Transformer-based Human Activity Recognition Calibration system

Stylianos Charalampous

MInf Project (Part 2) Report
Master of Informatics
School of Informatics
University of Edinburgh

2022

Abstract

The rapid growth of the Internet of Things (IoT) technologies observed in recent years has provided opportunities for innovative solutions in various domains. In this project, the machine learning research area of human activity recognition is explored utilising the RESpeck wearable monitor, which encapsulates tri-axial accelerometer and gyroscope sensors.

In this research, a domain-agnostic transformer-based calibration framework is implemented, enabling calibration of an already trained machine learning model to a particular subject. Furthermore, a transformer-based model is implemented for the purposes of this project, in an attempt to improve performance on human activity classification.

A series of model evaluation methods is then carried out to assess the performance of the implemented algorithms, including subject-independent cross-validation, as well as statistical testing, to evaluate the significance of the results.

Research Ethics Approval

This project obtained approval from the Informatics Research Ethics committee.
Ethics application number: 2019/27996
Date when approval was obtained: 2021-09-30
The participants' information sheet and a consent form are included in the appendix.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(Stylianos Charalampous)

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor DK Arvind for his continuous guidance and support throughout the academic year. I would also like to thank Teodora Georgescu who provided technical insights throughout this project, assistance with obtaining the ethics approval and helped with the data collection procedure. I would also like to thank Celina Dong for her cooperation in gathering the dataset.

Last but not least, I would like to thank my family and friends for their constant support throughout the year.

Table of Contents

Chapter 1

Introduction

1.1 Human Activity Recognition

Over the past decades, the interest in human activity recognition has seen rapid growth both in academia and industry, due to the wide range of applications it can applied for. In the past few years, a rapid development of wearable devices, as well as machine learning, has been observed, with remarkable advances emerging in the medical, fitness, military and security fields [32]. The human activity classification task focuses on extracting knowledge from data retrieved using Internet of Things (IoT) devices to accurately identify human activities. Methodologies presented in the literature are usually divided into two different categories, based on whether they use ambient sensors (extrinsic) or wearable devices (intrinsic) to acquire data for the classification model [32]. The former includes observing the behaviour of the subject along with the environment using extrinsic sensors fixed in predetermined positions and subsequently applying machine learning algorithms to determine the activity performed by the subject [60]. Even though the above-noted methods utilising data from sensors external to the subject have shown reliable performance even in diverse conditions, they pose numerous challenges related to privacy and applicability in smart environments which prevent them from being widely utilised [45]. In an attempt to eliminate the above-mentioned concerns, the intrinsic approaches employ machine learning algorithms applied to inertial signals gathered using sensors directly attached to the subject for the estimation of the physical activity [26]. This methodology, however, poses different challenges due to its dependence on data acquired by wearable sensors, which are significantly affected by noise thus deteriorating data quality and model accuracy. Most widely implemented solutions exploit several pre-processing algorithms and filtering techniques, to ensure that noise and sensor biases are eliminated prior to using the raw data in the classification framework. Therefore, effective data pre-processing is considered a crucial step in the classification pipeline, significantly impacting the performance of the machine learning model as well as the usability of the dataset [6].

1.2 Previous work carried out

In the MInf - Part 1 Project, an annotated dataset was gathered using linear acceleration data, recording 96 participants performing a range of physical activities. The dataset was then utilised for the development and training of an Auxiliary Classifier Generative Adversarial Network (AC-GAN) used for human activity classification. During the MInf - Part 1 Project a transfer learning methodology was also explored, in an attempt to improve the classifying capabilities of the activity recognition model for a specific subject, using linear acceleration samples of that subject performing a set of physical activities to train the model.

In this project, the annotated dataset is re-used for the training of the classification models implemented. The AC-GAN model is also used for comparison purposes in order to assess the performance of the transformer-based methodology implemented in this research. Finally, the transfer learning methodology as well as the findings of the experiments performed in the MInf - Part 1 Project are exploited for the development of a novel domain-agnostic transformer-based calibration framework presented in this project.

1.3 Research Objectives

Intra-subject dependencies have posed a significant challenge for the human activity classification task throughout the literature, requiring vast amounts of annotated data to achieve generalisability. As a result of the above-noted issue, it has been observed that models trained and evaluated on dissimilar subjects achieve inferior performance in comparison to models using a common pool of subjects for training and evaluation purposes [5].

This project aims at bridging the performance gap identified on account of intra-subject correlation by the implementation of a calibration framework that allows the neural network to learn subject-specific patterns using samples of a subject performing a range of activities; developing a proof-of-concept calibration system based on the work presented in the MInf - Part 1 Project [5].

In an attempt to resolve issues revealed in the MInf - Part 1 Project [5] while using the Auxiliary Classifier Generative Adversarial Network (AC-GAN) for transfer learning, as well as improve classification performance in activity recognition, a Transformer-based architecture is proposed in this project, improving the learning capacity of the implemented system.

1.4 Contributions

The main contributions of this research are outlined below:

- implementation of deep learning network architectures utilised for the evaluation of the human activity classification framework;

- design and implementation of a Transformer-based deep learning model for human activity recognition;
- development of a testing framework used to assess the impact of adding tri-axial angular velocity and increasing the sampling frequency of the initial tri-axial acceleration inputs;
- design and implementation of a novel domain-agnostic transformer-based calibration system utilising transfer learning methodologies.

1.5 Outline

The thesis structure is outlined below:

- **Chapter 1 - Introduction** provides an overview of this project, indicating the motivation instigating this research as well as an outline of the contributions demonstrated in this work.
- **Chapter 2 - Background** presents an assessment of related research applied for human activity recognition establishing the context of this project.
- **Chapter 3 - Data Collection and Pre-processing** details the data acquisition procedures carried out to form the dataset used for the classification task. Besides, it provides a detailed description of the operations applied to improve the quality and usefulness of the dataset for training and evaluation purposes.
- **Chapter 4 - Methodology** provides a comprehensive description of the methodology developed for the purposes of this project as well as justification of the decisions made during the design and implementation phases.
- **Chapter 5 - Results** demonstrates the performance yielded from the evaluation procedures carried out along with an analysis of the results extracted from each method.
- **Chapter 6 - Conclusions** summarises the contributions of this project, discussing limitations of the current implementation and identifies valuable trajectories for future research.

Chapter 2

Background

This chapter provides an overview of the approaches applied for human activity recognition found in the literature, as well as summarising concepts and methodologies applied in related projects utilising the RESpeck monitor, which have influenced the progression of this work.

2.1 Human Activity Recognition

The human activity recognition task has been a widely attempted machine learning problem, with numerous tackling strategies being demonstrated in the literature due to its wide applicability. A significant portion of the approaches demonstrated in the literature often involves manual extraction of domain-specific hand-crafted features from a particular dataset, that are utilised along with the original input by the machine learning model to make a prediction [46]. Machine learning models extracting information from manually generated features include but are not limited to: Decision Trees, Bayesian networks, Instance-based Learning, Support Vector Machines (SVM), Artificial Neural Networks and ensembles of classifiers [9]. Manual feature engineering, however, is often limited to domain-specific features which are less informative to other applications, or is even restricted to a subset of the most task-relevant features, due to the high dependence on human observance and expert knowledge. As a result, the implementation of machine learning based algorithms is often not feasible for every application due to the significant reliance on human expertise. To overcome the aforementioned problem, recent advances in deep learning have given rise to novel approaches of data classification, that automatically learn the most discriminative features of a particular dataset, resulting in improved performance [6, 9]. The deep learning methodology allows the model to learn representations of data with multiple levels of abstraction [33], which often surpass the performance of manually engineered algorithms, while also allowing domain-agnostic implementation. In the human activity classification domain, a range of methodologies has been presented, yielding state-of-the-art performance in multiple occurrences. Deep learning algorithms with noteworthy performance in the human activity recognition task found in the literature include but are not limited to: Convolutional Neural Networks (CNNs) [59], Recurrent Neural Networks (RNNs)

[16, 3], Long Short-Term Memory Networks (LSTMs) [24, 39], and Convolutional Long Short-Term Memory Networks (ConvLSTMs) [42].

2.2 Related work

In this section, previous approaches carried out for human activity recognition applied to datasets collected using the RESpeck monitor device are presented, in an attempt to highlight some of the limitations observed in those attempts and provide solutions to address them, thus improving the classification performance achieved in this research.

2.2.1 Simultaneous human activity and social signal classification

In the methodology presented by Wei in [55], the effect of combining a human activity recognition algorithm along with a social signal classifier was investigated, in an attempt to reveal potential benefits in the predictive capabilities of the model. In the paper, Wei presented a two-stage classification architecture, using two approaches introduced in previous research papers for human activity recognition [25] and social signal classification [18]. The algorithm utilised a binary classifier to differentiate between stationary and dynamic human activities, with the output of the model being fed into social signal and human activity recognition models, used to predict the social signal as well as the physical activity corresponding to the input sample, respectively. The two-step classification described, achieved a higher accuracy on the dataset used in the experiment in comparison to two standalone human activity and social signal classification architectures highlighting the benefits of a hierarchical configuration. The above observation was evaluated in the MInf - Part 1 Project [5], using a more complicated network architecture as well as an enhanced dataset, and it was seen that in cases when the numerosity of the dataset is large enough, distinct models for each task outperform the combined classifier. Therefore, in this work it was decided to continue using separate models for each task, focusing on human activity classification in this study.

2.2.2 Human Activity Classifier

The human activity recognition model used by Wei [55], was first proposed by Irsch [25], designed in an attempt to develop a data-efficient way of classifying human activity tackling the incredibly costly problem of acquiring enough annotated data. This approach included a Semi-Supervised Learning method using a Generative Adversarial Network (GAN) leveraging both labelled and unlabelled data. The GAN consisted of a *generator* and a *discriminator*, with the first generating fake samples based on the data, and the second distinguishing real and synthetic unlabelled data as well as the type of activities in the labelled data. As a result, after each epoch, the generator improved its generating capabilities, thus forcing the discriminator to improve as well. Irsch's method outperformed the compared Linear Regression and CNN models in cases when the amount of labelled data was limited, which widens the range of possible applications.

2.2.3 AC-GAN Human Activity Classifier with Transfer Learning

The Semi-Supervised GAN model was then implemented in the MInf - Part 1 Project [5], in an effort to address any limitations found previously, and consequently improve the performance of the neural network. During the investigation, it was revealed that the model was not learning patterns for all classes uniformly, thus showing significant imbalances in the performance between human activities. The solution implemented to mitigate this problem was to develop an Auxiliary Classifier Generative Adversarial Network (AC-GAN) that allowed the model to train uniformly for each physical activity class, thus preventing classes with a higher number of data samples from outweighing the remaining classes during training. Furthermore, a transfer learning methodology was exploited in the MInf - Part 1 Project [5], used to resolve the issue of limited availability of annotated datasets for human activity recognition as well as the problem of substantial diversity between data samples drawn from subjects with different body characteristics[32]. Transfer learning is a technique used to improve the classifying capabilities of a model by transferring information from a related domain [56], mimicking the way the human brain works, and applying knowledge gained from a learned task to a new one. In the past decade, several machine learning problems have been approached using transfer learning methods, due to their ability to utilise datasets from similar tasks for pattern recognition, with the availability of such datasets being significantly easier currently as big data repositories become more prevalent. Recent literature has shown promising results with the use of transfer learning methods in a wide spectrum of applications, including human activity recognition [20], image classification [15] and speech recognition [53].

Chapter 3

Data Collection and Pre-processing

The establishment of a comprehensive dataset that includes sufficient data samples for all categories in the class spectrum is a fundamental aspect of any machine learning implementation, vital for the subsequent development of the system. The predictive performance of a machine learning model, as well as its ability to generalise on unseen data, is significantly correlated to the quality of the dataset. The data samples need to be representative both in terms of validity and accuracy of labelling of the categories identified, allowing the model to develop the desired pattern recognition capabilities.

This chapter provides a detailed description of the procedures carried out in assembling the datasets used for the human activity recognition calibration system, as well as the processing techniques applied to improve the overall quality of the dataset and consequently its usefulness for the project, as well as any related future studies.

3.1 Data Collection Framework

The complete dataset used in this project was an amalgam of physical activity data collected for the purposes of this project, combined with datasets collected in related studies by participants wearing the RESpeck monitor using identical protocols. This section provides a detailed explanation of the components comprising the dataset, along with a specification of the data acquisition procedures followed to collect the new data required for the purposes of this study.

3.1.1 Hardware

The physical activity pattern monitoring involved in the study was conducted using the RESpeck monitor device, version 6, shown in Figure 3.1. The RESpeck monitor contains an encapsulated Freescale MMA7260QT tri-axial accelerometer, as well as three orthogonally mounted ADXRS300 angular rate sensors, transmitting data wirelessly using the Bluetooth Low Energy technology. The device is paired with a smartphone through an Android application and transmits accelerometer and gyroscope signals at a sampling frequency of 25 Hz. The RESpeck monitor device (3.5cm x 4.5cm;

18gms) is worn unobtrusively as a plaster in the lower costal margin monitoring the rotation of the chest wall.

Figure 3.1: RESpeck monitor device

Accelerometer Sensor

The integrated tri-axial accelerometer sensor measures linear acceleration across each of the orthogonal axes, which is the rate of change of the velocity of an object. The measurement of acceleration is in meters per second squared (m/s^2) or in G-forces (g). For this particular project, the latter was used, with a single unit of G-force on Earth being approximately equivalent to $9.8m/s^2$. The accelerometer detects both static and dynamic forces of acceleration. Static forces include gravitational pull from the planet, while dynamic forces can include vibrations and movement of the sensor.

Gyroscope Sensor

The gyroscope sensor is used to measure the rotational motion of an object. The rotational motion of an object is determined by measuring the angular velocity, which is the speed of rotation around an axis. The angular velocity is measured in radians per second (rad/s) and is detected using a vibrating mechanical element attached to the sensor through the physical phenomenon of Coriolis [17].

3.1.2 Sensor Placement

The sensor was mounted on the anterior side of the torso of each participant, on the left-hand quadrant of the abdominal cavity, just inferior to the rib margin, as shown in Figure 3.2. The sensor placement position was chosen based on literature research indicating the intersection of the lower costal margin and the midclavicular line as the optimal position for respiratory monitoring [37]. Furthermore, the mounting position was set near the centre of gravity of the body, as studies have identified areas closer to the centre of mass of the human body as the ideal position for human activity classification [43].

3.1.3 Dataset Composition

For the purposes of this project, three distinct datasets were established, with each fulfilling a discrete role in the realisation of the calibration framework. The above-mentioned datasets are described below, along with the purpose each serves in the system implementation:

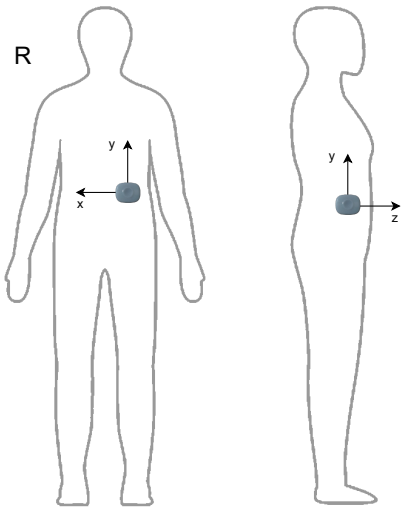

Figure 3.2: RESpeck monitor placement position

- **acceleration-only dataset sampled at 12.5Hz;** gathered from previously assembled datasets, used for the training of the base classification model;
- **acceleration and angular velocity dataset sampled at 25Hz;** collected for the purposes of this project, used to investigate the effect of sampling frequency and angular velocity in the performance of the classification model;
- **real-world dataset;** including acceleration and angular velocity information at a sampling rate of 25Hz, collected during the present study, and used for evaluating the performance of the calibration system on real-world data.

Data Acquisition

The data acquisition procedure included collecting data from a number of volunteers, who were asked to perform a series of everyday human activities and social signals. The eligibility criteria established, required each participant to be in a physical condition that would allow them to perform each of the designated activities without any difficulty, as well as have no history of severe chronic lung diseases that would cause discomfort when executing any of the activities. During the data gathering procedures, all the appropriate measures were taken by the participants and the investigators, to reduce the risk of catching and transmitting COVID-19. The data collection study was approved according to the Informatics Research Ethics Process (RT number 2019/17922). The study was conducted in collaboration with Celina Dong and Teodora Georgescu, with data collected from 14 participants in total. The data samples included acceleration and angular velocity information and were sampled at 25Hz frequency. The participant information sheet and the participant consent form can be found in Appendices A and B, respectively.

Each participant was invited to perform 19 everyday physical activities for at least 30 seconds each, although not necessarily in a single contiguous recording in deference for the participant's comfort. The hand-picked physical activities were consequently

categorised into static and dynamic, based on the movement involved in each of them. These included – static cases: sitting normally or bent forward and backwards; standing; and, lying down prone, or supine, or facing right or left; and dynamic cases: walking slowly and at a normal pace; running; ascending and descending stairs; swinging back and forth and left and right whilst sitting on a chair; standing up and sitting down on a chair; getting up from a lying down position and lying down from a seated position. Furthermore, a portion of the participants were asked to perform a series of social signal activities required for related projects running concurrently with the present study. The social signal activities were performed in combination with static activities and included: coughing; talking; eating or drinking; singing; laughing; breathing normally; breathing deeply; sighing; sobbing; yawning; hiccuping; and, hyperventilating.

Existing Datasets

For the training of the base classification model, the dataset presented in the MInf - Part 1 Project [5] was used, consisting of data from 96 participants, allowing direct comparison of the results demonstrated using the different methodologies presented in this work. The time sequence samples included accelerometer data information sampled at 12.5Hz frequency. The repository included data collected in the MInf - Part 1 Project [5] from a total of 14 participants using the above-mentioned protocol and activities; as well as annotated datasets collected by researchers working with the RESpeck monitor in previous years used to enhance the sample distribution [7, 18, 25, 41]. The data retrieved from related projects from previous years included subsets of the activities used in this project thus enabling seamless integration to the database. The uniform collection protocols and storage formats used for the experiments allowed seamless concatenation of the datasets without requiring any alterations to the original sampling frequency or sensor placement positioning.

Real-world Data

The newfound dataset introduced in this project involved congregating time sequences that simulated real-world circumstances, generating a dataset representative of the data that would be gathered in a live application of the classification framework. The data collection protocol included recording volunteers performing a sequence of physical activities and social signals continuously without halting the recording between activity changes. To ensure that the dataset would include a balanced mixture of physical activities and social signals simulating a real-life situation, each participant was asked to perform a predefined sequence of activities chosen between two established patterns. The first sequence included: sitting on a chair; sitting bent forward; getting up from the chair; walking at a slow pace; standing still; reaching for a glass of water; drinking while standing; talking; walking at a slow pace; standing; coughing; walking at a normal pace; sitting down on a chair; sitting bent backwards; and coughing. The second sequence included: lying down prone; coughing; lying down facing left; coughing; talking; sighing; sitting; sitting bent forward; talking; standing still; coughing; walking at normal speed; standing; reaching for a glass of water; and drinking while standing. The participants were video-taped while performing the activities, with the captured video being utilised for the annotation of the sensor signals. The duration of each

activity performed was left to the discretion of the participant with an upper limit of 10 seconds, to ensure that the recording was as realistic as possible. The above-mentioned decision was taken as each sample was evaluated separately thus not affecting other recordings, while the video input enabled accurate identification of the start and end of each activity, allowing annotation of activities of varying duration. Once both recordings were gathered for each volunteer, the video was temporally aligned with the sensor data, allowing manual labelling of the activities performed at any instance. The accelerometer and gyroscope signals were exploited for the development of this dataset, with information being sampled at a frequency of 25Hz. In this study, data were collected from 6 participants in total, with participant statistics being displayed in Table 3.1. Each subject in the real-world dataset is also part of the previously-mentioned activity recordings dataset, allowing the application of the calibration methodology to evaluate performance gain. The participant information sheet and the participant consent form can be found in Appendices A and B, respectively.

Participant ID	**Sex**	**Age**
XXD001	Female	26
XXD002	Male	23
XXD014	Male	20
XXD015	Male	22
XXD016	Male	21
XXD021	Male	21
Total Participants	**Ratio (Male:Female)**	**Average Age**
6	5:1	22.2

Table 3.1: Participant statistics of the real-world dataset.

3.2 Data Processing

Data pre-processing is a fundamental component of a machine learning methodology, contributing to the enhancement of the performance and efficiency of the algorithm. This phase involves the removal of irrelevant, redundant, noisy and unreliable information found in the unrefined data samples, which could often increase the complexity of knowledge extraction and pattern recognition from the dataset. Consequently, the model is able to achieve superior generalisation performance, thus enhancing the handling of unseen input in real-world circumstances [30]. This section provides a detailed description of the data processing procedures applied to the dataset aiming to enhance the quality of the input samples as well as maximise the generalisability of the implemented algorithms.

3.2.1 Data Cleaning

The primary stage of the pre-processing pipeline included adjusting the class distributions to maximise the usability of the dataset and of the machine learning model. Due to the datasets serving different purposes in the calibration framework pipeline, different data cleaning strategies were carried out for each of them.

The analysis of the concatenated accelerometer-only dataset sampled at 12.5Hz in the MInf - Part 1 Project [5] has revealed a significant similarity between samples drawn from the physical activity classes of sitting and standing. This issue is often referred to as inter-cluster similarity in the literature and is observed when the data distribution of a number of categories significantly overlap, eventuating in samples drawn from the above classes being indistinguishable [48]. The correlation observed was a result of the sensor placement, which yielded similar recordings for both activities due to the position selected. As a consequence, the magnitude and direction of the motion in regards to the sensor seem identical for both movements thus not allowing differentiation between them. The solution chosen in the MInf - Part 1 Project [5] for this particular issue, which is also reproduced in this study, was to combine the above-mentioned activities into a singular class characterising both categories. The data analysis also revealed a sizeable intra-class variation between classes representing different inclinations of the sitting activity. The dissimilarity was mainly caused by the absence of a specific angle definition for each of the orientations, that was originally aimed to induce noise in the dataset and thus improve generalisability. In order to resolve this issue, all variations of the sitting activity were annotated as one class to avoid misinterpretation, as in the MInf - Part 1 Project [5]. Furthermore, after utilising the models developed in the MInf - Part 1 Project [5] for real-time classification, it was revealed that the omission of transitory activities including: standing up whilst sitting on a chair; sitting down on a chair; getting up from a lying down position; and, lying down from a seated position; was hindering the usability of the methodology in real-world applications. Therefore, it was decided to congregate the above-mentioned activities into a singular ”Movement” class in an effort to allow the model to recognise *Out-of-distribution* activities from the remaining classes. The class distribution of the physical activities following the data cleaning procedures is illustrated in Figure 3.3.

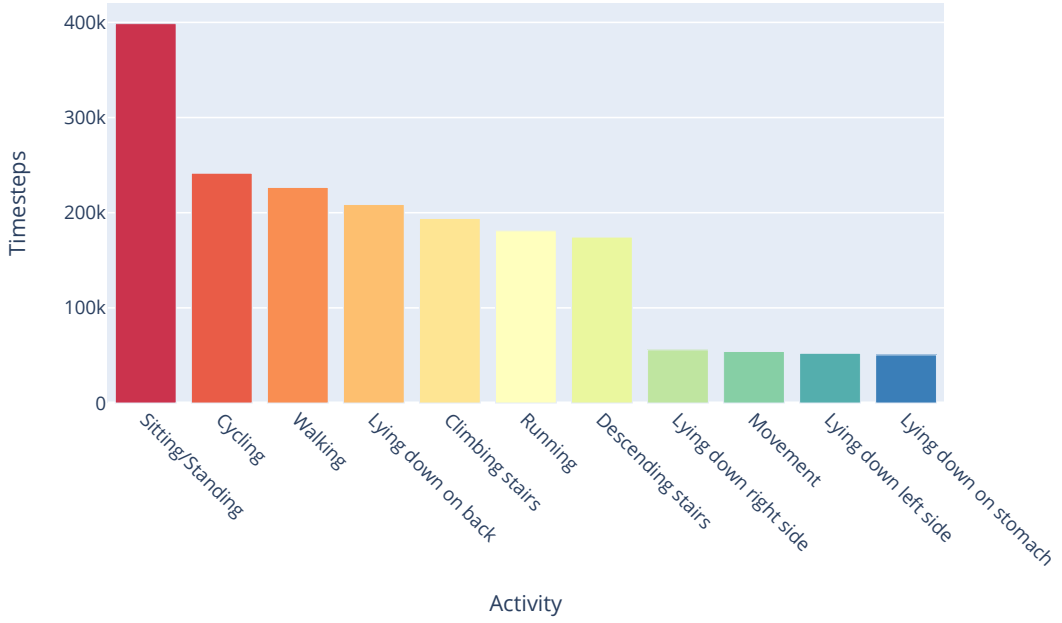

Figure 3.3: Distribution of the physical activity classes following data cleaning procedures

For the accelerometer and gyroscope dataset sampled at a frequency of 25Hz, a different cleaning strategy was utilised, in order to fulfil the objectives set for the usage of these data samples. A minimally invasive approach was adopted for the processing of the newly gathered dataset, allowing exploration of the potential benefits the introduction of angular velocity signals as well as the increase of sampling frequency may have on the performance of the classification model. In contrast to the procedure followed in the previously mentioned dataset, all distinct sitting and standing activities were kept unaltered, aiming to reveal potential improvement by the usage of more particularised categories. Similarly, to the above-noted methodology, transitory activities were grouped into a sole class enabling usage in real-time circumstances. The class distribution of the human activities deduced after the data cleaning procedures carried out on this dataset is displayed in Figure 3.4.

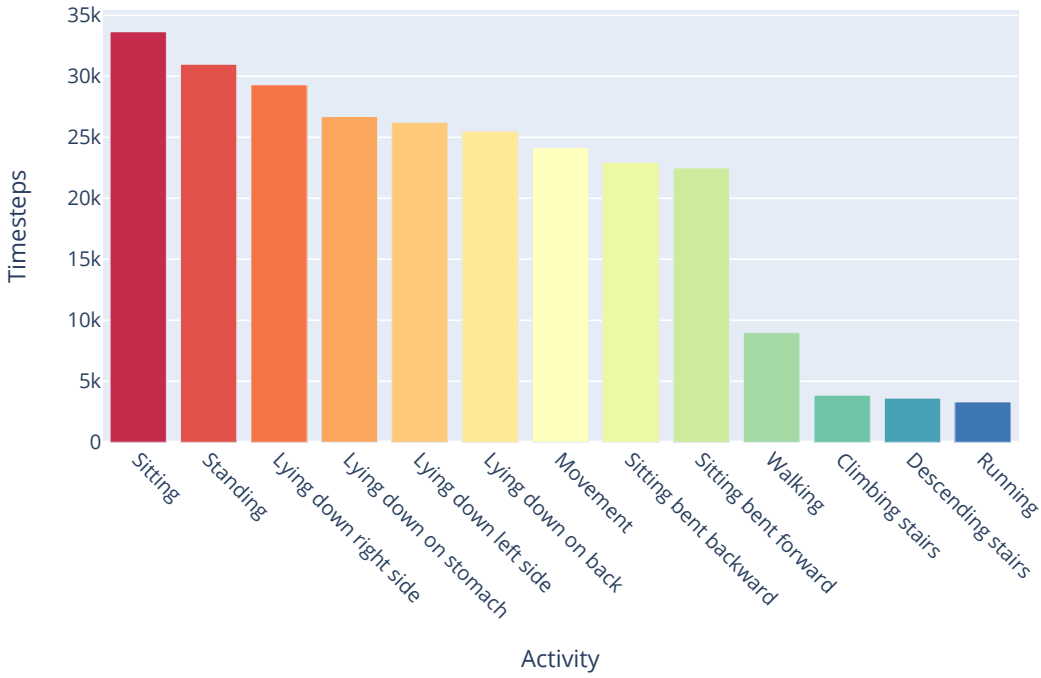

Figure 3.4: Distribution of the physical activity classes of the linear acceleration and angular velocity dataset following data cleaning procedures

In view of the real-world dataset being used for the evaluation of models trained on both datasets using the calibration methodology, two distinct iterations of the dataset were generated, in order to match the corresponding human activity ranges.

3.2.2 Temporal Alignment of Activity Boundaries

Data inspection carried out in the MInf - Part 1 Project [5] revealed a slight delay between the time the signal recording was initiated to the time the activity started being performed. The temporal delay was caused due to the recording procedure being manually initiated by the investigator which resulted in a deviation between the starting times of the recording and the activity initiation. The solution adopted to alleviate this particular issue was trimming 10% from the start of each recording, aimed at eliminating

this undesirable interval, hence resulting in a more accurately annotated dataset. A more aggressive approach was deferred in this particular case, as a way of keeping the numerosity of the dataset as close to the original as possible.

3.2.3 Noise Filtering

The following stage of the data pre-processing pipeline included identifying and removal of noisy signals and outlying elements that could affect the quality of the dataset. By virtue of using the activity recognition model on real-time data, no normalisation or standardisation techniques were applied to the data. The reason for this decision was that these methods require continuous updating of the mean and standard deviation of the dataset, thus requiring to load all the historical data for every time window, which introduces a significant computational cost. Alternately, the median filter algorithm was applied to de-noise the sensor data, which has been shown to offer excellent noise reduction as well as offering simplistic implementation allowing uncomplicated adoption in real-life applications [54].

A median filter applies a moving kernel along the data points, replacing each value with the median of the point itself and its direct neighbours. The recording is padded with zeros prior to the noise filtering to ensure the size of the output sample matches the length of the original input. The kernel size was set to 3 time-steps, minimising the added delay required before initiating real-time activity classification to approximately 0.16 seconds [5].

3.2.4 Sliding Windows

The following step in the data preparation methodology included framing the dataset into a collection of fixed-size time sequences establishing the framework for a supervised learning machine learning task. The above procedure was performed using the sliding window approach, which is the most widely employed segmentation technique applied to inertial sensor signals, and incorporates splitting the data into fixed-sized sliding windows with each being annotated by a specific physical activity. The sliding window methodology is also illustrated in Figure 3.5. In this project, a window size of approximately 3.84 seconds is used, which has been shown in the experiments performed in the MInf - Part 1 Project [5] to be the optimal balance between performance gain and responsiveness of the system, because as the window is prolonged the predictions become less dynamic. The sliding window segmentation technique is divided into two categories based on the step size between consecutive windows: Fixed-size Non-overlapping Sliding Windows (FNSW) and Fixed-size Overlapping Sliding Windows (FOSW) as described in [27]. Data overlap between adjacent windows aims at improving the detection of activities that may have been split between different windows resulting in significant loss of information. The overlapping among windows guarantees high numerosity of the dataset which may yield performance improvement, even though it raises the computational resources required for the training of the model [13]. For this project, a window overlap of 50% was chosen, balancing performance and increased computational demands of the system, while also demonstrating proven results in the literature [61].

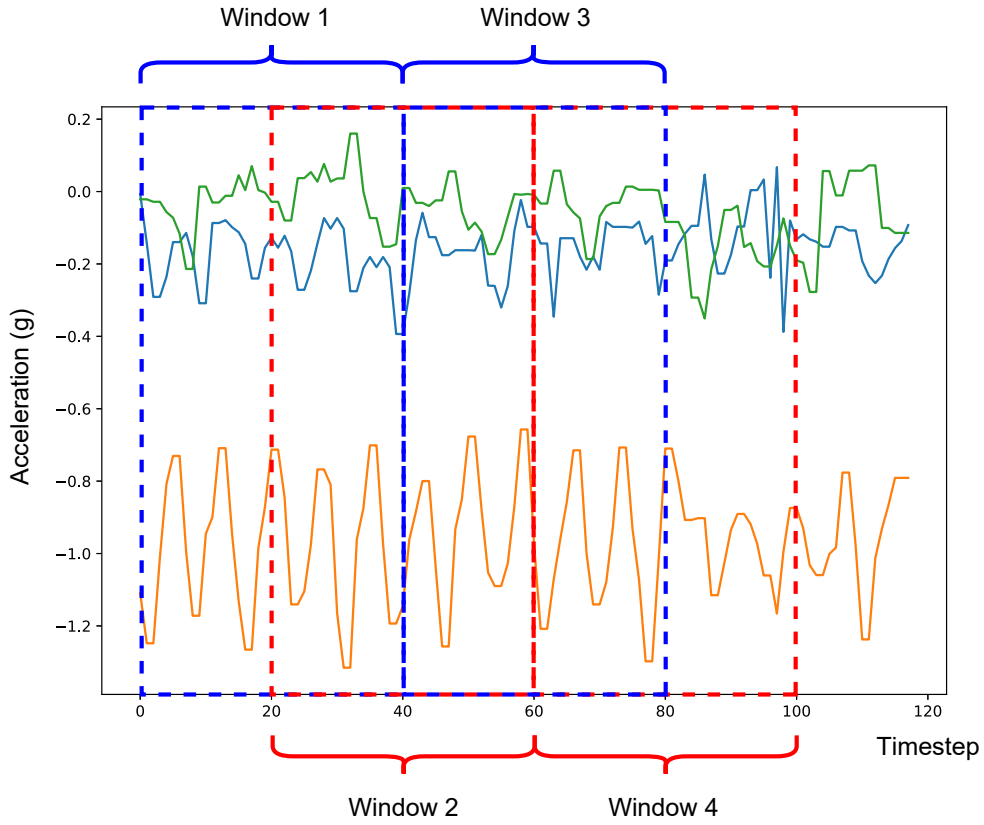

Figure 3.5: Sliding windows technique

3.2.5 Subject Independent Cross-Validation

Once the aforesaid processing operations were performed, a resampling procedure was implemented ensuring that the evaluation results reported were not dependent on a specific subset of the dataset, ensuring the model would generalise well on unseen data. A resampling technique widely used for human activity recognition in the literature, known as K-Fold Cross-Validation, involves randomly partitioning the dataset into k subsets, with $k-1$ of them used for the training of the model and the remaining fold used for testing purposes [21]. The testing fold is then iterated across all the partitioned subsets in order to evaluate the performance of the model for each combination of subsets. The random generation of subsets during the procedure means that the training and testing subsets may contain data from the same subject, therefore, this method is known as Subject-Dependent Cross-Validation [2]. This technique, however, assumes that the samples are *Independent and Identically Distributed (IID)*, even though this cannot be guaranteed for samples drawn from one subject, which are highly likely to show signs of correlation. This interdependence is often justified by two causes:

- *Intra-subject dependencies* in the data, signifying a higher correlation between samples extracted from the same subject in comparison to samples drawn from different subjects. This dependency indicates similarity in the way physical activities are conducted by a subject, which can be explained by the physiological characteristics of each person such as age and sex, which might affect the move-

ment patterns of an individual. In addition, the way an activity is performed may have been influenced by previous experience performing the movement, which can define its difficulty for each individual [4].

- *Temporal dependence* between activities performed by one subject, indicating similarity between samples drawn within a short time interval. This connection is likely caused by several factors such as physical and/or mental weariness which may cause deterioration in the quality of the recordings as time proceeds. [13].

As a consequence, k-fold Cross-Validation often leads to an overestimation of the performance of the model when used in human activity recognition, with an artificial increase in the classification accuracy caused by similarities between training and evaluation sets from samples drawn from the data distribution of a single subject [12]. To address the aforementioned problem, a *Subject-Independent Cross-Validation* technique is employed, dividing the dataset into subsets by subject [31]. This mechanism splits the subjects into folds, with each fold containing the full data of the allocated subjects. Therefore, intra-subject dependencies observed in k-fold Cross-Validation are no longer an issue during the evaluation procedure.

For the purposes of this project, the subjects are distributed into 5 folds dividing the dataset into partitions of 20%, influenced by the results shown in the MInf - Part 1 Project [5]. In order to evaluate the performance of the classification framework, explored in detail in the following chapter, reporting results on unseen data, the training set is further divided, with the last generated fold forming a validation set as shown in Figure 3.6. For every iteration, each model architecture is trained on the training set with the best model being extracted based on the performance on the validation set. Subsequently, the selected model is applied to the testing set to report performance as explained in detail in the next chapter.

3.3 Exploratory Data Analysis

Progressing the extensive exploratory data analysis carried out on the dataset in the MInf - Part 1 Project [5], in this project an investigation of the effects of the inclusion of angular velocity signals is carried out in order to improve the understanding of the dataset and guide the initial hypothesis to achieve optimal results. The analysis included applying a dimensionality reduction technique utilised to map the input samples into a more comprehensible two-dimensional space, thus allowing further study of the feature space. For this research, the Uniform Manifold Approximation and Projection (UMAP) [38] algorithm was exploited, which has demonstrated valuable results in a wide array of applications [50]. The UMAP dimensionality reduction technique works by firstly constructing a high dimensional graph representation of the dataset and then subsequently optimising a low-dimensional graph to be as structurally similar as possible to the original graph. For the purposes of this research, the newly concatenated acceleration and angular velocity dataset was exploited, by applying the UMAP method on the acceleration samples only on the one hand, and on acceleration and gyroscope signals on the other.

The projections of the two dataset instances into 2D spaces are shown in Figures 3.7

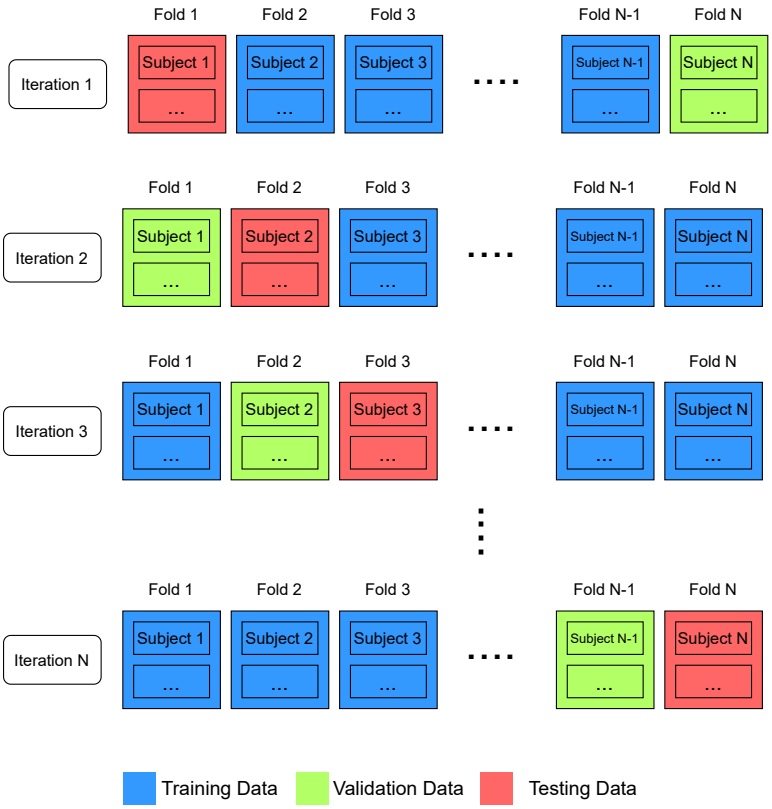

Figure 3.6: Subject-Independent Cross-Validation methodology

and 3.8. As it can be observed, the data samples in the dataset instance using only linear acceleration are more easily separable than the instance that includes angular velocity. This phenomenon occurs because the additional information retrieved for each time-step introduces increased complexity in the data thus increasing the difficulty in finding decision boundaries between the classes. It can also be noted that the stationary activities are projected into different spaces in the acceleration-only dataset forming uniform clusters, which indicates that their identification is easier in comparison to the dynamic activities. The closely related projections for non-stationary activities in both dataset instances as well as the broad scattering of projections sampled from the same classes, however, determine the deep learning methodology as the optimal configuration for this particular task. Deep learning methodology enables efficient feature extraction of the most relevant features in the dataset [11] in contrast to manual feature engineering that might potentially result in less informative features, while also allowing domain-agnostic implementation which is a requirement for this project.

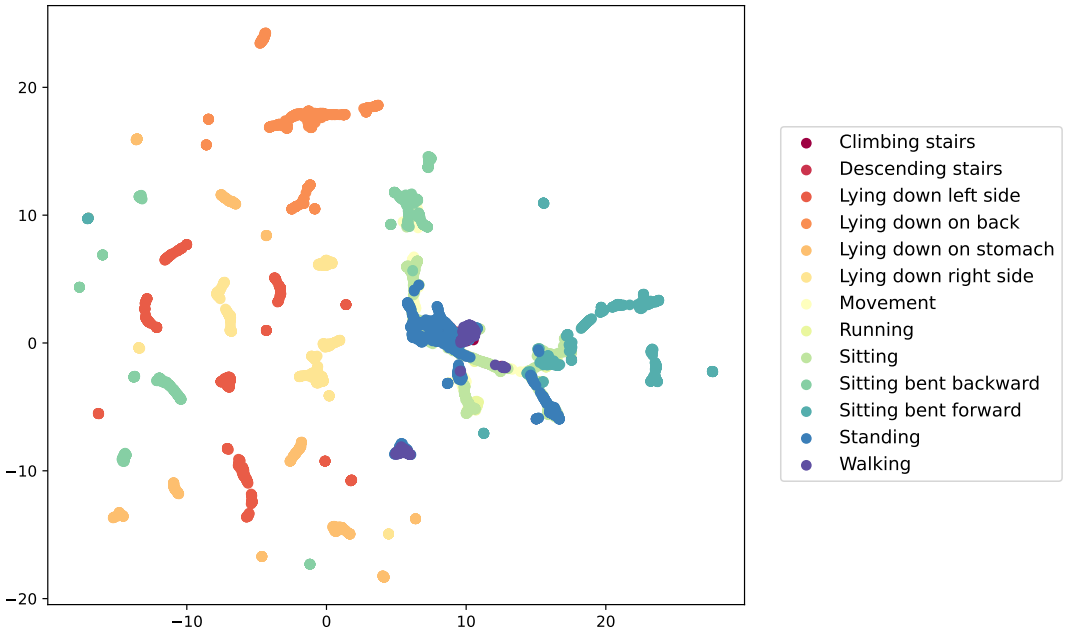

Figure 3.7: Projection of Acceleration Signals using UMAP

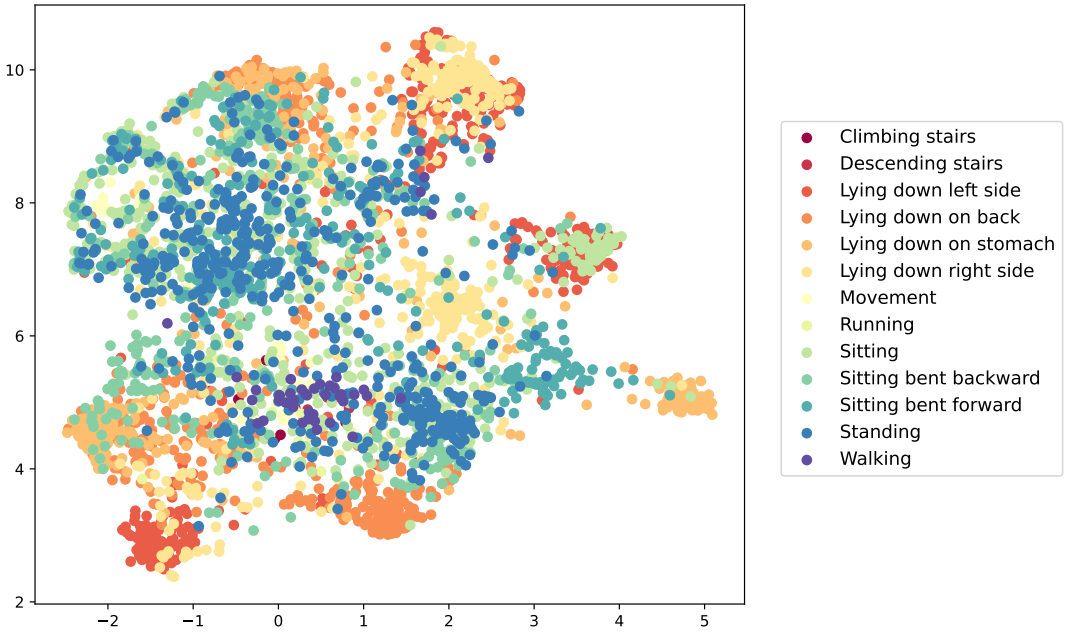

Figure 3.8: Projection of Acceleration and Gyroscope Signals using UMAP

Chapter 4

Methodology

Even though in recent years vast progress has been observed in the research area of human activity recognition; the significantly high cost required to acquire labelled data both in terms of human labour as well as in time has posed a great challenge in applying the classification methodology in real-time data or for applications where the range of activities includes uncommon activities, such as recognising a seizure event. The following chapter provides a detailed description of the methodology applied to develop the domain agnostic human activity recognition calibration system, aimed at alleviating the aforementioned problem, including justification of the decisions taken during the design and implementation phases. The contributions of the current project explored in the following chapter include:

- implementation of deep learning network architectures utilised for the evaluation of the human activity classification framework;
- design and implementation of a Transformer-based deep learning model for human activity recognition;
- development of a testing framework used to assess the impact of adding tri-axial angular velocity and increasing the sampling frequency of the initial tri-axial acceleration inputs;
- design and implementation of a novel domain agnostic Transformer-based calibration system utilising transfer learning methodologies.

4.1 Classification Algorithms

A critical component of the human activity calibration framework is the machine learning model used for the activity recognition which is fundamental for the performance on the data prior to the application of the transfer learning methodology as well as the pattern learning once the calibration samples are introduced. After a thorough examination of the literature, it was observed that a number of deep learning architectures performed similarly well in the human activity classification task, with the results varying based on the particular dataset being applied and the hyper-parameter configuration used, thus

showing that there is no single approach that is optimal for all cases.

This section presents a series of different model configurations applied for human activity recognition, assessing the benefits and limitations associated with each method and comparing their performance in order to find the most suitable algorithm for this setting.

The implementation of each of the neural networks was carried out using the TensorFlow framework [1], version 2.4.0, and the Keras application programming interface [8], version 2.2. The TensorFlow framework was chosen owing to the fact that it enables direct conversion of the trained models to a format that can be employed on edge devices through the TensorFlow Lite framework [49], which is important for the employment of the system in a real-time setting.

4.1.1 Evaluation Metrics

In order to evaluate the performance of each classification model, a range of metrics were used to ensure that the results would be useful and meaningful for real-world application. Due to the fact that the models are trained on a multi-class classification task, focusing solely on the accuracy of the model when using an imbalanced dataset, may jeopardise its ability to learn all the classes uniformly, by focusing on a single class to improve accuracy. To mitigate the above risk, the following performance metrics were implemented in this project:

- *accuracy*, measuring the number of correctly identified activities, denoted by

$$accuracy = \frac{True\,Positives + True\,Negatives}{True\,Positives + False\,Positives + True\,Negatives + False\,Negatives}$$

- *precision*, quantifying the number of predictions of a class that actually belong to that specific class, denoted by

$$precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

- *recall*, measuring the number of positive class predictions made out of all positive examples in the dataset, denoted by

$$recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

- *F1-Score*, which is the harmonic average of precision and recall metrics, denoted by

$$F1\text{-}Score = 2 * \frac{precision * recall}{precision + recall}$$

The above metrics were evaluated for every iteration of the Subject Independent Cross-Validation procedure, with the average accuracy and F1-Score metrics being employed for the analysis of the results demonstrated in the following chapters. Each model is then subsequently chosen based on the mean of the accuracy and F1-Score average scores, balancing the 2 objectives.

4.1.2 Baseline Models

In order to establish a performance threshold providing a point of reference for the assessment of the classification algorithms, two different baseline models were implemented, providing reasonable results as well as a simplistic set-up. A baseline model is vital for any machine learning task not only to evaluate whether the complex methods applied yield superior performance but also as a comparison metric to examine whether the trade-off between performance difference and computational cost can be justified.

For the purposes of this study, a *ZeroR* classification model was utilised, predicting the majority class of the training dataset for all inputs. Even though this model doesn't have any predictability power given that it completely ignores the input data and its output is based solely on the target labels, it can still provide a valuable benchmark when training on unbalanced data, which is the case in our dataset.

The second baseline model employed was a *Random Forest Classifier*, which is a machine learning approach that builds an assembly of decision trees based on the training dataset, and outputs the prediction class accumulating the highest number of votes gathered from all the individual trees. The advantage of using a random forest model in lieu of a single decision tree classifier is that the importance of each feature of the data in the prediction is proportional to the number of occurrences in the ensemble, thus features portraying noisy parts of the data are de-emphasised, resulting in a model less prone to overfitting. In this project, the Random Forest Classifier is implemented using the `scikit-learn` framework [44], version 0.24.1, using the default hyperparameters, setting the number of decision trees to 100.

4.1.3 Auxiliary Classifier Generative Adversarial Network

The first deep learning architecture implemented for this project was the Auxiliary Classifier Generative Adversarial Network (AC-GAN) which was introduced in MInf Project Part 1 and was able to outperform all other state-of-the-art approaches evaluated in the paper [5].

The class of Generative Adversarial Networks (GANs) was first introduced as a way of tackling the issue of limited availability of data that was inhibiting the performance of supervised machine learning algorithms, by generating synthetic samples to improve generalisability on unseen data [19]. GANs frame the unsupervised learning task of generating realistic synthetic data based on patterns seen in the training set, known as generative modelling, into a supervised learning task by combining a sample generation model with a classification model. The generation model, also known as the generator, is given a randomly generated vector in a high-dimensional space with the objective of producing realistic samples. The classification model, also known as the discriminator, is then supplied with data samples aiming to distinguish whether the input was drawn from the dataset distribution or from the distribution of the synthetic samples that were generated using the generator model. The above-mentioned models are trained together in a zero-sum game until a stable solution is reached, where the models cannot further improve their objectives, and the discriminator is deceived by the generator samples on half of the iterations.

The architecture of the Auxiliary Classifier GAN used for human activity classification in this project is outlined in Figure 4.1. This specialised extension of the GAN architecture includes a generator model that is given not only a vector randomly projected into a latent space but also a specific activity category, requesting the network to generate samples solely based on the patterns learned for that particular activity, thus generating windows giving the impression of being drawn from the distribution of that class. The output of the network is a three-dimensional time sequence window sample matching the shape of the actual input, that in combination with samples drawn from the original dataset is transferred as an input to the discriminator model. The discriminator is trained firstly on the task of distinguishing whether the input samples are real or synthetic, and secondly on the task of classifying the input window into the range of chosen human activities.

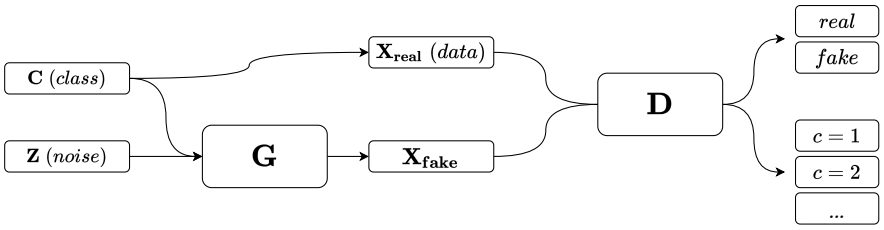

Figure 4.1: Auxiliary Classifier Generative Adversarial Network architecture

The configuration used for the model implementation was sourced from the parameter and hyper-parameter optimisation carried out in the MInf - Part 1 Project [5], and is outlined below. The architecture of the generator model used to map randomly drawn points from a dense high-dimensional space into time sequence windows that could have been plausible samples from the original data recordings distribution is displayed in Figure 4.2. The latent space has a dimensionality of 100 dimensions, and each variable is drawn from a unit Gaussian distribution. The class label is projected using a fully-connected layer into the shape of the latent space and subsequently concatenated to the random noise vector. The input is then processed using 5 Transpose Convolutional blocks, that are used to upsample the input feature map to the desired output shape. Except for the last block, each of the previous blocks consists of a Transpose Convolution, a Batch Normalisation layer and a Leaky Rectified unit (Leaky ReLU) layer which is used as an activation function [58] to inject non-linearity. The Leaky ReLU implemented, uses 0.2 as the slope of the activation function for negative values. The last block uses a tanh activation function to restrict the output of the generator to a range between -1 and 1 [40]. The activation functions are used to transform the representation learned from previous layers via non-linear combinations, in order to enable non-linear decision boundaries and thus allow the network to recognise more complex data structures.

The discriminator model configuration is also acquired from the architecture derived in the MInf - Part 1 Project [5], and is displayed in Figure 4.3. The neural network receives a sliding window time sequence that might have been drawn either from the dataset distribution or from the synthetic samples created by the generator, and using a series of convolutional blocks, recognises patterns in the input in order to determine the activity as well as the genuineness of the sample. The input window is processed

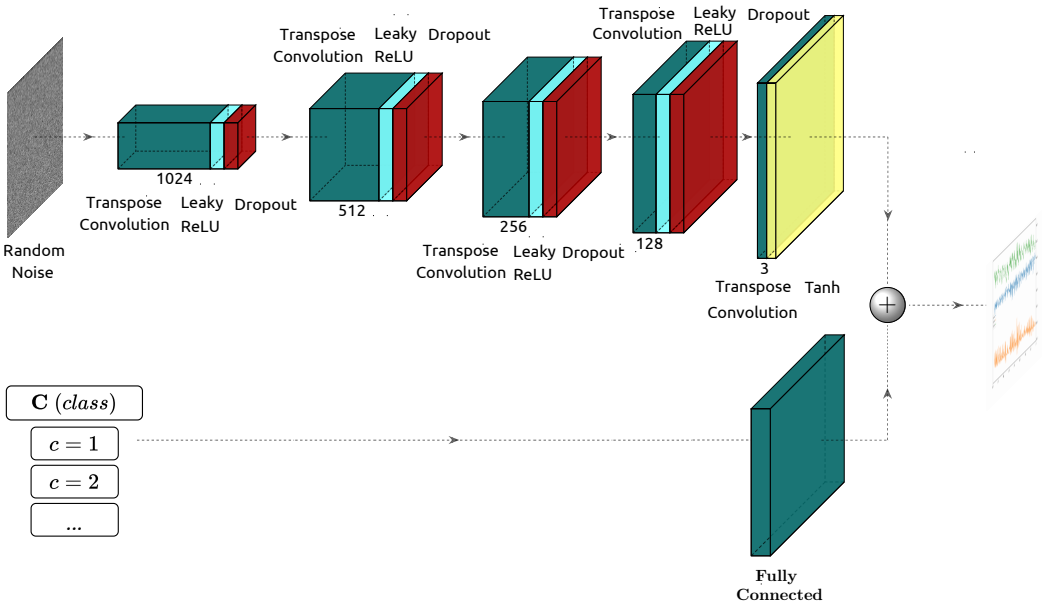

Figure 4.2: AC-GAN Generator Architecture

through 2 convolutional blocks, that consist of a convolutional layer, interleaved with a Leaky Rectified unit (Leaky ReLU) activation layer and a Dropout layer. The Leaky ReLU replaces negative gradient values with a negative of 0.2, and the possibility of omitting a neuron in the Dropout layer is set to 0.5, keeping the original settings presented in the MInf - Part 1 Project [5]. The dropout layer aims at preventing certain neurons in the network from overpowering the remaining neurons by diluting their weight, thus reducing dependency on specific neurons and improving robustness against noisy data. The output of the final convolutional block is then shared between two output layers, the first being utilised for sample authenticity detection and the second for human activity classification. The output layers use a sigmoid activation function and a softmax activation function, respectively, converting the feature map into a probability distribution with each being chosen based on the number of outputs required.

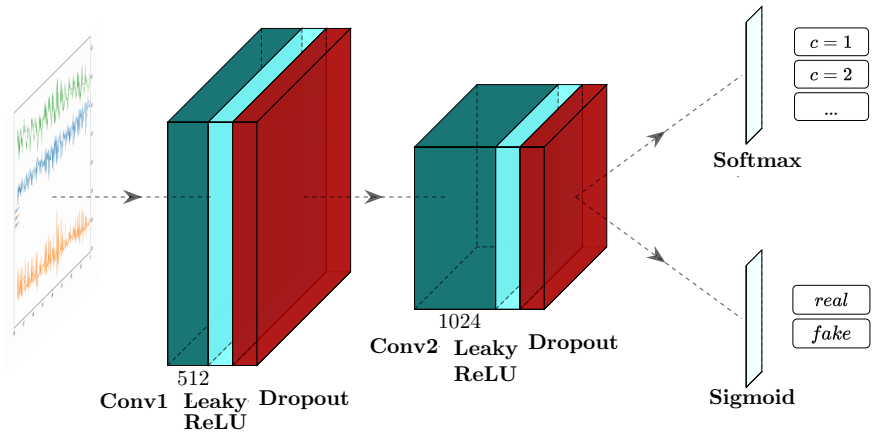

Figure 4.3: AC-GAN Discriminator Architecture

In each training epoch, the discriminator is trained using half batch of real samples and

half batch of generated samples, for both activity and realness classification. Subsequently, the discriminator is frozen and the model is trained using randomly generated class labels as well as noise vectors present in the latent space to improve the performance of the generator model. The samples forming each batch are drawn with a replacement on each iteration to induce randomness whilst shaping the weight coefficients.

The class label used as input to the generator is randomly drawn from a uniform distribution of the activity classes, to ensure that all categories are trained equally. The model is trained for 10,000 mini-batches, each consisting of 128 samples, using the Adam version of stochastic gradient descent to optimise the network [29]. The final performance of the model is reported using 5-Fold Subject Independent Cross-Validation.

4.1.4 Transformer Model

In recent years, transformers have achieved state-of-the-art performance in a wide array of machine learning problems [51], including natural language processing [57], time series forecasting [35] and computer vision [14]. In this project, the transformer architecture is utilised for human activity classification, motivated by the results shown in [47], in an attempt to utilise its predictive capacity, attempting to surpass the performance achieved using the previously-mentioned Auxiliary Conditional Generative Adversarial Network configuration presented in the MInf - Part 1 Project [5].

Self-Attention Mechanism

The key to the ground-breaking performance of the Transformer architecture lies in its use of the attention mechanism. Attention is a component within a neural network that manages and quantifies the interdependence between vectors within a sequence. The transformer network leverages the self-attention mechanism in order to capture context within the source sequence, by correlating different positions of the input in order to compute a representation of the same sequence. Self-attention methodology computes the similarity of a data point in comparison to all other data points, and transforms the representation of each time step using information from the remaining timesteps according to their importance.

Multi-Head Attention

A critical component of the implementation of the transformer methodology is the multi-head attention functionality. Multi-head attention refers to the implementation of distinct self-attention heads aiming to capture different perspectives of the input sequence, with each head being calculated independently, thus enabling parallel processing [36]. Each one of the attention heads uses distinct parameters to represent the relevance scores between the time-steps, thus capturing different relations within the input window. For instance, one self-attention head might be used for capturing shorter-term dependencies whilst another head might be utilised for attending longer-term dependencies. The

outputs from the distinct attention heads are subsequently concatenated and converted back to the dimension of a single attention head using a linear transformation.

Positional Embeddings

The self-attention mechanism, however, ignores the sequential nature of the input, considering the source as a set instead of a sequence. As a result, any permutation of the same input sequence will output exactly the same result, except permuted also, i.e. self-attention is *permutation equivariant*. In order to mitigate this issue, a positional embedding vector is introduced for each sequence, materialising the relative position of each token in the input. The positional embeddings are added to the input sliding window, once the latter is mapped into a dense vector space using convolutional layers, to ensure that the positional encoding remains unaffected after the feature mapping, enhancing the embeddings by injecting the order of each timestep.

The positional embedding needs to satisfy the following criteria to ensure applicability in multiple scenarios:

- output a unique encoding for each position;
- ensure that the time-step delta is consistent across sequences, to ensure that the distance between two time-steps is consistent across varying length sequences;
- ensure that the output values are bounded within a specified range to ensure generalisability in longer sequences;
- should be deterministic to allow reproducibility.

The positional embedding technique acquired for this project was firstly presented by the authors in [52] and satisfies all of the above-mentioned criteria. The encoding function is given by:

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{encoding}}) \tag{4.1}$$

$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{encoding}}) \tag{4.2}$$

where pos represents the desired position within the input sequence, $d_{encoding}$ is the dimensionality of the encoding and i is the index to each one of the dimensions of the encoding.

Latent Sequence Aggregation

The following component of the transformer architecture includes a latent sequence aggregation layer, which utilises the representation learned by the self-attention blocks for each time-step to rank each position according to its importance in determining the class of the input time sequence. The aggregation layer computes a weighted summation of the sequence in each time-step based on the relative importance of all other time-steps, which is the feature space extracted from this layer. The extracted feature mapping is then passed into a fully-connected layer, that outputs the probability distribution for activity categories, using the softmax activation function.

Architecture

Combining the above-mentioned components presented with the transformer architecture has introduced a range of benefits that enabled transformers to achieve state-of-the-art performance in multiple areas of machine learning. The advantages over other network configurations include:

- **improved performance in capturing long-range context dependencies**; in contrast to recurrent neural networks, they don't rely on a memory state, which is biased by the most recent inputs in the sequence, to convey temporal information. i.e. all time-steps are directly connected to every other time-step in the input;
- **reduced training time**; enabling parallel computation of each attention head thus processing more data at the same time;
- **minimal inductive biases** usage, allowing the models to learn without being restrictive [28].

The architecture of the transformer model applied for human activity classification is presented in Figure 4.4. The neural network receives a batch of sliding window time sequences and applies a series of 4 one-dimensional convolutions to embed the raw data into a higher dimensional vector space, generating a latent embedding of the inputs. The convolution block is aimed at replacing the word embeddings methodology found in the original transformer architecture applied for language processing [51], which projected each word token into an embedding space. Similarly, the convolutional layers project each time step from the input windows into a vector space, allowing usage of the subsequent components found in the transformer architecture. The GELU non-linearity is also applied after every convolutional layer as an activation function, inducing non-linear combinations to the data structure [23]. The positional embedding is then learned using Equations 4.1 and 4.2, and is subsequently added to the latent sequence representation. The combined embeddings are then processed through a Transformer Encoder block, which is repeated N times. Each transformer block consists of a multi-head self-attention sub-block which also includes a dropout layer as well as a layer normalisation and a feed-forward network sub-block, consisting of 2 one-dimensional convolutions, interleaved with a dropout layer and a layer normalisation, respectively, with GELU non-linearity also being applied to the first convolution. A residual connection is finally added around each sub-block, in order to improve model convergence and gradient flow [22]. The output of the ensemble of transformer encoders is then passed into a Global Average Pooling layer, that computes the weighted aggregation of the input sequence mapping the sliding window into a hidden space. The extracted feature mapping is then passed into a fully-connected layer that outputs the probability distribution for activity categories using the softmax activation function.

Class Balancing

In order to reconcile any imbalances observed in the number of time samples available for each activity class in the dataset, over-sampling, as well as under-sampling techniques, are employed in an attempt to improve and balance pattern recognition across all activities in the classification task. The class balancing techniques are applied

only to the training dataset, to ensure that the validation set is kept the same, enabling direct comparison of the methodologies. The over-sampling technique is carried out by generating new samples in the classes which are under-represented by sampling from the distribution of the category. The under-sampling method is carried out by randomly sampling window sequences from each class until the size of the category bin is equal to the size of the least represented class label. The implementation for both balancing strategies is implemented using the `imbalanced-learn` library, version 0.9.0 [34].

Hyper-parameter Tuning

The concluding stage in developing the transformer-based neural network included tuning a range of hyper-parameters found in the model, to ensure that the network is optimised for this particular task and dataset. The hyper-parameter tuning was conducted using the grid search tuning technique by exhaustively searching through all the combinations of hyper-parameters displayed in Table 4.1. The hyper-parameters explored include the number of attention heads used in each Multi-Head Attention layer in the Transformer Encoder, the number of Transformer Encoder blocks used, the dropout rate used within the transformer encoder, the embedding size of each attention head and finally the class balancing strategy. The final feed-forward layer was set to 512 hidden units, with a dropout rate of 0.1. Each iteration of the neural network was trained for 300 epochs, using the early stopping technique with patience set to 150 epochs, to stop training the model when the validation F1-Score does not improve in the last 150 epochs. For the purposes of hyper-parameter tuning, the dataset was split into subject folds using 60% of the subjects for the training set, 20% of the subjects for the validation set and 20% of the subjects for the testing set. The final performance of the optimised model is reported using 5-Fold Subject Independent Cross-Validation.

Hyper-parameter	Values
Number of Attention Heads	[8, 16]
Number of Transformer Encoder Blocks	[4, 8]
Dropout Rate	[0.1, 0.2]
Attention Head Size	[64, 128]
Class Balancing Strategy	[None, Under-Sample, Over-Sample]

Table 4.1: Hyper-parameter Tuning Configurations

4.2 Data Augmentation

Once the classification algorithm based on the transformer methodology was optimised, the following set of experiments was carried out, aiming at investigating how increasing the sampling frequency as well as incorporating tri-axial angular velocity to the linear acceleration readings affected the performance of the model. For the purposes of this experiment, the newly collected dataset containing accelerometer and gyroscope readings at a 25Hz sampling frequency was processed into the following versions:

- **acceleration-only data sampled at 12.5Hz**; generated by down-sampling the data recordings and removing angular velocity information;

- **acceleration-only data sampled at 25Hz**; created by removing angular velocity information and keeping original sampling frequency;
- **acceleration and angular velocity sampled at 25Hz**; keeping the original sampling rate and all the information captured.

For each version of the dataset, the transformer model was trained using an exhaustive search of all possible combinations of hyper-parameters shown in Table 4.1, to accommodate for differences in data that might have required a different model configuration for optimised performance. Due to the small size of the dataset, it was decided to use the over-sampling class balancing technique for all permutations to ensure that no information was lost during pre-processing. In order to ensure that the time sequences represented the same activity samples, a window size of 96 timesteps was used for the datasets sampled at 25Hz, instead of the 48 timesteps used for the dataset sampled at 12.5Hz, allowing them to capture inputs of 3.84 seconds, which was shown to be the most suitable for this project in the MInf - Part 1 Project [5].

4.3 Calibration System

Aggregating the experimental results accumulated through the afore-mentioned studies, the models yielding the best performance for the acceleration-only dataset presented in the MInf - Part 1 Project [5] were selected as the foundation for the calibration system development, based on transfer learning.

For the implementation of the calibration framework, the models trained using 5-Fold Subject Independent Cross-Validation are initially employed. The real-world data collected from subjects in this project are then utilised to evaluate the performance improvement yielded by the use of the calibration methodology. For each subject in the real-world data, the appropriate base classification model is selected, ensuring that no data samples used for either training or validation purposes were drawn from the data distribution of the particular subject tested. Thereafter, a number of sliding windows from the subject's training data are used to train the base classification model in an attempt to recognise patterns unique to the subject, with the remaining data used for validation purposes. The number of sliding windows selected is based on the experiments carried out in the MInf - Part 1 Project [5], showing that approximately 3 sliding windows with 50% overlap for each activity are optimal for this particular application, balancing performance improvement and data requirement. The base classifier is trained for a further 30 epochs using the calibration samples, extracting the configuration with the best performance on the validation set.

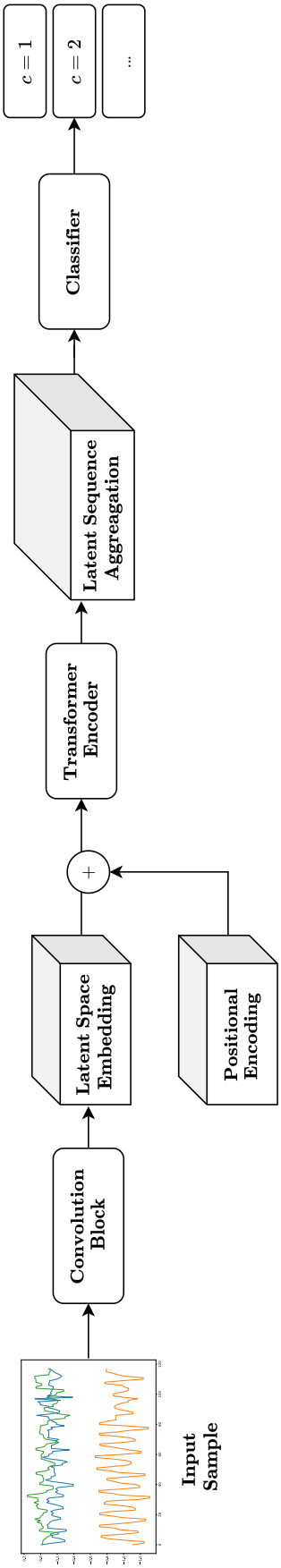

Figure 4.4: Transformer Network Architecture

Chapter 5

Results

Once the evaluation framework was established, the hyper-parameter tuning, as well as the subject-independent cross-validation testing were performed, assessing the performance of each of the methodologies described in the preceding chapters. This chapter encapsulates the results accumulated throughout the experimenting phase along with an interpretation of their significance to this research.

5.1 Human Activity Recognition

The initial component of the evaluation procedure involved analysing the classification performance of the implemented machine learning models in the aforenamed human activity recognition task, using the acceleration-only dataset sampled at 12.5Hz frequency.

5.1.1 Hyper-parameter Tuning

In pursuance of achieving the optimal results with the newly designed transformer-based methodology, the evaluation procedure was commenced with a hyper-parameter tuning step, ensuring that the optimised version of each algorithm is used for comparison, revealing the capabilities of each system.

Following the grid search parameter tuning technique explained in the Methodology chapter, the set of parameters achieving the highest F1-Score was selected as the optimal configuration for the transformer model. The best performing configuration achieved an accuracy of 87.02%, and an F1-Score of 86.97% on the test set. The aforementioned hyper-parameters are displayed in Table 5.1. It is observed that a higher number of attention heads enables the model to capture inter-dependencies within the time sequence better, thus improving the classification performance. On the other hand, increasing the number of transformer encoder blocks resulted in the models overfitting to the dataset, as the network consisted of a significantly higher number of parameters. Class balancing strategies did not yield any improvements to the predictive performance, which indicates that even though the dataset is imbalanced, the deviation between the class sizes is not large enough to lead to the training of one class overshadowing the

remaining categories. The use of a larger attention head has also deteriorated the classification accuracy possibly due to the increase of the network neurons which led to overfitting to the training data samples.

Hyper-parameter	Optimal Value
Number of Attention Heads	16
Number of Transformer Encoder Blocks	4
Dropout Rate	0.1
Attention Head Size	64
Class Balancing Strategy	None

Table 5.1: Hyper-parameter Tuning Configurations

5.1.2 Subject-Independent Cross-Validation

Once the hyper-parameter tuning was concluded, the different classification algorithms introduced in the Methodology chapter were evaluated using subject-independent cross-validation, in order to obtain an indicator of the average performance of each model, along with a confidence interval used to assess the discriminating power of each architecture. The results obtained from the 5-fold subject-independent cross-validation are displayed in Table 5.2.

Model Architecture	Accuracy	F1-Score
ZeroR	0.2583 ± 0.0083	0.0373 ± 0.0009
Random Forest	0.7646 ± 0.0261	0.7209 ± 0.0253
AC-GAN	0.8419 ± 0.0143	0.8401 ± 0.0160
Transformer	0.8426 ± 0.0232	0.8417 ± 0.0232

Table 5.2: Subject-Independent Cross-Validation Evaluation Results

Test Statistic

In order to assess whether the differences in performance between the Transformer model and the AC-GAN model are statistically important, a statistical test was carried out, with the results shown below. For the hypothesis testing, a two-tailed t-test was utilised, with the assumption that the difference between the scores of the two systems was drawn from a Normal Distribution. The hypothesis examined for the t-test is shown below:

$$H_0 : \overline{A} = \overline{B}$$
$$H_1 : \overline{A} \neq \overline{B}$$

where, H_0 is the Null Hypothesis and H_1 is the Alternative Hypothesis, A is the random variable representing the scores of the Transformer model, and B is the random variable representing the scores for the AC-GAN model, respectively. The test statistic is given by:

$$t = \frac{\overline{A-B}}{\sigma_{(A-B)}}\sqrt{N}$$

where, $\sigma_{(A_B)}$ is the standard deviation of $A - B$ and N is the number of samples.

The p-values calculated for each evaluation metric are displayed in Table 5.3. As it can be observed, none of the p-values is less than the significance level $\alpha = 0.05$, thus there is no evidence to reject the null hypothesis H_0, which means there is no evidence that the Transformer model is, statistically, significantly better than the AC-GAN model on any of the evaluation scores used.

	Accuracy	F1-Score
p-value	0.9325	0.8467

Table 5.3: The p-value from the two-tailed t-test for the Transformer and AC-GAN models for each metric score.

The cross-validation results, as well as the statistical test, have shown that the deep learning methodologies exploited have performed similarly well on the human activity classification task with the confidence intervals being comparable in magnitude. This conclusion lies in agreement with the observations seen in the MInf - Part 1 Project [5], where the above-noted similarity was also found in the experiments. The inability of any of the implemented methodologies to significantly outperform the remaining algorithms can be justified by the size of the dataset used, which is possibly not large enough to exploit the capabilities of the Transformer model to its fullest extent.

5.2 Data Augmentation

The following set of experiments included transforming the newly gathered accelerometer and angular velocity dataset into 3 distinct formats, in an attempt to assess how increasing the sampling frequency or incorporating tri-axial angular velocity to the accelerometer readings affected the performance of the model.

5.2.1 Hyper-parameter Tuning

The primary step in the investigation included hyper-parameter tuning performed separately for each model trained on each dataset instance, in order to ensure that specific requirements raised by the needs of each dataset configuration were met prior to testing, securing equal opportunities for all the models.

Acceleration - 12.5Hz

The first experiment was carried out on the dataset including acceleration data only, under-sampled at 12.5Hz frequency. Following the grid search parameter tuning, the optimal set of hyper-parameters was chosen and is displayed in Table 5.4.

Acceleration - 25Hz

The second experiment was carried out on the dataset including acceleration data only, sampled at the original 25Hz frequency, with the optimal hyper-parameter configuration being displayed in Table 5.5.

Hyper-parameter	Optimal Value
Number of Attention Heads	16
Number of Transformer Encoder Blocks	8
Dropout Rate	0.1
Attention Head Size	64

Table 5.4: Hyper-parameter Tuning Configuration for acceleration-only at 12.5Hz

Hyper-parameter	Optimal Value
Number of Attention Heads	16
Number of Transformer Encoder Blocks	8
Dropout Rate	0.2
Attention Head Size	64

Table 5.5: Hyper-parameter Tuning Configuration for acceleration-only at 25Hz

Acceleration and Angular Velocity - 25Hz

The final hyper-parameter tuning experiment was carried out on the complete dataset including acceleration and angular velocity data, sampled at 25Hz frequency, with the optimal hyper-parameter configuration being displayed in Table 5.6.

Hyper-parameter	Optimal Value
Number of Attention Heads	8
Number of Transformer Encoder Blocks	8
Dropout Rate	0.2
Attention Head Size	128

Table 5.6: Hyper-parameter Tuning Configuration for acceleration and angular velocity at 25Hz

The selection of hyper-parameters demonstrated a tendency of the temporally-dense datasets sampled at 25Hz, towards a higher dropout rate to achieve better performance. This is justified by the model's tendency to overfit into data samples with higher feature numerosity, thus increasing regularisation enables better generalisation to unseen data. Furthermore, it is observed that the model using accelerometer and gyroscope signals benefits from a larger hidden space within each attention head, which indicates that more complicated patterns can be deduced through the angular velocity inputs, thus a more complex embeddings space is required to capture them. Finally, it is seen that the model utilising the complete dataset does not benefit from an increased number of attention heads, even though the difference in performance is not substantial, which might be an indication that the lower number of attention heads is sufficient to capture the data structure thus increasing it does not provide any notable improvements.

5.2.2 Subject-Independent Cross-Validation

Once the optimal hyper-parameter configuration was obtained for each model, the methodologies were evaluated using subject-independent cross-validation, to compute

the average performance of each model as well as the dispersion of the results relative to the respective mean, reported using the value of one standard deviation.

The results exerted from the 5-fold subject-independent cross-validation are displayed in Table 5.7. It can be observed increasing the sampling frequency has improved the model's ability to recognise patterns, with an improvement in performance in terms of both accuracy and F1-Score. On top of that, the introduction of angular velocity signals in the dataset has allowed the model to improve further, achieving an improvement of approximately 4.51% in accuracy and 4.55% in F1-Score, from the model that used accelerometer-only data sampled at 12.5Hz. This observation indicates that both data enhancements can potentially yield substantial improvements in classification performance; however, the comparatively high standard deviation that results in overlapping between the confidence intervals of the algorithms, suggests that further data samples need to be collected to make a definite conclusion on the benefit of each augmentation.

Dataset	Accuracy	F1-Score
Accelerometer at 12.5Hz	0.6855 ± 0.0786	0.6867 ± 0.0872
Accelerometer at 25Hz	0.7169 ± 0.0604	0.7150 ± 0.0600
Accelerometer & Gyroscope at 25Hz	0.7306 ± 0.0621	0.7322 ± 0.0632

Table 5.7: Subject-Independent Cross-Validation Results for Data Augmentation

5.3 Calibration System

Even though the transformer methodology did not yield a statistically significant improvement in performance in comparison to the AC-GAN technique, it provided a number of advantages critical for this particular application. In contrast to the AC-GAN model, the transformer classifier is able to learn new information through transfer learning which is crucial for allowing the calibration framework to work as expected; resolving the issue found in the AC-GAN configuration in the MInf - Part 1 Project, where the model was adhered in a local optimum and was unable to learn any information through the calibration samples [5]. Furthermore, the multi-head attention methodology enables parallel encoding of intra-sample interdependencies, thus reducing model training time. Finally, the architecture offers uncomplicated training, as the model consists only of one component, contrary to the AC-GAN model which consists of 4 different modules.

The evaluation of the calibration framework was carried out by firstly selecting a classifier model trained on the accelerometer-only dataset sampled at 12.5Hz for each subject, ensuring that the subject was not included in the training or validation sets used for the training of that model. Thereafter, for each subject, the model is trained using a number of sliding windows from the selected subject's training data as described in the Methodology chapter. The results before and after the calibration are then reported enabling assessment of any differences in performance, as shown in Table 5.8. It is observed that in the results reported, the calibration technique has yielded an average improvement of 4.74% in accuracy and 4.02% in F1-Score, using only a

small number of samples from the subject. Another important remark from the above-noted results is the relatively low performance of the base classifier initially utilised, in comparison to the results reported in the subject-independent cross-validation for the development of the model. The drop in performance can be interpreted by the different nature of the data, as the real-world dataset includes a significantly higher ratio of transition movements between different activities which often confuses the classifier. Furthermore, the different social signals performed simultaneously with the physical activities introduced more uncertainty in the dataset, as a highly active social signal such as coughing may mislead the classifier to a more active physical activity as well even though the subject might be stationary. Finally, in the real-time dataset, the sliding windows are generated on a rolling basis, assigning the mode class label as the target of the time sequence, in contrast to the base classifier which is trained on windows containing only one activity. As a result, this disparity diminishes the reported evaluation scores as the classifier might be affected by a segment of the input samples which is not shown in the class label. The above-mentioned concerns, however, highlight the importance of the improvement gained using the calibration framework, which has shown performance increases even when used in a domain with a larger deviation in input.

	Base Classifier		Calibration Model	
Participant ID	Accuracy	F1-Score	Accuracy	F1-Score
XXD001	0.7632	0.8594	0.7632	0.8594
XXD002	0.6364	0.7460	0.6818	0.7884
XXD014	0.5676	0.5979	0.6216	0.6167
XXD015	0.6429	0.6177	0.6857	0.6391
XXD016	0.6747	0.6760	0.7470	0.7385
XXD021	0.4483	0.4211	0.5172	0.5172
Average	0.6222 ± 0.0970	0.6530 ± 0.1353	0.6694 ± 0.0824	0.6932 ± 0.1144

Table 5.8: Calibration Framework Evaluation Results

Test Statistic

In order to assess whether the difference in performance between the base classifier and the calibration model is statistically important, a statistical test was carried out, with the results shown below. For the hypothesis testing, a two-tailed t-test was utilised, with the assumption that the difference between the scores of the two systems was drawn from a Normal Distribution. The hypothesis examined for the t-test is shown below:

$$H_0 : \overline{A} = \overline{B}$$
$$H_1 : \overline{A} \neq \overline{B}$$

where H_0 is the Null Hypothesis and H_1 is the Alternative Hypothesis, A is the random variable representing the scores of the calibration model and B is the random variable representing the scores for the base classifier respectively. The test statistic is given by:

$$t = \frac{\overline{A-B}}{\sigma_{(A-B)}}\sqrt{N}$$

where $\sigma_{(A_B)}$ is the standard deviation of $A - B$ and N is the number of samples.

The p-values calculated for each evaluation metric are displayed in Table 5.9. As it can be observed, the p-value for both metrics is less than the significance level $\alpha = 0.05$, which indicates that there is significant evidence to reject the null hypothesis H_0. Therefore, based on the two-tailed t-test, there is evidence that the calibration model is, statistically, significantly better than the base classifier.

	Accuracy	F1-Score
p-value	0.0068	0.0367

Table 5.9: The p-value from the two-tailed t-test for the base classifier and calibrated model for each metric score.

Chapter 6

Conclusions

In this project, a human activity classification framework is explored, utilising accelerometer and gyroscope signals gathered using the wearable RESpeck monitor for the development of a range of machine learning methodologies. This project presents a human activity calibration methodology, enabling improved performance on unseen subjects by training a human activity recognition model using annotated accelerometer samples of them performing the activities, allowing the classifier to calibrate to their specific characteristics. In an endeavour to establish a proof-of-principle study of the work presented in the MInf - Part 1 Project [5], a real-world dataset is collected and utilised for the purposes of this project, enabling the demonstration of the complete calibration framework in a real-life setting. Furthermore, a Transformer neural network is implemented in this research, addressing the issues encountered in the MInf - Part 1 Project [5] in applying the Auxiliary Classifier Generative Adversarial Network (AC-GAN) methodology for transfer learning. The findings of this research have also highlighted the potential utility of the novel domain-agnostic transformer-based transfer learning methodology in various data-rich tasks involving intra-subject dependencies within the data, indicating the importance of this research in future work.

6.1 Contributions

For the purposes of this research, an ensemble of previously collected datasets along with newly gathered annotated datasets is used for the development of a human activity classification model using deep learning methodology. In addition, a newly gathered dataset using accelerometer and gyroscope sensor signals is exploited for this study, showing that the introduction of angular velocity in addition to linear acceleration input signals, as well as the increase of sampling frequency, can ameliorate the classification performance. Furthering the research on transfer learning methodology presented in the MInf - Part 1 Project [5], a real-world dataset is assembled and used to demonstrate the feasibility of the calibration framework in a real-life scenario. The calibration system achieved an average improvement of 4.74% in accuracy and 4.02% in F1-Score in the real-world dataset, calibrating on samples of approximately 7.68 seconds for each activity. Furthermore, the introduced transformer architecture has enabled pattern

recognition through transfer learning, which is a critical component of the calibration framework, which was not achieved using the AC-GAN model in the MInf - Part 1 Project [5].

6.2 Limitations

Due to resource availability and time constraints, the hyper-parameter tuning was evaluated using a fixed dataset split instead of using subject-independent cross-validation in an effort to experiment with a wider search space. Ideally, the hyper-parameter tuning on the base classifier as well as on the data augmentation models would be carried out using subject-independent cross-validation to ensure that the results represent all the dataset subjects.

Furthermore, the experiments carried out during this project have shown that a larger dataset, especially on the data augmentation experiment, would allow the model to exploit its full capabilities without overfitting. Moreover, an enhanced dataset would result in a lower deviation between the cross-validation results for each fold, decreasing the error margin of the result interpretation. The preferable data collection process would have included data from a larger number of participants with more diverse demographic characteristics enhancing generalisation to unseen data.

6.3 Further Research

Future iterations of this research should aim at extending the already congregated dataset with participants with a wider range of characteristics, enabling a thorough examination of the effect of angular velocity and sampling frequency in the classification performance of the algorithms. The range of annotated activities could also be expanded, improving the usability of the model in real-life circumstances.

The domain-agnostic implementation of the transformer-based calibration framework enables the adoption of the methodology in any data-rich classification task where the availability of annotated data is limited. Future work may implement the methodology for the detection of events of different nature, such as social signal classification. Future research could also work on investigating the effect of calibration using different data distributions. For instance, the effect of using social signal samples for calibrating a model trained on the human activity classification task could be investigated, with a successful application unlocking numerous possibilities for the calibration system in various domains.

Moreover, further research extending the present work may focus on enhancing the transformer classifier in an attempt to improve the overall performance of the implemented system. A possible advancement could include implementing the modifications proposed in the Transformer-XL architecture [10], allowing the model to capture dependencies between different sliding windows thus learning longer-term dependencies in the dataset.

Bibliography

[1] Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, 2016.

[2] Fadi Al Machot, Ali Elmachot, Mouhannad Ali, Elyan Al Machot, and Kyandoghere Kyamakya. A deep-learning model for subject-independent human emotion recognition using electrodermal activity sensors. *Sensors*, 19(7):1659, Apr 2019.

[3] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.

[4] Andreas Bulling, Ulf Blanke, and Bernt Schiele. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Comput. Surv.*, 46(3), January 2014.

[5] Stylianos Charalampous. Human activity and social signal classification using auxiliary classifier generative adversarial network and transfer learning. Master's thesis, University of Edinburgh, 2021.

[6] Stylianos Charalampous and Andreas Ramsoy. Human activity recognition using deep learning methods. Coursework 3, Principles and Design of IoT Systems (PDIoT), School of Informatics, University of Edinburgh.

[7] Pinzhen Chi. Cough detecting using an integrated tri-axial accelerometer and gyroscope. Master's thesis, University of Edinburgh, 2018.

[8] François Chollet et al. Keras. `https://keras.io`, 2015.

[9] Oscar Cleve and Sara Gustafsson. *Automatic Feature Extraction for Human Activity Recognitionon the Edge*. PhD thesis, 2019.

[10] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context, 2019.

[11] S. Dara and P. Tumma. Feature extraction by using deep learning: A survey. In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 1795–1801, 2018.

[12] Akbar Dehghani, Tristan Glatard, and Emad Shihab. Subject cross validation in human activity recognition, 2019.

[13] Akbar Dehghani, O. Sarbishei, Tristan Glatard, and Emad Shihab. A quantitative comparison of overlapping and non-overlapping sliding windows for human activity recognition using inertial sensors. *Sensors*, 19:5026, 11 2019.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[15] Lixin Duan, Dong Xu, and Ivor Tsang. Learning with augmented features for heterogeneous domain adaptation, 2012.

[16] Samuel Dupond. A thorough review on the current advance of neural network structures. *Annual Reviews in Control*, 14:200–230, 2019.

[17] Steven C. Frautschi, Richard P. Olenick, Tom M. Apostol, and David L. Goodstein. *The Mechanical Universe: Mechanics and Heat, Advanced Edition*. Cambridge University Press, 1986.

[18] Teodora Georgescu. Classification of coughs using the wearable respeck monitor. Master's thesis, University of Edinburgh, 2019.

[19] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[20] Maayan Harel and Shie Mannor. Learning from multiple outlooks, 2011.

[21] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[23] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2016.

[24] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997.

[25] Martin-Philipp Irsch. Semi-supervised human activity recognition with the wearable respeck sensor using gans. Master's thesis, University of Edinburgh, 2019.

[26] Artur Jordão, Jr Nazare, Jessica Sena de Souza, and William Schwartz. Human activity recognition based on wearable sensor data: A standardization of the state-of-the-art. 06 2018.

[27] E. Keogh, S. Chu, D. Hart, and M. Pazzani. An online algorithm for segmenting time series. In *Proceedings 2001 IEEE International Conference on Data Mining*, pages 289–296, 2001.

[28] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM Computing Surveys*, jan 2022.

[29] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[30] Sotiris Kotsiantis, Dimitris Kanellopoulos, and P. Pintelas. Data preprocessing for supervised learning. *International Journal of Computer Science*, 1:111–117, 01 2006.

[31] Atesh Koul, Cristina Becchio, and Andrea Cavallo. Cross-validation approaches for replicability in psychology. *Frontiers in Psychology*, 9:1117, 2018.

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

[33] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature Cell Biology*, 521(7553):436–444, May 2015.

[34] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.

[35] Bryan Lim, Sercan O. Arik, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting, 2020.

[36] Saif Mahmud, M Tanjid Hasan Tonmoy, Kishor Kumar Bhaumik, A K M Mahbubur Rahman, M Ashraful Amin, Mohammad Shoyaib, Muhammad Asif Hossain Khan, and Amin Ahsan Ali. Human activity recognition from wearable sensor data using self-attention, 2020.

[37] J. Mann, R. Rabinovich, A. Bates, S. Giavedoni, W. MacNee, and D. K. Arvind. Simultaneous activity and respiratory monitoring using an accelerometer. In *2011 International Conference on Body Sensor Networks*, pages 139–143, 2011.

[38] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

[39] Abdulmajid Murad and Jae-Young Pyun. Deep recurrent neural networks for human activity recognition. *Sensors*, 17(11), 2017.

[40] Ashkan Namin, Karl Leboeuf, Roberto Muscedere, Huapeng Wu, and Majid Ahmadi. Efficient hardware implementation of the hyperbolic tangent sigmoid function. pages 2117 – 2120, 06 2009.

[41] Nikita Nikolajev. Multiclass classification of social signals using the respeck sensor. Master's thesis, University of Edinburgh, 2020.

[42] Francisco Javier Ordóñez and Daniel Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1), 2016.

[43] Jorge Luis Reyes Ortiz. *Smartphone-based human activity recognition*. Springer, 2015.

[44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[45] Sreenivasan Ramasamy Ramamurthy and Nirmalya Roy. Recent trends in machine learning for human activity recognition—a survey. *WIREs Data Mining and Knowledge Discovery*, 8(4):e1254, 2018.

[46] Daniele Ravì, Charence Wong, Benny Lo, and Guang-Zhong Yang. Deep learning for human activity recognition: A resource efficient implementation on low-power devices. pages 71–76, 06 2016.

[47] Yoli Shavit and Itzik Klein. Boosting inertial-based human activity recognition with transformers. *IEEE Access*, 9:53540–53547, 2021.

[48] Julien Soler, Laurent Gaubert, Fabien Tencé, and Cédric Buche. Data clustering and similarity. pages 492–495, 05 2013.

[49] TensorFlow. Tensorflow lite: Ml for mobile and edge devices. `https://www.tensorflow.org/lite`, 2021.

[50] Laurens van der Maaten, Eric Postma, and H. Herik. Dimensionality reduction: A comparative review. *Journal of Machine Learning Research - JMLR*, 10, 01 2007.

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[53] D. Wang and T. F. Zheng. Transfer learning for speech and language processing. In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1225–1237, 2015.

[54] W. Wang, Y. Guo, B. Huang, G. Zhao, B. Liu, and L. Wang. Analysis of filtering methods for 3d acceleration signals in body sensor network. In *International Symposium on Bioelectronics and Bioinformations 2011*, pages 263–266, 2011.

[55] Tianze Wei. Simultaneous classification of human activity and social signals using the wearable respeck device. Master's thesis, University of Edinburgh, 2020.

[56] Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big Data*, 3(1):9, May 2016.

[57] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020.

[58] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network, 2015.

[59] Rikiya Yamashita, Mizuho Nishio, Richard Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, 9, 06 2018.

[60] Shugang Zhang, Zhiqiang Wei, Jie Nie, Lei Huang, Shuang Wang, and Zhen Li. A review on human activity recognition using vision-based method. *Journal of Healthcare Engineering*, 2017:1–31, 07 2017.

[61] Óscar D. Lara, Alfredo J. Pérez, Miguel A. Labrador, and José D. Posada. Centinela: A human activity recognition system based on acceleration and vital sign data. *Pervasive and Mobile Computing*, 8(5):717 – 729, 2012.

Appendix A

Participants' information sheet

Participant Information Sheet

Project title:	Classification of Physical Activities and Social Signals using a wearable Respeck monitor
Principal investigator:	D.K. Arvind
Researcher collecting data:	Celina Dong/ Stylianos Charalampous/ Shuai Shi Teodora Georgescu

This study was certified according to the Informatics Research Ethics Process, RT number 2019/27996. Please take time to read the following information carefully. You should keep this document for your records.

Who are the researchers?

The three students, Celina Dong, Stylianos Charalampous and Shuai Shi, will collect data as part of their undergraduate projects. They are all 4th/5th year Masters in Informatics students at the School of Informatics, University of Edinburgh.

The main researcher is Teodora Georgescu, a Research Associate at the School of Informatics, University of Edinburgh. Other researchers involved in the project include Andrew Bates and Sharan Maiya who will provide technical support during data collection. The project is being supervised by Professor D K Arvind as the Principal Investigator, under the aegis of the Centre for Speckled Computing, University of Edinburgh.

What is the purpose of the study?

The aim of the project is to identify physical activity and social signals in people by analysing data from the Respeck monitor worn as a plaster on their chest. Examples include walking, running and climbing stairs for physical activities, and social signals such as coughing, speaking and swallowing (due to eating or drinking). You will be invited to wear the Respeck device as a plaster on the chest and perform instances of the examples listed previously. You will be filmed during one part of the data collection for the purpose of correct data labelling – in the post-processing part of your data we will use the video as a guide to correctly label the data with the appropriate activities you performed. Your data will be collected and added to a mix of similar data collected from other volunteers which will be analysed to classify

accurately the different activities. The labelled data collected will be used to train machine learning models trained to distinguish accurately between them.

Why have I been asked to take part?

You have been invited to take part in this study because you are either a student at the University of Edinburgh, or because you belong to an age group that our research is interested in.

Do I have to take part?

No – participation in this study is entirely up to you. You can withdraw from the study at any time without giving a reason. After this point, personal data will be deleted and anonymised data will be combined such that it is impossible to remove individual information from the analysis. Your rights will not be affected. If you wish to withdraw, contact the PI. We will keep copies of your original consent, and of your withdrawal request.

What will happen if I decide to take part?

You will be invited to wear the Respeck device encased in a small disposable bag with the blue, flat surface against the skin just below your ribcage and secured to your chest with the medical tape provided.

Please ensure the device is the right way up, i.e. you can read the text on the flat side of the device.

A mobile phone with a specially designed application will automatically collect data from the Respeck device.

You will be asked to perform a series of gentle activities as listed below. The optional activities will be only be administered for the students,

Physical activities:

- Sitting down (straight, bent forward, bent backward)
- Standing up
- Lying down (back, front, left, right)
- Walking at three different speeds (slow, medium and fast)

- Ascend/Descend a set of stairs
- (Optional) Wear when travelling in a bus/car/train
- (Optional) Riding a bike
- Moving your body at the waist from left to right and repeat 5 times.
- Swinging your body to the front and back and repeat 5 times
- Running

Social signals:

- Coughing
- Talking
- Eating/Drinking
- Singing
- Laughing
- Breathing normally
- Hyperventilating

You might be asked to perform some of these activities at the same time, such as coughing when you are lying down. The intensity of these activities will be adjusted to your comfort level. Each activity and social signal will be recorded for at least 30 seconds, and tiring activities, such as forced coughing, will be divided into shorter segments of 10-15 seconds of continuous coughing.

For the second part of the data collection, you will be asked to perform a sequence of activities, uninterrupted, in order to simulate the real data we might be getting from Respeck wearers. During this time you will also be filmed using a simple phone camera operated by the data collector. We ask for your permission to film you so that, in the post-processing phase of the collection, we can accurately label the actions you performed.

At any point in time, if you feel that you do not wish to continue with the study, then please feel free to let me know and the study will be stopped immediately.

Are there any risks associated with taking part?

You'll be invited to wear the Respeck device which has undergone the necessary safety tests. Participants with known plaster/plastic allergy will be excluded. The

device is enclosed in a disposable plastic bag and is not in direct contact with the skin. The Respeck device is cleaned and sterilised once returned. There are no significant risks associated with participation. The researchers will maintain at least 2m social distance and will wear masks and safety visor.

Are there any benefits associated with taking part?

No.

What will happen to the results of this study?

The results of this study may be summarised in published articles, reports and presentations. Quotes or key findings will always be anonymous. With your consent, information can also be used for future research. Your data may be archived for a minimum of 5 years.

With your consent, the research team might share the fully anonymised data of this study with other researchers outside of the University of Edinburgh as part of publications.

Data protection and confidentiality.

Your sensor data will be processed in accordance with Data Protection Law. All information collected about you will be kept strictly confidential. Your data will be referred to by a unique participant number rather than by name.

Your sensor data will only be viewed by the research team: Teodora Georgescu, Andrew Bates and Professor D K Arvind for this project. Your anonymised data may be used in other ethically approved research projects supervised by Professor D K Arvind or be made available to other researchers outside of the University of Edinburgh as part of publications. By signing the consent form, you agree to such usage.

Summaries of the anonymised sensor data is stored on the University's secure encrypted cloud storage services *datasync* (https://www.ed.ac.uk/information-services/computing/desktop-personal/datasync), for which the research team has writing access and MInf and Year 4 project students supervised by Professor Arvind will have reading access. We only store summaries of accelerometer data, and not personal information such as name, age or address.

Your consent information will be kept separately from your responses in order to minimise risk.

What are my data protection rights?

The University of Edinburgh is a Data Controller for the information you provide. You have the right to access information held about you. Your right of access can be exercised in accordance Data Protection Law. You also have other rights including rights of correction, erasure and objection. For more details, including the right to lodge a complaint with the Information Commissioner's Office, please visit www.ico.org.uk. Questions, comments and requests about your personal data can also be sent to the University Data Protection Officer at dpo@ed.ac.uk.

Who can I contact?

If you have any further questions about the study, please contact Teodora Georgescu (tgeorges@ed.ac.uk).

If you wish to make a complaint about the study, please contact:
Professor D K Arvind (dka@inf.ed.ac.uk) or the Informatics Ethics Panel (inf-ethics@inf.ed.ac.uk).

When you contact us, please provide the study title and detail the nature of your complaint.

Updated information.

If the research project changes in any way, an updated Participant Information Sheets will be made available on request from Teodora Georgescu (tgeorges@ed.ac.uk).

Alternative formats.

To request this document in an alternative format, such as large print or on coloured paper, please contact Teodora Georgescu (tgeorges@ed.ac.uk).

General information.

For general information about how we use your data, go to: edin.ac/privacy-research

Appendix B

Participants' consent form

Participant number:_____

Participant Consent Form

Project title:	Classification of Physical Activities and Social Signals using a wearable Respeck monitor
Principal investigator (PI):	D.K. Arvind
Researcher:	Celina Dong/ Stylianos Charalampous/Shuai Shi/ Teodora Georgescu
PI contact details:	dka@inf.ed.ac.uk

By participating in the study you agree that:

- I have read and understood the Participant Information Sheet for the above study, that I have had the opportunity to ask questions, and that any questions I had were answered to my satisfaction.
- My participation is voluntary, and that I can withdraw at any time without giving a reason. Withdrawing will not affect any of my rights.
- I consent to my anonymised data being used in academic publications and presentations.
- I understand that my anonymised data will be stored for the duration outlined in the Participant Information Sheet.

Please tick yes or no for each of these statements.

		Yes	No
1.	I agree to my physical activity being recorded using the Respeck monitor.	☐	☐
2.	I agree to being video recorded.	☐	☐
3.	I allow my data to be used in future ethically approved research.	☐	☐
4.	I agree to take part in this study.	☐	☐

Name of person giving consent	Date dd/mm/yy	Signature

Name of person taking consent	Date dd/mm/yy	Signature

THE UNIVERSITY of EDINBURGH informatics