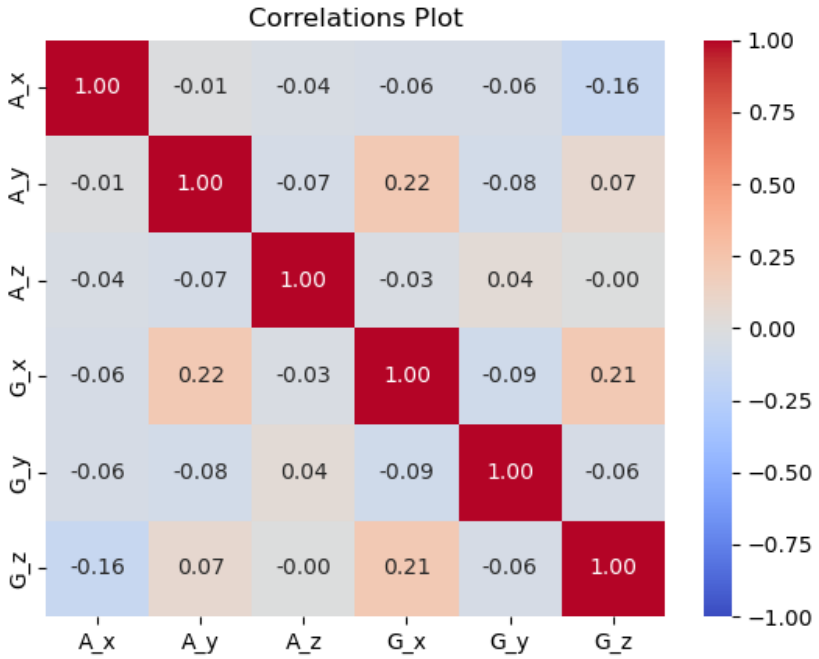


Figure 3.6: Correlations plot of features in entire dataset.

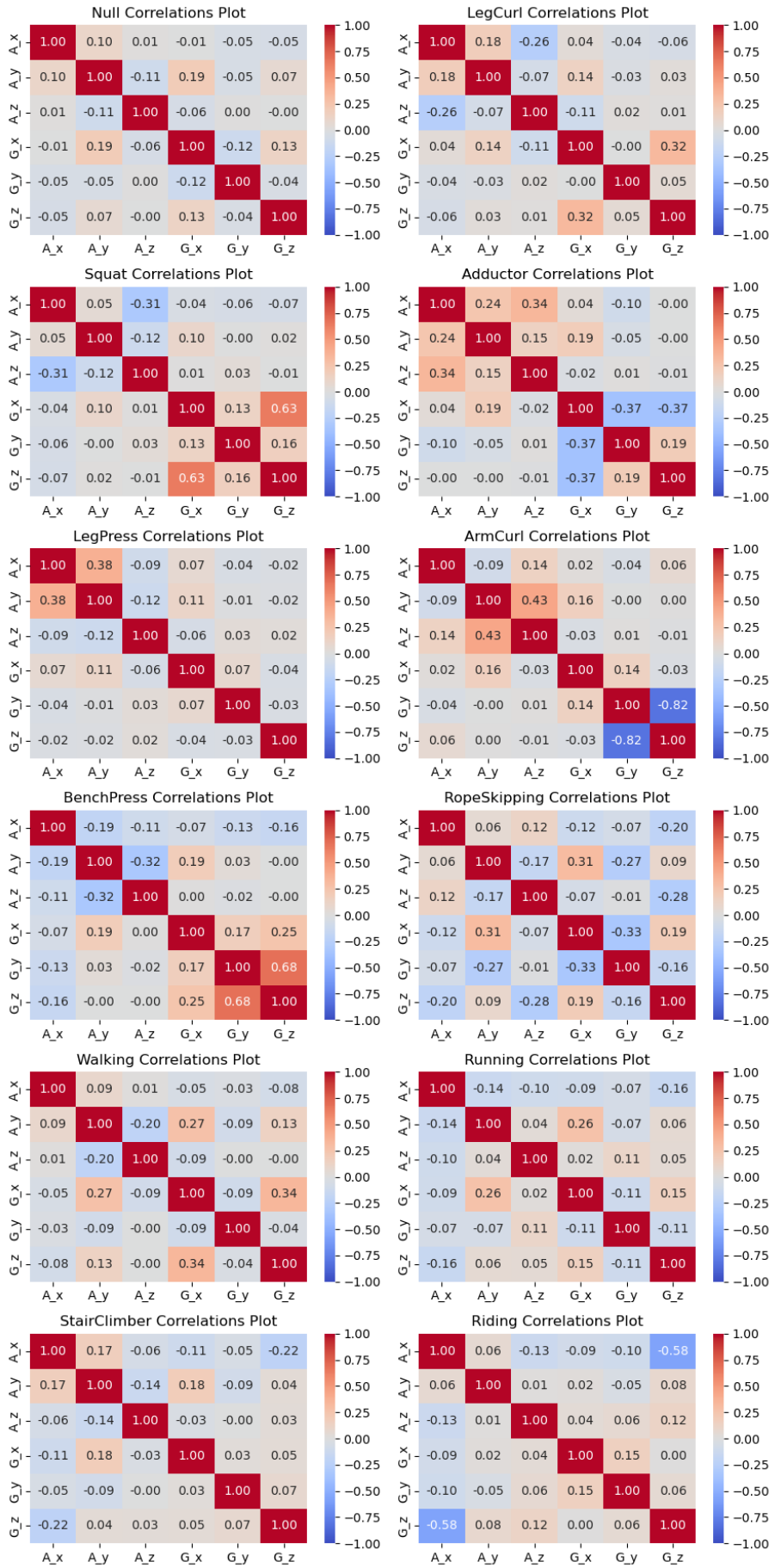


Figure 3.7: Correlations plot of features for each separate workout.

When broken down into each individual workout class the features had different correlations. For the most part the features were not correlated however, a few workout classes exhibited high correlation between features. These were G_x and G_z for Squat, G_y and G_z for ArmCurl, G_y and G_z for BenchPress and A_x and A_z for Riding. This indicated that for these workout classes those features may have be introducing redundant information that may have lead to overfitting. Comparing test set accuracy with training set accuracy is key in assessing if this was the case. Therefore, particular attention was paid towards those workout classes as they may have affected model performance. Nonetheless, for the most part the features were not highly correlated and therefore model performance was likely not affected by these specific high correlations.

3.3.4 Time Series Analysis

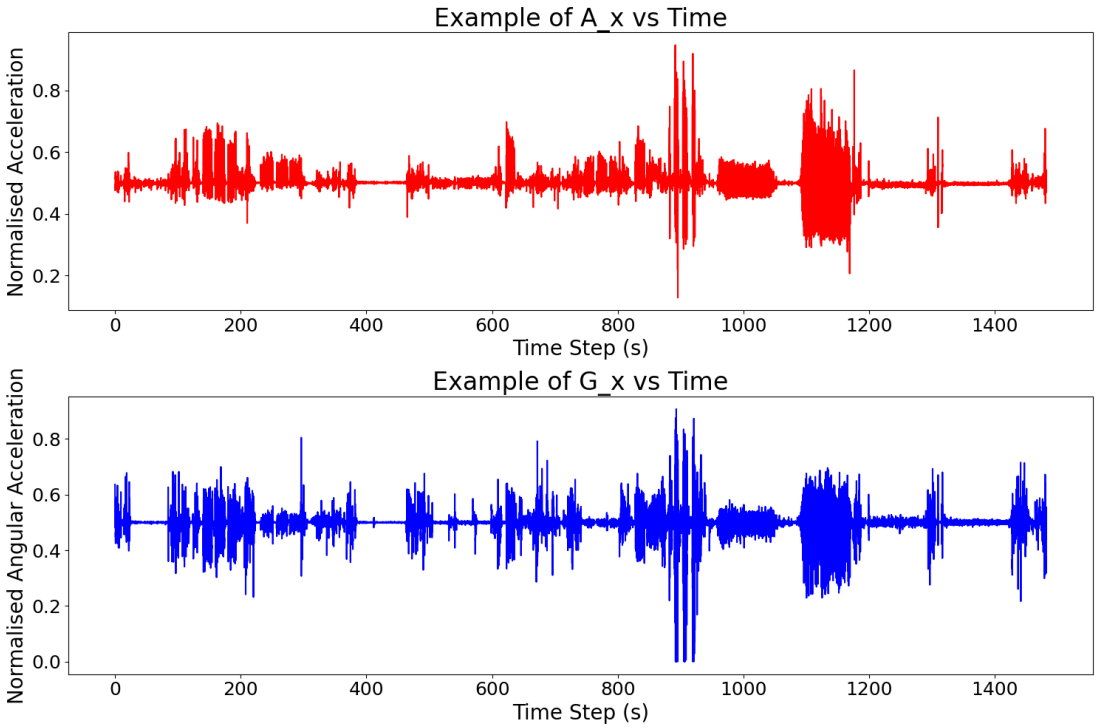


Figure 3.8: Time series graphs of participant 8 on day 4 for the wrist IMU.

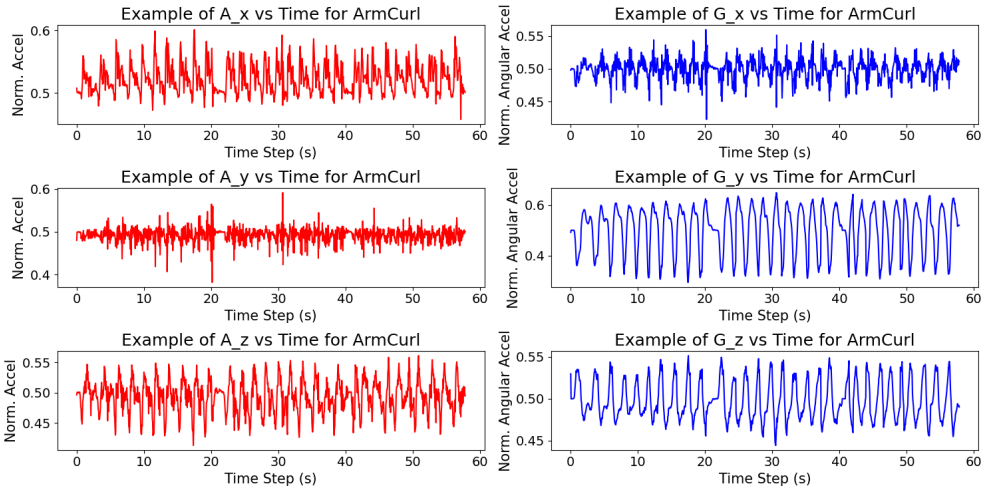
This thesis involved sequential acceleration and angular acceleration data. Therefore, it was important to visualise these on a time-series graph. The patterns observed can uncover outliers that could potentially negatively impact the performance of the models. In addition, any trends identified through this analysis can help to explain the results observed from the models. Figure 3.8 shows an example of the data collected from one session. It displays the acceleration and angular acceleration data in the x direction for participant 8 on day 4 for

the wrist IMU. Different workouts can be seen from different groups of signals. It can also be seen that there were some similarities and differences between the two measures. For example, in some areas the signal follows a similar pattern for the acceleration and angular acceleration. Whilst in other areas the signal for one measure is very different from the other. This only provides a general view of the signal trends for the features and a detailed close-up view was therefore necessary.

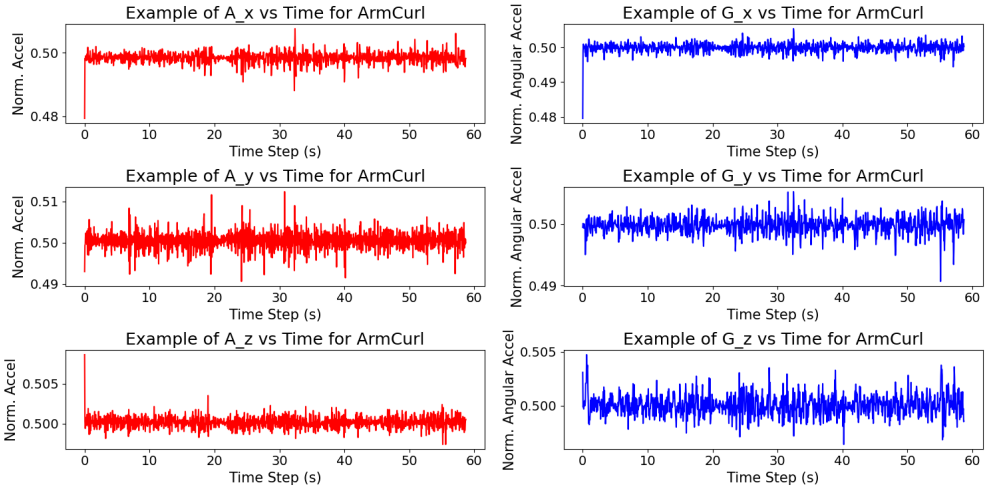
Figure 3.9 displays a close-up view of the data collected from participant 8 on day 4 for the ArmCurl workout. In Figure 3.9a there are clear periodic peaks and troughs. This is especially visible for the G_y and G_z plots, as 30 peaks and 30 troughs can be counted for each of them, with a slight gap in-between each set of 10. This clear periodic signal was due to the IMU being attached to the wrist which moves periodically in an ArmCurl workout. This suggests that IMUs attached to the wrist are likely to facilitate the recognition of workouts that use the arms. In contrast, Figures 3.9b and 3.9c, do not display a clear periodic signal. This is because the IMUs were attached to the pocket and leg which were not involved in the ArmCurl workout. When observing the y values on the plots (normalised acceleration and normalised angular acceleration) it can be deduced that Figures 3.9b and 3.9c mainly display noise. There are outliers at the start of a few plots for Figures 3.9b and 3.9c. However, the absolute deviation of these outliers is not as large as the periodic signals seen in Figure 3.9a, so they were ignored.

Figure 3.10 displays a close-up view of the data collected from participant 8 on day 4 for the Running workout. In contrast to the figures displaying the ArmCurl workout, these figures all display a high frequency signal. This accurately represents running as it is a whole-body movement that involves a high frequency of strides.

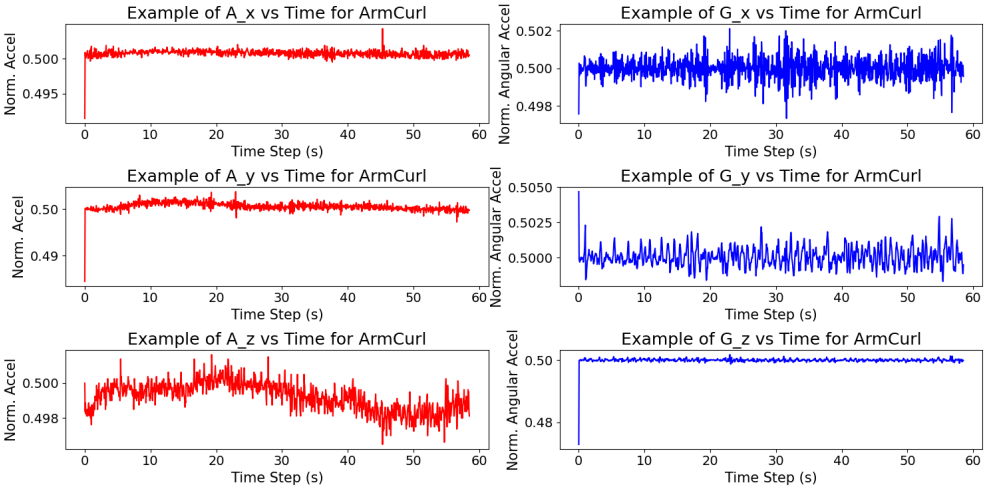
Together Figures 3.9 and 3.10 show how the position of the IMU affects the signal observed within the same exercise and between different exercises. Although these differences are clearly visible to the human eye, DL models construct features that cannot be interpreted by humans. Therefore, the models to be constructed may find patterns in the sequential data that enable them to recognise activities from sensors that are attached to limbs that are not involved in the primary movements of the workouts (e.g, the legs during an ArmCurl workout). Consequently, this time-series signal analysis provided an insight into what the results of the models could look like but they were taken with caution.



(a) Wrist IMU.

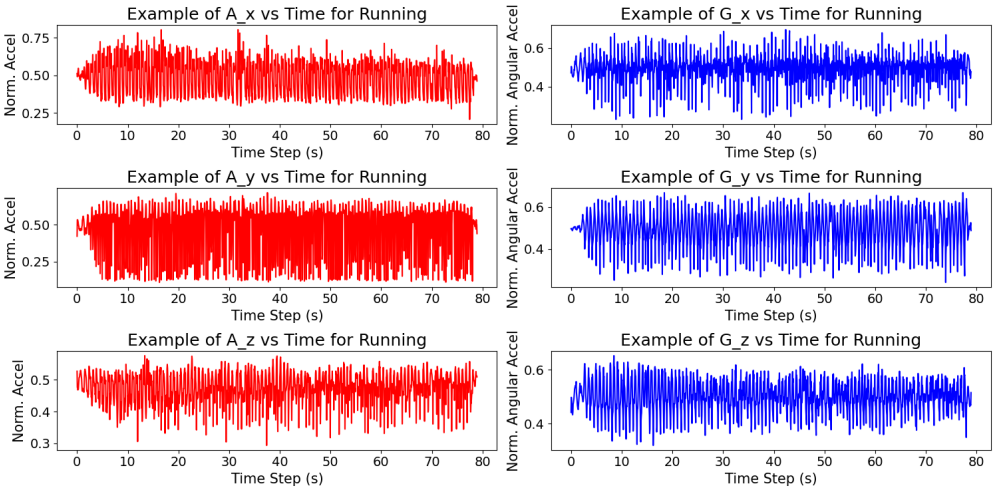


(b) Pocket IMU.

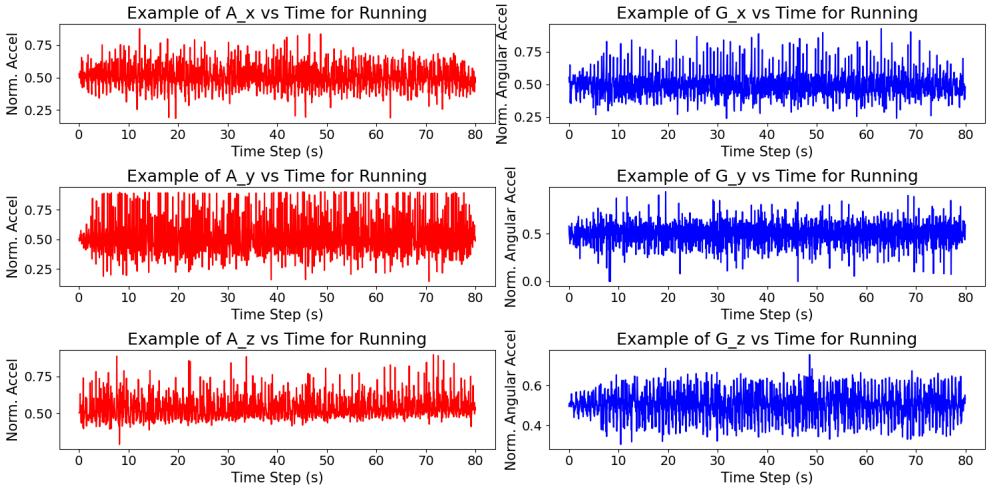


(c) Leg IMU.

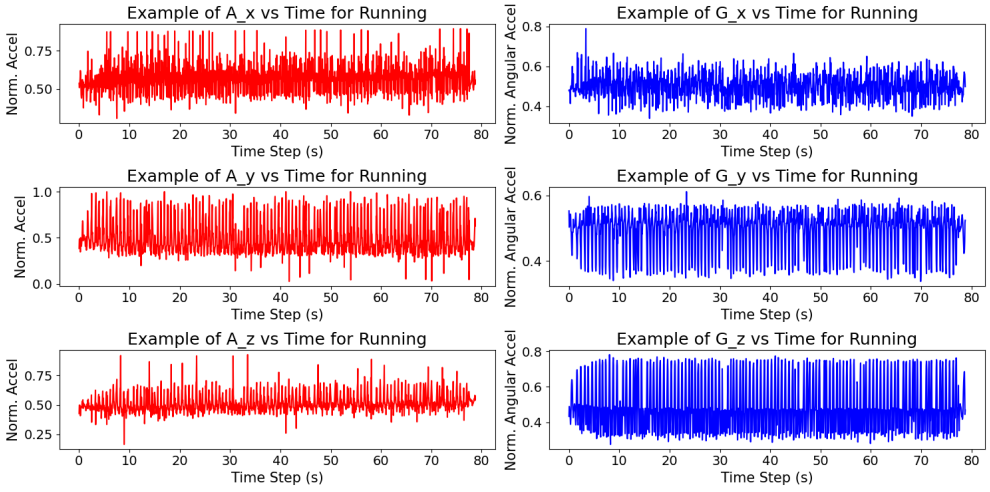
Figure 3.9: Time series graphs of participant 8 on day 4 for the ArmCurl workout.



(a) Wrist IMU.



(b) Pocket IMU.



(c) Leg IMU.

Figure 3.10: Time series graphs of participant 8 on day 4 for the Running workout.

3.4 Hardware

To collect the data the researchers integrated two sensing devices (IMU and HBC) onto a small, printed circuit board (PCB). The PCB was powered using a 3.7V lithium battery. For the purposes of this study the HBC aspect of the dataset was ignored. The data was collected at a frequency of 20 Hz. This is reflective of other studies in the literature that have collected IMU data between 20 and 50 Hz [4]. This had implications for the segmentation of the data (i.e., the window size of the sequences and the percentage they overlapped) when building the LSTM models.

3.5 Data Preprocessing

As discovered in the EDA, the dataset had already been normalised. This meant that there were fewer preprocessing steps required. Before the models were trained, two key preprocessing steps were performed. Firstly, the Null class was randomly under sampled to the same number of samples as the next highest class count (Walking). This is because the Null class was identified to have a significantly greater count that the other classes. In addition, the Null class did not follow any particular sequential pattern so it was deemed fine to under sample. This ensured a more even spread of class counts (Figure 3.11) compared to the original dataset (Figure 3.5) and therefore reduced the bias towards the Null class.

Nonetheless, there was still a large difference between the lowest and highest class counts (Walking and RopeSkipping) (Figure 3.11). This meant that there was still bias present towards the classes with higher counts. This was taken into consideration when analysing the results of the models. No further class balancing techniques were performed on the other classes as randomly under or over sampling will have damaged the sequential nature of the data.

The second preprocessing step performed was encoding the class labels into categorical numbers. This was completed using the .LabelEncoder() method from the Sci-kit Learn library. Encoding the classes is a crucial step as DL models require the classes to be in a numerical form so that the model can run efficiently. However, training a model on encoded data in this way may result in weaker performance. This is because the encoded labels in this initial encoding step were from 0 – 11. This implies that there was an order to these categories and the model may have therefore tried to learn an ordered relationship in the categories when they did not have this relationship.

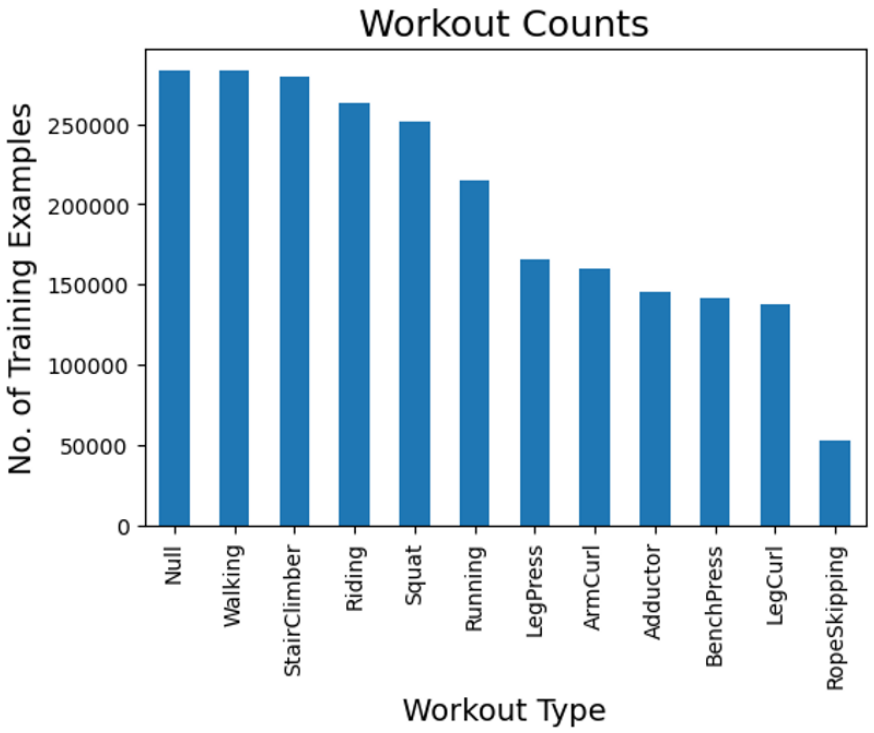


Figure 3.11: Bar graph displaying the distribution of different categories in the dataset after randomly under sampling the Null class.

To address this problem, a second encoding step called one-hot encoding was required. This step was performed after the data was split and segmented. This essentially created a new column for each class (in this case 12 columns). A row was classified by having the value of 1 in only one of the 12 class columns (indicating which class it was), whilst the rest of the class columns were 0.

3.6 Model Design

After completing the EDA to a satisfactory level, the model architecture was designed, and a model optimisation strategy was implemented. As highlighted in Chapter 1, a separate LSTM model was built for each experiment. Each experiment used different input data to train the model but produced the same output (the combined accuracy at classifying each of the workouts and the loss). Each model had the same task of classifying the workouts as accurately as possible.

This section of the thesis is divided into the following parts: data splitting, sequence preparation, hyperparameter tuning, regularisation, evaluation and experiments. To ensure optimal performance of the final models each of the parts was careful implemented.

3.6.1 Data Splitting

Before training an ML model on data, the most important step is to divide the data into partitions for training and testing. Crucially, partitioning the data ensures that a section of the data is left out during training, in the aim to avoiding overfitting. Overfitting is when the model performs exceedingly well on the dataset it was given, however when a new dataset (that the model has not seen before) is introduced it performs very badly. Additionally, splitting the dataset into relevant partitions allows the model to be evaluated effectively. It is essential to evaluate the performance of the model on completely unseen data to obtain an unbiased estimate of its generalisation ability. The testing set acts as this unseen data for assessing the model's performance in real-world scenarios.

For this thesis the data was split into training, validation and testing sets using the holdout method. The holdout method is typically used for its simplicity and speed and is advantageous when the dataset is very large and when training is very slow. Additionally, in ML, splitting the data into training, validation and testing sets is best practice. The training set contained the data from participants 1 to 7, the validation set contained the data from participant 8 and the testing set contained the data from participants 9 and 10. Usually instances are randomly placed into each partition and split by a certain percentage (e.g., 70%/10%/20%). However, this could not be done using this dataset as it was sequential, and the order of each subsequent data point was therefore of high importance.

Another method that is typically used when splitting the data is cross validation, particularly k-fold cross validation. This method involves dividing the dataset into k partitions where one partition is used for testing and the rest are used for training. The process is repeated k times where each partition is left out to be used as the testing set once. The performance of the model is then taken as the average of the k number of models (usually k is 10). This method produces more reliable results when compared to the holdout method, however it requires more processing time. Consequently, this method was not chosen as this thesis involved training multiple DL models using a large dataset and slow training.

3.6.2 Sequence Preparation

The next step in preparing the data for training was specific to the LSTM network. This step involved dividing the dataset again, however this time into smaller sequences or windows. This step was necessary as the full-length accelerometer or gyroscope signal could not be used to classify the activity as it was too long, and

it therefore could not capture the characteristics of the activity. In addition, the LSTM network requires separate sequences for it to work and to create its long- and short-term memories. When deciding on how to divide the data, the window size and the percentage of window overlap are both very important parameters as they greatly influence the performance of the model. For example, selecting window sizes that are too short or too long may significantly decrease model performance and therefore these parameters must be selected/tuned carefully. In addition, having a percentage overlap is useful for continuous activity data as the overlap guarantees that the succeeding windows carry some information from the preceding windows.

In order to select an optimal window size and percentage overlap the following values were tested/tuned: window sizes of 2.5s, 5s, 7.5s and 10s and percentage overlaps of 0%, 25%, 50% and 75%. This tuning was performed on the model using all three sensors (wrist, pocket and leg). This resulted in an optimal window size of 2.5s and an optimal percentage overlap of 50%. This agrees with typical values used for these parameters in the literature [4]. This optimal window size and percentage overlap were then assumed to also be optimal for the individual sensors, so they were carried over and used for the models built using the data from each individual sensor.

3.6.3 Hyperparameter Tuning

In ML, hyperparameter tuning is a fundamental step to achieving high performing models. Each ML model has multiple hyperparameters that control various aspects of the training process and selecting the correct hyperparameters is crucial to ensure efficient resource utilisation and to avoid overfitting and underfitting. For classical ML, hyperparameters are often tuned using grid searches, however for DL this is often not possible because of the high computational demands. For this reason, in this thesis, the hyperparameters were systematically tuned using trial and error. Importantly, each parameter was tuned on its own, whilst the other parameters were kept the same. This ensured that any changes observed in model performance could be directly associated with the corresponding change in that particular hyperparameter. Additionally, during the hyperparameter tuning process, a random seed was set so that models using the same hyperparameters could be reproduced.

The following hyperparameters were kept the same throughout the training process as they were widely used in previous literature discussed in Chapter 2 [4]:

- loss function = categorical cross entropy
- optimizer = Adam
- $B_1 = 0.9$
- $B_2 = 0.999$
- output layer = 12 hidden units + softmax function
- Activation function for the LSTM layers = ReLU

Each model was tuned in the same systematic order. Initially the following hyperparameters were set:

- learning rate (alpha) = 0.001
- LSTM layer number = 1
- No. hidden units = 16
- Mini-batch size = 128

Subsequently, the learning rate (alpha) was tuned for the values 0.01, 0.001 and 0.0001. Then the mini-batch size was tuned for the values 32, 64, 128, 256 and 512. The layer number was then tuned between 1 to 5 LSTM layers. Finally, the number of hidden units were tuned for the values 16, 32, 64, 128 and 256. The values tuned for the mini-batch size and the number of hidden units were all exponentials of 2 because this facilitated a more efficient use of computer memory. Each LSTM layer contained the same number of hidden units and used the ReLU activation function. Finally the model was tuned for 1 or 2 dense layers (positioned after the LSTM layers) containing each of the same hidden unit values (16, 32, 64 and 128).

3.6.4 Regularisation

Regularisation steps are often used in DL model training in order to prevent overfitting and improve the model's ability to perform well on unseen data. The key regularisation technique that was used in this thesis was early stopping. Early stopping, as the name suggests, stops the training early when the model reaches a certain threshold. Early stopping also reduces the time required for training and therefore it was implemented during the hyperparameter tuning process.

When tuning the models, the maximum epoch number was set to 300, the early stopping patience was set to 15 and the early stopping evaluation metric was set to validation loss. This essentially means that if the validation loss did not decrease for 15 epochs the model would stop training. The validation loss was chosen as the evaluation metric for early stopping because in the literature it has been noted that the error in the validation dataset can be used as a proxy for the generalisation error [92]. Therefore, this error in the validation dataset gave a good indication of when overfitting had begun.

To be able to access the results of models again, without having to retrain them, each model was saved using model checkpoints. The model checkpoints were set at each epoch and the evaluation metric set was the model accuracy. This means that each model was saved after each epoch only if the accuracy score had increased from the previous highest accuracy score. Each model was initially saved as a .h5 file and when a higher accuracy score was reached this file was overwritten.

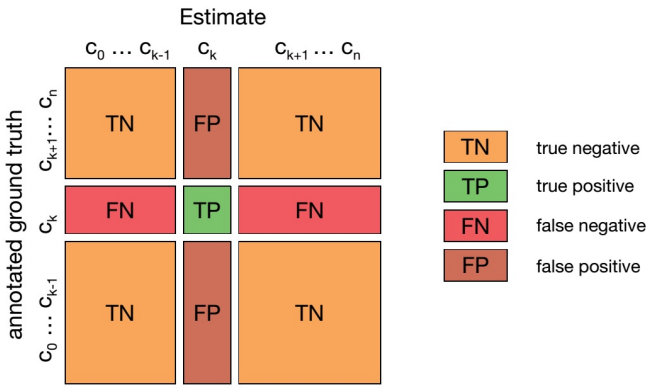


Figure 3.12: Example confusion matrix for a multi-classification with n classes, highlighting that there are multiple TN, FP and FN values for every TP value. Source [3]

3.6.5 Evaluation

After training and tuning the models on the training and validation datasets to reach optimal model performance, the models were evaluated using the testing datasets that were held out. In order to effectively evaluate whether these trained models performed well on unseen data, appropriate evaluation metrics were used. As mentioned previously, the aim of these models was to optimise for the combined accuracy at classifying all of the workouts. Accuracy is simply the proportion of correct predictions out of all the data instances. However, as alluded to in the EDA, the dataset was imbalanced which means that accuracy could have been

a misleading evaluation metric on its own. Therefore, the following evaluation metrics were also used: precision, recall and F1 score. The definitions and relevant abbreviations of each evaluation metric used can be found in Table 3.1. In addition to these evaluation metrics, for multi-classification tasks it is typical to use a confusion matrix to visualise and evaluate the model's performance for each class (Figure 3.12).

Table 3.1: Definitions of evaluation metrics

Evaluation Metric	**Definition**
True Positive (TP)	The number of instances where the model correctly predicts the positive class
True Negative (TN)	The number of instances where the model correctly predicts the negative class
False Positive (FP)	The number of instances where the model incorrectly predicts the positive class
False Negative (FN)	The number of instances where the model incorrectly predicts the negative class
Precision	The proportion of TPs out of the total number of positively predicted instances. It measures the model's ability to avoid false positives
Recall	The proportion of TPs out of the total number of correctly predicted instances. It measures the model's ability to avoid false negatives
F1 Score	The harmonic mean of Precision and Recall

The below equations show how each evaluation metric is calculated:

$$accuracy = \frac{TP + TN}{TP + FP + TB + FN} \tag{3.2}$$

$$precision = \frac{TP}{TP + FP} \tag{3.3}$$

$$recall = \frac{TP}{TP + TN} \tag{3.4}$$

$$f1\ score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{3.5}$$

3.6.6 Experiments

To address the aims and objectives, four experiments were conducted. Each experiment essentially consisted of training an LSTM model using a different section of the dataset.

1. Wrist sensor alone
2. Pocket sensor alone
3. Leg sensor alone
4. All three sensors (wrist, pocket and leg)

Chapter 4

Results and Discussion

This chapter displays the results of each experiment in this thesis using the methods described in Chapter 3. For each experiment the results are displayed using tables and figures. The parameters used, final accuracy and loss and testing dataset results are displayed in tables. The training progress and the confusion matrices are displayed using figures. This chapter also discusses the results in relation to the literature and the two key objectives set out in Chapter 1.

4.1 Experiment 1: Wrist IMU Data

The first experiment only used wrist IMU data to build an LSTM model. Table 4.1 shows the optimal hyperparameters that were chosen after tuning this model. Figure 4.1 shows the training progress over time for every epoch. Table 4.2 shows the final accuracy and loss numbers for each split of data. The test dataset was then used to produce two confusion matrices, one regular version (Figure 4.2) and one normalised version (Figure 4.3), and the evaluation metrics (Table 4.3). These tables and figures were repeated for each experiment.

Table 4.1: Hyperparameters for Experiment 1.

Hyperparameter	**Value**
Learning Rate	0.0001
Number of Layers	2 LSTM
Hidden Units Per Layer	32
Batch Size	32
Number of Epochs	190* (early stopping)

From Table 4.1 it can be seen that this model was made up of only two LSTM layers of 32 hidden units each. This is in fact quite a shallow DL model. This means that the model was lightweight, which suggests that wrist IMU strength exercise based HAR may be practical to implement into smart watches. This was a positive finding as it reflects the future need of lighter weight models in the literature [4, 93]. However, this model was trained offline and for effective real-time HAR the model must be tested with live data.

Figure 4.1 displays the training loss and accuracy curve for the training of this LSTM model. It follows the typical pattern of training accuracy, increasing rapidly at the start and then plateauing as it gets closer to optimal performance. This is the same for training loss, but it decreases quickly then plateaus. The validation accuracy and loss curves follow a similar overall pattern however they are much noisier. This is typical as they are always based on a smaller dataset. A model demonstrating these patterns is usually a high performing model.

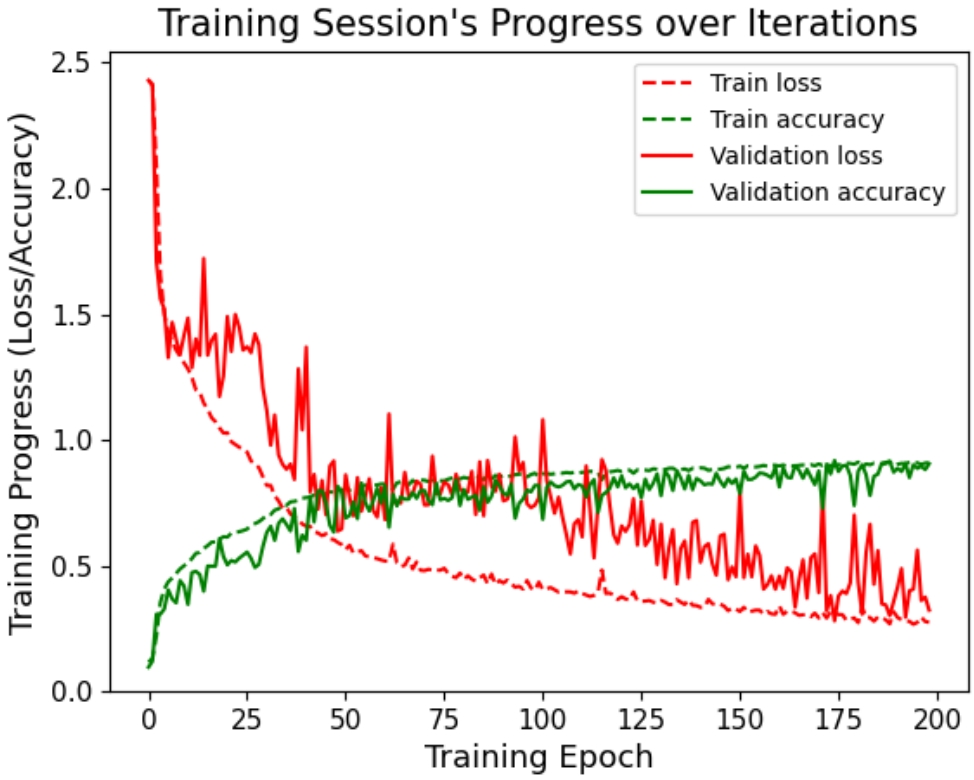


Figure 4.1: Accuracy and loss over epochs for Experiment 1.

Table 4.2 shows great final accuracy and loss results. The overall accuracy for training, validation and testing was 0.91, 0.92 and 0.91 respectively. This suggests that the model was not only trained well but that it generalises well on unseen data. The high testing accuracy of 0.91 means that the model is likely to perform

well on unseen data that is collected in a similar fashion. The accuracy score of 91% is also similar to a study conducted by Bayat et al [27]. They also achieved an accuracy of around 91% with 1 IMU sensor. Overall this suggests that the wrist IMU alone may be suitable for strength exercise based HAR. Hence, a more detailed analysis was performed to verify this.

Table 4.2: Accuracy and loss for Experiment 1.

Data Split	**Accuracy**	**Loss**
Training	0.91	0.27
Validation	0.92	0.30
Testing	0.91	0.39

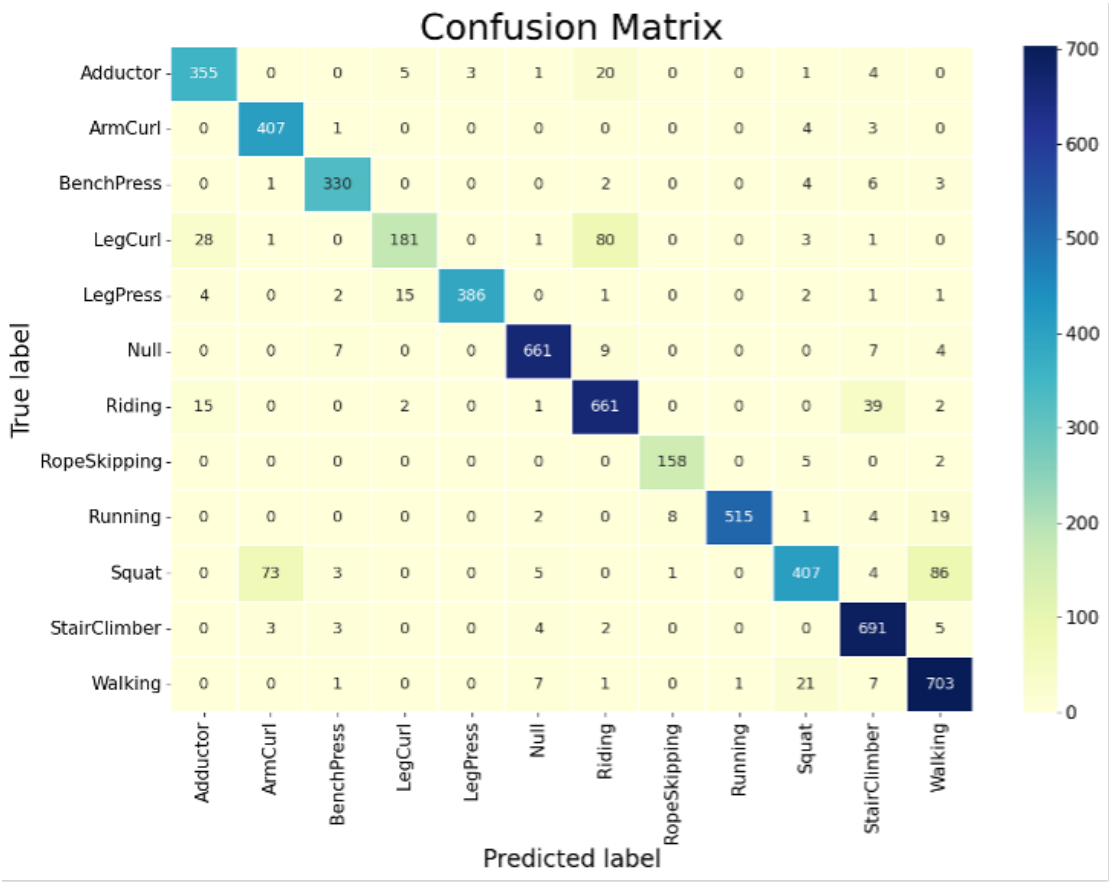


Figure 4.2: Confusion matrix of Experiment 1.

Figure 4.2 shows that the model had the highest number of TP's for Walking, Stairclimber, Riding and the Null class. However, because the classes were imbalanced this was not a fair representation of the model's performance. Figure 4.3 gives a fairer comparison of each class as the confusion matrix has been normalised. For this reason, this chapter focused on the normalised confusion matrices.

Specifically these normalised confusion matrices calculate the recall score for each workout class. Figure 4.3 shows that 10 out of the 12 classes had recall scores of above 0.9 for this model (all of them except LegCurl and Squat). The other 2 classes were LegCurl and Squat and had recall scores of 0.61 and 0.7 respectively. Surprisingly, this model was the highest performing model out of the 4 experiments.

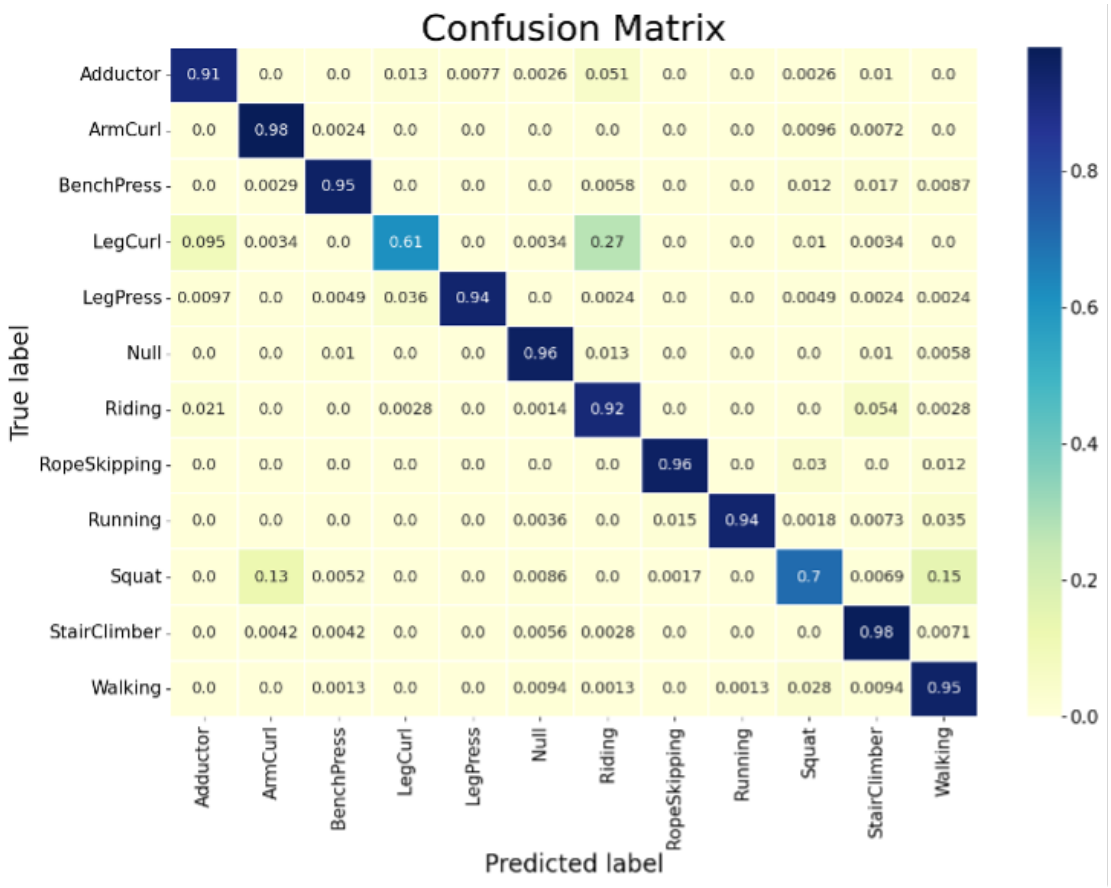


Figure 4.3: Normalised confusion matrix of Experiment 1.

A likely reason why LegCurl and Squat had lower recall scores could be due to the wrist being stationary during the movement or the wrist not moving a great deal (Figure 3.1). Therefore, the acceleration signals would have just been noise, as demonstrated in the EDA (Figure 3.9), so harder for the algorithm to detect a movement pattern. However, for Adductor, LegPress and Riding the wrist was also stationary or had very little movement (Figure 3.1), and these had good recall scores (Figure 4.3). This suggests that a wrist IMU on its own may be good enough for effective strength exercise HAR. This mirrors findings by Mannini et al [44], Valarezo et al [49] and Depari et al [85]. Therefore further tuning of the model, using more granular hyperparameter alterations, may have resulted in a higher performing model. However, this likely creates deeper models which require more processing and as a result may not be practical to implement in real-time.

A reason why LegCurl had the lowest recall score could be that it had the second lowest class count (Figure 3.11). ML models will always be biased towards classes with a higher number of instances. However, Adductor had the lowest class count, a high recall score (0.91) and it was mostly stationary during movement. Consequently, the poor performance at recognising LegCurl in this model was likely due to the lack of wrist movement when performing the LegCurl not the model's bias.

Lastly, Table 4.3 displays the precision, recall and F1 scores for each workout in Experiment 1. It can be seen that all the F1 scores were above or close to 0.9 for all workouts except LegCurl (0.73) and Squat (0.79). On closer inspection this was due to lower recall scores (0.61 and 0.70) compared to precision scores (0.89 and 0.91). Therefore, for these two workout classes the model had a poorer ability to avoid false negatives. Meaning that for these two classes the model was incorrectly predicting the negative classes more often compared to the other classes.

Table 4.3: Testing data results of each workout for Experiment 1.

Workout	**Precision**	**Recall**	**F1 Score**
Adductor	0.88	0.91	0.90
ArmCurl	0.84	0.98	0.90
BenchPress	0.95	0.95	0.95
LegCurl	0.89	0.61	0.73
LegPress	0.99	0.94	0.96
Null	0.97	0.96	0.96
Riding	0.85	0.92	0.88
RopeSkipping	0.95	0.96	0.95
Running	1.00	0.94	0.97
Squat	0.91	0.70	0.79
StairClimber	0.90	0.98	0.94
Walking	0.85	0.95	0.90

4.2 Experiment 2: Pocket IMU Data

Experiment 2 only used pocket IMU data to build an LSTM model. Table 4.4 displays the optimal hyperparameters chosen to train the model. The model contained 3 LSTM layers with 64 hidden units in each layer. This was a slightly deeper model that required more processing power compared to Experiment 1.

Table 4.4: Hyperparameters for Experiment 2.

Hyperparameter	**Value**
Learning Rate	0.0001
Number of Layers	3 LSTM
Hidden Units Per Layer	64
Batch Size	64
Number of Epochs	237* (early stopping)

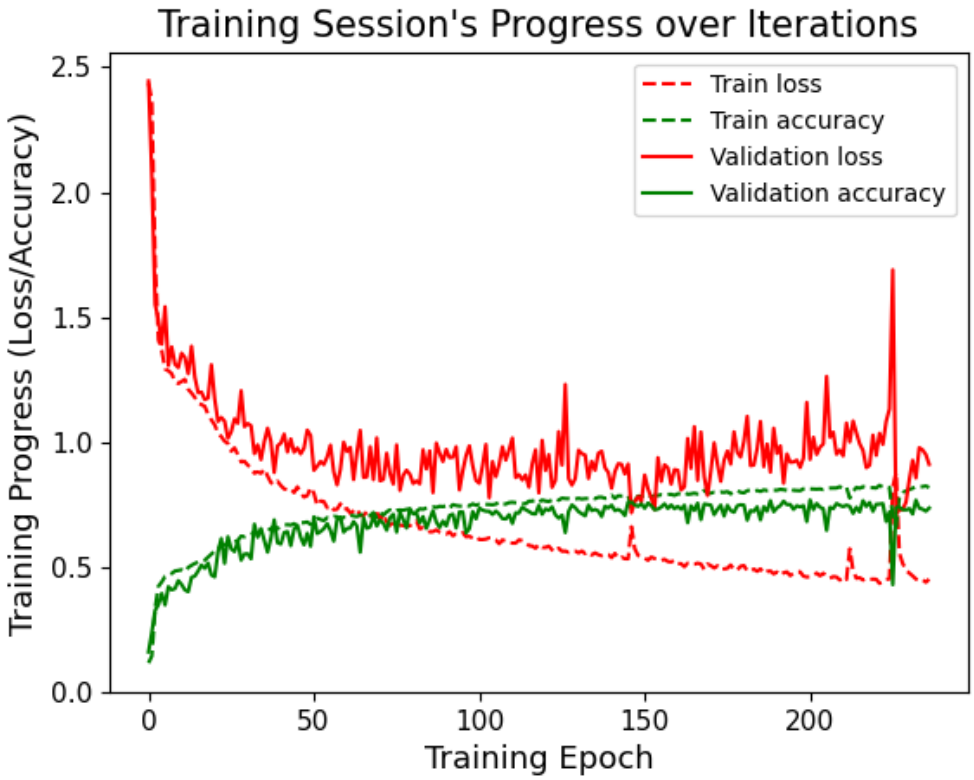


Figure 4.4: Accuracy and loss over epochs for Experiment 2.

Figure 4.4 displays the training accuracy and loss curves for Experiment 2. The train accuracy and loss curves follow the characteristic training curve for DL

models however the validation accuracy and loss curves do not. The validation accuracy and loss seem to plateau much earlier than the train accuracy and loss. This is much more pronounced in the loss lines. This suggests that the model was beginning to become overfit to the training data. Thanks to early stopping this was cut short, preventing the model from overfitting and compromising the testing accuracy. This also suggests that this model was not as high performing as the validation accuracy and loss plateaued earlier than in other Experiments (1 and 3).

Table 4.5: Accuracy and loss for Experiment 2.

Data Split	**Accuracy**	**Loss**
Training	0.83	0.44
Validation	0.75	1.04
Testing	0.67	1.61

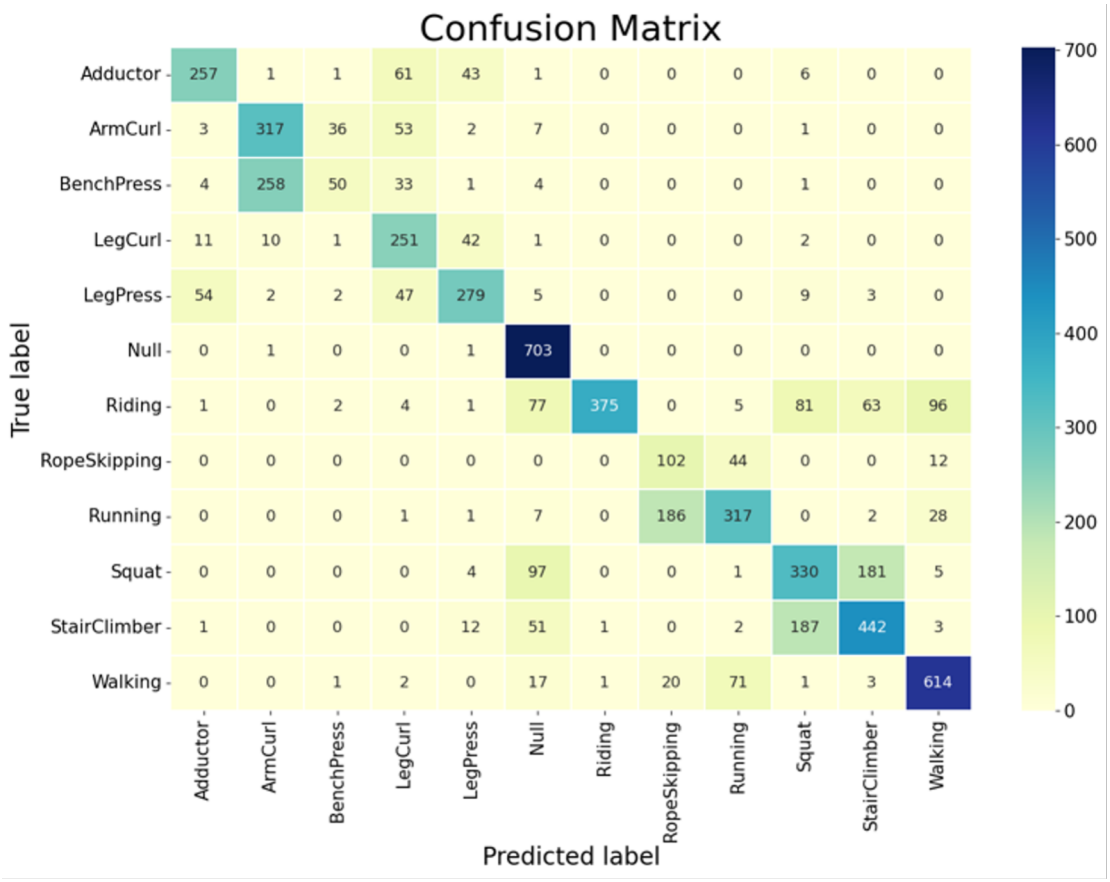


Figure 4.5: Confusion matrix of Experiment 2.

Table 4.5 shows the poor overall accuracy and loss for this model. The model was trained to an accuracy of 0.83 however this only resulted in a validation

accuracy of 0.75 and a testing accuracy of 0.67. These overall scores suggest that the model was not good at generalising on unseen data. However, accuracy scores alone are not always suitable evaluation metrics. As the data was biased towards other classes with higher counts, further analysis was performed and evaluated.

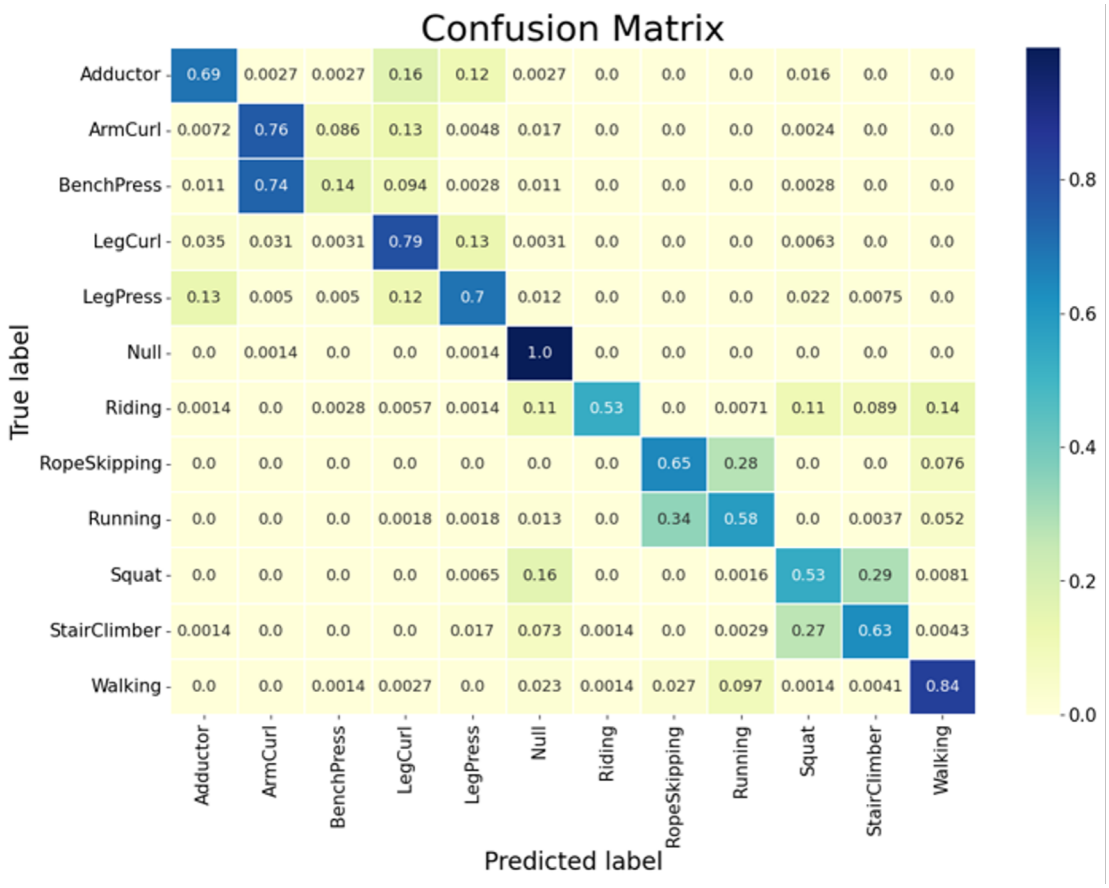


Figure 4.6: Normalised confusion matrix of Experiment 2.

Figure 4.6 highlights the recall scores of this poor performing model. Only the Null class had a recall score above 0.90 and the next 4 highest were Walking (0.84), LegCurl (0.79), ArmCurl (0.76) and LegPress (0.7). Notably BenchPress had a very low recall score of 0.14. Table 4.6 tells a similar story. The F1 scores ranged from 0.84 (Null) all the way down to 0.23 (BenchPress). Only four F1 scores were above 0.7 (Adductor, LegPress, Null and Walking).

A likely reason for the poor performance of this model is the placement of the IMU sensor. The pocket is close to the centre of the body and, although it is on the thigh, it is not involved in as much movement as the distal arms or legs. Therefore the acceleration signals recorded for this IMU were likely much fainter than the other two sensors as they were both attached to distal parts of the arms and legs (wrist and lower leg). Another reason for the poor performance of this model is that the pocket IMU sensor may not have been fixed in position. This

means the data collected may not have been suitable to build a DL model. On the one hand the IMU not being fixed in the pocket represents real-world scenarios. This is because when performing workouts people tend to have mobile phones in their pockets which are free to move and not fixed in position. However, as demonstrated by the results of this model, this creates much more variability in the data and therefore it is more difficult to build an accurate DL model with the data. As a result, an IMU positioned in the pocket may not be suitable for strength exercise based HAR.

Table 4.6: Testing data results of each workout for Experiment 2.

Workout	**Precision**	**Recall**	**F1 Score**
Adductor	0.78	0.69	0.73
ArmCurl	0.54	0.76	0.63
BenchPress	0.54	0.14	0.23
LegCurl	0.56	0.79	0.65
LegPress	0.72	0.70	0.71
Null	0.72	1.00	0.84
Riding	0.99	0.53	0.69
RopeSkipping	0.33	0.65	0.44
Running	0.72	0.58	0.65
Squat	0.53	0.53	0.53
StairClimber	0.64	0.63	0.63
Walking	0.81	0.84	0.83

4.3 Experiment 3: Leg IMU Data

The third experiment only used leg IMU data to build an LSTM model. Table 4.7 displays the tuned hyperparameters that produced the optimal model. It can be seen that there were 3 LSTM layers each of 32 hidden units. This is a model with less hidden units than Experiment 2 and the same batch size as Experiment 1.

Table 4.7: Hyperparameters for Experiment 3.

Hyperparameter	**Value**
Learning Rate	0.0001
Number of Layers	3 LSTM
Hidden Units Per Layer	32
Batch Size	32
Number of Epochs	114* (early stopping)

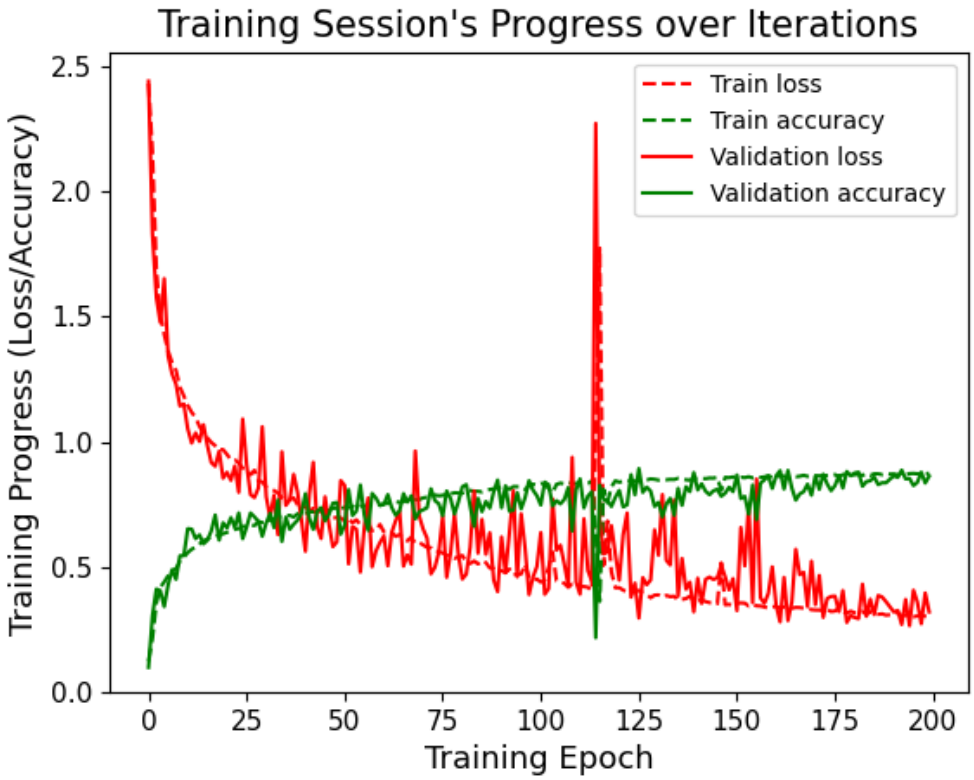


Figure 4.7: Accuracy and loss over epochs for Experiment 3.

Table 4.8: Accuracy and loss for Experiment 3.

Data Split	**Accuracy**	**Loss**
Training	0.91	0.21
Validation	0.87	0.34
Testing	0.87	0.37

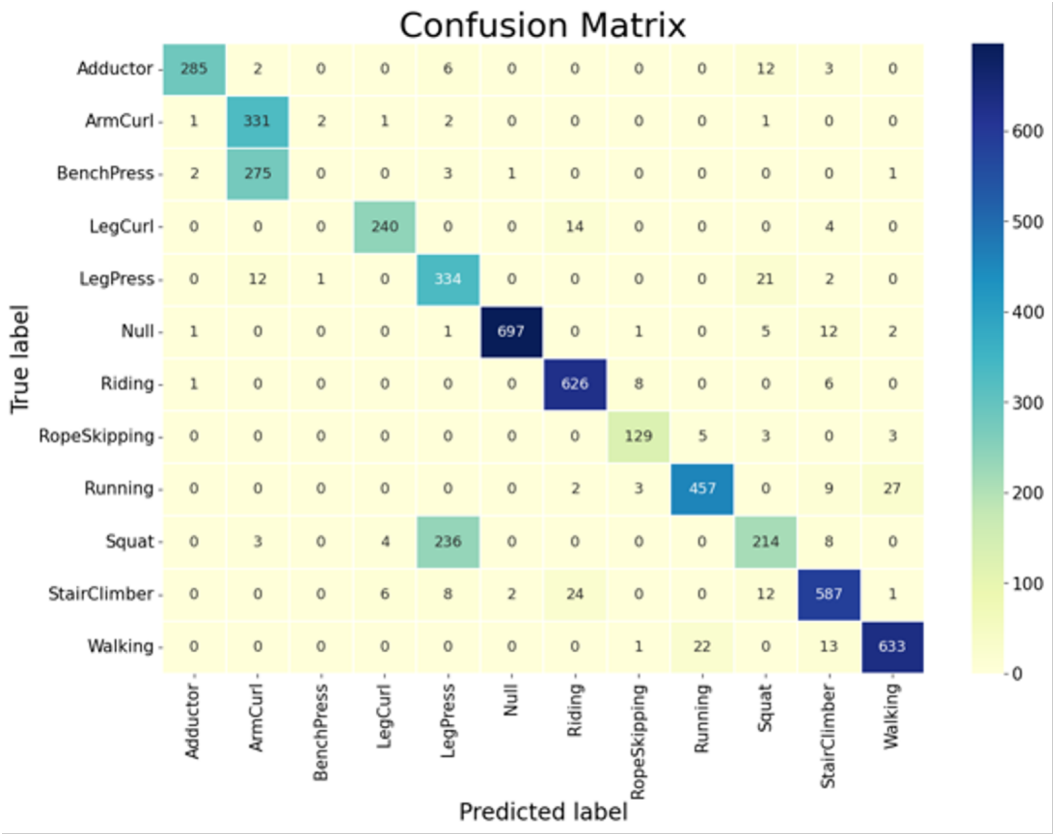


Figure 4.8: Confusion matrix of Experiment 3.

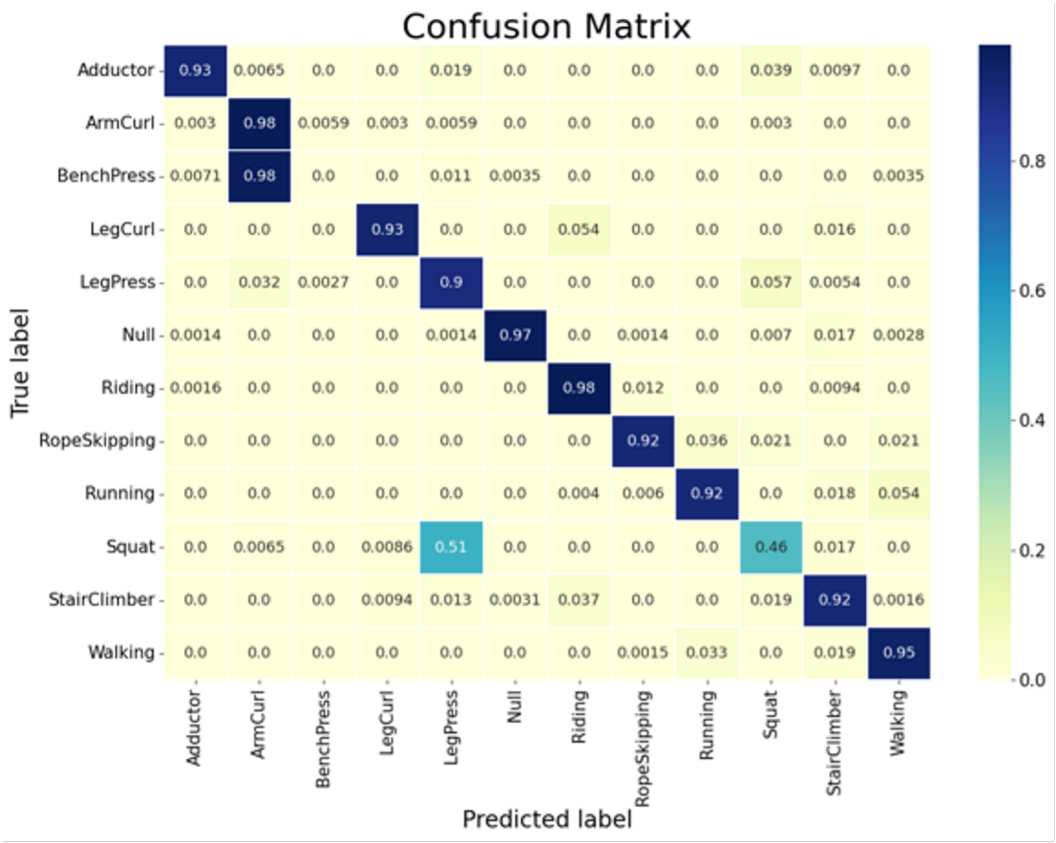


Figure 4.9: Normalised confusion matrix of Experiment 3.

Figure 4.7 displays the training loss and accuracy curves for this LSTM model. It follows a similar pattern to the model produced in Experiment 1. This is the typical pattern of training accuracy and training loss. The validation accuracy and loss also follow typical patterns that oscillate more due to the smaller dataset they were derived from. However it can also be seen that the validation accuracy and loss follow the training and accuracy and loss curves much more closely that in Experiment 1. This suggests that this is also a high performing model.

Table 4.8 shows great final accuracy and loss results. The overall accuracy for training, validation and testing was 0.91, 0.87 and 0.87 respectively. The validation and testing accuracy were only slightly lower compared to the results from Experiment 1. Therefore this model was also a high performing model and it generalises well on unseen data.

Echoing results from Figure 4.3, Figure 4.9 shows that 10 out of the 12 classes had recall scores of above 0.9 for this model. However, in this model the two low scoring classes were BenchPress and Squat. BenchPress had a recall score of 0.0 and Squat and had recall scores of 0.46. This model was the 2nd highest performing model out of the 4 experiments.

Table 4.9: Testing data results of each workout for Experiment 3.

Workout	**Precision**	**Recall**	**F1 Score**
Adductor	0.98	0.93	0.95
ArmCurl	0.53	0.98	0.69
BenchPress	0.00	0.00	0.00
LegCurl	0.96	0.93	0.94
LegPress	0.57	0.90	0.70
Null	1.00	0.97	0.98
Riding	0.94	0.98	0.96
RopeSkipping	0.91	0.92	0.91
Running	0.94	0.92	0.93
Squat	0.80	0.46	0.58
StairClimber	0.91	0.92	0.91
Walking	0.95	0.95	0.95

Interestingly in Table 4.9 all the F1 scores were above 0.9 except for ArmCurl (0.69), BenchPress (0.0), LegPress (0.70) and Squat (0.58). ArmCurl had a low F1 score due the low precision score it had (0.53). This can be explained by Figure 4.8, as it can be seen that most of the BenchPress workouts were being predicted as ArmCurl workouts. This is likely due to the leg positioning in both being very similar (Figure 3.1) and therefore it made it harder for the DL model to distinguish between the two. Similarly, LegPress had a low F1 score due to its low precision score of 0.57. Figure 4.8 shows that roughly half of the Squat workouts were predicted as LegPress workouts. Again the reason for this is that the movement of a Squat and a LegPress were very similar and it is not possible to infer the direction of motion from normalised acceleration signals.

4.4 Experiment 4: All 3 IMUs

The fourth experiment used data from all three sensors (wrist, pocket and leg) to build an LSTM model. Table 4.10 displays the optimal hyperparameters that were chosen after tuning the model. It follows a similar theme to the previous experiments with 2 layers and 64 hidden units per layers. It was one of the lighter weight models (it only has 2 layers). However it used 3 times the amount of data (compared to Experiment 1 which also had 2 layers), so it took longer to train.

Table 4.10: Hyperparameters for Experiment 4.

Hyperparameter	**Value**
Learning Rate	0.0001
Number of Layers	2 LSTM
Hidden Units Per Layer	64
Batch Size	32
Number of Epochs	229* (early stopping)

Figure 4.10 displays the training loss and accuracy curve for Experiment 4. Comparably to Experiments 1 to 3, the train accuracy and loss follows the characteristic training curve for DL models. Although this figure has a higher final accuracy and a lower final loss compared to Figure 4.4, it follows the same general pattern. This is that the validation loss plateaus, hinting the start of overfitting. This was again haulted by early stopping. It is likely that this model experienced this due it containing pocket IMU data. As the only two models that experienced the onset of overfitting contained pocket IMU data. This further reinforces the

notion that the data collected from the pocket IMU either was not reliable or was not suitable for strength exercise based HAR.

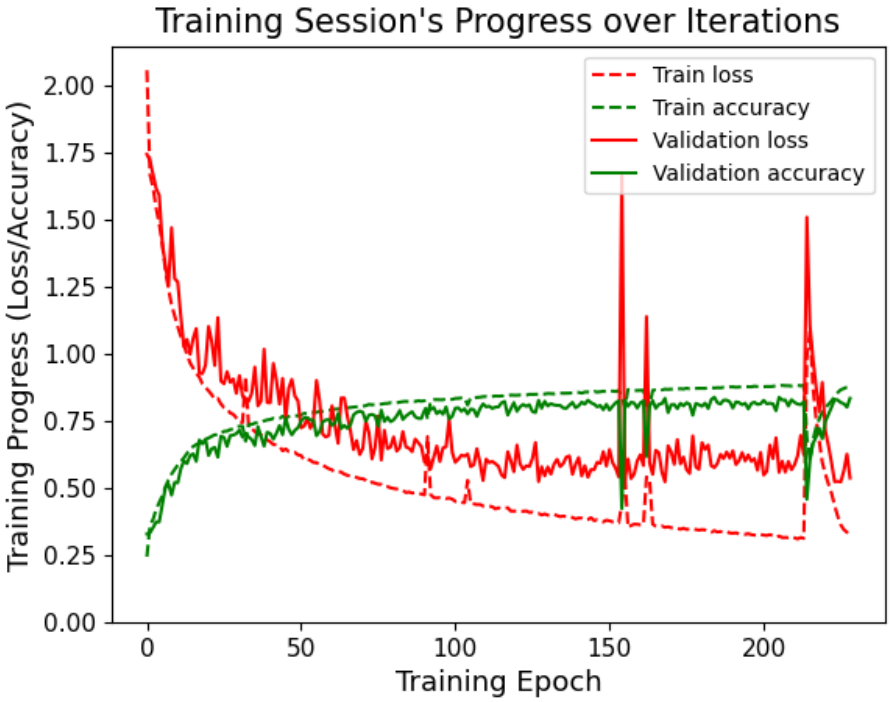


Figure 4.10: Accuracy and loss over epochs for Experiment 4.

Table 4.11 shows the final accuracy and loss results. The overall accuracy for training, validation and testing was 0.88, 0.80 and 0.79 respectively. The validation and testing accuracy were a significant step lower compared to the results from Experiments 1 and 3. This suggests that this model did in fact experience some slight overfitting. However, the confusion matrices and the testing data results give positive signs towards a model with a combination of IMUs.

Firstly, in Figure 4.12 the recall score for all workouts was above 0.54, with most values being above 0.72, except for Arm Curl (0.66), BenchPress (0.6) and Squat (0.55). There was only one workout class that had an F1 score below 0.6 (BenchPress: 0.56). All the other F1 scores were 0.65 and above, two of them were 0.90 and above (Null and Walking) (Table 4.12). Overall the results of this experiment contradict previous research that demonstrates that more than one IMU gives significantly higher performing models [4, 47, 48]. However it may have been the wrong combination of IMUs. The pocket IMU, as discussed previously, may have had flaws to its positioning and the data that it collected. According to the findings from Experiments 1 and 3, a combination of the wrist and leg IMU may yield better results than a combination of all three IMUs.

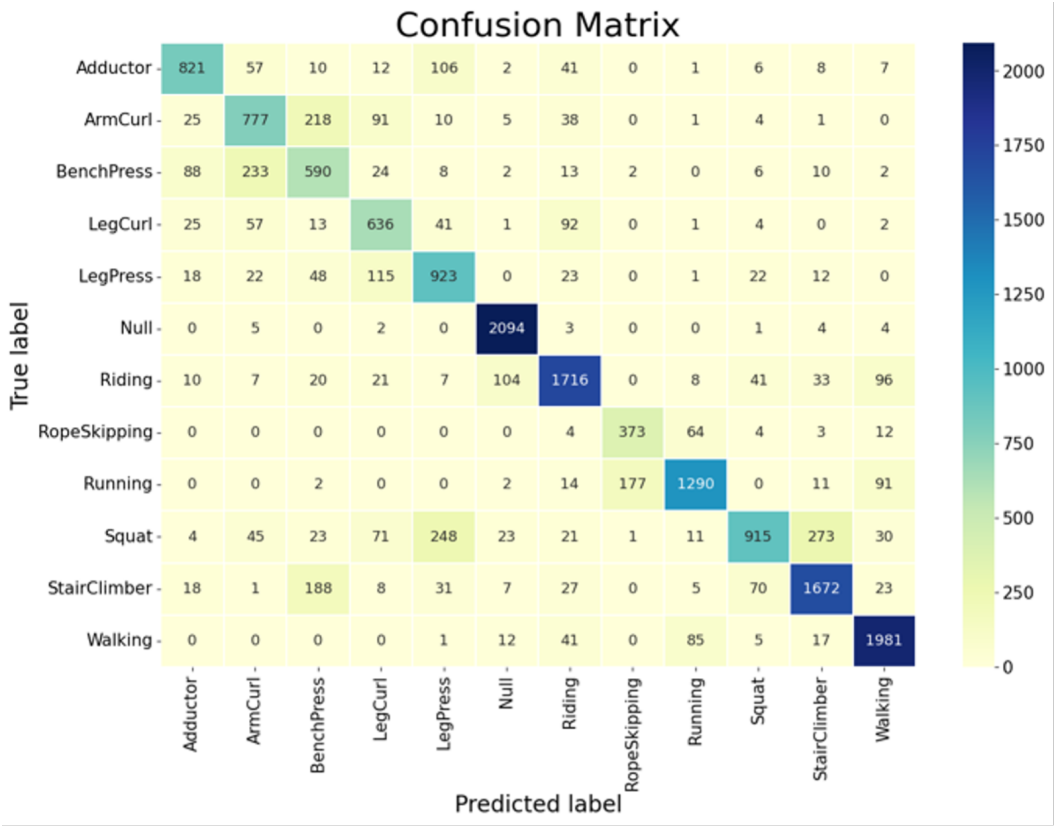


Figure 4.11: Confusion matrix of Experiment 4.

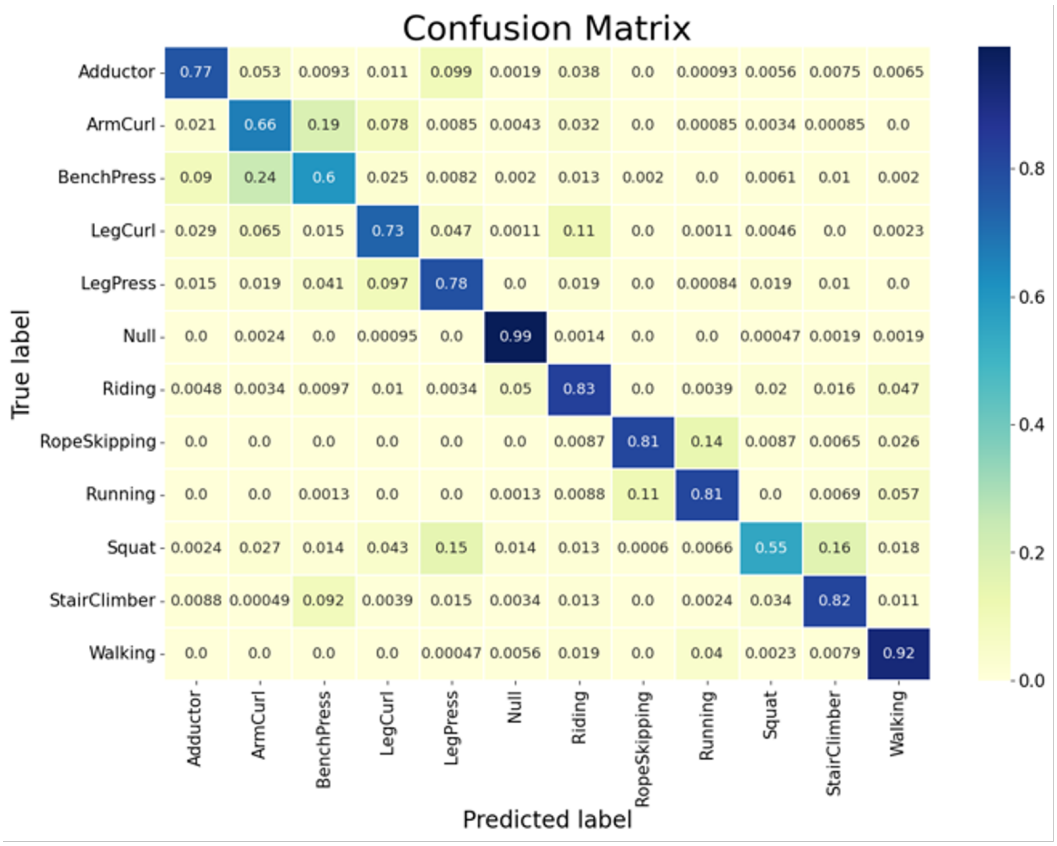


Figure 4.12: Normalised confusion matrix of Experiment 4.

Table 4.11: Accuracy and loss for Experiment 4.

Data Split	Accuracy	Loss
Training	0.88	0.34
Validation	0.80	0.63
Testing	0.79	0.77

Table 4.12: Testing data results of each workout for Experiment 4.

Workout	Precision	Recall	F1 Score
Adductor	0.81	0.77	0.79
ArmCurl	0.65	0.66	0.65
BenchPress	0.53	0.60	0.56
LegCurl	0.65	0.73	0.69
LegPress	0.67	0.78	0.72
Null	0.93	0.99	0.96
Riding	0.84	0.83	0.84
RopeSkipping	0.67	0.81	0.74
Running	0.88	0.81	0.84
Squat	0.85	0.55	0.67
StairClimber	0.82	0.82	0.82
Walking	0.88	0.92	0.90

4.5 Experiment Comparison

This subsection of this chapter compared the results of the experiments in more detail to answer the two key objectives of this thesis. Firstly, the results of each experiment were compared to answer the first objective, as each experiment was based upon data from IMUs in different or multiple positions. Secondly, the testing data results tables were used to compare between the recognition accuracy of aerobic vs strength training exercises.

The accuracy and loss curves for each experiment gave an indication of what the final performance of each model might be. For instance, Experiment 1 and 3

exhibited accuracy and loss curves that are characteristic of high performing DL models (Figures 4.1 and 4.7). This resulted in great final accuracy and loss scores (Tables 4.2 and 4.8). However, Experiment 2 and 4 exhibited slightly different accuracy and loss curves (Figures 4.4 and 4.10). Even though Experiment 4 resulted in better final accuracy and loss scores than Experiment 2, it followed a similar curve trend. This trend was that the validation curves (especially validation loss) plateaued much earlier than the accuracy curves. As discussed, this implies the onset of overtraining, which was halted by the early stopping regularisation technique employed. The common denominator between these two Experiments was that they both included data from the pocket IMU device. As alluded to, this suggests that this IMU position may not be suitable for strength training based HAR due to 1) the possibility that the IMU was not secured in position so the data may not be accurate and 2) the fact that there was relatively less movement in the pocket position (upper leg) compared to the wrist and lower leg.

This disparity between the experiments containing the pocket IMU device was also evident in the confusion matrices and testing data results. In Experiments 1 and 3, most of the precision, recall and F1 scores were 0.9 and above (Table 4.3 and 4.9). Whilst in Experiment 2 there were no F1 scores above 0.9 (Table 4.6) and in Experiment 4 there were only 2 F1 scores above 0.9 (Table 4.12). In Experiments 1 and 3 there were a few exceptions due to the wrist or the leg being almost stationary during that particular workout class. Interestingly, the low scoring workout classes in Experiment 3 were high scoring in Experiment 1 and vice versa. For example, the BenchPress F1 score for Experiment 3 was 0.0 and for Experiment 1 it was 0.95. In addition, the LegCurl F1 score for Experiment 1 was 0.73 and for Experiment 3 it was 0.94. This is likely due to the BenchPress being an arm related exercise and the LegCurl being a leg related exercise. This suggests that the IMU on the moving limb was able to detect the exercise much better. Therefore, a combination of these two IMU devices (wrist and leg) may generate better results than each in isolation as they account for movement the other sensor lacks.

The testing data evaluation metrics were used to compare the classification accuracy of aerobic vs anaerobic (strength) exercises for each LSTM model. The precision, recall and F1 score values were grouped together and then averaged. This was the mean average rounded to 2 decimal places. The results of the averaging for each experiment are displayed in Tables 4.13, 4.14, 4.15 and 4.16 respectively. It can be clearly seen that on average all three evaluation metrics were consistently higher for the aerobic workouts compared to the anaerobic workouts.

Only the mean precision for Experiment 1 had an equal score of 0.91 for both the aerobic and anaerobic workouts (Table 4.13). The rest of the mean anaerobic workouts scores were lower compared to the aerobic workouts. This suggests that DL models do find it harder to recognise anaerobic exercises compared to aerobic exercises.

Table 4.13: Mean testing results for aerobic and anerobic workouts for Experiment 1.

Workout	**Precision**	**Recall**	**F1 Score**
Aerobic Workouts	0.91	0.95	0.93
Anaerobic Workouts	0.91	0.85	0.87

Table 4.14: Mean testing results for aerobic and anerobic workouts for Experiment 2.

Workout	**Precision**	**Recall**	**F1 Score**
Aerobic Workouts	0.70	0.65	0.65
Anaerobic Workouts	0.61	0.60	0.58

Table 4.15: Mean testing results for aerobic and anerobic workouts for Experiment 3.

Workout	**Precision**	**Recall**	**F1 Score**
Aerobic Workouts	0.93	0.94	0.93
Anaerobic Workouts	0.64	0.70	0.64

Table 4.16: Mean testing results for aerobic and anerobic workouts for Experiment 4.

Workout	**Precision**	**Recall**	**F1 Score**
Aerobic Workouts	0.82	0.84	0.83
Anaerobic Workouts	0.69	0.68	0.68

A reason for this may be due to there being many more anaerobic exercise movements compared to aerobic exercise movements. This means that more data would need to be collected for these movements, which would increase the resources

needed. In addition, anaerobic movements are much more complex and less repetitive (there is often a lot of rest time in between anaerobic exercise bouts). This means that the variability in the acceleration signal pattern is higher. Therefore, the models may have struggled to recognise anaerobic movement as the pattern was less consistent and noisier. To tackle these possible issues, they must be investigated and addressed in more depth.

Chapter 5

Conclusions

5.1 Summary

In short, this thesis aimed to recognise and classify strength training exercises using acceleration and angular acceleration from wearable IMUs and LSTM algorithms. This aim was crafted as a result of the literature review and Chapter 1 highlighted the key opportunities this thesis aimed to tap into. These were the lack of research in strength exercise based HAR [33], the plethora of benefits strength training provides [22, 34, 35] and the state of the fitness and wearable industries [8, 9, 10]. The broad aim was then broken down into two testable objectives to narrow the focus of this thesis. These were 1) to compare the effects that different IMU sensor positions have on the classification accuracy of an LSTM model and 2) to compare the classification accuracy of an LSTM model between aerobic vs strength training exercises. To answer the objectives, four experiments were conducted, as detailed in Chapter 3. The results of these experiments were then displayed and discussed in Chapter 4.

Regarding Objective 1, it was found that the best performing LSTM models were built from IMUs located on the wrist and the lower leg. The worst performing LSTM model was built from the pocket IMU device. Finally, the model built from a combination of all three sensors (wrist, pocket and leg) only performed 3rd best. These results conflict with most of the literature [4, 47, 48] as it has often been found that a combination of IMU devices gives the best results. This was likely due to the pocket position being unsuitable for strength exercise based HAR. The pocket IMU may have been loose in the pocket, causing the data to be more variable and unreliable. The pocket IMU may have also performed worse due to its anatomical position (the upper thigh) not moving as much as the other two positions (wrist and lower leg). Smaller movements may have led to smaller

acceleration and angular acceleration signals that may have made it harder for the DL algorithm to extract clear patterns.

Conversely this finding supports research that demonstrates the effectiveness of using single IMU devices in HAR [44, 49, 85]. This suggests that 1 IMU may be suitable to effectively recognise strength training exercises. This would make HAR systems for strength exercise recognition, simpler, easier and cheaper to implement. However, not all possible combinations of IMU sensors were tested in this thesis. Only a combination of all three IMU sensors was tested for. Whilst a combination of pairs of sensors were not tested for (wrist and pocket, wrist and leg, pocket and leg). A combination of the wrist and leg IMUs may have yielded even better results than each of them in isolation as they both produced high performing models. Furthermore, the wrist and leg IMUs were both attached to the distal end of the limbs which tend to move more than the proximal end of limbs during exercise. The LSTM built from the wrist IMU was better at detecting arm-based exercises whilst the LSTM built from the leg IMU was better at detecting leg-based exercises. A model built with just the wrist and leg IMU may have taken the best from both IMUs and yielded even better results. This indicates a clear future direction for this research.

Regarding Objective 2, it was distinctly found that each LSTM model was better at recognising and classifying aerobic workouts compared to anaerobic workouts. This was evident from the mean testing results tables (Tables 4.13, 4.14, 4.15 and 4.16). This supports the idea that anaerobic exercises are more difficult for DL algorithms to recognise. The likely reasons for this are that anaerobic exercise movements are more complex, less repetitive and there are more of them (compared to aerobic exercise movements). This means that more resources are required to get the same level of recognition performance. For example, more data is required to be collected to account for the large rest periods and the higher number of different workout classes. This larger volume of data requires computers with higher processing capacity. This suggests that a limiting factor in strength exercise based HAR may be the current processing power of wearables. Until wearables are capable of processing the amount of data required, strength based HAR may not be feasible in real-world scenarios. In addition, the complexity of the movements may make the signal created by the movements noisier and therefore harder for DL algorithms to recognise. To tackle these possible issues, anaerobic exercise based HAR must be investigated in more depth.

In summary this thesis, agreed and disagreed with the literature but provided valuable insights into the possible optimal strategies for strength exercise based HAR. It also unearthed the disparity between aerobic and anaerobic exercise recognition capability. These findings provide a clear direction that future research can embark on, with the aim of taking advantage of the opportunities outlined in Chapter 1.

5.2 Future Considerations

This thesis provides a springboard for future research with a few different possibilities available. The main findings emphasized two key routes for future research. Firstly, it was found that the wrist and lower leg IMU positions were the best for strength exercise based HAR. The wrist is a practical IMU position as there are many wearable smartwatches on the market that already have IMUs integrated in them [32]. This leverages the expanding commercial wearable market [8, 9]. However, the lower leg position is not practical, as individuals do not tend to where devices on their lower legs during exercise as this is uncomfortable and restrictive. To address this issue future research should look to embed IMUs in apparel or accessories that are already part of everyday life. In this instance shoes offer that opportunity. Shoes are a frictionless alternative that are customary for people to wear during exercise. Therefore, future research should aim to investigate strength based HAR using IMUs attached to the wrist and the feet/shoes.

Secondly, the LSTM models found it more difficult to classify anaerobic exercises compared to aerobic exercises. This suggests that upcoming research should focus on uncovering why this is the case. Understanding the limitations to recognising and classifying anaerobic exercises will enable research efforts to address these limitations and improve upon them. Ultimately leading to the improvement of anaerobic/strength exercise based HAR systems, which could be highly beneficial to society and also highly lucrative.

In terms of the methodology used in this thesis, there are a few alternatives and developments used in previous research that may be worth considering. To start with, the classes were imbalanced. This is because large rest periods are common in strength exercise workouts. Therefore, future research should aim to ensure that all workout classes (excluding the null class) have similar class counts to ensure bias is kept to a minimum. In addition, some of the models were prone to overfitting and each model was only executed once for each combination of hyperparameters and data partitions. To address this, cross-validation should be used

in future research as it has been used in previous research [2]. Cross-validation ensures a more reliable assessment of the model's performance as the model is tested with different partitions of data and the performance scores are then averaged. Lastly, the learning rate for each model was kept static. This is often the case, but in recent years adjustable learning rates have led to models that converge faster and perform better [94]. Future research on strength exercise based HAR may benefit from using and testing different learning rate decay techniques to achieve higher model performance.

Although hybrid models were outside of the scope of this thesis, due to the added complexity and resources required, future research should consider utilising them. Hybrid DL models combine the advantages of each DL algorithm used to build better performing models [30, 77, 78, 79]. Additionally, alternative models or novel DL techniques should be investigated to see if they can be applied and benefit this area of study. For example, a novel technique called transfer learning (TL) may be very suitable for the difficulties associated with strength exercise based HAR. TL re-uses information learned from a previous task to increase the performance of a similar task. As there are many different types of similar anaerobic/strength exercises, TL may be able to address this and reduce the amount of labelled data that needs to be collected.

However, all of these improvements to model performance come at a cost to resources, processing requirements and time. If the aim is to create HAR systems that can be integrated into people's daily lives, then higher complexity and processing requirements only makes this more difficult. Instead, future research efforts may focus on constructing models that are lightweight and flexible [13]. Additionally, this thesis trained and tested the models offline. If the aim is to create a real-time HAR system that works on smartwatches, then future research should focus on training and testing models online and in real-time.

In summary, the findings of this thesis provide a progressive step in the field of strength exercise based HAR. To take this study forward, future research should focus on 1) training DL models from IMUs attached to the wrist and to the feet/shoes and 2) understanding the limitations of anaerobic/strength exercise based HAR. Furthermore, future researchers should consider their aims and objectives and apply the recommendations of this thesis where appropriate.

References

[1] G. V. Houdt, C. Mosquera, and G. Nápoles, "A review on the long short-term memory model," *Artificial Intelligence Review*, vol. 53, pp. 5929–5955, 12 2020. [Online]. Available: http://dx.doi.org/10.1007/s10462-020-09838-1

[2] S. Bian, V. F. Rey, S. Yuan, and P. Lukowicz, "The contribution of human body capacitance/body-area electric field to individual and collaborative activity recognition," 10 2022. [Online]. Available: http://dx.doi.org/10.48550/arXiv.2210.14794

[3] J. Smith, "Activity, context, and plan recognition withcomputational causal behaviour models," PhD thesis, University of Rostock, Rostock, Germany, 2018. [Online]. Available: http://dx.doi.org/10.18453/rosdok_id00002015

[4] S. K. Chaurasia and S. R. Reddy, "State-of-the-art survey on activity recognition and classification using smartphones and wearable sensors," *Multimedia Tools and Applications*, vol. 81, pp. 1077–1108, 1 2022. [Online]. Available: http://dx.doi.org/10.1007/s11042-021-11410-0

[5] M. G. Morshed, T. Sultana, A. Alam, and Y.-K. Lee, "Human action recognition: A taxonomy-based survey, updates, and opportunities," *Sensors*, vol. 23, 2 2023. [Online]. Available: http://dx.doi.org/10.3390/s23042182

[6] S. Bian, M. Liu, B. Zhou, and P. Lukowicz, "The state-of-the-art sensing techniques in human activity recognition: A survey," *Sensors*, vol. 22, 6 2022. [Online]. Available: http://dx.doi.org/10.3390/s22124596

[7] F. Demrozi, G. Pravadelli, A. Bihorac, and P. Rashidi, "Human activity recognition using inertial, physiological and environmental sensors: A comprehensive survey," *IEEE Access*, vol. 8, pp. 210 816–210 836, 2020. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2020.3037715

[8] IDC, "Wearables deliver double-digit growth for both q4 and the full year 2021, according to idc," 3 2022. [Online]. Available: https://www.idc.com/getdoc.jsp?containerId=prUS48935722

[9] I-SCOOP, "Wearables market outlook 2020: drivers and new markets," 2016. [Online]. Available: https://www.i-scoop.eu/wearables-market-outlook-2020-drivers-new-markets/

[10] W. R. Thompson, "Worldwide survey of fitness trends for 2023," *ACSM's Health & Fitness Journal*, vol. 27, pp. 9–18, 2023. [Online]. Available: http://dx.doi.org/10.1249/FIT.0000000000000834

[11] F. Foerster, M. Smeja, and J. Fahrenberg, "Detection of posture and motion by accelerometry: a validation study in ambulatory monitoring," *Computers in Human Behavior*, vol. 15, pp. 571–583, 1999. [Online]. Available: http://dx.doi.org/10.1016/S0747-5632(99)00037-0

[12] K. V. Laerhoven and O. Cakmakci, "What shall we teach our pants?" 2000. [Online]. Available: http://dx.doi.org/10.1109/ISWC.2000.888468

[13] S. O. Slim, A. Atia, M. M. A. Elfattah, and M.-S. M. Mostafa, "Survey on human activity recognition based on acceleration data," *International Journal of Advanced Computer Science and Applications*, vol. 10, pp. 84–98, 2019. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2019.0100311

[14] A. Prati, C. Shan, and K. I.-K. Wang, "Sensors, vision and networks: From video surveillance to activity recognition and health monitoring," *Journal of Ambient Intelligence and Smart Environments*, vol. 11, pp. 5–22, 2019. [Online]. Available: https://dl.acm.org/doi/abs/10.3233/AIS-180510

[15] G. Ogbuabor and R. La, "Human activity recognition for healthcare using smartphones." Association for Computing Machinery, 2 2018, pp. 41–46. [Online]. Available: http://dx.doi.org/10.1145/3195106.3195157

[16] C. Debes, A. Merentitis, S. Sukhanov, M. Niessen, N. Frangiadakis, and A. Bauer, "Monitoring activities of daily living in smart homes: Understanding human behavior," *IEEE Signal Processing Magazine*, vol. 33, pp. 81–94, 3 2016. [Online]. Available: http://dx.doi.org/10.1109/MSP.2015.2503881

[17] Y.-L. Hsu, S.-C. Yang, H.-C. Chang, and H.-C. Lai, "Human daily and sport activity recognition using a wearable inertial sensor network," *IEEE Access*, vol. 6, pp. 31 715–31 728, 5 2018. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2018.2839766

[18] M. A. Nystoriak and A. Bhatnagar, "Cardiovascular effects and benefits of exercise," *Frontiers in Cardiovascular Medicine*, vol. 5, 9 2018. [Online]. Available: http://dx.doi.org/10.3389/fcvm.2018.00135

[19] S. R. Colberg, R. J. Sigal, B. Fernhall, J. G. Regensteiner, B. J. Blissmer, R. R. Rubin, L. Chasan-Taber, A. L. Albright, and B. Braun, "Exercise and type 2 diabetes: The american college of sports medicine and the american diabetes association: Joint position statement," *Diabetes Care*, vol. 33, pp. 147–167, 12 2010. [Online]. Available: http://dx.doi.org/10.2337/dc10-9990

[20] L. Mandolesi, A. Polverino, S. Montuori, F. Foti, G. Ferraioli, P. Sorrentino, and G. Sorrentino, "Effects of physical exercise on cognitive functioning and wellbeing: Biological and psychological benefits," *Frontiers in Psychology*, vol. 9, 4 2018. [Online]. Available: http://dx.doi.org/10.3389/fpsyg.2018.00509

[21] P. T. Katzmarzyk, C. Friedenreich, E. J. Shiroma, and I.-M. Lee, "Physical inactivity and non-communicable disease burden in low-income, middle-income and high-income countries," *British Journal of Sports Medicine*, vol. 56, pp. 101–106, 1 2022. [Online]. Available: http://dx.doi.org/10.1136/bjsports-2020-103640

[22] "Physical activity," World Health Organization, 10 2022, accessed Apr. 25, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/physical-activity

[23] A. Garcia-Hermoso, J. F. López-Gil, R. Ramírez-Vélez, A. M. Alonso-Martínez, M. Izquierdo, and Y. Ezzatvar, "Adherence to aerobic and muscle-strengthening activities guidelines: A systematic review and meta-analysis of 3.3 million participants across 32 countries," *British Journal of Sports Medicine*, vol. 57, pp. 225–229, 2 2023. [Online]. Available: http://dx.doi.org/10.1136/bjsports-2022-106189

[24] D. M. Bravata, M. C. Smith-Spangler, V. Sundaram, M. L. A. Gienger, B. N. Lin, S. R. Lewis, M. D. C. Stave, M. I. Olkin, and J. R. Sirard, "Using pedometers to increase physical activity and improve health: A systematic review," *JAMA*, vol. 298, pp. 2296–2304, 2007. [Online]. Available: http://dx.doi.org/10.1001/jama.298.19.2296

[25] D. Merom, C. Rissel, P. Phongsavan, B. J. Smith, C. V. Kemenade, W. J. Brown, and A. E. Bauman, "Promoting walking with pedometers in the community: The step-by-step trial," *American Journal of Preventive Medicine*, vol. 32, pp. 290–297, 4 2007. [Online]. Available: http://dx.doi.org/10.1016/j.amepre.2006.12.007

[26] M. Ermes, J. Pärkkä, J. Mäntyjärvi, and I. Korhonen, "Detection of daily activities and sports with wearable sensors in controlled and

uncontrolled conditions," *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, pp. 20–26, 1 2008. [Online]. Available: http://dx.doi.org/10.1109/TITB.2007.899496

[27] A. Bayat, M. Pomplun, and D. A. Tran, "A study on human activity recognition using accelerometer data from smartphones," vol. 34. Elsevier B.V., 2014, pp. 450–457. [Online]. Available: http://dx.doi.org/10.1016/j.procs.2014.07.009

[28] M. A. Alsheikh, A. Selim, D. Niyato, L. Doyle, S. Lin, and H.-P. Tan, "Deep activity recognition models with triaxial accelerometers," 11 2016, pp. 8–13. [Online]. Available: http://arxiv.org/abs/1511.04664

[29] M. A. Ayu, S. A. Ismail, T. Mantoro, and A. F. A. Matin, "Real-time activity recognition in mobile phones based on its accelerometer data," 2016. [Online]. Available: http://dx.doi.org/10.1109/IAC.2016.7905732

[30] F. J. Ordóñez and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, 1 2016. [Online]. Available: http://dx.doi.org/10.3390/s16010115

[31] A. Ignatov, "Real-time human activity recognition from accelerometer data using convolutional neural networks," *Applied Soft Computing*, vol. 62, pp. 915–922, 1 2018. [Online]. Available: http://dx.doi.org/10.1016/j.asoc.2017.09.027

[32] A. Ometov, V. Shubina, L. Klus, J. Skibińska, S. Saafi, P. Pascacio, L. Flueratoru, D. Q. Gaibor, N. Chukhno, O. Chukhno, A. Ali, A. Channa, E. Svertoka, W. B. Qaim, R. Casanova-Marqués, S. Holcer, J. Torres-Sospedra, S. Casteleyn, G. Ruggeri, G. Araniti, R. Burget, J. Hosek, and E. S. Lohan, "A survey on wearable technology: History, state-of-the-art and current challenges," *Computer Networks*, vol. 193, 7 2021. [Online]. Available: http://dx.doi.org/10.1016/j.comnet.2021.108074

[33] J. Qi, P. Yang, M. Hanneghan, S. Tang, and B. Zhou, "A hybrid hierarchical framework for gym physical activity recognition and measurement using wearable sensors," *IEEE Internet of Things Journal*, vol. 6, pp. 1384–1393, 4 2019. [Online]. Available: http://dx.doi.org/10.1109/JIOT.2018.2846359

[34] T. J. Suchomel, S. Nimphius, and M. H. Stone, "The importance of muscular strength in athletic performance," *Sports Medicine*, vol. 46, pp. 1419–1449, 10 2016. [Online]. Available: http://dx.doi.org/10.1007/s40279-016-0486-0

[35] W. L. Westcott, "Resistance training is medicine: Effects of strength training on health," *Current Sports Medicine Reports*, vol. 11, pp. 209–216, 2012. [Online]. Available: http://dx.doi.org/10.1249/JSR.0b013e31825dabb8

[36] L. M. Dang, K. Min, H. Wang, M. J. Piran, C. H. Lee, and H. Moon, "Sensor-based and vision-based human activity recognition: A comprehensive survey," *Pattern Recognition*, vol. 108, 12 2020. [Online]. Available: http://dx.doi.org/10.1016/j.patcog.2020.107561

[37] F. Gu, M. H. Chung, M. Chignell, S. Valaee, B. Zhou, and X. Liu, "A survey on deep learning for human activity recognition," *ACM Computing Surveys*, vol. 54, 11 2021. [Online]. Available: http://dx.doi.org/10.1145/3472290

[38] X. Fan, F. Wang, F. Wang, W. Gong, and J. Liu, "When rfid meets deep learning: Exploring cognitive intelligence for activity identification," *IEEE Wireless Communications*, vol. 26, pp. 19–25, 6 2019. [Online]. Available: http://dx.doi.org/10.1109/MWC.2019.1800405

[39] J. Zhang, F. Wu, B. Wei, Q. Zhang, H. Huang, S. W. Shah, and J. Cheng, "Data augmentation and dense-lstm for human activity recognition using wifi signal," *IEEE Internet of Things Journal*, vol. 8, pp. 4628–4641, 3 2021. [Online]. Available: http://dx.doi.org/10.1109/JIOT.2020.3026732

[40] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 221–231, 2013. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2012.59

[41] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," 2018, pp. 7444–7452. [Online]. Available: http://dx.doi.org/10.48550/arXiv.1801.07455

[42] P. Wang, W. Li, Z. Gao, C. Tang, and P. Ogunbona, "Depth pooling based large-scale 3d action recognition with convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 20, 3 2018. [Online]. Available: http://dx.doi.org/10.1109/TMM.2018.2818329

[43] F. Fereidoonian, F. Firouzi, and B. Farahani, "Human activity recognition: From sensors to applications," 2020. [Online]. Available: http://dx.doi.org/10.1109/COINS49042.2020.9191417

[44] A. Mannini, S. S. Intille, M. Rosenberger, A. M. Sabatini, and W. Haskell, "Activity recognition using a single accelerometer placed at the wrist

or ankle," *Medicine and Science in Sports and Exercise*, vol. 45, pp. 2193–2203, 11 2013. [Online]. Available: http://dx.doi.org/10.1249/MSS.0b013e31829736d6

[45] M. Shoaib, H. Scholten, and P. J. M. Havinga, "Towards physical activity recognition using smartphone sensors," 2013, pp. 80–87. [Online]. Available: http://dx.doi.org/10.1109/UIC-ATC.2013.43

[46] P. P. Ariza-Colpas, E. Vicario, A. I. Oviedo-Carrascal, S. B. Aziz, M. A. Piñeres-Melo, A. Quintero-Linero, and F. Patara, "Human activity recognition data analysis: History, evolutions, and new trends," *Sensors*, vol. 22, 5 2022. [Online]. Available: http://dx.doi.org/10.3390/s22093401

[47] I. Cleland, B. Kikhia, C. Nugent, A. Boytsov, J. Hallberg, K. Synnes, S. McClean, and D. Finlay, "Optimal placement of accelerometers for the detection of everyday activities," *Sensors*, vol. 13, pp. 9183–9200, 1 2013. [Online]. Available: http://dx.doi.org/10.3390/s130709183

[48] V. X. Rahn, L. Zhou, E. Klieme, and B. Arnrich, "Optimal sensor placement for human activity recognition with a minimal smartphone–imu setup." SciTePress, 2021, pp. 37–48. [Online]. Available: http://dx.doi.org/10.5220/0010269100370048

[49] E. Valarezo, P. Rivera, J. M. Park, G. Gi, T. Y. Kim, M. A. Al-Antari, M. Al-Masni, and T.-S. Kim, "Human activity recognition using a single wrist imu sensor via deep learning convolutional and recurrent neural nets," *Journal of ICT, Design, Engineering and Technological Science*, vol. 1, pp. 1–5, Jun. 2017. [Online]. Available: https://jitdets.com/ojs/index.php/jitdets/article/view/30

[50] T. Hayashi, M. Nishida, N. Kitaoka, and K. Takeda, "Daily activity recognition based on dnn using environmental sound and acceleration signals," 2015, pp. 2351–2355. [Online]. Available: http://dx.doi.org/10.1109/EUSIPCO.2015.7362796

[51] P. Bharti, D. De, S. Chellappan, and S. K. Das, "Human: Complex activity recognition with multi-modal multi-positional body sensing," *IEEE Transactions on Mobile Computing*, vol. 18, pp. 857–870, 2019. [Online]. Available: http://dx.doi.org/10.1109/TMC.2018.2841905

[52] Q. Zou, L. Ni, Q. Wang, Q. Li, and S. Wang, "Robust gait recognition by integrating inertial and rgbd sensors," *IEEE Transactions*

on Cybernetics, vol. 48, pp. 1136–1150, 4 2018. [Online]. Available: http://dx.doi.org/10.1109/TCYB.2017.2682280

[53] E. Kanjo, E. M. Younis, and C. S. Ang, "Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection," *Information Fusion*, vol. 49, pp. 46–56, 9 2019. [Online]. Available: http://dx.doi.org/10.1016/j.inffus.2018.09.001

[54] A. K. Jain and J. Mao, "Artificial neural networks: A tutorial," *Computer*, vol. 29, pp. 31–44, 1996. [Online]. Available: http://dx.doi.org/10.1109/2.485891

[55] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bulletin of Mathematical Biophysics*, vol. 5, pp. 115–133, 1943. [Online]. Available: http://dx.doi.org/10.1007/bf02478259

[56] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986. [Online]. Available: http://dx.doi.org/10.1038/323533a0

[57] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, 2012. [Online]. Available: http://dx.doi.org/10.1145/3065386

[58] O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, A. M. Umar, O. U. Linus, H. Arshad, A. A. Kazaure, U. Gana, and M. U. Kiru, "Comprehensive review of artificial neural network applications to pattern recognition," *IEEE Access*, vol. 7, pp. 158 820–158 846, 2019. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2019.2945545

[59] W. Fang, P. E. Love, H. Luo, and L. Ding, "Computer vision for behaviour-based safety in construction: A review and future directions," *Advanced Engineering Informatics*, vol. 43, 1 2020. [Online]. Available: http://dx.doi.org/10.1016/j.aei.2019.100980

[60] H.-C. Li, Z.-Y. Deng, and H.-H. Chiang, "Lightweight and resource-constrained learning network for face recognition with performance optimization," *Sensors*, vol. 20, 11 2020. [Online]. Available: http://dx.doi.org/10.3390/s20216114

[61] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten

zip code recognition," *Neural Computation*, pp. 541 – 551, 1989. [Online]. Available: http://dx.doi.org/10.1162/neco.1989.1.4.541

[62] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, pp. 2278–2324, 1998. [Online]. Available: http://dx.doi.org/10.1109/5.726791

[63] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F.-F. Li, "Large-scale video classification with convolutional neural networks." IEEE Computer Society, 9 2014, pp. 1725–1732. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2014.223

[64] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," 2014, pp. 568–576. [Online]. Available: http://dx.doi.org/10.48550/arXiv.1406.2199

[65] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 3200–3225, 3 2023. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2022.3183112

[66] A. Diba, M. Fayyaz, V. Sharma, M. M. Arzani, R. Yousefzadeh, J. Gall, and L. V. Gool, "Spatio-temporal channel correlation networks for action classification," 6 2019, pp. 299–315. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-01225-0_18

[67] H. Salehinejad, S. Sankar, J. Barfett, E. Colak, and S. Valaee, "Recent advances in recurrent neural networks," 12 2017. [Online]. Available: http://dx.doi.org/10.48550/arXiv.1801.01078

[68] J. L. Elman, "Finding structure in time," *Cognitive Science*, vol. 14, pp. 179–211, 1990. [Online]. Available: http://dx.doi.org/10.1016/0364-0213(90)90002-E

[69] A. Shewalkar, D. Nyavanandi, and S. A. Ludwig, "Performance evaluation of deep neural networks applied to speech recognition: Rnn, lstm and gru," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 9, pp. 235–245, 10 2019. [Online]. Available: http://dx.doi.org/10.2478/jaiscr-2019-0006

[70] R. K. Pathan, M. Biswas, and M. U. Khandaker, "Time series prediction of covid-19 by mutation rate analysis using recurrent neural network-based lstm

model," *Chaos, Solitons and Fractals*, vol. 138, 9 2020. [Online]. Available: http://dx.doi.org/10.1016/j.chaos.2020.110018

[71] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, 1997. [Online]. Available: http://dx.doi.org/10.1162/neco.1997.9.8.1735

[72] H. Sak, A. Senior, K. Rao, and F. Beaufays, "Fast and accurate recurrent neural network acoustic models for speech recognition," 2015. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2015-350

[73] S. Mekruksavanich and A. Jitpattanakul, "Lstm networks using smartphone data for sensor-based human activity recognition in smart homes," *Sensors*, vol. 21, 3 2021. [Online]. Available: http://dx.doi.org/10.3390/s21051636

[74] M. Z. Uddin and A. Soylu, "Human activity recognition using wearable sensors, discriminant analysis, and long short-term memory-based neural structured learning," *Scientific Reports*, vol. 11, 12 2021. [Online]. Available: http://dx.doi.org/10.1038/s41598-021-95947-y

[75] S. W. Pienaar and R. Malekian, "Human activity recognition using lstm-rnn deep neural network architecture," 2019. [Online]. Available: http://dx.doi.org/10.1109/AFRICA.2019.8843403

[76] G. Weiss, "WISDM Smartphone and Smartwatch Activity and Biometrics Dataset ," UCI Machine Learning Repository, 2019. [Online]. Available: https://dx.doi.org/10.24432/C5HK59

[77] M. S. Singh, V. Pondenkandath, B. Zhou, P. Lukowicz, and M. Liwickit, "Transforming sensor data to the image domain for deep learning - an application to footstep detection," vol. 2017-May. Institute of Electrical and Electronics Engineers Inc., 6 2017. [Online]. Available: http://dx.doi.org/10.1109/IJCNN.2017.7966182

[78] C. Liu, L. Zhang, Z. Liu, K. Liu, X. Li, and Y. Liu, "Lasagna: Towards deep hierarchical understanding and searching over mobile sensing data." Association for Computing Machinery, 10 2016, pp. 334–347. [Online]. Available: http://dx.doi.org/10.1145/2973750.2973752

[79] Y. Zheng, Q. Liu, E. Chen, Y. Ge, and J. L. Zhao, "Exploiting multi-channels deep convolutional neural networks for multivariate time series classification," *Frontiers of Computer Science*, vol. 10, pp. 96–112, 2016. [Online]. Available: http://dx.doi.org/10.1007/s11704-015-4478-2

[80] D. Roggen, A. Calatroni, L.-V. Nguyen-Dinh, R. Chavarriaga, and H. Sagha, "OPPORTUNITY Activity Recognition," UCI Machine Learning Repository, 2012. [Online]. Available: https://dx.doi.org/10.24432/C5M027

[81] T. Stiefmeier, D. Roggen, G. Ogris, P. Lukowicz, and G. Tröster, "Wearable activity tracking in car manufacturing," *IEEE Pervasive Computing*, vol. 7, pp. 42–50, 2008. [Online]. Available: http://dx.doi.org/10.1109/MPRV.2008.40

[82] J. Reyes-Ortiz, D. Anguita, A. Ghio, L. Oneto, and X. Parra, "Human activity recognition using smartphones," UCI Machine Learning Repository, 2012. [Online]. Available: https://dx.doi.org/10.24432/C54S4K

[83] D. Roggen, M. Plotnik, and J. Hausdorff, "Daphnet Freezing of Gait," UCI Machine Learning Repository, 2013. [Online]. Available: https://dx.doi.org/10.24432/C56K78

[84] K. H. Chang, M. Y. Chen, and J. Canny, "Tracking free-weight exercises," vol. 4717 LNCS. Springer Verlag, 2007, pp. 19–37. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-74853-3_2

[85] A. Depari, P. Ferrari, A. Flammini, S. Rinaldi, and E. Sisinni, "Lightweight machine learning-based approach for supervision of fitness workout." Institute of Electrical and Electronics Engineers Inc., 5 2019. [Online]. Available: http://dx.doi.org/10.1109/SAS.2019.8706106

[86] H. Koskimaki and P. Siirtola, "Recognizing gym exercises using acceleration data from wearable sensors." Institute of Electrical and Electronics Engineers Inc., 1 2014, pp. 321–328. [Online]. Available: http://dx.doi.org/10.1109/CIDM.2014.7008685

[87] D. Morris, T. S. Saponas, A. Guillory, and I. Kelner, "Recofit: Using a wearable sensor to find, recognize, and count repetitive exercises." Association for Computing Machinery, 2014, pp. 3225–3234. [Online]. Available: http://dx.doi.org/10.1145/2556288.2557116

[88] zhaxidele, "Toolkit-for-hbc-sensing." [Online]. Available: https://github.com/zhaxidele/Toolkit-for-HBC-sensing/tree/main

[89] S. Bian, S. Yuan, V. F. Rey, and P. Lukowicz, "Using human body capacitance sensing to monitor leg motion dominated activities with a wrist worn device," in *Sensor- and Video-Based Activity and Behavior Computing*. Springer, 2022, pp. 81–94. [Online]. Available: http://dx.doi.org/10.1007/978-981-19-0361-8_5

[90] S. Bian and P. Lukowicz, "A systematic study of the influence of various user specific and environmental factors on wearable human body capacitance sensing." Springer, 2021, pp. 247–274. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-95593-9_20

[91] N. Sikder and A.-A. Nahid, "Ku-har: An open dataset for heterogeneous human activity recognition," *Pattern Recognition Letters*, vol. 146, pp. 46–54, 2021. [Online]. Available: http://dx.doi.org/10.1016/j.patrec.2021.02.024

[92] L. Prechelt, "Early stopping - but when?" in *Neural Networks: Tricks of the Trade - Second Edition*, ser. Lecture Notes in Computer Science, G. Montavon, G. B. Orr, and K. Müller, Eds. Springer, 2012, vol. 7700, pp. 53–67. [Online]. Available: https://doi.org/10.1007/978-3-642-35289-8_5

[93] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognition Letters*, vol. 119, pp. 3–11, 3 2019. [Online]. Available: https://doi.org/10.1016%2Fj.patrec.2018.02.010

[94] Y. Ding, "The impact of learning rate decay and periodical learning rate restart on artificial neural network," 1 2021, pp. 6–14. [Online]. Available: http://dx.doi.org/10.1145/3460268.3460270