UNIVERSITY OF ST ANDREWS

DISSERTATION

Supervised Learning Methods for Measuring Information Transfer in Human Activity Recognition Time-Series Data

Anastasia Akchurina
230012908
MSc Applied Statistics and Datamining
School of Mathematics & Statistics

Supervised by:
Dr. Giorgos Minas
School of Mathematics & Statistics

August 2024

Declaration

I hereby certify that this dissertation, which is approximately 15,000 words in length, has been composed by me, that it is the record of work carried out by me and that it has not been submitted in any previous application for a higher degree. This project was conducted by me at the University of St Andrews from May 2024 to August 2024 towards fulfilment of the requirements of the University of St Andrews for the degree of MSc Applied Statistics and Datamining under the supervision of Dr. Giorgos Minas.

Date: 13 August 2024

Acknowledgements

I am finishing writing my dissertation today at the best University in the UK... Still can't believe it but I wanted to say how entirely grateful I am to my family, especially to my parents, for supporting my decision to apply to the University of St Andrews and making my dream come true!

Many many thanks to my lovely supervisor Dr. Giorgos Minas for being empathetic, patient and kind to me! Thank you so much for inspiration and your invaluable help during these 3 months! It's been a great pleasure working with you and I appreciate everything you've done for me!

Last but not least, I was blessed to have my dear friends I met here in St Andrews who would always stay by my side to cheer me up or calm me down. Thank you so much for your love, care and support!

Lots of love

Contents

Abstract

This research project attempts to apply supervised learning methods to execute Human Activity Recognition (HAR) in senior citizens using the HAR70+ datasets comprising of time-series data of participants aged between 70 and 95 years. The paper deals with the issues of identifying a number of day-to-day activities, including the problems of non-stationarity in the data gathered by sensors and class imbalance in the datasets. The supervised learning models used are Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Decision Trees, Random Forest, XGBoost, Multilayer Perceptron (MLP), Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) on individual and population HAR70+ data.

Key research findings showed that Random Forest has the highest accuracy score on individual HAR70+ datasets, while RNN exerted great performance in handling population training data and poorer performance on unseen data due to the high variability in human activities and the absence of seasonality in the time-series data. Resampling and averaging sampling techniques were considered optimal strategies for tackling the issue of class imbalance. Altogether, this research contributes to the improvement and development of Human Activity Recognition systems for elderly care proposing the ways of a better functioning of supervised learning methods and selection of sampling strategies. The outline for future research suggests focusing on the extension of the deep learning algorithms and inclusive sampling methods which consider both participants using walking assistance and the ones who do not require walking aids for diverse movement patterns recognition in the senior citizens HAR.

Chapter 1

Introduction

1.1 Motivation

Human Activity Recognition (HAR) has become a skyrocketing area of research in the sphere of data science and machine learning. The ability to identify and classify human activities based on sensor data has various applications from healthcare (Gu Zhanzhong, 2024) and sports activity monitoring (Xiaochun, 2024) to smart environments (Monica-Andreea Dragan, 2013) (Ranjit Kolkar, 2021) and elderly care (Kevin Yeap Chen Keng, 2018). There is a growing need to monitor and support the basic daily-life activities of senior citizens, making them maintain independent and high-quality life as the population ages worldwide, especially, in the developing countries. HAR systems play an important role in providing real-time monitoring and alerts assisting in accidents prevention and providing insights into the health and well-being of elderly individuals (Han Sun, 2024).

Typically, HAR entails pooling of data from one or more sensors such as accelerometer which are often found in wearable devices or placed in the environment (Liming Chen, 2019). These sensors collect information regarding a person's movements which can then be analysed in real-time to identify activities being performed by the user such as walking, sitting, standing, lying, etc. The design of an efficient HAR system involves assembling of data which calls for effective data analysis that can handle non-linearity and variability in human movements. This is even more challenging when working with senior citizens since their movements may be slower and may have less variation compared to those of youths (Di Wang, 2015).

To recognise and classify human daily-life activities efficiently supervised learning methods may come in handy as they have become the cornerstone of HAR research (Qiancheng Tan, 2023). Supervised learning involves training algorithms on labelled datasets where the input data (sensor readings) is mapped to the correct output (activity labels). Once trained, these models can be used to predict activities on new, unseen data. The choice of supervised learning algorithms is important as it directly influences accuracy and, hence, reliability of the Human Activity Recognition system.

Supervised learning methods have been discussed in the context of HAR and each has its own strength and weaknesses. An advantage of Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) is that they are not computationally heavy or time-consuming but rather simple and easy to interpret. The downside of these models, however, is the fact that they do not perform well on data with complex non-linear patterns. Decision Trees, Random Forest and XGBoost are less rigid, more flexible and generally more precise in computing accuracy scores, especially when working with high-dimensional datasets. Some of the deep learning models, for instance, Multilayer Perceptron (MLP) and Convolutional Neural Networks (CNNs) have been found to deliver exceptional results when identifying activities in the sensor data.

However, one of the biggest problems of HAR is the temporal character of the data. The human

movements are inherently temporal and, therefore, the order of the movements is the key element that should be captured by the recognition process. For instance, when actions are being carried out, the order in the walking sequence is different from that of sitting. Time-series data consists of ordered sequences of sensor readings and requires models that can keep these temporal dependencies. Compared to many traditional classifiers that treat each observation independently models that respect the time component such as Recurrent Neural Networks (RNNs) are a better choice for the problems in which sequence of actions matters.

The main emphasis of this work is on the use of the supervised learning algorithms for time-series human activity recognition analysis in senior citizens based on the HAR70+ dataset. It includes sensor data recordings from 18 participants aged 70 to 95 performing a variety of basic daily-life activities. The research aims at comparing various supervised learning techniques including both models that respect and those that ignore temporal dependencies.

The first chapter will, therefore, present the concepts and issues related to HAR and later on discuss in detail various supervised learning techniques for HAR. After that, it will focus on time series data analysis in the context of HAR and discuss traditional and deep learning methods. Furthermore, the research will give recommendations on practice-oriented approaches to the creation of HAR systems for senior citizens and explain the application of machine learning in time-series HAR data analysis.

1.2 Research Questions

Over the course of this research we will attempt to address the following three questions listed in *Table 1* which will be approached in *Chapter 3 : Methodology* and *Chapter 4 : Results and Discussion*.

1. What insights can exploratory data analysis provide about the variability and structure of the HAR70+ dataset and how can these insights assist with or work against the development of robust human activity recognition (HAR) machine learning algorithms?
2. Which supervised learning algorithm performs better under specific conditions for each individual HAR70+ dataset and across a population of HAR70+ datasets? How do different models, namely, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Decision Tree, Random Forest, XGBoost, Multilayer Perceptron (MLP) and Convolutional Neural Networks (CNN) compare in terms of prediction power? Does a model (Recurrent Neural Networks (RNN)) that respects the time component in the time-series HAR data have higher performance scores than the models which ignore it?
3. What cross-validation method provides the most optimal performance for the selected classifiers (LDA, QDA, Decision Tree, Random Forest, XGBoost, MLP and CNN) for each individual dataset and across a population of HAR70+ datasets? How does the choice of cross-validation

method impact the accuracy and other metrics' scores such as Cohen's Kappa for senior citizens activity recognition HAR70+ datasets?

Table 1. List of Research Questions

1.3 Goals and Objectives

In this research we explore the application of supervised learning methods for Human Activity Recognition (HAR) using the HAR70+ dataset. The structure of the dissertation is organised in the following way:

Chapter 2: Literature Review gives a precis of all the literature reviewed in relation to the research questions of the study. *Section 2.2 Human Activity Recognition (HAR)* which entails the identification and analysis of human actions from sensor information presents various approaches of HAR such as sensor-based models, ontology-based, hybrid models and data segmentation techniques all of which have their advantages and disadvantages. Furthermore, the chapter also discusses the use of HAR in smart homes for analysing activities of daily life ADLs and the opportunities and issues associated with the practical use of such systems. Section *2.3 Supervised Classification Methods* describes various supervised learning algorithms, namely, Linear and Quadratic Discriminant Analyses (LDA, QDA), Decision Trees, Random Forest, XGBoost, Multilayer Perceptron (MLP), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) as well as their use in Human Activity Recognition (HAR). In addition, Section *2.5 Time-Series* is dedicated to the time-series data and its analysis with the focus on its sequential nature and the use in various domains including HAR70+ data to recognise human activities. Thus, we describe the theoretical background and prior research related to this work concerning the algorithms and methods employed in HAR and time-series data analysis.

In *Chapter 3 : Methodology* we address the three main research questions. We begin with introducing the data in *Section 3.2 Data HAR70+*, then continue with the discussion of *3.3 Exploratory Data Analysis* to understand the variability and patterns in the HAR70+ dataset. We then approach the detection of stationarity in *Section 3.4 Detecting Time-Series Stationarity* and the necessity of *3.5 Data Cleaning and Preparation.* The core of the methodology involves fitting various supervised learning models (LDA, QDA, Decision Tree, Random Forest, XGBoost, MLP, CNN, RNN) to both individual and combined population of HAR70+ datasets. A range of cross-validation methods such as resampling to a class with a maximum number of observations with replacement, downsampling to a class with a minimum number of observations, and averaging sample sizes aross labels are used to evaluate models' performance which is measured using accuracy, precision, recall, F1 score and Cohen's Kappa metrics in *Section 3.6 Fitting Supervised Learning Models*.

Chapter 4 : Results and Discussion presents the results of the data analysis and the evaluation of the models in detail. *Section 4.1* is based on the findings of the exploratory data analysis of the variability in the HAR70+ dataset. *Section 4.2* provides a comparison of results of the chosen supervised learning

algorithms (LDA, QDA, Decision Tree, Random Forest, XGBoost, MLP, CNN, RNN). *Section 4.3* introduces the most optimal approach to data sampling to enable the correct recognition of human activities.

Chapter 5 : Conclusion & Outline for Future Research provides the summary of the key findings of the research, highlights the contribution of the dissertation to the field of human activity recognition and proposes potential areas for the future direction that can be taken to enhance the methods and models of HAR.

Overall, the purpose of this research is to improve the performance of the supervised machine learning techniques in the HAR for seniors using the HAR70+ dataset. This dataset offers real and diverse physical activities of older adults and can be considered as a great source of training data for machine learning models in HAR domain. To achieve this goal, we addressed three target research questions which guide the methodology and subsequent analysis. The goals and objectives that have been linked to them and how they have been addressed in this research are as follow:

The following steps were taken to answer question 1:

Objective 1: Perform the Exploratory Data Analysis (EDA) to understand the characteristics and the nature of the HAR70+ dataset and how these characteristics can either enhance or hamper the performance of the human activity recognition (HAR) machine learning algorithms.

Chapter 3: Methodology (Section 3.3): In this chapter, the techniques of exploratory data analysis are used on HAR70+ dataset. It defines the process of understanding the patterns, trends, and anomalies of the data; this includes the examination of feature distributions, correlation, and other issues that may affect the performance of the machine learning models.

Results Chapter (Section 4.1): In this chapter, the main results of EDA are described and discussed, focusing on variability and structure of the data. The chapter also discusses on how these insights affected the selection of the algorithms and model development, how they helped in the generation of better HAR models or created issues that had to be solved during the machine learning process.

To answer question 2 the following steps were undertaken:

Objective 2: Compare the performance of the selected supervised learning algorithms applied to individual and a population of HAR70+ datasets.

Chapter 3: Methodology (Section 3. 6. 1): This chapter also outlines the use of different supervised learning algorithms to individual HAR70+ datasets including the training, cross-validation and testing of the models and their assessment.

Chapter 3: Methodology (Section 3. 6. 2): This section proposes the extension of the evaluation to a population of HAR70+ datasets to understand the behaviour of the different algorithms across a wider range of datasets. It takes into account the reasons why different models are employed in a population and how these models are compared in terms of their capability to produce results that are generalisable across several datasets.

Results Chapter (Section 4. 2): In this chapter, a comparison of the performance of each of the

supervised learning algorithm is presented in terms of accuracy, precision, recall, F1 score and Kappa. It shows which models work best for specific datasets and which are more effective for the overall population; the focus is made on the performance of RNNs compared to other models that may not consider temporal dependencies in the data.

The following steps were taken to answer question 3:

Objective 3: Determine the impact of different sampling methods on models' performance for individual HAR70+ datasets and populations.

Chapter 3: Methodology (Section 3.6.1): The chapter discusses the application of various sampling methods to individual datasets. These include the application of resampling of maximum label values with replacement, downsampling to minimum label values, and averaging values divided by the number of labels.

Chapter 3 : Methodology (Section 3.6.2): This section investigates the application of different sampling methods to HAR70+ populations and some reasons for using different combinations of train and test datasets.

Results Chapter (Section 4.3): The impact of each sampling method on models' performance is analysed highlighting which methods provide the most optimal performance in terms of accuracy, precision, recall, F1 score, and Kappa metrics for each classifier both for individual HAR70+ datasets and populations.

By answering each of the research questions the paper seeks to establish the supervised learning algorithms with higher performance and optimal sampling techniques for the recognition of senior citizens' activities using the HAR70+ data. This is meant to provide valuable insights into the reliability and transferability of these models for further research and real-world usage of HAR.

Chapter 2

Literature Review

2.1 Introduction

In this chapter, we focus on the current and emerging domain of Human Activity Recognition (HAR) to analyse different approaches and strategies that can be employed in the detection and analysis of human actions from sensor data. We will discuss various methods which are used to identify human activities such as sensor-based, ontology-based, hybrid and segmentation and how these approaches contribute to the accurate recognition of human activities. Furthermore, we will introduce supervised learning techniques: Linear Discriminant Analysis, Quadratic Discriminant Analysis, Decision Trees, Random Forest, XGBoost and neural networks, namely, MLP, CNNs and RNNs. We will touch upon the implementation of these methods in the analysis of time-series data to understand the temporal dynamics

of human behaviour.

2.2 Human Activity Recognition (HAR)

Human Activity Recognition (HAR) is a skyrocketing field focused on recognition of human actions and movements with the help of sensor data (Liming Chen, 2019). This area of research has been gaining popularity in recent years due to application in a wide range of spheres such as health monitoring, elderly care, sports analytics, and smart homes. HAR systems are used to recognise physical activities that can be effective for monitoring and controlling human behaviour at various conditions and environments.

Concerning Human Activity Recognition several methods have been developed. In this work, we will describe seven HAR approaches: Sensor-Based, Ontology-Based, Hybrid, Semantic-Based, Time-Window Based, Composite HAR, and Smart Homes for HAR and ADL. In this research, we shall endeavour to apply sensor-based activity recognition and time-window based data segmentation since the data is collected through sensors and is of a time-series nature.

2.2.1 Sensor-Based Activity Recognition Review

Sensor-based activity recognition has been one of the most popular approaches in HAR. This approach employs data from different sensors like the accelerometers, gyroscopes, and magnetometers to record the movement of the people. The data is then analysed to identify activities that have been undertaken. The foremost difficulty of the sensor-based activity recognition is the efficient and correct acquisition of sensors' information.

The analysis of the sensor-based HAR systems conveyed that the quality of the activity recognition depends on the kind of sensors used and their location. For example, having sensors on different parts of the body could give different dynamics of the movement and, therefore, different accuracy levels. Also, the integration of data from several sensors can improve the reliability of the recognition system. The intricacy of the human movements and the requirement of real-time processing is also an issue in developing efficient HAR systems.

2.2.2 An Ontology-Based Approach to Activity Recognition

Ontology-based methods in HAR entail the use of well-defined knowledge models to describe and classify activities. Ontologies are a formal manner of defining the relationship between a set of activities and the environment in which they are undertaken. This approach allows the system to put the data from the sensors into a particular context, thus increasing the efficiency of activity recognition.

Ontology based HAR system can be divided into three layers: the sensor layer, the reasoning layer, and the application layer. The first one contains the raw data in the form of sensor readings, the second one is responsible for the data interpretation according to the defined ontologies and the last one is the application layer which gives the final activity recognition output (see Figure 1). This approach is helpful in identifying complex patterns of activities that may include a series of steps or conditions.

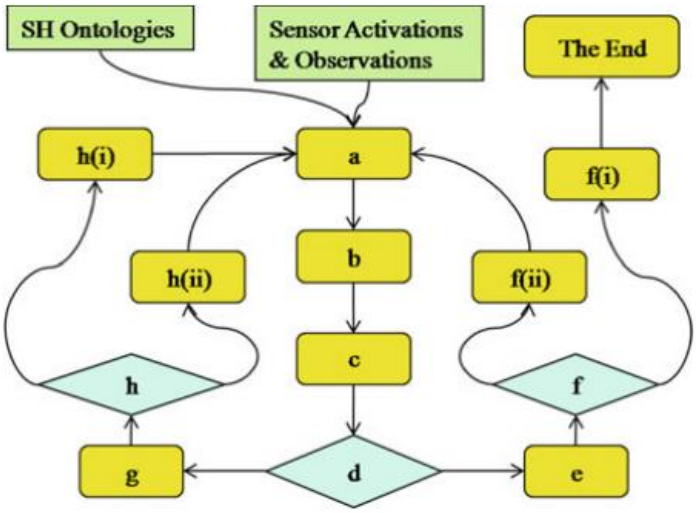

Figure 1. The Ontology-Based Activity Recognition Algorithm. This figure illustrates the flow of the ontology-based activity recognition algorithm, showing the process from sensor activations and observations to the final outcomes, as guided by SH ontologies. (Liming Chen, 2019).

One of the biggest strengths of ontology-based HAR is its capability to incorporate contextual information which is often very important for the recognition of human activities. For instance, differentiating between two quite similar static actions such as sitting and lying is rather problematic if one relies only on the data collected by sensors. But adding contextual information like the time of the day or the place can greatly improve the overall results of the system.

2.2.3 A Hybrid Approach to Activity Modelling

A hybrid approach is the combination of the two or more methods for activity modelling that can help in enhancing the performance of HAR systems. This approach can combine the sensor data with ontological reasoning, machine learning schemes and other components of activity recognition.

Hybrid models may take advantages of all the approaches combined. For instance, data collected using sensors can give out the physical movement's details, while ontology-based reasoning can incorporate context. The data collected can then be processed by machine learning algorithms to determine relation between the data and certain activities.

The HAR framework for the use of hybrid models can be split into several stages, namely, data acquisition, feature selection, modelling and testing. This approach means that there is need to consider how different components will be integrated in the system.

2.2.4 Semantic-Based Sensor Data Segmentation

Semantic-based sensor data segmentation is the process of partitioning the sensor data into a

number of elements depending on the activities being done. This process is very important for finding the right activities and their further classification. It divides the data into different segments of activities based on certain rules or by using machine learning techniques.

Human activities are diverse and, therefore, the segmentation process can be complicated. The problem arises from the fact that there can be similarities between the activities which makes the classification of the data not very precise. Also, the requirement of real-time segmentation brings extra challenges into the problem.

Semantic-based segmentation can enhance the performance of HAR systems since the data that is fed into the recognition models is relevant and labelled appropriately. This process can be further improved by using contextual information and knowledge to set the segmentation rules.

2.2.5 Time-Window Based Data Segmentation

Time-window based data segmentation in HAR is a method where the collected sensor data is divided into fixed-length windows. Each window represents a period within which a certain activity is presumed to take place. This technique helps manage the continuous data streams by dividing them into smaller units.

Among these techniques time-window based segmentation is one of the most common methods and the choice of window length is a crucial factor. A window size that is too small may not contain enough information to recognise an activity, while a window size that is too large may contain more than one activity and hence the recognition becomes difficult. Researchers often experiment with different window lengths to find the optimal one for the given application.

Time-window based segmentation is widely used in HAR due to it is effective and not hard to implement. It enables efficient management of the data from sensors and can be applied in real-time systems. However, it also has some constraints, for example, it assumes that only one activity takes place in each window which may not always be true in most real-world scenarios.

2.2.6 Composite Activity Recognition

Composite activity recognition involves recognising complex activities that are composed of a number of simpler actions. This approach is useful, especially when one needs to model a higher level of behaviour, for example, a daily routine or a sequence of tasks. Composite activities often involve temporal and contextual dependencies making their recognition more difficult than the identification of simple activities.

To recognise composite activities HAR systems has to analyse sequences of actions and understand the relationships between them. It is possible to use hierarchical models to make this process easier, where simple components of the activity are defined first and then integrated to define more complex behaviour. Other machine learning algorithms which can work with sequential data include Hidden Markov Models (HMMs) and Long Short-Term Memory (LSTM) networks.

In an composite activity recognition the main issue is modelling the relations between the different actions. This is something that has to be trained on a huge data and with algorithms that can model human behaviour patterns. Furthermore, the variability in how people perform the same composite activity in unique ways complicates recognition even further.

2.2.7 Smart Homes for HAR and ADL (Activities of Daily Life)

Smart homes equipped with sensors present a perfect environment for HAR and monitoring of Activities of Daily Life (ADL). These environments are capable of recognising residents' activities and then performing health monitoring, care of elderly and improving their quality of life.

In smart homes, sensors are usually placed into different objects and places, interior and furniture, devices and rooms. These sensors gather data on the interactions and movements of residents which is then used to identify activities. The integration of HAR systems in smart homes enables the monitoring to be done without having to wear or carry any devices.

The most significant advantages of smart homes for HAR are the availability of a naturalistic setting for monitoring and the possibility to identify and manage various incidents that may occur, including falls or changes in an individual's behaviour. However, these systems also have their problems to solve, including privacy issues, how to deal with the big data and how to identify the activities correctly in a complicated and frequently changing environment.

2.3 Supervised Classification Methods

2.3.1 Linear and Quadratic Discriminant Analyses (LDA, QDA)

Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) are traditional techniques for classification. While both methods assume the data points of each class has to be normally distributed, they have different assumptions about the covariance of the data. It has been postulated that different classes have an identical covariance matrix in the LDA model. This leads to having straight lines as decision surfaces between the classes. The decision rule in LDA involves computing the linear discriminant function for each class and assigning the observation to the class with the highest score:

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log(\pi_k)$$

Where *x* is the feature vector of the input data point. In other words, it is a vector that contains the values of the variables (features) which characterise a single instance of the data;

Sigma is a symbol that stands for the covariance matrix. In LDA it is assumed that all the classes have the same covariance matrix which describes the variability of the features and the relationships between them;

mu_k is the mean vector for the class *k* which is the average of the data points in the class *k* and is also an *n*-dimensional vector where *n* is the number of features;

pi_k is prior probability of class *k* – it is the probability density of *k*-th class and is the ratio of number of instances of class *k* with total number of instances. It shows the extent to which an observation would be expected to belong to class *k* prior to consideration of the features.

Quadratic Discriminant Analysis, on the other hand, allows each class to have its own covariance matrix, resulting in quadratic decision boundaries. The decision rule in QDA involves a quadratic discriminant function:

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log(\pi_k)$$

LDA is preferred more when the training data set is small since it has fewer parameters than QDA. Nonetheless, when the assumption of identical covariance matrices is violated, QDA may show a better performance as it allows for different covariance structures for each class. It is also important to bear in mind that, whereas LDA and QDA can work on data that is not normally distributed, their efficiency may still be affected.

2.3.2 Decision Tree

Decision Tree is a model used for both classification and regression problems. It is based on the splitting of the data into subsets depending on the feature values creating a tree-like structure of the decisions made. Each node corresponds to a decision based on the feature, each branch represents the outcome of the decision, and each leaf node denotes a class label. Decision Trees are easy to understand and explain which makes them a popular choice used for many applications. Nevertheless, they tend to overfit, particularly when a deep tree is used and the training data is noisy.

2.3.3 Random Forest

Random Forest is one of the ensemble classifiers which constructs several decision trees during the training process and the final output of the model is the class that occurs most frequently (classification) or the average of the individual trees (regression). It introduces randomness in two ways: The first method is the bootstrapping technique in which each tree is trained on a bootstrap sample of the original data (bagging), the second technique is the feature selection in which at each split a random subset of features is chosen. This helps to avoid overfitting and enhance the predictability of the model in comparison to a single decision tree model. Random Forests are specialised for the cases where the data has a great number of features and a high level of feature interactions which is the case for time-series classification. In HAR Random Forests can deal with variability of sensor data and model complex dependencies between variables. The algorithm's ability to compute feature importance also helps in understanding which features contribute most to the classification performance scores.

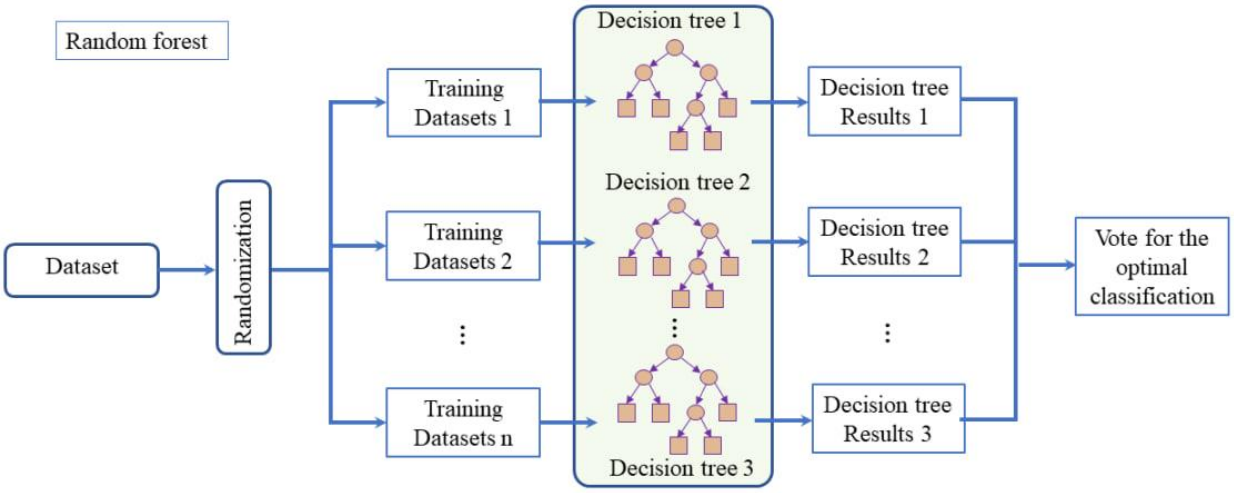

Figure 2. Random Forest Structure (Qiancheng Tan, 2023)

In a random forest, the ultimate classification decision is made by training over a period of *k* rounds producing a series of classification models {h1(X), h2(X), … , hk(X)} and implementing them to build a multi-classification model system. The final classification outcome of this system is chosen by means of a majority voting procedure:

$$H(x) = \underset{Y}{\operatorname{argmax}} \sum_{i=1}^{k} I(h_i(x) = Y)$$

where *Y* is the output variable, h_i is a single decision tree classification model, and *H(x)* is a multi-classification model.

2.3.4 XGBoost

XGBoost is the more sophisticated version of the gradient boosting algorithm that is intended to work faster and more efficient. It constructs a set of 'weak learners,' which are usually decision trees in a serial fashion. Every tree also has the ability to correct the error of the previous trees, and the model is trained with gradient descent (Grus, 2015). XGBoost can be used in order to model the patterns in sensor data and is more effective in handling non-linearity and feature interactions in HAR.

2.3.5 Multilayer Perceptron (MLP)

A Multilayer Perceptron (MLP) is a form of Artificial Neural Network which contains several layers of nodes in which all the nodes of one layer are connected to all those of the next layer. MLPs are able to handle non-linear problems through backpropagation where the model's parameters are adjusted to minimise on the prediction error.

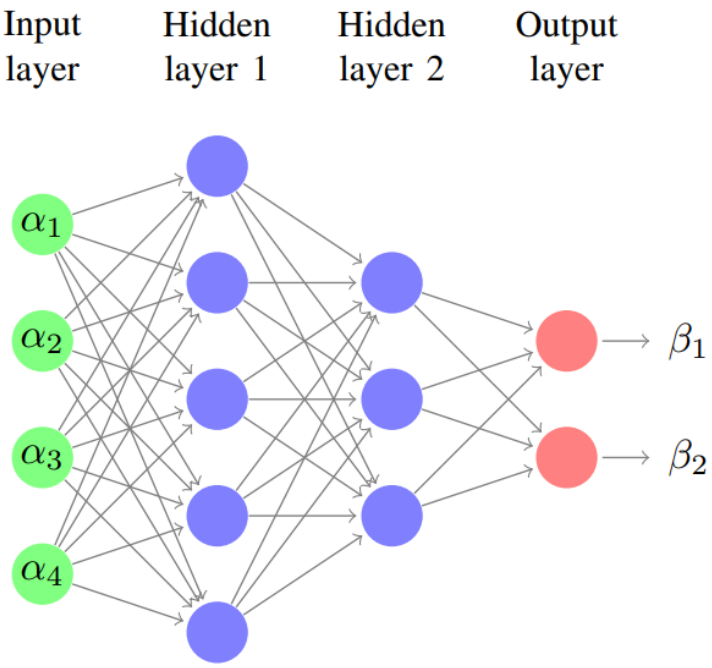

Figure 3. Multilayer Perceptron (MLP) Structure. The input layer (green circles) receives input features which are then processed through two hidden layers (blue circles) that apply learnt transformations. Each neuron in a layer is fully connected to neurons in the subsequent layer and the final output layer (red circles) produces the network's predictions. (Konstantinos Benidis, 2022)

MLPs are very flexible and can be employed for classification problem of almost any kind. They can be either trained on the raw time-series data or on some feature vectors that have been derived from the time-series data. However, they require tuning of hyperparameters like the number of layers, number of nodes per layer, the learning rate, and the regularisation. MLPs can model complex relationships in the sensor data and give precise classification of activities in terms of HAR.

2.3.6 Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNN) are specialised neural networks designed to process structured grid data such as images and time-series. CNNs use convolutional layers to automatically learn spatial hierarchies of features from the input data. Each convolutional layer applies a set of filters to the input capturing local patterns and reducing dimensionality through pooling layers.

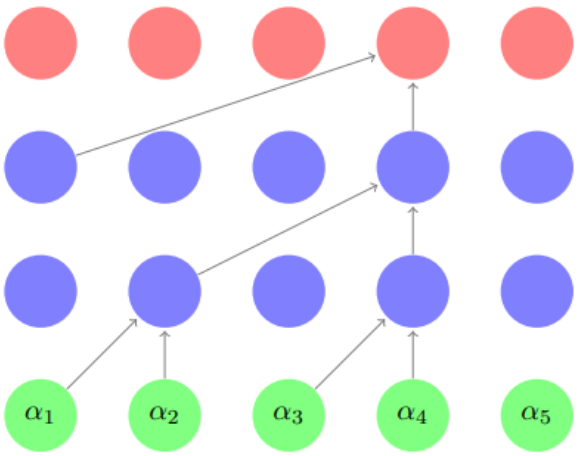

Figure 4. Structure of Convolutional Neural Networks (CNN). In every layer of a CNN there are neurons which are connected to a small portion of the input or the previous layer and thus preserve spatial relationships and decrease the number of parameters. The red circles at the top are the final layers of the network where the extracted features are combined to produce the network's output or classification. The arrows show the flow of information from one layer to the next illustrating how CNNs construct increasingly complex representations of the input. (Konstantinos Benidis, 2022)

CNNs can come in handy when dealing with spatial or temporal dependencies which is the case with time-series classification. For instance, CNN can detect features such as the repeated actions over time like walking where steps are consistent and different from shuffling or descending the stairs. The algorithm can pick up intricate patterns from raw sensor data with no need for feature engineering. The proposed system for HAR using CNNs can learn the relevant features from the sensor data and classify the activities based on the local and global patterns associated with the labels.

2.3.7 Recurrent Neural Networks (RNN)

Recurrent Neural Networks (RNNs) are a type of neural network that are effective to work with data that is sequential in nature, for instance time-series, natural language processing, and speech recognition. Unlike other neural networks that assume the input to be independent, RNNs incorporate feedback loops which keep memory of previous inputs making the model powerful for tasks where order of data is important.

In an RNN the hidden units are connected by a time delay as shown in the Figure 5 creating a recurrent structure. This repeated connection enables the network to keep information and pass it on from one time step to the next and therefore the network has memory. It works because the same set of weights or parameters is used at every step meaning that one can train the network on sequences of varying size and the network will still be able to process other sequences on unseen data.

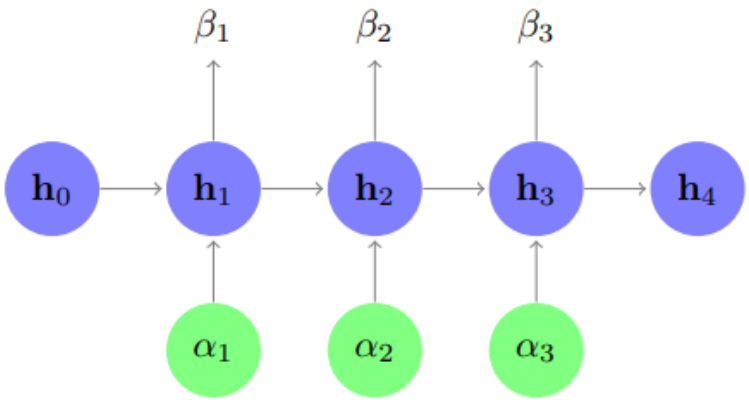

Figure 5. Recurrent Neural Networks (RNN) Structure. The hidden state at each time step denoted as h, is connected to the next time step with each hidden state also receiving an input from the previous

time step's hidden state. In addition, external inputs (alpha) influence the hidden states and there are parameters beta that guide the connections between the steps. (Konstantinos Benidis, 2022)

However, training of RNNs is quite complicated, especially when training on long sequences. The problem that stems from this is that sometimes the gradients can either vanish or explode which makes the learning process and performance poorer. Nevertheless, RNN is a fundamental algorithm for tasks in which sequential data prevails since it can learn the temporal dependencies in the data.

2.4 Measure of Performance

The performance of each model is assessed using the following metrics:

Accuracy is the percentage of correct predictions made over the total number of instances.

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

Precision is the ratio of the number of true positives to the total number of positives.

$$\text{Precision} = \frac{TP}{TP+FP}$$

Sensitivity or recall calculates the number of true positive predictions out of all actual positive cases.

$$\text{Recall} = \frac{TP}{TP+FN}$$

F1 score is the mean of precision and recall.

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Cohen's Kappa measures the extent of the agreement between the predicted and the actual labels with the consideration of the chance agreement.

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

22.5 Time-Series

Time-series data is a critical aspect in the HAR research as human actions are sequential and temporally related series of events. Time-series data consists of sequences of observations recorded over time at regular intervals. In HAR, this information is often recorded by using wearable devices, for example, accelerometers and gyroscopes which keep track of movements. This is especially important on the

temporal dimension of this data, because the same sensor readings may reflect different activities if taken in different order. For instance, a sharp rise in sensor time-series variance could mean walking or climbing the stairs depending on what comes before and after it.

2.5.1 Time-Series HAR Data and Analysis

There are some challenges we might face when analysing time-series data in the context of HAR. Firstly, the temporal dependencies in the given data should be modelled well enough so that it is possible to distinguish between different activities. This paper deals with time-series non-stationarity where the statistical properties of the data are not constant over time, and thus application of some traditional statistical techniques might be challenging.

In the HAR domain, time-series analysis helps in determining the patterns and trends related to certain activity. For instance, sensor readings for walking, sitting or standing may be defined as having specific sequences in a particular order in which they occur. In this way, through time-series analysis, it is possible to obtain attributes such as periodicity and trends that characterise various human activities.

2.5.2 Supervised Classification Methods for Time-Series HAR Data

Supervised classification methods play an important role in time-series analysis in HAR. These methods are quite different in terms of the way they address the temporal aspect of the data which is critical for applications in HAR. They include the traditional classifiers which are easy to understand and explain such as LDA and QDA. Each of these methods suggests that the observations are independent of each other which is not always the case for the time-series data since the order and temporal relations between the observations play a crucial role. This can be a disadvantage in HAR since the time component and the sequences of the activities as well as transitions are important. Under LDA and QDA, 'independent observations' denotes a situation where each of the observation points in the dataset is treated as being mutually exclusive and not connected in any way to the other points. These models suppose that a given target variable (or label) of one data point is irrelevant to the target variable of another data point. For instance, if the activity is being predicted in a HAR dataset, LDA and QDA assume that every activity label like 'walking', 'sitting', is independent of activities in the past or future time-points. In time-series data, though, the observations are often not independent. The value or label of a certain time is normally related to the values or labels of other time points in the past or future (for example, at a certain moment the person is "walking" and it increases the probability that at the previous moment the person was also "walking" and at the next moment will also be "walking"). This is an important aspect of dependency in time-series data which requires predictive models that respect the time component. Thus, LDA and QDA can fail to capture these dependencies when applied to time-series data and, hence, are likely to provide less accurate or meaningless results.

Advanced algorithms like Decision Trees and ensemble models as Random Forests and XGBoost are more flexible as they can capture non-linear patterns present in the data. Nevertheless, similar to LDA

and QDA, these techniques do not explicitly take into consideration temporal dependencies of the observations. To overcome it, other methods such as feature engineering, creating lagged features or moving averages are used to capture time dependence in the model.

On the other hand, there are deep learning models that can handle time-series data - MLPs and CNNs which are more suitable for HAR tasks. While MLPs are quite strong, they sometimes need to be fine-tuned to incorporate temporal information effectively. CNNs usually used in image processing have been applied to time series data with techniques such as time-distributed convolution which allows the model to recognise patterns in movements over time.

Since we want to make the most of the sequential nature of time-series data, using models that can cope with the time component in time-series data like Recurrent Neural Networks (RNNs) should be efficient for application in HAR. RNNs are a type of neural network designed to work with sequences of data, they have a memory which allows them to consider temporal relations between different timesteps. This puts RNNs in a very good position for tasks where the time order of events matters.

2.6 Summary

Human Activity Recognition (HAR) is an area of study that aims at recognising and understanding human actions from sensor data. It covers different ways such as sensor-based recognition, ontology-based methods, hybrid methods, and segmentation techniques. The chapter also describes supervised classification techniques including LDA, Decision Trees, Random Forest and others including XGBoost, MLP, CNN, and RNN in the context of HAR. Last but not least, we focused on the role of time-series data in HAR and underlined that the algorithms can be effective for human activity recognition if they consider temporal dependencies in the modelling process.

Chapter 3

Methodology

3.1 Introduction

In this chapter, we outline the methodology used to analyse the HAR70+ dataset for human activity recognition (HAR). This includes a detailed description of the dataset, the exploratory data analysis (EDA) performed to understand its characteristics, and the data preparation steps undertaken to ensure its quality. We also discuss the sampling methods and machine learning models applied to the data, aiming to achieve accurate and robust activity recognition among senior citizens.

3.2 Data HAR70+

The HAR70+ dataset is a collection of time-series data designed for human activity recognition (HAR). It includes recordings of 18 older adults aged 70 to 95 and performing various daily activities. The data is valuable and curious to work with because it captures the real-life physical activities of senior citizens in a semi-structured free-living environment providing realistic and diverse data for training and evaluating machine learning models.

During data collection, five participants used walking aids. To capture a wide range of daily activities in natural contexts, data was gathered in a semi-structured free-living setting both indoors (participants' houses) and outdoors. Two Axivity AX3 accelerometers were employed. One was attached to the lower back and the other to the right front thigh. These are triaxial sensors that are lightweight and record acceleration data at a 50 Hz sampling rate. Each participant had a GoPro Hero 8 camera placed to their chest to record video data that was utilised to annotate actions (see Figure 6). The lower body movements were captured by the camera's orientation which offered a visual reference for precise activity labelling.

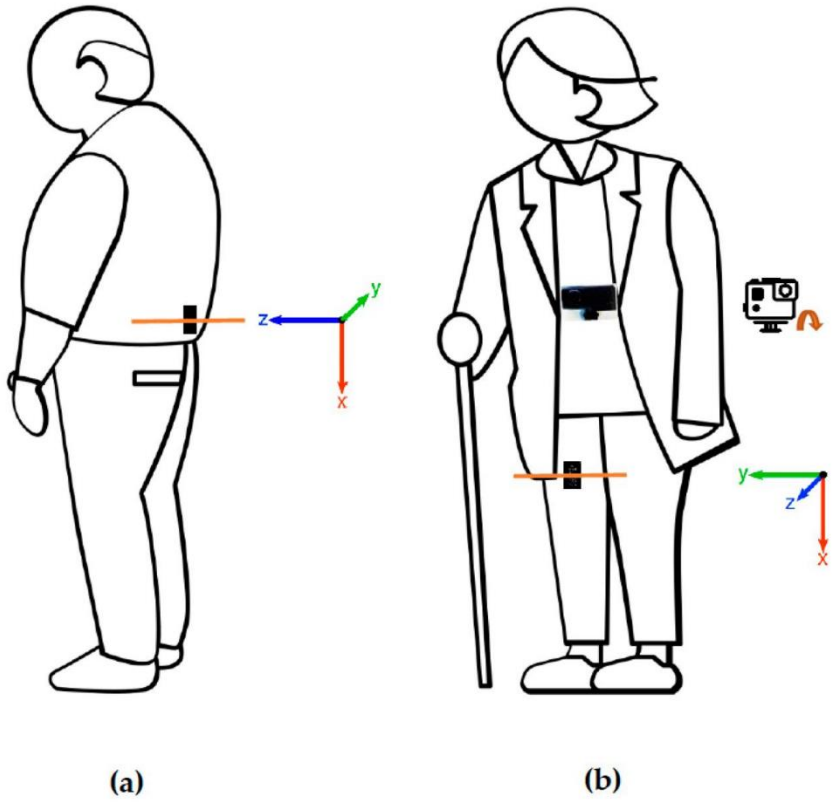

Figure 6. Sensor and Camera Placement for HAR70+ Dataset Collection. This picture shows the locations of the chest-mounted camera used for the HAR70+ datasets as well as the two skin-attached accelerometers (highlighted with orange lines). (a) The third lumbar vertebrae served as the central location for the back accelerometer. The coordinate system's z-axis was orientated forward. (b) The accelerometer on the thigh was positioned about 10 cm above the patella's top border. The z-axis was

orientated in reverse. The camera was affixed to the exterior of the garment at the chest, orientated downwards. (Astrid Ustad, 2023)

Both inside and outside participants repeated a range of daily life activities. The protocol's design ensures that standing, sitting, and lying last for at least five minutes, and walking for at least fifteen minutes. In order to synchronise the accelerometer data with the video recordings, individuals were asked to do heel drops. If this was not possible, the accelerometers were clapped together in front of the camera by the researchers prior to connection. Data from each participant is kept in different CSV files. The following columns are part of the files' structure:

- *timestamp*: The day and time of every sample that was taken;
- as *back_x*, *back_y*, and *back_z*: acceleration data from the back sensor along the x (down), y (left), and z (forward) axes;
- *thigh_x*, *thigh_y*, and *thigh_z*: the thigh sensor's acceleration information along the x (down), y (right), and z (backward) axes.
- *label* indicates an annotated code of an activity being carried out, namely: 1: Lying, 2: Sitting, 3: Standing, 4: Shuffling, 5: Walking, 6: Stairs (descending), 7: Stairs (ascending).

Activities were annotated frame by frame using video recordings. For consistency, predefined descriptions of when each sort of activity started and stopped were determined. To guarantee the accuracy of the data representation walking aids were considered throughout the categorisation process. The HAR70+ datasets were collected over a period of approximately 40 minutes for each participant. The data collection took place at different times of the day for different participants. Each participant's data was recorded continuously with accelerometers sampling at 50 Hz, meaning that data points were recorded every 20 milliseconds. However, in some instances, the data was recorded at slightly varying intervals due to the natural variations in daily activities and sensor performance. The HAR70+ dataset provides extensive and well-annotated time-series data that reflect a variety of physical activities of senior citizens in real-world settings making it a valuable tool for researchers working on human activity recognition.

3.3 Exploratory Data Analysis

The Exploratory Data Analysis (EDA) has been done on the HAR70+ datasets extracted from multiple CSV files including ‘senior501’ – ‘senior518’. Each dataset comprises timestamped recordings of six sensor variables: back_x, back_y, back_z, thigh_x, thigh_y, thigh_z and an activity label. The first part of the analysis deals with the structure of the HAR70+ datasets, the unique values, the descriptive statistics and the assessment of missing values.

We begin with the analysis of the structure of the data by looking at the datasets individually. Each dataset consists of eight columns: timestamp, back_x, back_y, back_z, thigh_x, thigh_y, thigh_z and label. The number of samples in each dataset varies with ‘senior501’ having 103860 items while ‘senior518’ has 141714 items which is good enough for patterns analysis and modelling later.

As shown by the unique value count the data is quite diverse when it comes to the sensor readings. For instance, the participant ‘senior501’ shows 103,860 distinct timestamps meaning that there are no duplicates in this data set. It was observed that the sensor readings for back_x, back_y and back_z have distinct values of 6424, 4052 and 5594, respectively whereas thigh_x, thigh_y and thigh_z have higher variability with unique values of 11007, 6545 and 9539, respectively. The actions recorded 7 distinct activities in the ‘senior501’ dataset, although the number of classes in all the datasets ranges between 5 and 7 classes.

This is also supported by the descriptive statistics which also illustrate dispersion of the sensor readings. For example, in ‘senior501’ the back_x averages were about -0. 88 with the standard deviation of 0. 34 and values vary from -2. 90 to 0. 24. The same tendencies in mean, standard deviation, and range are seen in back_y, back_z, thigh_x, thigh_y, and thigh_z. The comparison of the real mean and variance values for every patient showed that every person has slightly different average values, but the variance patterns are similar for all the patients. This means that, despite the variation in the average sensor outputs across participants, the variability in the results is uniform across all participants.

As part of Exploratory Data Analysis for the HAR70+ we checked data on missing values and the results showed that there are no missing values in the datasets. This ensures that the data is ready for the next step in the analysis without having to undertake data imputation.

The initial exploratory analysis of the HAR70+ datasets is presented characterising their structure, variability and completeness. The data has no missing values, unique value counts are high, descriptive statistics are provided in detail and the structure of the databases is uniform which mean the datasets are ready for further descriptive and predictive analyses.

3.3.1 Data Distribution - Histograms

The data distribution histograms of the sensor readings for participants ‘senior501’-‘senior518’ provide an understanding of variability of the data. Different patterns can be seen from the histograms of each sensor variable within the HAR70+ datasets which can help to identify trends that will assist in understanding the recorded activities.

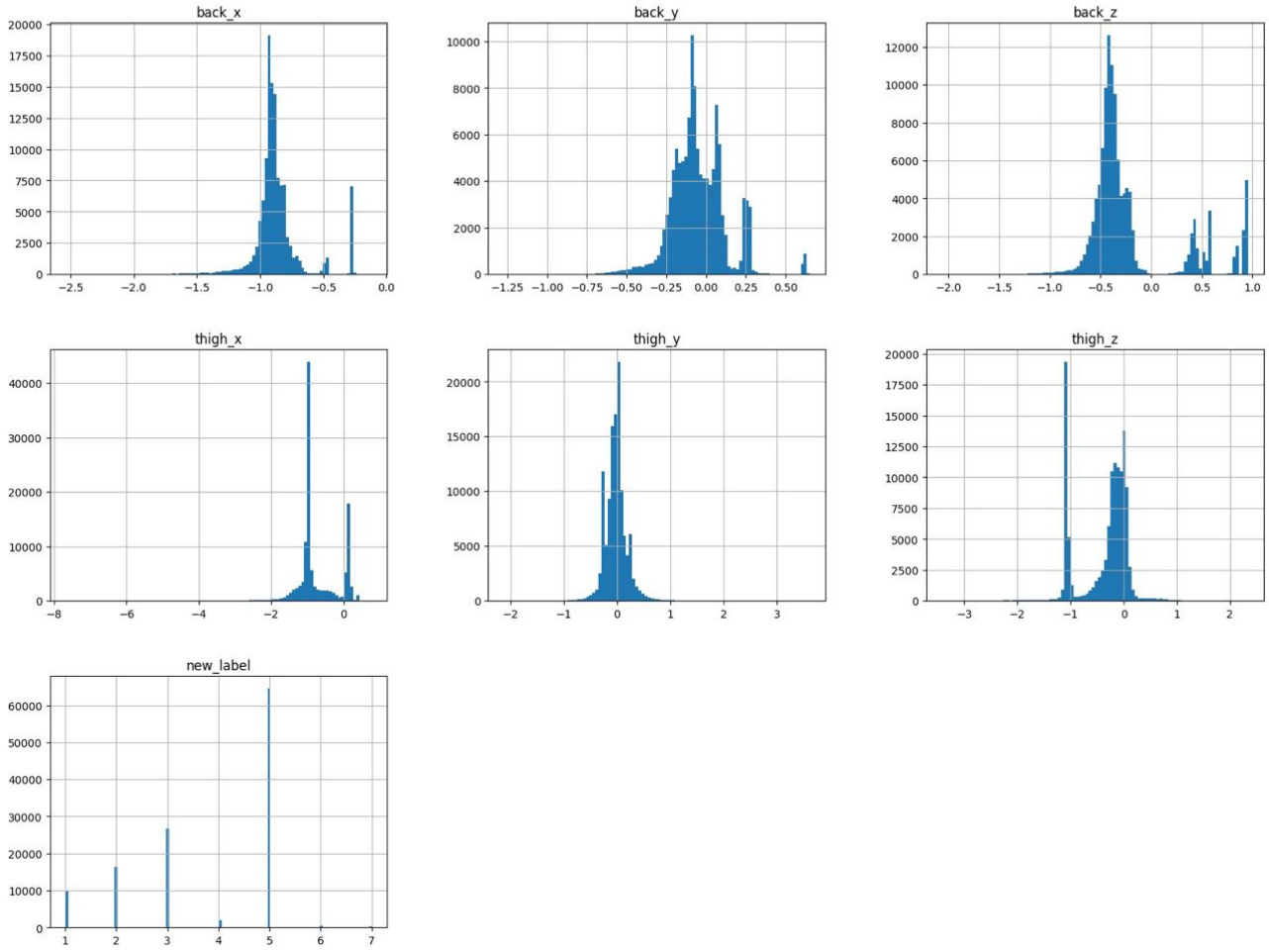

Figure 7. Senior507 Data Distribution Histograms. This figure shows the histograms of sensor variables back_x, back_y, back_z, thigh_x, thigh_y, thigh_z of the single participant senior507 from the HAR70+ dataset. Each subplot contains the frequency of each sensor value, it allows for understanding of the data dispersion and possible central tendencies. On the x-axis shows the data from the sensor, whereas on the y-axis, there is the frequency of the values for every bin. The last subplot is the activity label which shows the number of times each activity was captured in the data. These histograms convey the visual representation of the data distribution in relation to the sensors axes and labels.

The histograms for the back_x, back_y, and back_z of 'senior507' show the data has a tendency of clustering in certain intervals. The back_x readings are mainly located in the range of -1. 5 and -0.5 with a clear peak at -1.0. Just as with back_x values, back_y values are also clustered around -0.25 while back_z values show a broader distribution with multiple peaks indicating variability in movements values or orientations of the back sensor during the experiment.

The same can also be observed with regard to the thigh_x, thigh_y, and thigh_z as these variables also exhibit different distribution patterns. The histogram of the thigh_x shows that the data is rather bell-shaped with two distinct peaks at about -4.0 and slightly above 0.0 which shows different leg positions or movements. Thigh_y values are rather small and tightly concentrated around -0.5, whereas thigh_z values have a dominant peak at -1.0 indicating particular movements of the legs.

The new_label histogram represents the number of activities that were recognised in ‘senior507’. The most frequently encountered label is 5 – ‘walking’ while other labels have relatively high counts for ‘standing’, ‘lying’, and ‘sitting’, while ‘shuffling,’ stair ‘descending and ‘ascending’ have much fewer occurrences meaning that when training the model, we are likely to end up with imbalanced classes.

Likewise for datasets ‘senior501’ – ‘senior518’, using the same visualisation approach, similar conclusions can be drawn. All the datasets produce different distribution patterns which are a result of the different activities and movements that are captured. For instance, datasets with a high concentration of ‘walking’ activities will have more spread-out thigh_x and thigh_z values which represent leg movements. In contrast, datasets that include ‘sitting’, ‘lying’ or ‘standing’ may have more density towards some certain values which mean static body positions (see Figure 8).

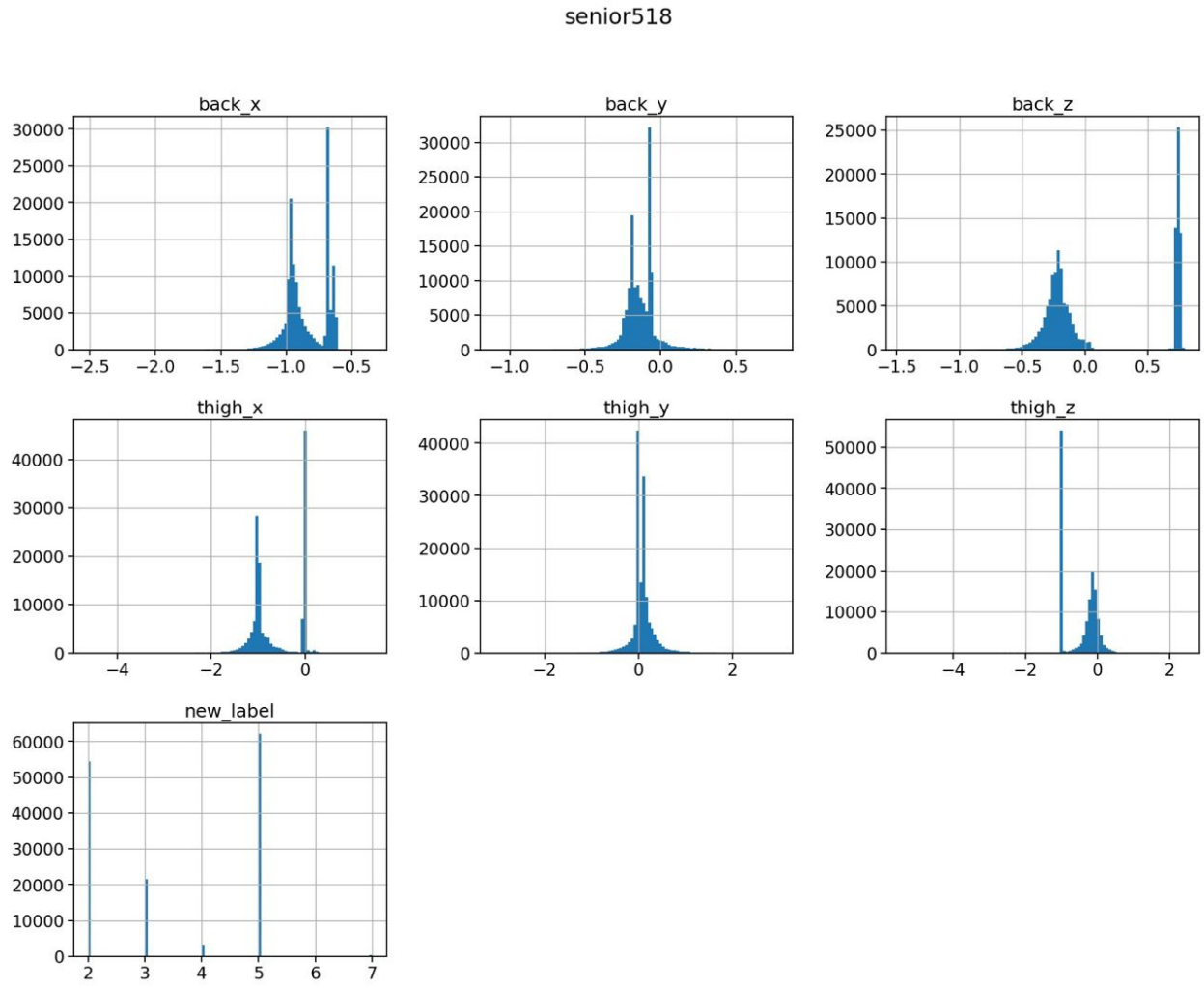

Figure 8. Senior518 Data Distribution Histograms. This figure displays histograms representing the distribution of sensor variables back_x, back_y, back_z, thigh_x, thigh_y, thigh_z for a single participant senior518 in HAR70+ dataset. Each subplot shows the frequency of different sensor values, providing insights into the distribution and central tendency of the data. The x-axis represents the sensor readings, while the y-axis shows the count of occurrences for each bin. The final subplot presents the distribution of activity labels showing a good amount of static activity data, namely, sitting and standing which convey a more concentrated distribution of data.

The data distribution histograms are a useful tool to visualise the data and understand the sensor behaviour as well as to prepare the data for further analysis. It assists in checking the quality and reliability of the data ensuring the sensor readings match the expected activity patterns.

3.3.2 Scatterplots

Figure 9 conveys a pairplot of the entire 'senior513' dataset with variables back_x, back_y, back_z, thigh_x, thigh_y, thigh_z and activities standing, shuffling, walking, sitting, lying. The pairplot matrix illustrates both the individual distributions of the sensor readings and their pairwise relationships colour-coded by the activity labels: blue – lying, yellow – shuffling, green – sitting, orange – standing and pink – walking.

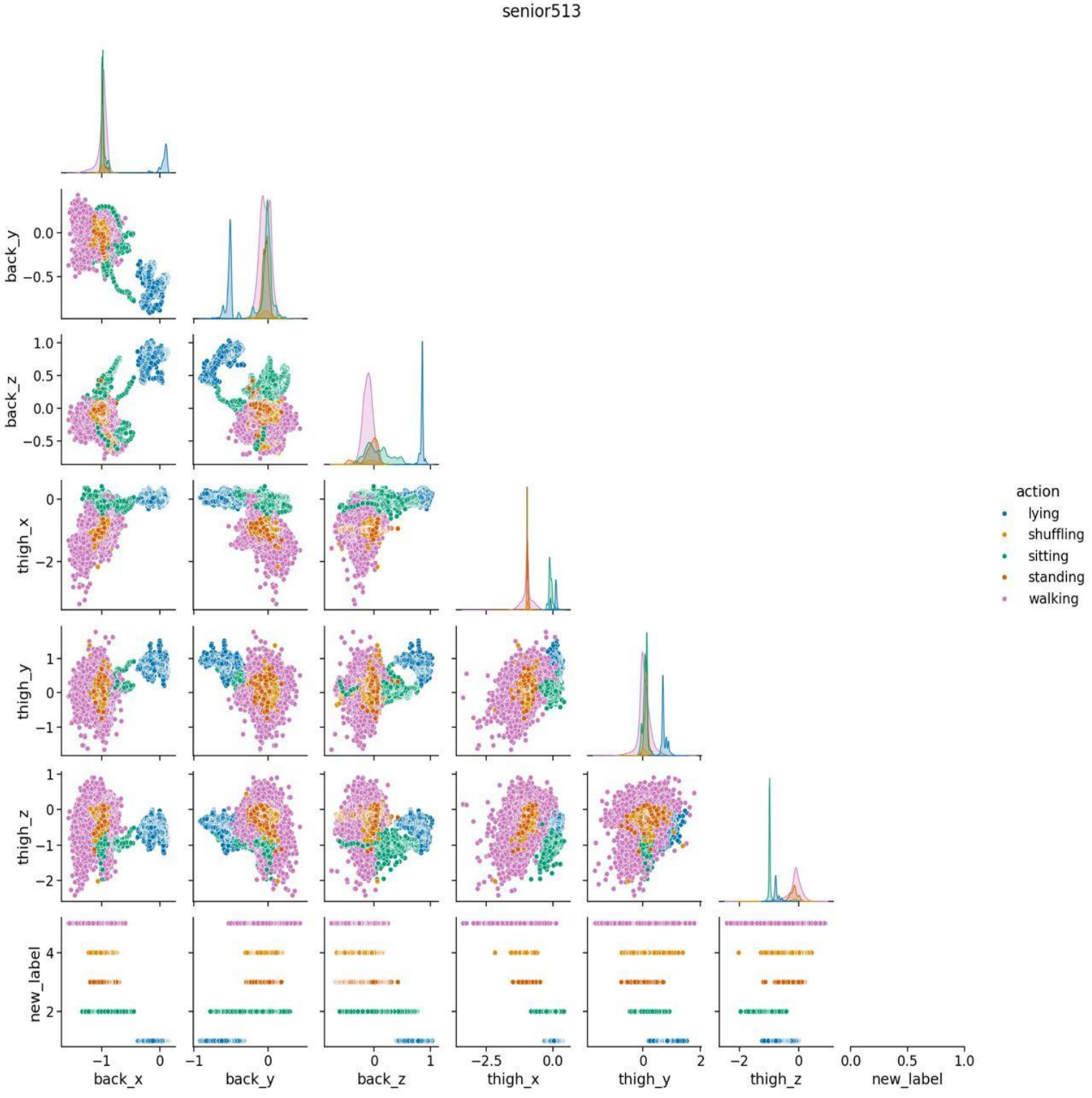

Figure 9. Senior513 Pairplots. The matrix of scatterplots shows the correlation between values from the following sensor features back_x, back_y, back_z, thigh_x, thigh_y, thigh_z for one participant senior513 in HAR70+ dataset across various activities, including standing, shuffling, walking, sitting, and lying. The diagonal plots show the distribution of each individual sensor feature, while the off-diagonal plots illustrate the scatterplots and the corresponding correlations between pairs of features. Each colour represents a different activity, helping to visualise the separation and overlap of activities.

In the diagonal subplots, there are the kernel density estimates of each sensor reading which gives insights into the distribution of the values of each variable. For instance, back_x has a bimodal distribution with two peaks at -1.0 and -0.5 representing positions or movements that are frequently taken or made by subject's back sensor. A similar behaviour can be observed on thigh_x having a peak at around -0.5 which gives an insight of frequent leg orientation or movement.

The off-diagonal subplots display the sensor reading against each other giving the idea about how the two variables related. Colour coding of the data points by activity makes it easier to identify trends related to certain activities. For example, standing (orange) and sitting (green) activities seem to cluster in certain regions in the back_x vs. back_y plot, while lying (blue) is more spread out which could mean that the positions or orientations are different.

When comparing the back_y vs. thigh_x plot, the separated groups can be seen for every activity. The activities of standing and walking have more defined clusters for blue and green respectively and this shows that the sensors data is more consistent during these activities as compared to lying down where the pink data spread is more dispersed.

Some important correlations of the variables can be seen from the pairplot. For instance, the back_x has a negative relationship with back_y for all the activities suggesting that movements in one axis are often accompanied by movements in another. Likewise, we see a strong positive correlation between the thigh_x and thigh_y variables for every activity. The separation and location of the activities in the scatter plots mean that certain sensor readings are indicative for specific actions.

3.3.3 Line Plots for Features and Labels Individually

The line plots of each feature and their corresponding labels in the 'senior504' dataset in Figure 10 is useful for identifying and understanding the temporal behaviour of the data and the relations between them. The sensor readings back_x, back_y, back_z, thigh_x, thigh_y, thigh_z and activity labels are plot against timestamp. The plot gives insights on how the sensor data changes with different movements and how it shifts from one activity and another.

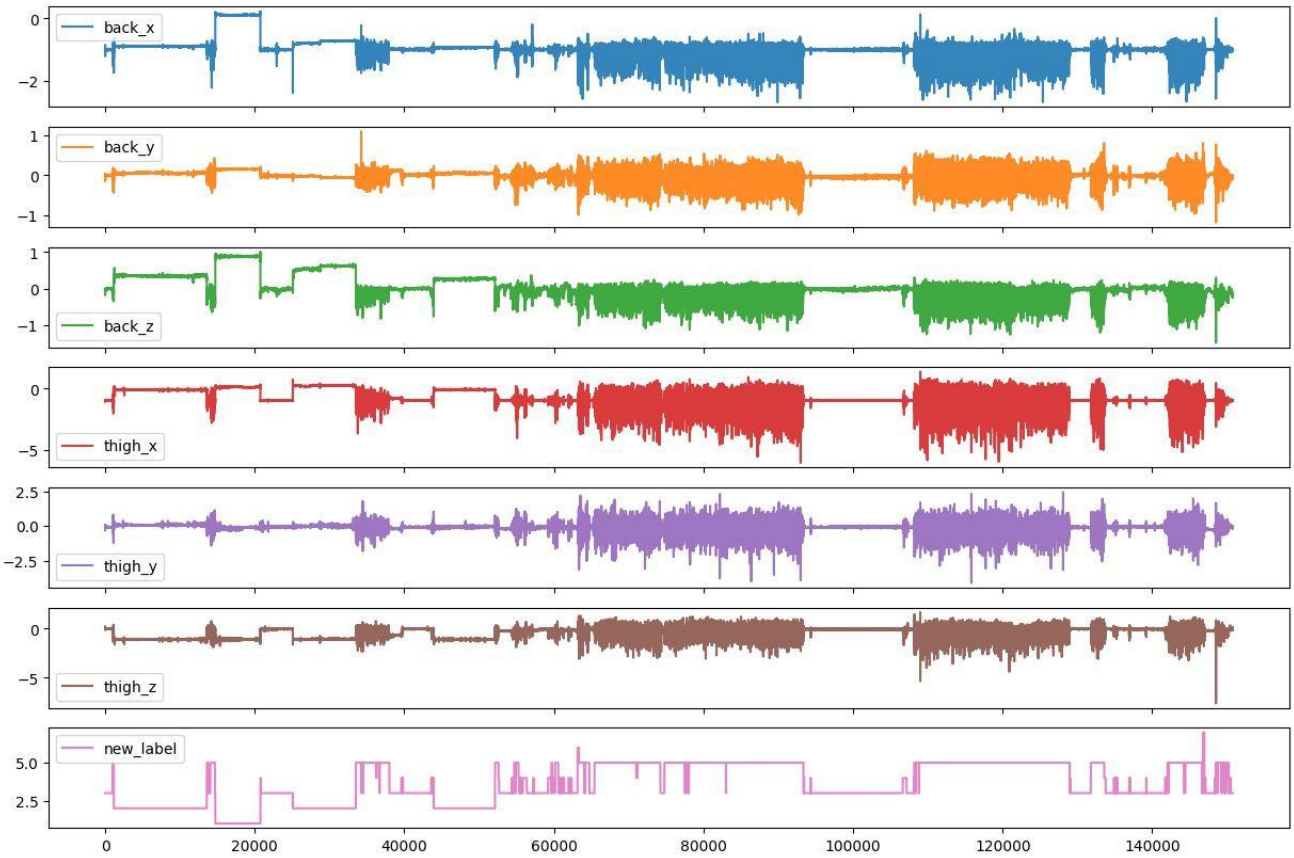

Figure 10. Senior504 Line Plots for Features and Labels Individually. This figure presents line plots showing the time-series data for sensor features back_x, back_y, back_z, thigh_x, thigh_y, thigh_z for a single participant senior504 in HAR70+ dataset and the corresponding activity labels. Each subplot represents sensor axis, with the final subplot displaying the time-series data for the activity labels. The x-axis represents the time or sequence of data points, while the y-axis represents the sensor values or the activity label at each point in time. These plots help visualise how sensor readings change over time and how this variability corresponds to different activities.

From the 'senior504' dataset line plot several observations can be made. Each subplot displays the sensor readings' values in the time domain and the last subplot shows the activity labels transitions.

In the 'back_y' and 'back_z' subplots, the patterns are quite stable with sharp rises from time to time, showing periods of activity or changes in posture. These patterns are important in the recognition of walking, standing, and sitting among other activities. The 'thigh_x,' 'thigh_y,' and 'thigh_z' subplots have more oscillations, especially the 'thigh_x' which is in conjunction with movement of the legs. These variations in the values of features can be easily related to the activities shown in the new_label subplot.

The 'new_label' subplot at the bottom of the figure shows the actual labels of the activities over time. When correlating these labels with the sensor data it is obvious what sensor data fluctuations correspond to a particular activity. For instance, stable levels of back_x, back_y, back_z, thigh_x, thigh_y, thigh_z reflect low activity such as standing or sitting while varying levels reflect activities such as walking

or stair ascending.

Line plots are very useful in the analysis of data and have the following functions. They give a graphical representation of the data variability over time and trends and outliers, which assist in the understanding of the relationship between the sensor data and the physical activities that are being undertaken. Moreover, it is easy to identify anomalies in the data such as sensor failure or mislabelling in the line plot, thus enhancing data quality and accuracy.

3.3.4 Time-Series Plots for Time Intervals of Variables and Actions Combined

The time-series plots created for the HAR70+ datasets represent sensor data across four time intervals: the first 10 minutes, 10-20 minutes, 20-30 minutes and 30-40 minutes. All plots contain variables back_x, back_y, back_z, thigh_x, thigh_y, and thigh_z with the timestamp and the corresponding data sensor readings on the y-axis. Also, the colour bar along the-top of each plot shows the action/label associated with the sensor data value for the activity performed.

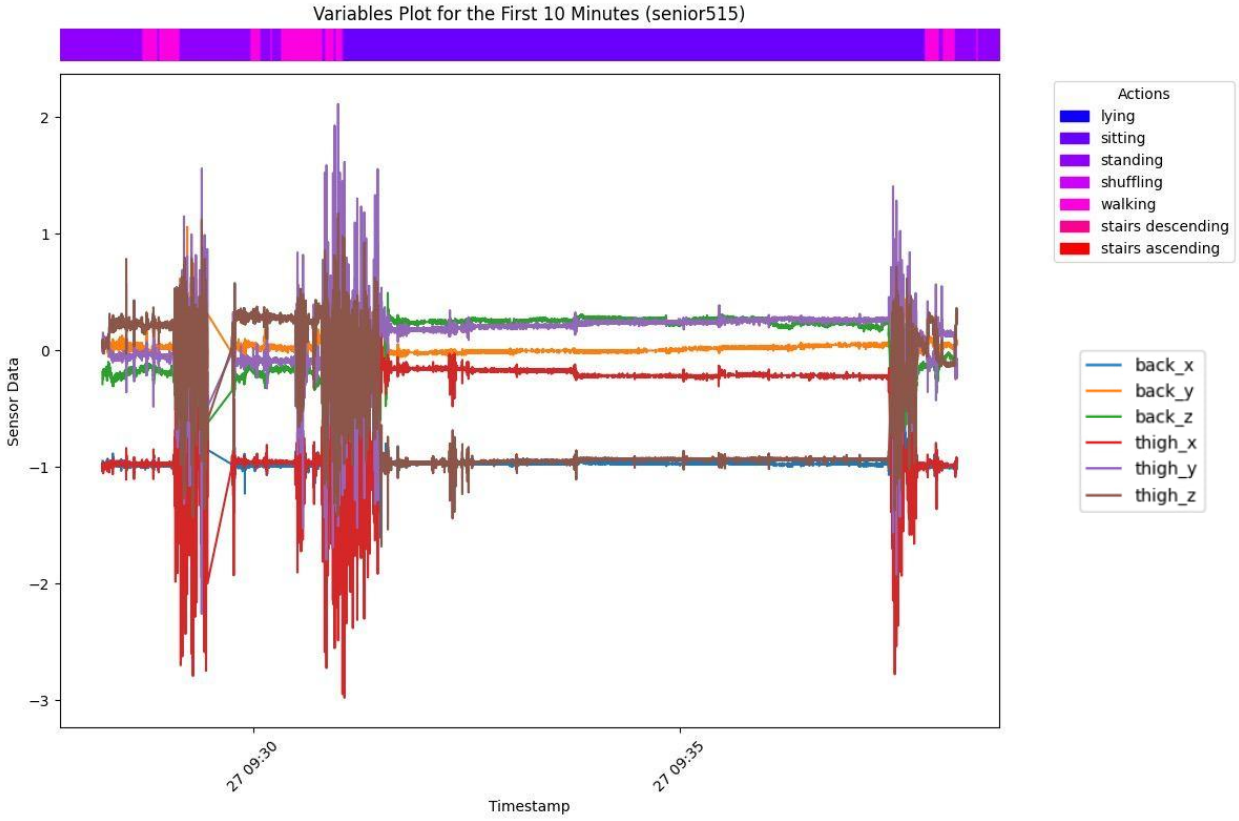

Figure 11. The Variables and Actions plot for the first 10 minutes (senior515) The following are the sensor data variable back_x, back_y, back_z, thigh_x, thigh_y, thigh_z which were recorded during the first 10 minutes of activities done by participant senior515. Here, the plot shows a time-series of the different actions that have been captured by the different sensors (lying, sitting, standing, shuffling, walking, stairs down, stairs up). On the x-axis - time sequence, while on the y-axis, there are the readings of the sensors. Colour coding is used on the right side of the legend to match every action to a particular colour which makes it easier to determine the actions taken and variables. The top bar gives an overview

on which action follows another and how long the actions take.

For instance, in the plot for 'senior515' dataset (see Figure 11) during the first 10 minutes there is a significant variability in the sensor data in the beginning with fluctuations around the -3 to 2 range which indicate a change from high intensity activities to activities with lower intensity and back - standing and walking as depicted by the colour bar. After that, both the colour bar and the features values fluctuations show a drastic change from walking state to sitting state and then once again back to the frequent change between the walking and standing.

The second plot (see Figure 12), for the next 10-20 minutes, likewise shows much fluctuation, particularly in the areas of intense activities. This interval has several periods of standing, walking and sitting and then the last few minutes consist of lying down as depicted in the blue colour bar.

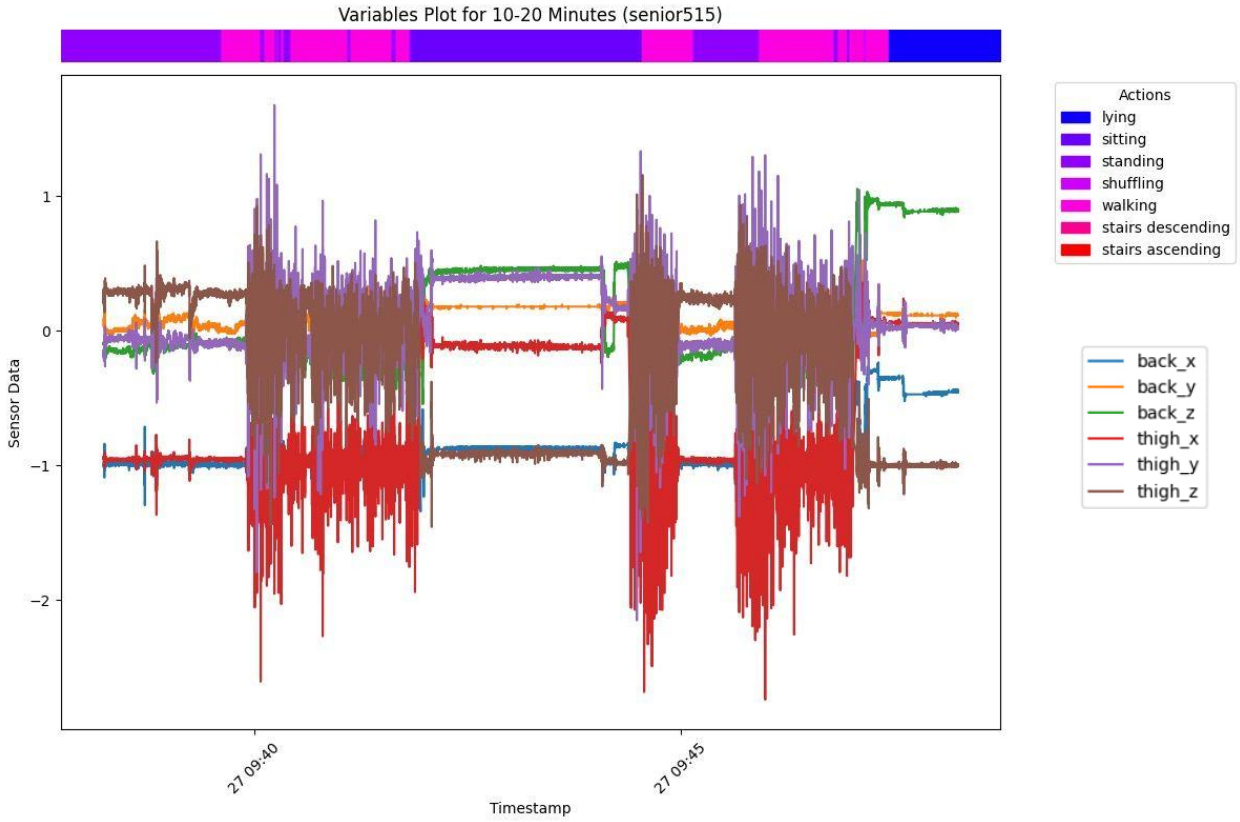

Figure 12. The Variables and Actions Plot for the 10-20 Minutes Interval of the 'senior515'. This figure illustrates the sensor data variables, back_x, back_y, back_z, thigh_x, thigh_y, thigh_z and the actions for the first 10–20 minutes for participant senior515. The plot contains time-series of the different activities where on the x-axis presents the timestamp and on the y-axis presents the sensor data. The legend on the right hand side of the figure connects each action to a colour so that the temporal and spatial distribution of these activities and variables during this period. The top bar shows the general trends of actions with time.

In the third plot (Figure 13) based on the 20-30 minutes of the HAR70+ senior515 time-series dataset the sensor data shows both stable and fluctuating phases. Having analysed the variables' trends, one can notice that they are stable, which specifies the period of low activity with static posture – lying which

is the most common activity at this time interval. But the fluctuations turn to more active movements later in the plot. The colour bar also illustrates these trends – long time of lying are followed by short shifts between walking, standing, and sitting. The data changes from the low to higher activity levels.

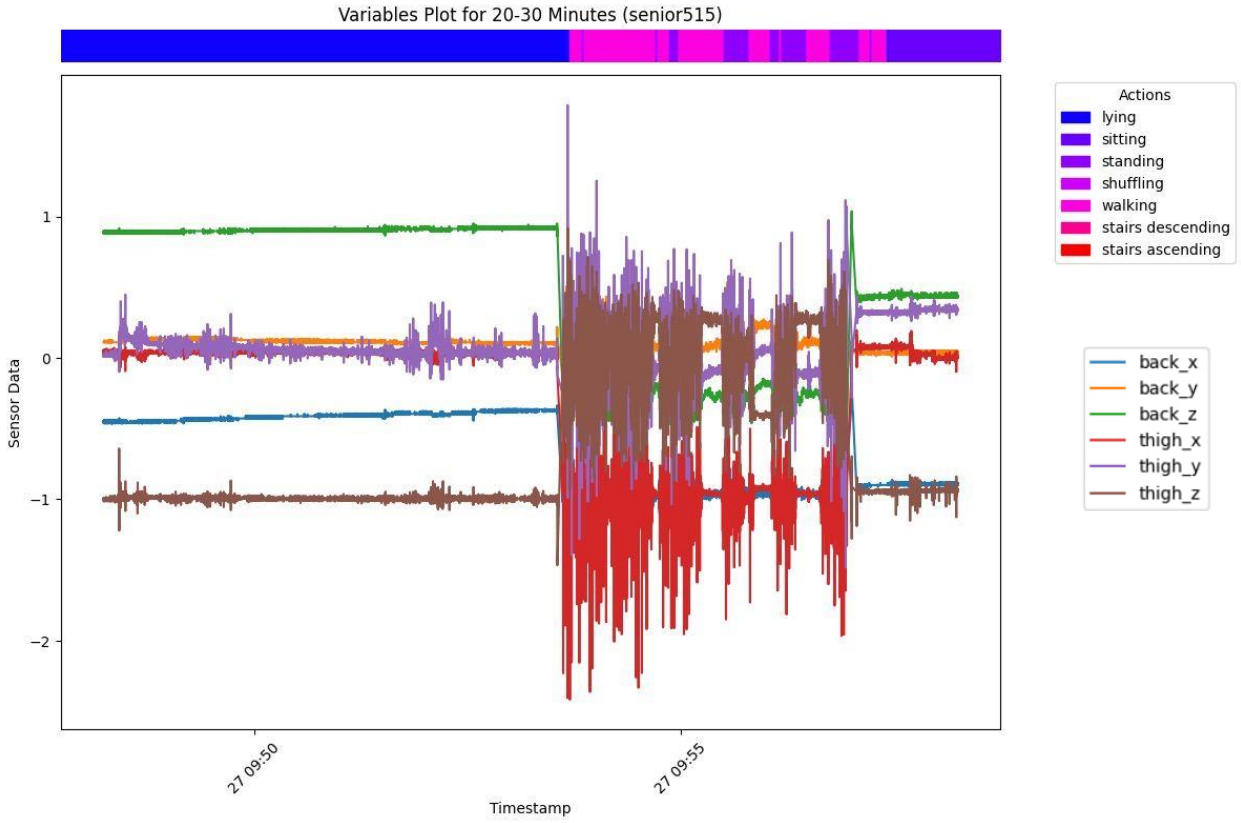

Figure 13. Variables and Actions Plot for the 20-30 Minutes Interval (senior515). This figure displays the sensor data variables, back_x, back_y, back_z, thigh_x, thigh_y, thigh_z and corresponding actions during the 20–30-minute interval for participant senior515. The plot shows the time-series data of various activities, with the x-axis representing the timestamp and the y-axis displaying sensor values. The colour-coded legends indicate the actions and variables, while the top bar provides an overview of action transitions throughout this time period.

Finally, the fourth plot (see Figure 14), when picking 30-40 minutes of the senior515 data, shows a non-changing state of the participant sitting over the first 7 minutes of this period. Then the monitoring period ends with a big range of varying values for standing and walking. The colour bar corresponds with these findings.

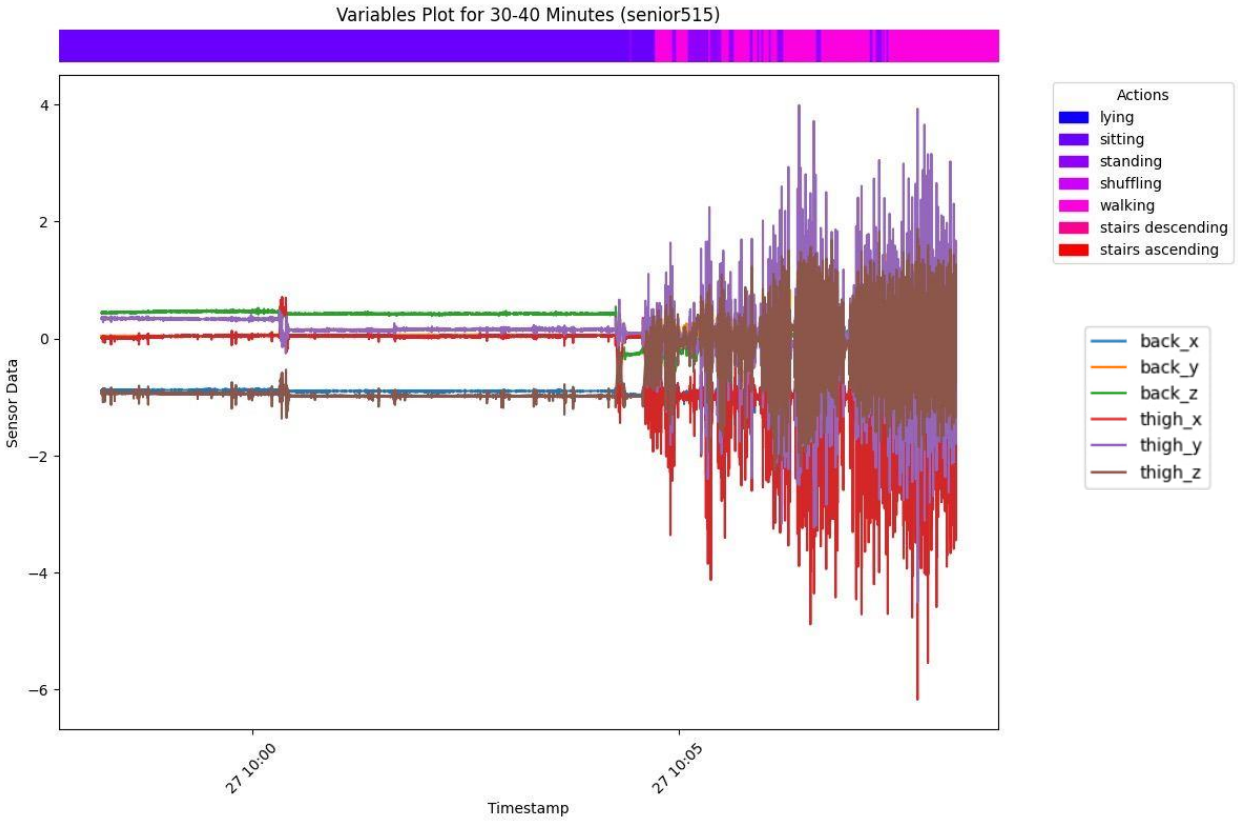

Figure 14. Variables and Actions Plot for the 30-40 Minutes Interval (senior515). The figure shows the sensor data variables, back_x, back_y, back_z, thigh_x, thigh_y, thigh_z and the actions performed during the 30–40 minute interval for participant senior515. The plot represents the time-series data, the x-axis contains the timestamps and the y-axis contains the sensor values. The colour-coded legends links each activity and variable with a particular colour, the top bar shows the sequence of actions during this period.

These plots are important for data visualisation and analysis as they help give a graphical representation of the data collected by the sensors to enable one to observe trends, relations and phases between different activities in the day. They assist in the analysis of the data and provide information about the daily patterns of activity of the participants and could also be used to confirm the efficiency of the data collection and labelling process. The first point that is to be discussed is that the main difference between the activities is in the variability of the measurements rather than the means, the correlation between the different sensor variables is the most crucial factor. For example, dynamic activities such as walking will result to high variability across all the sensor readings while in activities like lying, the variability will be low in each of the sensor reading. This pattern is repeated with all the participants which shows that the variation of the data is the best predictor of the type of activity that is being undertaken.

3.3.5 Line Plots – Variation of Variables' Means Across Actions

For further exploratory data analysis we plot Line plots to conduct a comparative analysis of the variation of mean sensor values across different actions for all the participants. These plots are part of a more general goal of visualising and analysing the data in the HAR70+ dataset for all participants, and

particularly the effects of different activities on average sensor readings.

The plot of the mean values for six actions including lying, sitting, standing, shuffling, walking and stairs descending for the following sensor variables back_x, back_y, back_z, thigh_x, thigh_y, and thigh_z is given in Figure 15 for the senior502 dataset. Figure 16 shows a line plot of the mean values of all the variables in five actions: lying, sitting, standing, shuffling and walking for the senior514 dataset. All datasets senior501-senior518 are different, however exhibit similar trends that can be seen in these two plots. All the datasets in the graph seem to exhibit a similar trend as the one observed for the senior502 participant's data, only five datasets, senior515 dataset is one of them, convey slightly different trends. This may be since 5/18 of the HAR70+ participants used walking aids during the experiment and thus might have affected the response of the sensors and the values obtained.

Figure 15 shows the line plot for senior502 dataset where a similar pattern of sensor readings is visible across the activities. The back_x and thigh_z have a negative mean value when lying which might suggest that the sensor is always facing in the same direction during this action. As the participant moves to sitting and standing, the mean of these values shifts towards zero with standing showing the least variance among the sensor readings. The back_x and thigh_x variables have the relatively constant values in most actions except for sitting when values show more variation. The shuffling and walking actions have little change in variation of the sensor readings which suggest that the sensor readings are constant during these movements. The variance in the sensor values is also somewhat higher in the stairs descending action which is an indication of the complexity of this action and the range of movement required to perform it.

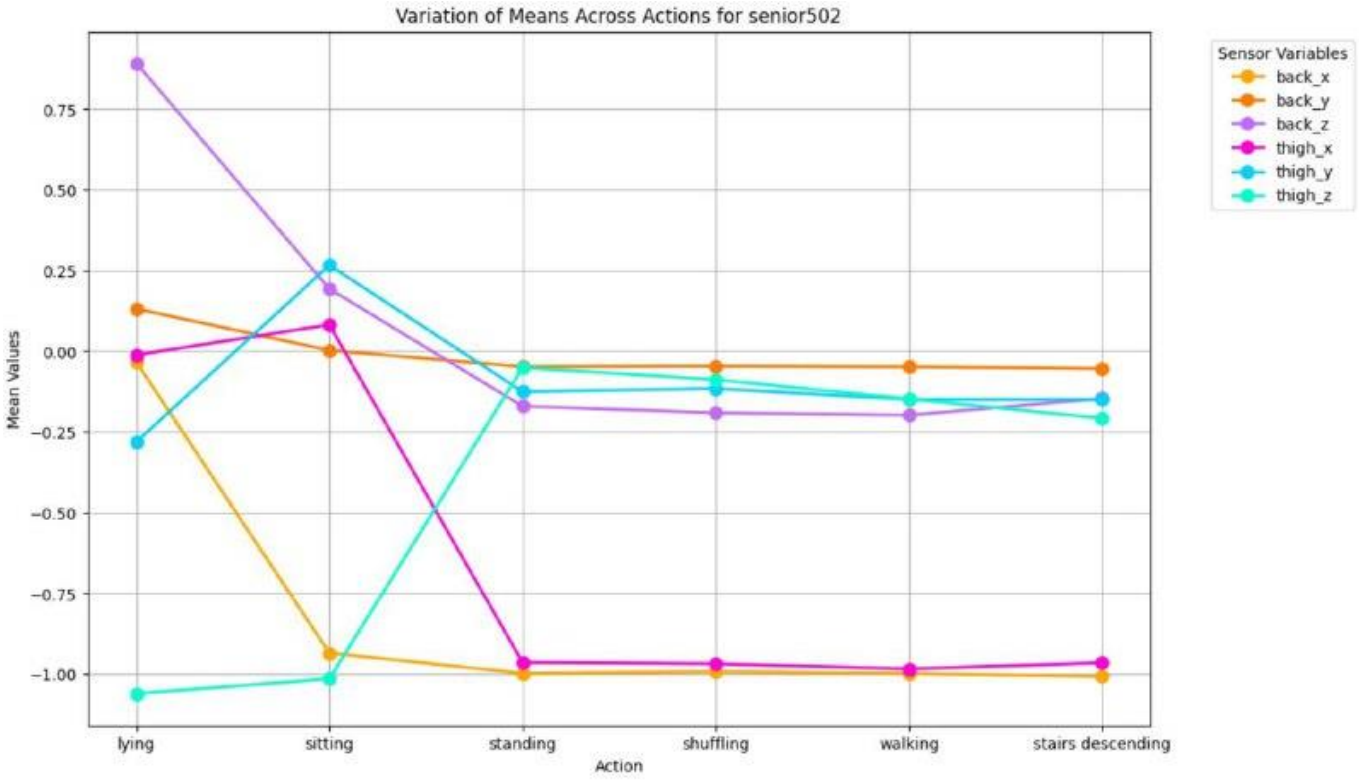

Figure 15. The Senior502 Line Plot for the Variation of Means Across Actions. This figure portrays the comparison of the mean sensor values of senior502 for the various action taken. The x-axis is the action that include lying, sitting, standing, shuffling, walking, descending stairs while the y-axis is the mean sensor value for each action. Each of the coloured lines represents sensor variables namely back_x, back_y, back_z, thigh_x, thigh_y, thigh_z as indicated on the right side of the figure. This plot shows how the average sensor readings vary with the different activities to enable one to distinguish between different activities from the sensor data.

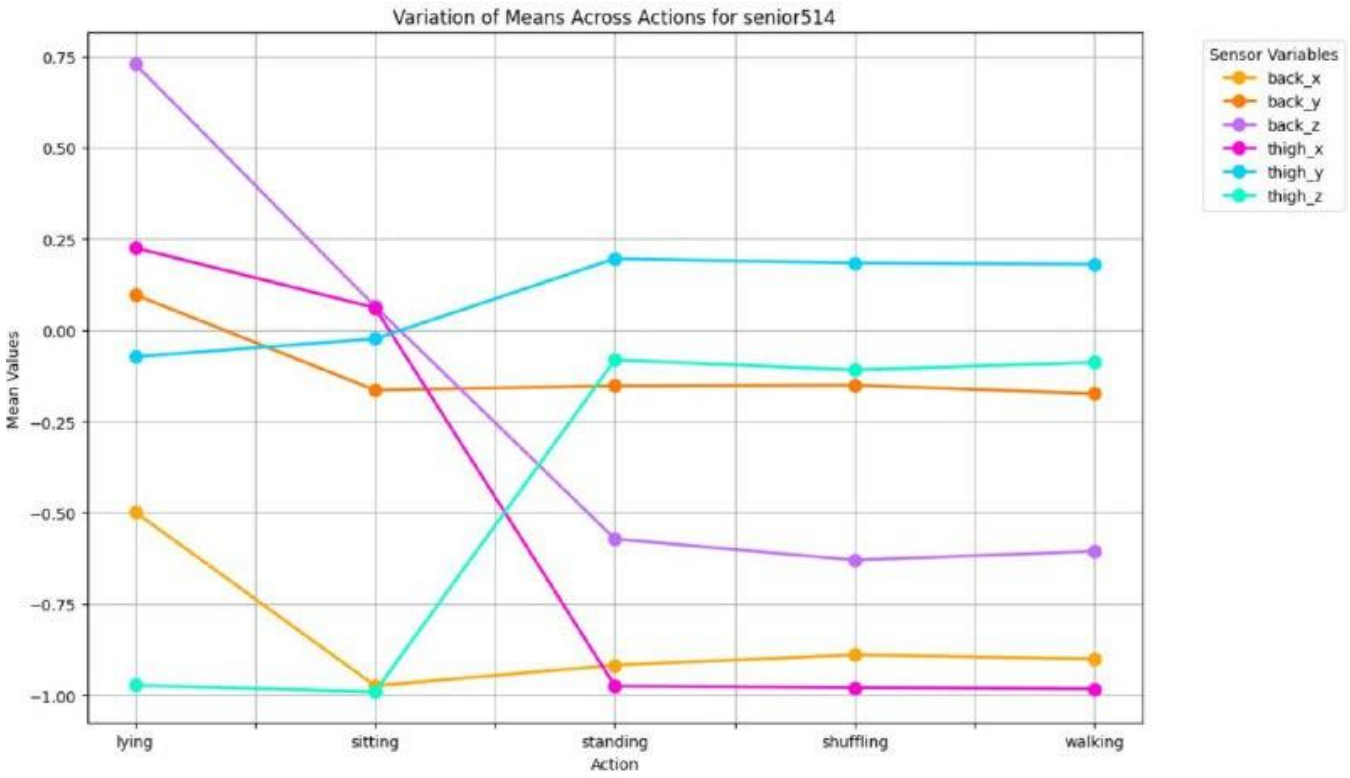

Figure 16. Senior 514 Line Plot – Variation of Means Across Actions. This figure shows how the mean of the sensor values for each action varies for participant senior514 across actions. The x-axis represents different activities (lying, sitting, standing, shuffling, walking) and the y-axis the average sensor output for the movements. However, patterns of the data are different from participant senior502 which might be due to the use of walking aids of this participant which influences the readings of sensors and changes in the sensor mean during the various activities.

In contrast, the line plot of senior514 in Figure 16 presents a different trend. The back_z and thigh_z variables also have a negative mean when lying as with the senior502 line plot. But the changes in all the six features' values are more evident when the subject moves from lying down to sitting and then to standing. The sitting action for senior514 has relative similar trend in variation to senior502 with little fluctuation in the sensor data. However, the standing, shuffling and walking actions have a very high variation and we see from the sensor readings that senior502 may be having differences in posture due to the use of walking aids. The stairs descending action was never performed by the senior514 participant

during the monitoring period and as such cannot be compared.

3.3.6 Standard Deviation of Time Difference

The exploratory analysis of the HAR70+ dataset includes the examination of the standard deviation of the differences in time when the data was recorded for each participant. This investigation helps to understand the variability of the data that can be a source of concern for the performance of the supervised learning models.

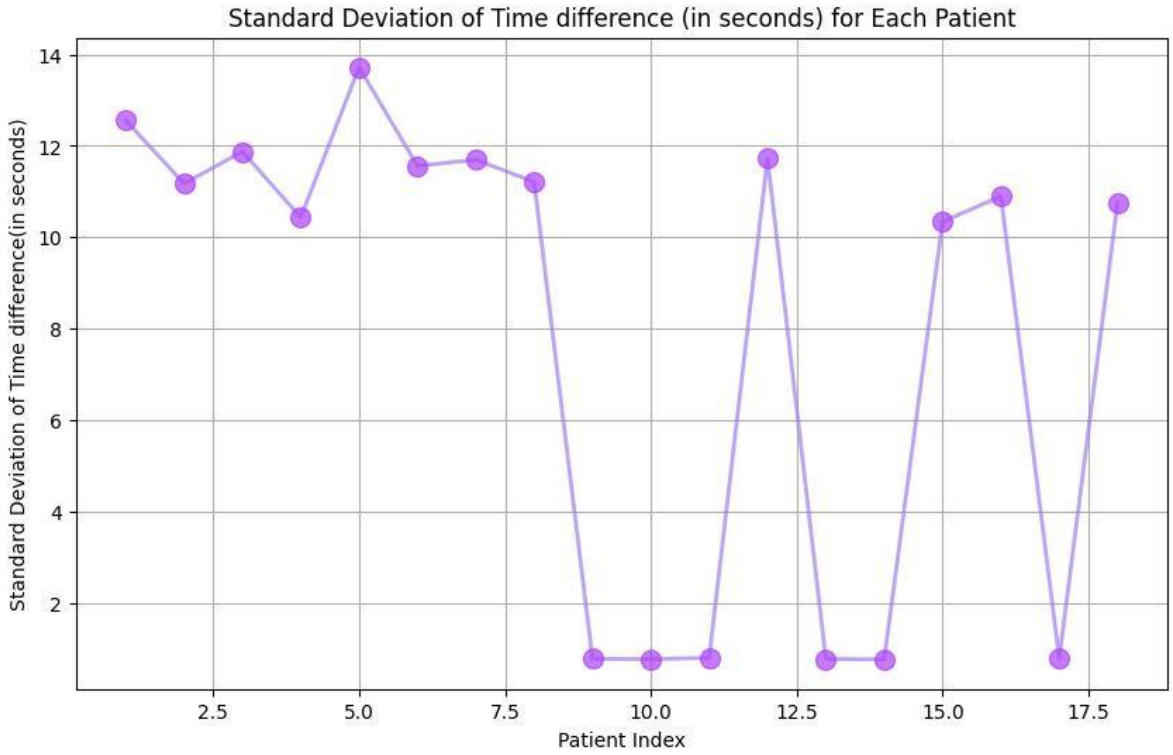

Figure 17. Standard Deviation of Time Difference (in seconds) of all Participants. This figure illustrates the standard deviation of the time between the recorded points for each of the 18 participants. The x-axis is patient index while the y-axis is standard deviation of time differences in seconds. Larger standard deviations imply that the time intervals are not taken at regular intervals while small values indicate more consistent time intervals.

The plot in Figure 17 shows the standard deviation of the time difference between consecutive data point for each of the participants. This metric is very important because it helps in identifying the regularity of data recording at certain intervals. Generally, the data should be captured at a regular interval, but the inconsistency may be seen due to the sensor's characteristics or other conditions.

In the case of participants 1-18, the standard deviation of time differences is quite different. The most flare is seen in participant 5 who has the highest standard deviation of time of about 14 seconds. This shows that the recording intervals are not regular, and this may cause a problem in the temporal analysis of the activity recognition models. On the other hand, there are some participants especially those with index

numbers 9 to 11, 13, 14 and 17 who seem to have least variability thus having a more reliable system of record.

While the detailed plots of individual sensor variables back_x, back_y, back_z, thigh_x, thigh_y, thigh_z are not included (see Appendix D: Code) a general analysis shows that the significance of standard deviation for these variables is quite different across various datasets. This is perhaps due to variability in the distribution of the sensor variables but the standard deviation is consistent across the different sensor variables. For instance, if one sensor variable has high variability for a certain participant, there is usually high variability in other sensor variables. This is an indication that the movement patterns and the activity types are in some ways still similar for all the participants. This general trend of variability across the different sensor variables gives an assurance on the validity of the data provided. It shows that, although the sensor values may vary from one individual to another, the data has a certain structure that can be used for training models. Consequently, these findings do not give any particular concerns as to the supervised learning modelling on the HAR70+ data.

3.4 Detecting Time-Series Stationarity

In time-series analysis, stationarity is a crucial property since it influences the results of the statistical models. Stationarity means that there is no change in the parameters of a time-series including the mean, variance and autocorrelation over time. In the context of real-world time-series datasets, such as those used in HAR, the data is usually non-stationary by nature. This section defines what stationarity means, how stationarity can be checked and why non-stationarity does not have to be dealt with in the HAR70+ dataset.

Firstly, we plotted the original sensor data to provide a comparison with the data that was changed later in the project. The plot in Figure 18 shows the raw time-series data. This visualisation exhibits trends and oscillations that are typical of the data, and which can be associated with various human activities.

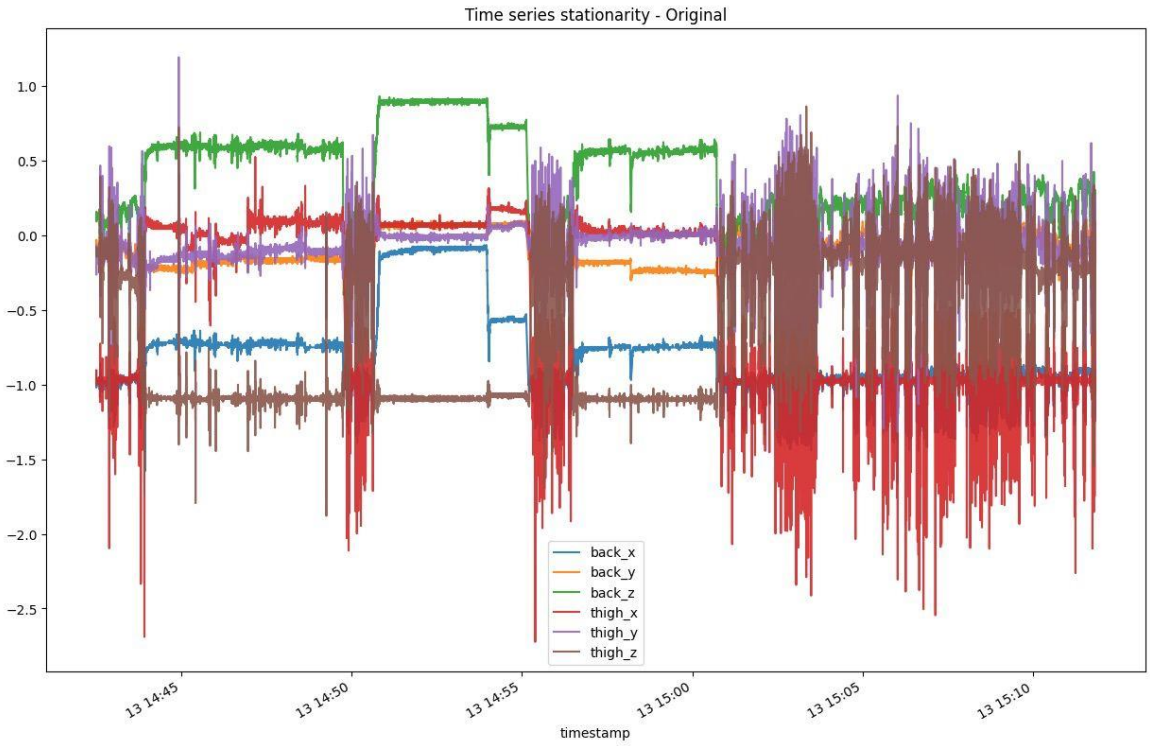

Figure 18. Detecting Time-Series Non-Stationarity – Original Data. This figure presents the time-series data from various sensors (back and thigh sensors on the x, y, and z axes) to detect non-stationarity in the original data. The x-axis represents the timestamp, and the y-axis shows the sensor values. Each coloured line corresponds to a different sensor axis as indicated in the legend. The plot shows fluctuations and shifts in the data, indicating non-stationarity meaning there are changes in mean or variance over time.

To quantitatively assess the stationarity of the time-series data we applied two well-established statistical tests: the Augmented Dickey-Fuller (ADF) test and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test.

The ADF test is used to test the null hypothesis that a time-series has a unit root implying non-stationarity. The test regression equation is the following:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^{p} \delta_i \Delta y_{t-i} + \epsilon_t$$

The KPSS test, on the other hand, tests the null hypothesis that a time-series is stationary. The test is based on the following model:

$$\text{KPSS} = \frac{1}{T^2} \sum_{t=1}^{T} S_t^2$$

The initial application of the ADF and KPSS tests conveyed that the time-series data from the HAR70+ dataset is non-stationary. To address this issue the first-order differencing method was applied which involves subtracting each data point from its preceding value. This transformation aims to stabilise the mean of the time-series by removing trends and seasonality thus converting the data into a stationary form.

The plot of the first-order differenced data (see Figure 19) shows the transformed time-series. The ADF and KPSS tests were reapplied, and the tests conveyed that the data had become stationary after differencing.

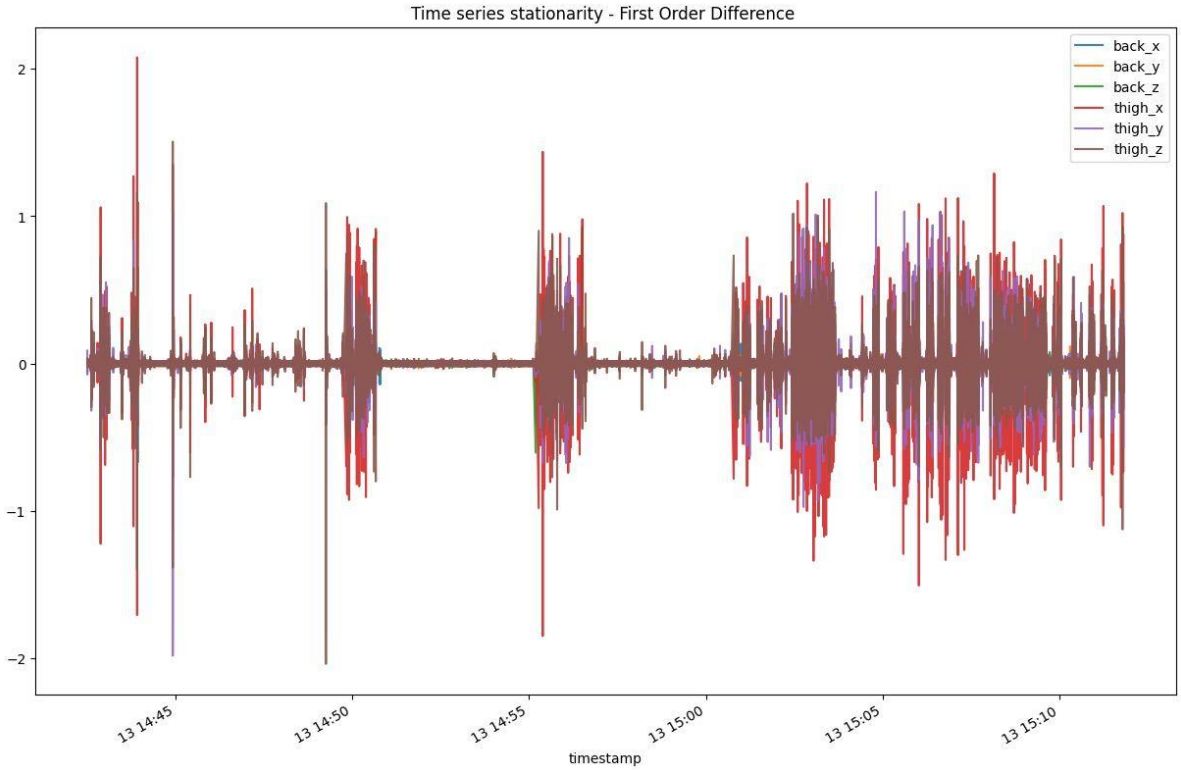

Figure 19. Detecting Time-Series Non-Stationarity – First Order Difference. This figure shows the time-series data after applying the first-order differencing to detect non-stationarity in the data. The x-axis represents the timestamp, and the y-axis displays the differenced sensor values. Each line corresponds to a different sensor axis (back and thigh sensors on the x, y, and z axes) as indicated in the legend.

While first-order differencing seems to have made the time-series stationary, it introduced a significant drawback. By differencing, we essentially made the mean of the values almost constant which is problematic for our predictive models. The goal in human activity recognition is to differentiate between various activities based on sensor readings. By making the data mean constant, we lose the variability that is crucial for distinguishing between different activities.

To illustrate this, we examined the sensor data for the "lying" activity before and after differencing (see Figures 20 and 21). Although "lying" is one of the most stationary activities the differenced data did not show reduced variability, instead it made the mean for all variables constant which diminishes the distinct characteristics needed for accurate classification.

‘

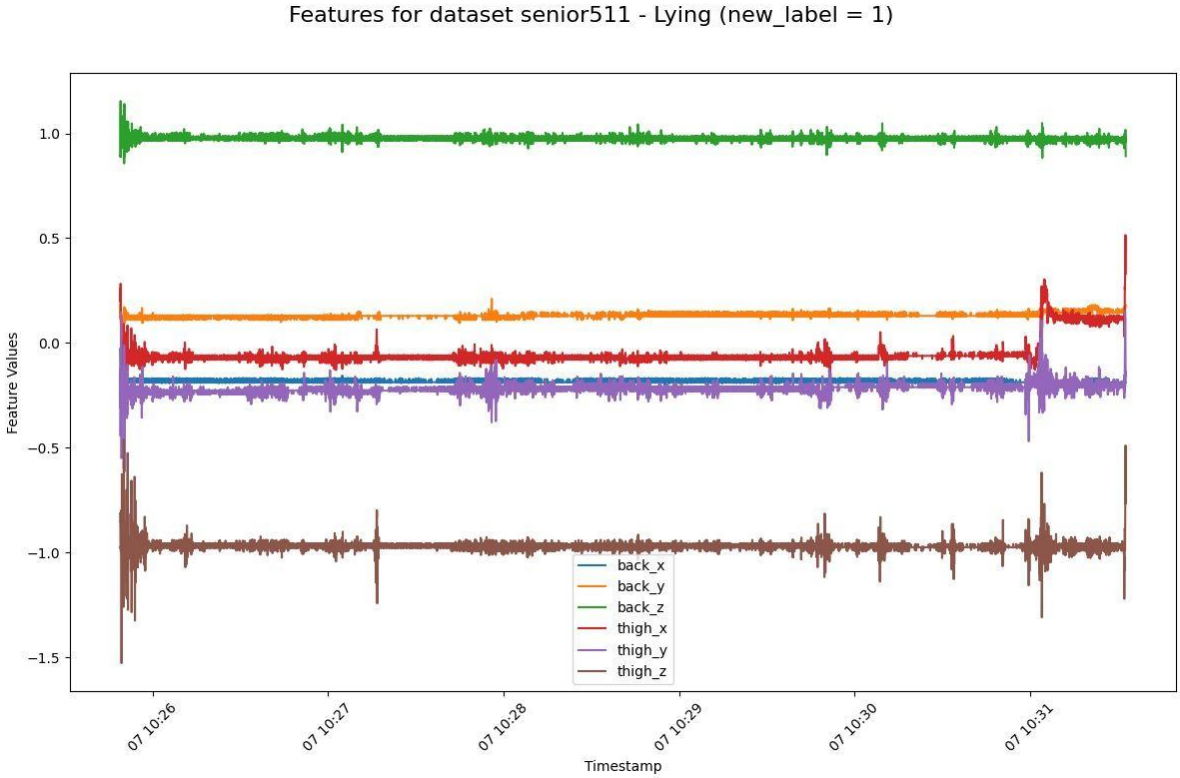

Figure 20. Detecting Time-Series Non-Stationarity – Original Senior511 – Lying. This figure illustrates the time-series sensor data for participant senior511 for the activity of lying down which new_label is assigned as 1. The x-axis represents the timestamp, and the y-axis shows the feature values from various sensors (back and thigh sensors on the x, y, and z axes). Each coloured line line is related to a certain sensor axis, as indicated in the legend. The plot highlights the stability of sensor readings during lying with minimal fluctuations indicating a relatively stationary state for this activity.

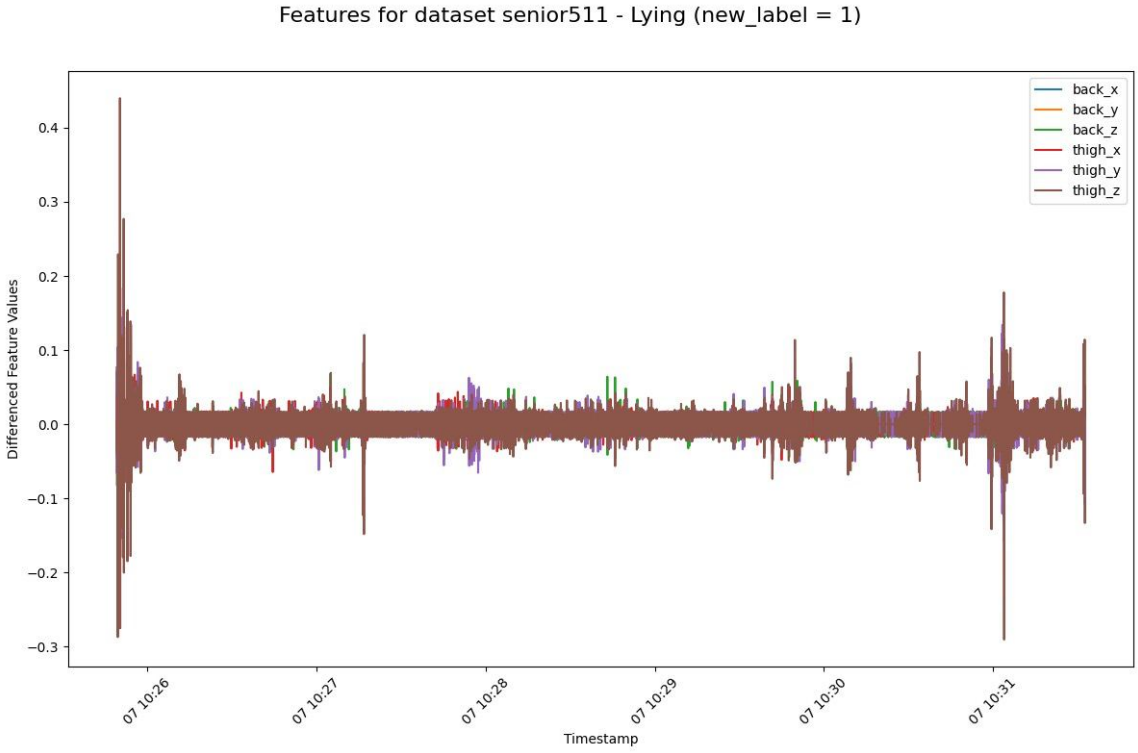

Figure 21. Detecting Time-Series Non-Stationarity – First Order Difference Senior511 – Lying. This figure shows the time-series data for participant senior511 during the lying activity (new_label = 1) after applying first-order differencing. The x-axis represents the timestamp and the y-axis shows the differenced feature values from various sensors (back and thigh sensors on the x, y, and z axes). Each coloured line corresponds to a different sensor axis, as indicated in the legend. First-order differencing has been applied to remove trends and stabilise the mean, highlighting changes in the sensor data that indicate non-stationarity during the lying activity.

Human activities are non-stationary and do not have a seasonal or a trending characteristic as other types of time-series do such as financial or climate data. Such actions as walking, sitting and lying include dynamic changes and fluctuations that show variations and shifts of the activities essential for supervised learning modelling in HAR.

Thus, using first order differencing to make the series stationary would remove the variability that the series has to convey. The presence of non-stationarity in the dataset should not make the performance of machine learning models any poorer but should rather provide the necessary variability for distinguishing between different activities.

HAR focuses on the recognition of activities using the information provided by the sensors, and the patterns and variations contained in the data. To preserve the variability we might want to keep the data in its original form so that the models can distinguish between each activity.

Although the use of first order differencing and stationarity tests such as ADF and KPSS is quite common when dealing with time series data, they are not beneficial for supervised learning methods application to the HAR70+ dataset. Thus, we will not make further stationarity adjustments and will use the original data for the analysis and modelling.

3.5 Data Cleaning and Preparation

The data cleaning and preparation process is important for the data used for fitting supervised learning models. This section includes renaming of action labels, splitting features and labels into two different datasets and dropping unnecessary columns.

First, the raw datasets were relabelled numerically according to the type of actions that the participants have done from 1 to 7 with the increasing intensity of the activity. For easier interpretation the above numeric labels were mapped to the action labels: ‘lying’, ‘sitting’, ‘standing’, ‘shuffling’, ‘walking’, ‘stairs descending’ and ‘stairs ascending’.

Second, the variables used for modelling comprise accelerometer signals from different axis back_x, back_y, back_z, thigh_x, thigh_y, thigh_z. The labels that are now descriptive activities were moved to a separate dataset to become the target values/labels the models try to predict.

The next step, timestamp field which is helpful during the data collection and the initial study of the data is not needed when it comes to modelling, thus, it was deleted from the datasets. This step also

assists in the reduction of dimensionality of the data making it easier to handle while applying models.

Preprocessing steps that have been carried out prepare a solid ground for the following predictive analysis of data and supervised learning modelling.

3.6 Fitting Supervised Learning Models

In this section, we delve into details of the sampling methods implemented for models fitting to the HAR70+ datasets. The target algorithms comprise of Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Decision Trees, Random Forests, XGBoost, Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). The models were applied to both individual and population-level data to assess and compare various techniques of data sampling and combining in order to identify the most suitable one based on the results of performance measures. However, since MLP, CNN, and RNN tend to be very time-consuming models, they were applied only once for individual datasets using a single sampling method 1 or 2 as they return the same scores for all classifiers (see Section 3.6.2), and once for the population combination 2 (see Section 3.6.2) with a single sampling method.

3.6.1 Sampling Methods: Individual Datasets

Individual datasets sampling involves working with each participant's data separately. This method helps in understanding how well models perform on a per participant basis capturing individual variances in activity patterns.

Sampling Method 1: Train-Test Split

The first way of sampling is splitting the data using train-test split method for every dataset. Features and labels are split into training and testing sets, the class distribution is preserved by using stratified sampling. This helps the model to be trained on 80 % of the data and the model is tested with the remaining 20%.

Sampling Method 2: Resampling of Maximum Label Values with Replacement

To address class imbalance this resampling method boosts the number of instances of minority classes by randomly sampling with replacement. This is done in order to make sure that each class has the same number of samples as the class with maximum instances. It reduces the chances of the model being over fitted on the majority class hence improving the models ability to identify the minority class activities.

Despite this, there are some drawbacks of using this resampling method to handle class imbalance. By replicating the instances of the minority class, the model may recognise the noise or specific features of such copied samples and, therefore, overfit. This means that the model could be good at predicting on the training set but bad at predicting on the test set as it may not be able to generalise to unseen data. In addition,

it can make the dataset larger which would increase the training time and computational cost.

Sampling Method 3: Downsampling to Minimum Label Values

Downsampling is reducing the number of instances of the majority classes to match the number of instances in the minority class. This approach entails that training data classes are balanced and the prevents the algorithm from being skewed towards the majority class. It aids in increasing the level of model's ability to generalize and decrease overfitting. The drawback of downsampling is that it reduces the number of instances and, therefore, information which could be used for making predictions. This gives rise to the notion that the downsampling technique can make the models performance poorer.

Sampling Method 4: Averaging of Values + Method 2 & Method 3

In this method we find the average number of all observations in the dataset and divide it by the total number of classes.

We also tried the following method on one of the models (LDA): calculating averages of instances of each label, then for those labels that have less than average we apply method 2 resampling of maximum label values with replacement and for those above average – method 3 downsampling to minimum label values. This ensures that there is no incline to either sampling method 2 or sampling method 3 in the labels.

These two averaging methods produced the same results.

3.6.2 Sampling Methods: Population

Population sampling is the process of integrating data from several participants for training and testing and thereby creating a real-world environment.

Sampling Method: Train-Test Split for Population

The data from 14 subjects is employed for training and the data from the other 4 subjects is used for testing. This makes the models to be trained on the whole datasets which could return interesting results which could then be compared to the ones received from individual dataset and see if population-based one can improve the models. This is because when the data from several participants are used the models learn from a wide variety of the activity. We use the three methods described below, fit models to each and evaluate performance using chosen performance metrics.

Population Splitting Combinations

Three different combinations were used to split the HAR70+ data:

1. *Combination 1:* Two test datasets are from senior subjects who had no history of using walking aids in their daily lives whilst two other test datasets are from patients who had to use walking aids to some extent. This combination assesses the models' capacity to transfer the results from participants with and

without aids.

2. *Combination 2:* All four test datasets are from senior citizens who do not need any assistance in walking. This combination checks the capacity of the models in a homogenous group of patients, which do not include the variability related to walking aids.

3. *Combination 3:* All four test datasets contain senior individuals who utilised walking aids during the experiment. This combination is useful in establishing the ability of the models in identifying activities for participants with walking aid who may have different movement pattern as evidenced by the data captured by the sensors.

To estimate the performance of the models we have selected the sampling methods, and the performance measure techniques mentioned above. The findings will provide insights to the applicability and validity of the proposed models for human activity recognition for older adults both individually and at population level. The detailed analysis will help with the choice of the most accurate model and optimal sampling method for HAR70+ data.

The application of these models to the HAR70+ dataset involves the following steps:

1. *Preprocessing:* Making sure the data is clean, consistent, and well-prepared by relabelling the action labels, splitting the features and labels into two different datasets and cleaning the data by removing unnecessary columns like timestamps.

2. *Sampling Methods:* Applying the aforementioned sampling techniques to cope with class imbalances. Methods include resampling to the maximum label values, downsampling to the minimum label values, and averaging the values both for individually split datasets and populations.

3. *Model Training and Evaluation:* Training and testing the models as well as measuring performance by calculating accuracy, precision, recall, F1 score and Kappa scores.

4. *Analysis of Results:* Comparing the performance metrics of each model in terms of different sampling methods and different combinations of data splitting. This analysis assists in identifying the best model as well as the sampling approach that may provide the most accurate senior citizens activity recognition.

3.7 Summary

In Chapter 3 we presented the methodology for analysing the HAR70+ dataset emphasising the importance of data exploration and preparation in achieving reliable results. We applied various sampling methods and trained multiple machine learning models and neural networks to recognise human activities evaluating their performance across individual and population-based datasets.

Chapter 4

Results and Discussion

4.1 Problem 1: HAR70+ Data Variability

The Exploratory Data Analysis (EDA) of the HAR70+ dataset gave valuable information on the variability and organisation of the data that is vital in the creation of Human Activity Recognition machine learning algorithms. Based on the results of the analysis of the structure and distribution of sensor readings across the participants, we determined patterns and variability of data that are critical for the predictive models' performance.

HAR70+ datasets are diverse and the sensors variability can range dramatically from one activity to another. Diversity is a main characteristic of the HAR70+ dataset which reflects human movements in real life. Furthermore, we found out that the datasets are non-stationary and it cannot and does not need to be. The non-stationarity of this time-series data is advantageous for the HAR70+ dataset since it enables the models to have the chance of operating on full variability of the sensor outputs during the various activities. This non-stationarity should be preserved and it proved to be important since it makes the models capable of recognising and learning variations in sensor readings across labels that results in better and more reliable activity recognition.

The EDA also showed differences in the distribution of sensor variables, and their variability between various activities. For instance, the back_x, back_y and back_z were often observed to have readings that were clustered within certain ranges, which suggested that the participants' postures or movements were not random. On the other hand, thigh (x, y, z) sensors had higher variability, especially during the walking activities which can be attributed to difference in leg movements. This variability is very important for model training because it gives the required distinction between different activities which are crucial when classifying the activities.

Pairplots highlighted important correlations between sensor variables. For instance, back_x has a negative correlation with back_y, this shows that movement in one axis is often associated with movement in the other axis thus being a good predictor of participants activities. This leads us to the conclusion that some sensors like back_x are more important than other features and influence the models outcome the most. The scatterplots also showed well-defined regions of the activity clusters as well as poor ones due to the imbalanced classes.

From the EDA it can be inferred that the readings from the sensors in the HAR70+ datasets are of high variability in the sensor readings which enables recognition of human activities and assist the supervised learning methods in their modelling on a high level. However, the problem of having imbalanced activity classes such as datasets where walking activity prevails while other activities like ascending or descending the stairs are minority classes hinders the creation of good HAR algorithms. The problem of imbalanced classes should be addressed if the aim is to improve the models performance.

4.2 Problem 2: Supervised Learning Methods

The performance of the selected supervised learning algorithms was evaluated across individual HAR70+ datasets and a population of datasets to investigate which models perform best under certain conditions.

Table 2 presents accuracy and Cohen's Kappa scores for individual datasets and population Combination 2 which points out the difference of models' performance. Since the classifiers' performance scores are the same for both resampling to the maximum label values and averaging methods, we calculated mean accuracy and kappa scores across all 18 datasets within each model for one of the sampling methods and obtained the following results:

Model	*Individual Datasets*		*Population – Combination 2*	
	Accuracy	*Kappa*	*Accuracy*	*Kappa*
LDA	81.1	69.4	80.3	67.6
QDA	87.3	81.6	85.4	78.6
Decision Tree	92.2	88.3	88.4	82.4
Random Forest	94.9	92.3	91.9	87.4
XGBoost	94.7	92.1	91.9	87.6
MLP	94.1	90.9	91.7	87.4
CNN	93.8	85.4	93.6	90.1
RNN	94.3	91.3	94.1	91

Table 2. Supervised Learning Methods Performance on HAR70+ Data. The table presents mean accuracy and kappa scores in % (rounded to one decimal place) for various machine learning models and neural networks applied to both individual datasets and a population-based Combination 2. The table conveys the averaged scores across all individual datasets and populations within each combination for a single sampling method – resampling to max() with replacement.. The highlighted red cells represent the highest performance scores – Random Forest for Individual datasets and RNN for Population Combination 2.

For individual datasets the Random Forest model was found to be the best classifier with a kappa score of 92.3% and accuracy of 94.9% for both the training and testing data. These high scores suggest that Random Forest deals with overfitting. We also tried to overcome overfitting using gradient descent which did not show any improvement and struggled to solve the problem meaning that more adjustments or training time is needed since gradient descent can be very time-consuming.

For the population analysis on Combination 2 the Recurrent Neural Network at first gave good results on the training data with Kappa score of 91% and accuracy of 94.1% . However, its performance on unseen data was worse with Kappa score of 56% and the accuracy of 71.5%. The decline of the RNN

performance on the test dataset might be attributed to the lack of stationarity in the time-series HAR70+ dataset. Moreover, the HAR70+ datasets do not have any trends or seasonality, instead it has high variance with no specific pattern to follow which makes it challenging to RNN to learn and excel when exposed to new data. Thus, to enhance the outcome of the RNN, it might be beneficial to work with some of the approaches modifying the architecture of the RNN, for instance, LSTM or GRU which are more intricate forms of RNNs but very time-consuming and computationally costly.

In conclusion, Random Forest has been found to be the most accurate on individual datasets but there is work needs to be done on ways of avoiding or reducing overfitting. The RNN model was able to do well on the population Combination 2 training data but scored lower on the test data set which may be due to the lack of appropriate pattern recognition and therefore needs more fine tuning or application of RNNs extensions.

4.3 Problem 3: Optimal Sampling Strategy

The aim of the analysis was to determine which cross-validation approach would be the most suitable for each classifier (LDA, QDA, DecisionTree, Random Forest, XGBoost, MLP, CNN) as well as for each individual HAR70+ dataset and the entire population. The effect of these methods was compared in terms of accuracy and other measurements like Cohen's Kappa.

For different datasets, various sampling strategies have been tried such as train-test split, resampling to maximum labelled data, downsampling to the minimum labelled data and averaging observations within labels. We found out that the methods that preserved class balance (resampling and averaging) achieved the same high accuracy and Kappa scores as the simple 80/20 splitting method meaning there is no difference whether we apply those sampling strategies or not which was an interesting and unexpected outcome that requires further investigation. On the other hand, the downsampling method had the worst performance of all due to the loss of data that caused underfitting.

In addition, three different combinations of populations were evaluated. Second combination where testing was done on a group of subjects without walking aids produced the highest accuracy for all classifiers. This setup might have been advantageous owing to the fact that the movement patterns were consistent, thus allowing the models to easily make generalisations. On the other hand, the combinations that involved participants with walking aids (Combinations 1 and 3) brought more variability and thus caused slightly worse results.

To conclude, the resampling and averaging methods are not better than a simple train test splitting method and Combination 2 is the most suitable for the population-based modelling. However, this combination has its disadvantage as it does not involve all the participants who used walking aids, and thus the variability of data from the participants is different from those who did not require walking assistance, meaning the combination is not inclusive and fails to generalise on data with different variability.

Chapter 5

Conclusion & Outline for Future Research

In conclusion, this research aimed to examine the use of supervised learning techniques in HAR among senior citizens using the HAR70+ dataset. The research was organised into five chapters starting with the introduction to HAR and difficulties of identifying activities in people within the age of 70 to 95. The literature review chapter described in detail the HAR approaches, supervised classification methods and handling time-series data in HAR.

In the methodology chapter, we described each step in detail that we used to prepare the HAR70+ dataset for the analysis, including the EDA, stationarity test, and the application of sampling techniques. The application of the supervised learning models such as LDA, QDA, Decision Tree, Random Forest, XGBoost, MLP, CNN, and RNN and the performance of the models in the form of accuracy and Cohen's Kappa scores for both individual and population-level datasets was also discussed.

In the results and discussion chapter, we sought to respond to the research questions that were defined in chapter one of this research. The EDA showed that there is a high level of data variability in the HAR70+ dataset and the non-stationarity of the data should be maintained for better activity identification., although there is an issue of imbalanced classes which works against the modelling process. Of all the supervised ML models, the individual dataset classification accuracy was best performed by Random Forest, although the RNN model for population-based analysis showed potential on the training data but was found to be poor when applied to test data. The best sampling techniques were then found to be resampling and averaging techniques within the datasets and a population-based Combination 2 which proved to perform the best with all the supervised learning methods, meaning the models do not generalise well on data containing participants using walking aids.

All in all, in this research, we tried to bring light to the issues of Human Activity Recognition in senior citizens and attempted to show how some of the supervised learning methods and sampling strategies could improve the recognition rate. However, the problems like overfitting, imbalance of classes and models struggling to deal with different data variability for seniors with and without walking aids are still present. We believe the study provides important theoretically and practically driven implications for the design of HAR system for application in elderly care.

Possible future work could include the investigation of more complex models whereby the use of LSTM layers can be applied for modelling time-series data and coping with imbalanced classes in the HAR dataset. In addition, more inclusive sampling strategies should be developed to account for diverse movement patterns, particularly in populations with reduced mobility such as those using walking assistance which would enhance the performance of Human Activity Recognition systems as a whole.

References

Abhijitt Dhavlle, S. M. P. D., 2020. *A Comprehensive Review of ML-based Time-Series and Signal Processing Techniques and their Hardware Implementations,* Fairfax.

Astrid Ustad, A. L. S. Ø. T., 2023. Validation of an Activity Type Recognition Model Classifying Daily Physical Behavior in Older Adults: The HAR70+ Model. *sensors.*

Atwan, T. A., 2022. *Time Series Analysis with Python Cookbook.* ISBN 978-1-80107-554-1. Birmingham: Packt.

Beran, J., 2010. *Mathematical Foundations of Time Series Analysis.* ISBN 978-3-319-74378-3. Konstanz: Springer.

Changquan Huang, A. P., 2022. *Applied Time Series Analysis and Forecasting with Python.* ISBN 978-3-031-13583-5. Cham: Springer.

Di Wang, A.-H. T. D. Z., 2015. *Non-Intrusive Robust Human Activity Recognition for Diverse Age Groups.*

Grus, J., 2015. *Data Science from Scratch.* 978-1-491-90142-7. O'Reilly.

Gu Zhanzhong, X. H. G. F. C. X. F. X. W. J., 2024. *MILLIMETER WAVE RADAR-BASED HUMAN ACTIVITY RECOGNITION FOR HEALTHCARE MONITORING ROBOT.* arXiv:2405.01882v1.

Han Sun, Y. C., 2024. *Real-Time Elderly Monitoring for Senior Safety by Lightweight Human Action Recognition.* Binghamton, EEE 16th International Symposium on Medical Information and Communication Technology (ISMICT).

Kevin Yeap Chen Keng, L. Y. H. K. S. N. S. B. R. K. M. G. W. W., 2018. *A Review of Ambient Intelligence Based Activity Recognition for Ageing Citizens.* 978-1-5386-7167-2/18. Subang Jaya.

Konstantinos Benidis, S. S. R. V. F., 2022. *DEEP LEARNING FOR TIME SERIES FORECASTING: TUTORIAL AND LITERATURE SURVEY,* Berlin: arXiv:2004.10240v2.

Liming Chen, C. D. N., 2019. *Human Activity Recognition and Behaviour Analysis.* ISBN 978-3-030-19407-9. Cham.

Monica-Andreea Dragan, I. M., 2013. *Human Activity Recognition in Smart Environments.*

Pascal A. Schirmer, I. M., 2024. PyDTS: A Python Toolkit for Deep Learning Time Series Modelling. *entropy.*

Peter J. Brockwell, R. A. D., 1996. *Introduction to Time Series and Forecasting.* DOI 10.1007/978-3-319-29854-2. Springer.

Qiancheng Tan, Y. Q. R. T. S. W. J. C., 2023. A Multi-Layer Classifier Model XR-KS of Human Activity Recognition for the Problem of Similar Human Activity. *sensors.*

Ranjit Kolkar, G. V., 2021. *Human Activity Recognition in Smart Home using Deep Learning Techniques.* OI: 10.1109/ICTS52701.2021.9609044. Manglore.

Robert H. Shumway, D. S. S., 1999. *Time Series Analysis and Its Applications.* DOI 10.1007/978-3-319-52452-8. Springer.

Spiliotis, E., 2022. Decision Tree for Time-Series Forecasting. *ResearchGate.*

Trevor Hastie, R. T. J. F., 2009. *The Elements of Statistical Learning.* DOI: 10.1007/b94608. Stanford: Springer.

Xiaochun, C., 2024. *Research on entertainment robots based on artificial intelligence interaction for human posture recognition and sports activity monitoring.*

Appendices

A. List of Figures

B. List of Tables

C. HAR70+ Data

Activity Recognition in Senior Citizens (kaggle.com)

HAR70+ - UCI Machine Learning Repository

D. Code – Python – Google Collaboratory Environment

All_Datasets_HAR70+.
ipynb