Universidad Politécnica
de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

Master's Degree in Artificial Intelligence

Master Thesis

Daily Activity Classification Based on Wearable Sensors for Enhancing Tremor Assessment

Author: Ainhoa Ruiz Vitte
Tutors: Javier de Lope Asiaín and Álvaro Gutiérrez Martín

Madrid, February 2025

This Master Thesis has been deposited at the ETSI Informáticos de la Universidad Politécnica de Madrid for its defense.

Master Thesis
Master's Degree in Artificial Intelligence

Title: Daily Activity Classification Based on Wearable Sensors for Enhancing Tremor Assessment

February 2025

Author: Ainhoa Ruiz Vitte
Tutors: Javier de Lope Asiaín
Departamento de Inteligencia Artificial
ETSI Informáticos
Universidad Politécnica de Madrid

Álvaro Gutiérrez Martín
Departamento de Tecnología Fotónica y Bioingeniería
ETSI Telecomunicación
Universidad Politécnica de Madrid

Acknowledgements

I would like to express my sincere gratitude to all those who have supported me in the completion of this thesis. In particular, I am deeply thankful to my supervisors, Álvaro Gutiérrez Martín and Javier de Lope Asiaín, whose valuable feedback and contributions have greatly enriched my learning. Thanks to their guidance, I have not only gained deeper knowledge of human activity recognition using inertial sensors but also developed a better understanding of conducting a research project. Their support has also helped me appreciate the significance of research and its potential to make a meaningful impact on improving the quality of life for those in need, particularly patients with Parkinson's disease and essential tremor in the healthcare domain.

Resumen

Esta tesis de máster explora la clasificación de actividades de la vida diaria mediante sensores inerciales para mejorar la evaluación del temblor en pacientes con Enfermedad de Parkinson y temblor esencial. Las escalas clínicas tradicionales suelen ser subjetivas y no capturan completamente la variabilidad de los síntomas en la vida diaria. Para abordar esta limitación, se investiga el uso de sensores inerciales y aprendizaje automático para el monitoreo continuo de los pacientes.

La recolección de datos se realizó en dos fases. Primero, los participantes realizaron diez actividades en un entorno controlado mientras usaban un smartwatch con acelerómetro y giroscopio. Luego, se recopilaron datos continuos en entornos domésticos para reflejar mejor sus actividades diarias. Este enfoque permitió explorar tanto la clasificación segmentada como la continua.

Se emplearon técnicas de aprendizaje automático como Random Forest, Máquinas de Soporte Vectorial y Extreme Gradient Boosting. Además, se evaluó el impacto de diferentes tamaños de ventana temporal y se propuso un método de conjunto. Para la clasificación continua, se utilizaron Máquinas de Soporte Vectorial y enfoques de aprendizaje profundo, como redes Bidireccionales de Memoria a Corto y Largo Plazo y combinaciones con redes neuronales convolucionales.

Los resultados mostraron que las Máquinas de Soporte Vectorial fueron las más precisas en el enfoque segmentado, superando el 80% de precisión, mientras que las redes de Memoria a Largo y Corto Plazo lograron el mejor desempeño en la clasificación continua, reduciendo los falsos positivos. Además, se creó un nuevo conjunto de datos con información de pacientes con Parkinson y temblor esencial, aportando un recurso valioso para futuras investigaciones.

La tesis concluye que los sensores inerciales, junto con el aprendizaje automático y profundo, pueden mejorar la evaluación objetiva del temblor y permitir tratamientos más precisos y personalizados.

Abstract

This Master's thesis investigates the classification of Activities of Daily Living using wearable inertial sensors to enhance tremor assessment in Parkinson's Disease and Essential Tremor patients. Traditional clinical scales are often subjective and fail to capture symptom variability in real-life settings. To address this, the study explores Inertial Measurement Units and machine learning techniques for continuous movement monitoring.

Data collection was conducted in two stages: first, participants performed ten activities in a controlled environment while wearing a smartwatch with accelerometers and gyroscopes; second, continuous data was gathered in home environments for a more realistic representation of daily activities. This approach enabled the exploration of both segmented and continuous activity recognition.

Machine learning techniques such as Random Forest, Support Vector Machines, and Extreme Gradient Boosting were applied to segmented data. The study also examined the impact of different time window sizes and proposed an ensemble method combining various window lengths. For continuous classification, a modified ensemble method with SVM and deep learning approaches, including Bidirectional Long Short-Term Memory and convolutional neural networks, was evaluated.

Results showed that Support Vector Machines achieved the highest accuracy (over 80%) in segmented classification, while the ensemble method improved performance over single-window models. Long Short-Term Memory networks performed best for continuous classification, reducing false positives but offering only a slight improvement over traditional machine learning methods. Additionally, a new dataset incorporating Parkinson's and Essential Tremor patient data was developed as a resource for future research.

The thesis concludes that wearable sensors, combined with both machine learning and deep learning techniques, hold great potential for improving the objective assessment of tremor, enabling more accurate and personalised treatment plans for patients with tremor.

Contents

List of Figures

List of Tables

List of Abbreviations

ADL Activities of Daily Living

ET Essential Tremor

PD Parkinson's Disease

IMU Inertial Measurement Units

TETRAS Tremor Research Group Essential Tremor Rating Assessment Scale

WHIGET Washington Heights Inwood Genetic Study of Essential Tremor

UPDRS Unified Parkinson's Disease Rating Scale

MDS-UPDRS Movement Disorders Society Unified Parkinson's Disease Rating Scale

HAR Human Activity Recognition

ML Machine Learning

DL Deep Learning

RF Random Forest

SVM Support Vector Machines

XGB Extreme Gradient Boosting

BiLSTM Bidirectional Long Short-Term Memory networks

CNN Convolutional Neural Networks

RNN Recurrent Neural Networks

LSTM Long Short-Term Memory networks

PCA Principal Component Analysis

LDA Linear Discriminant Analysis

kNN k-Nearest Neighbors

DT Decision Trees

NB Naive Bayes

GRU Gated Recurrent Units

SDAE Stacked Denoising Autoencoder

NT No Task

RBF Radial Basis Function

RMS Root Mean Square

SWMI Sensor Window Mutual Information

CH Combing Hair

BT Brushing Teeth

BB Buttoning

EF Eating with Fork

CK Cutting with Knife

SD Simulate Drinking

OB Open/Close Box

WW Writing

TP Turning Pages

TD Open/Close Door

Chapter 1

Introduction

Pathological tremor is a prevalent movement disorder characterized by involuntary and rhythmic oscillations, primarily affecting the hands [5]. These symptoms significantly reduce the quality of life for those affected, impairing their ability to perform daily tasks from the early stages of the disease and throughout its progression [6]. Tremor is associated with several pathological conditions, such as essential tremor (ET) and Parkinson's disease (PD). ET is frequently encountered in general medical practice and affects approximately 4% of people aged 65 and older [7]. In comparison, PD, which typically impacts people over the age of 60, is the second most prevalent neurodegenerative disorder after Alzheimer's disease [8, 9]. ET typically presents with a tremor frequency between 4 and 8 Hz, which is most prominent when the hands are held in a still posture, whereas PD is characterized by a resting tremor with similar frequencies. The significant variability among patients complicates the diagnosis and classification of tremor, posing challenges for clinicians in accurately diagnosing and effectively treating patients [5, 10].

Motor symptoms associated with pathological tremor are generally assessed using mechanical demonstration and clinical scales such as the Tremor Research Group Essential Tremor Rating Assessment Scale (TETRAS) [11], Washington Heights Inwood Genetic Study of Essential Tremor (WHIGET) [12, 13], Fahn-Tolosa-Marin Tremor Rating Scale [14], or Unified Parkinson's Disease Rating Scale (UPDRSIII) [14], which rate tremor on a qualitative scale. These scales are vital for adjusting medications and evaluating their efficacy in clinical trials. However, because they depend on a clinician's assessment during a patient visit, they often only provide a snapshot of the condition and may not adequately reflect the tremor fluctuations that occur throughout the day during daily activities [6, 15, 16]. The primary challenge with current treatments is the trial-and-error approach to medication selection, often associated with various side effects [6]. Moreover, managing tremor is further complicated by issues such as drug intolerance and varying treatment responses [5]. These limitations hinder a comprehensive, objective evaluation of tremor severity during activities of the daily living (ADLs). Therefore, it is crucial to collect high-quality data on ET and PD patients in their home environments to further improve quantitative tremor assessment.

To reduce subjectivity in assessing movement disorders and address the challenges of current therapeutic approaches, motion analysis has been extensively developed, particularly for recognizing postural movements. There is increasing interest in us-

ing Inertial Measurement Units (IMUs) to monitor patients with neurological disorders and to identify postural movements and activities. IMUs are compact, portable sensors capable of tracking movement in three dimensions, making them ideal for continuous monitoring of patients in their daily environments. Some research has focused on using wearable systems to monitor tremor [17, 18, 19]. These studies measure tremor during task performance without prior activity classification, often using video recordings for task identification, which can be time-consuming and invasive in continuous monitoring scenarios [20, 21, 22, 23]. However, most of these systems do not provide information on the tasks performed during monitored tremor episodes, which could enhance understanding of the patient's clinical condition. Since there are still limitations in understanding how most therapies influence tremor throughout the day, having such a system could allow clinicians to better determine the most effective medications and dosages for patients in their daily environment.

Despite efforts to apply task identification methods to patients with tremor-related movement challenges, clear results have not been achieved yet [6, 15, 16, 17, 18, 24, 25]. Although some studies have explored the identification of ADLs, they are often limited to basic tasks such as sitting, walking, or running. These limitations indicate a need for more advanced approaches to accurately identify a wider range of activities in real-world settings. Many studies involving patients focus on basic postural and transitional activities rather than identifying complex tasks, and those that do address complex task identification often use multiple IMUs, which can be burdensome for patients.

Moreover, regarding task identification, accurate selection of window size is crucial for effective activity classification. However, it remains highly ambiguous and undefined yet [26, 27, 28], with most designs relying on estimates from empirical studies without substantial theoretical support.

Primary Objective

The main objective of this work is to design and develop a model capable of accurately classifying complex ADLs in PD and ET patients. With this work, it is aimed to achieve performance that is comparable to or exceeds the current state of the art. To reach this goal, a thorough study and analysis of existing methods will be essential to identify the most effective classification approach. This method does not only seek to meet the proposed objective but also to explore new, previously unexplored avenues for future research. Ultimately, this method is intended to enhance the ability of neurologists to monitor tremor progression and evaluate the effects of medication on patients' quality of life by providing accurate measurements of tremor during ADLs.

Specific Objectives

In addition to the overall goal of the thesis, several specific objectives can be established to support the main task:

- Review and analyze the state of the art in ADLs classification, both as discrete and continuous action recognition, using IMU signals, with a focus on patients with tremor.
- Study and assess the limitations of current ADLs classification systems.

- Design and carry out a battery of exercises relevant for evaluating tremor in home environments.
- *Develop a machine learning classification model for short-duration, discrete sequences of ADLs as an initial approach to continuous identification.*
- *Extend the classification models to handle continuous, long-duration signals.*
- *Explore the use of neural networks for the continuous classification of ADLs.*
- Develop a classification model for the identification of ADLs, with a focus on both discrete and continuous signal types, including the exploration of neural networks for improved performance.

Methodology and Work Plan

First, a review of the state of the art in the field of ADL identification and classification is conducted, with a particular focus on methods applied to patients with tremor. This is to understand the starting point of this work and to identify promising research directions, taking into account the limitations of the current methods.

Once the formal basis of the problem has been reviewed and the limitations of current state-of-the-art methodologies have been identified, the methodology followed for the development of the project is addressed, described through different phases. The first phase involves selecting the tasks to be classified. The next phase is the recruitment of patients for acquiring IMU signals from their wrists, which was conducted in two stages: an initial ambulatory stage performed in the clinic where patients performed tasks discretely and each task sequence was recorded individually, and a second stage in which patients took the device home to record their daily activities in a continuous way.

After the data collection, the ADL classification problem was addressed through several key steps. Initially, a machine learning classification model was developed to classify short, discrete sequences of ADLs, serving as a foundational approach for continuous identification. Following this, efforts were made to extend the classification models to accomodate continuous, long-duration signals. Additionally, as more data was gathered, the potential of neural networks was explored to enhance the continuous classification of ADLs.

Document Structure

Regarding the structure of this document, it is divided into five chapters:

1. Chapter 1 introduces the problem and main topic of this thesis, outlining the objectives and describing the methodology to be employed.
2. Chapter 2 presents a review of the state of the art on previous publications in the classification of ADLs from IMU signals, with a special emphasis on those aimed at patients with tremor.
3. Chapter 3: The methodology followed is presented. The first sections detail the design of the ambulatory protocol, which includes the selection of ADLs to be studied and performed by the patients, as well as the data collection and preprocessing steps. The next sections explain the methodology used to develop

classification models, covering both discrete sequence classification and continuous signal classification with traditional machine learning and deep learning approaches. The final classification pipeline is also described.

4. Chapter 4: The main results of the thesis are presented.
5. Chapter 5: Conclusions about the work carried out are included and potential avenues for future experimentation are suggested.

Contributions

Some of the findings presented in this Master's Thesis have been shared at various conferences and will be submitted for publication:

- A. Ruiz-Vitte, A. Comesaña, A. Muñoz-Arcentales, B. Larraga-García, Á. Alonso, E. Rocon, Á. Gutiérrez. "Activity Recognition in Patients with Tremor: Integrating Time-Length Windows for Enhanced Detection," *Computers in Biology and Medicine*, **Manuscript submitted for publication**.

  The methodology and results of the classification of ADLs using ML approaches in a segmented signal setting are presented.

- A. Ruiz-Vitte, A. Comesaña, B. Larraga-García, E. Rocón and Á. Gutiérrez, "Modelo de Aprendizaje Automático para el Reconocimiento Continuo de Actividades de la Vida Diaria," *Actas del XLII Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB)*, Seville, Spain, 2024.

  The results of the classification of ADLs using ML approaches in a continuous signal setting are presented.

- A. Ruiz-Vitte, A. Comesaña, B. Larraga-García, E. Rocón and Á. Gutiérrez, "Deep Learning for Continuous Recognition of Activities of Daily Living," *2024 E-Health and Bioengineering Conference (EHB)*, IASI, Romania, 2024, pp. 1-4, doi: 10.1109/EHB64556.2024.10805639.

  The results of the classification of ADLs using DL approaches in a continuous signal setting are presented.

Chapter 2

State of the Art

This chapter reviews the state of the art in activities of daily living (ADLs) classification and its application in tremor evaluation for conditions like PD and ET. In the biomedical field, Human Activity Recognition (HAR) is crucial, with wearable devices, particularly IMUs, playing a pivotal role to motion analysis. Combining HAR with tremor quantification improves clinical assessment, providing a robust framework for evaluating these conditions, which is central to this work.

2.1 The Role of ADLs Recognition in Healthcare

ADLs refer to essential self-care tasks that people perform routinely in their daily lives, such as eating, dressing, toileting, grooming, or interacting with items around the house [29]. These activities are fundamental to maintain independence and serve as key indicators of an individual's motor and cognitive abilities, especially among elderly adults or patients with motor or neurodegenerative diseases [30, 31]. In the medical context, ADLs assessments provide valuable insights into functional capacity and help predict the need for healthcare interventions. Furthermore, the recognition and monitoring of ADLs play a crucial role in applications like care centers and disease management, allowing caregivers to effectively track health conditions and provide on-time support, enhancing care and quality of life [32].

In recent years, significant advancements have been made in developing systems to recognise human daily activities, leveraging three main approaches. **Vision-based methods** use cameras to analyze activities through video and image data, offering rich contextual information and enabling the detection of various activities such as human gait analysis, and sitting, standing, or other behaviors [30]. **Wearable devices**, such as inertial sensors (e.g., accelerometers and gyroscopes), temperature sensors, or heart rate monitors, are typically used to gather physiological and movement data. These devices provide portable, non-invasive solutions ideal for continuous monitoring by attaching sensors to the body or clothing. Lastly, **device-free systems**, such as passive RFID tags placed on objects, enable the detection of activities through environmental interactions without requiring the user to wear any device [29]. Each approach has its advantages and limitations, but wearable sensors have emerged as a practical and widely adopted solution due to their versatility, affordability, and ease of integration into everyday life [33].

2.1.1 Challenges in Tremor Assessment

The increasing prevalence of pathological tremor, which can manifest in conditions such as PD and ET, has driven the growing application of wearable sensors, particularly IMUs, to monitor motor symptoms. IMUs are especially effective in capturing motor fluctuations, including tremor, dyskinesia (involuntary muscle movements), bradykinesia (slow movements and a decreased ability to move the body), and mobility impairments, which can fluctuate over time in response to both disease progression and treatment interventions [34, 35].

In clinical practice, the evaluation of tremor and related symptoms often relies on standarised scales such as the Unified Parkinson's Disease Rating Scale (UPDRS) [36], Movement Disorders Society (MDS)- UPDRS [37], or the Essential Tremor Rating Assessment Scale (TETRAS) [38]. These tools provide structured assessments of tremor severity, frequency, and impact on daily life, serving as benchmarks for diagnosing and tracking disease progression. However, these scales are typically limited to periodic evaluations in clinical settings, which may fail to capture the variability of motor symptoms in daily life. These tools rely on physical examinations, offering a momentary view of symptom severity observed during clinic visits, with patient-reported information that can be subject to bias. Moreover, healthcare professionals need to rely on retrospective data collected through patients' self-filled diaries and administered questionnaires [19]. These inaccurate evaluations and their episodic nature limits their ability to reflect the full range of symptom variability that patients suffer in their daily routines [21]. Patients should be evaluated, when possible, in normal living conditions in their home environments. In this context, a more objective system for symptom monitoring in a patient's daily life is claimed.

In this regard, wearable sensors allow for continuous monitoring, offering valuable insight into symptom changes over different time frames. As a result, IMUs, consisting of accelerometers, gyroscopes, and sometimes magnetometers, are increasingly used in remote monitoring settings, providing crucial support to both patients and healthcare professionals by linking recorded data to relevant clinical aspects, thereby improving disease management strategies and improving health outcomes. This growing popularity of IMUs reflects their potential to transform the way tremor and other motor symptoms are assessed, ultimately contributing to more personalised and effective treatments [39, 18].

2.2 Wearable Sensor Technology for Tremor Monitoring

In recent years, significant advances have been made in wearable sensor technology to monitor the effects of PD and ET. There are multiple initiatives and research works on the identification of motor symptoms [40, 41], with accelerometers being the most widely used sensors, followed by gyroscopes [42]. These systems provide continuous and objective measurements that address the limitations of clinical rating scales.

As reviewed by [43], several wearable devices have been developed to monitor symptoms of pathological tremor in home-like settings. Among these, the *STAT-ON* device uses accelerometers on a waist sensor to assess motor dysfunctions and gait parameters over a period of seven days, although it does not directly assess tremor [19]. Similarly, *Personal KinetiGraph*® (PKG), a wrist-worn device equipped with a

three-axis accelerometer, quantifies tremor, among other motor symptoms, offering extended monitoring periods [44, 35].

The *PDMonitor*® system integrates multiple sensors worn on the wrists, calves, and waist, allowing a comprehensive analysis of motor symptoms, including tremor, bradykinesia, and dyskinesia [34] . Despite its detailed analysis, the complexity of the system and the need for multiple sensors hinder its broad applicability.

Other systems reviewed, such as *Kinesia 360*™ [45], employ accelerometers and gyroscopes on both wrists and ankle to capture tremor during daily activities. A mobile application guides patients through the use of the motion sensor and provides instructions to complete assessment tasks based on common rating scales (e.g., UPDRS for Parkinson's disease [36], ETRS for essential tremor [38]). Once an assessment has been completed, motion data is transmitted to a secured database and several algorithms are used to calculate severity scores on a 0 (no signs)– 4 (severe signs) rating scale shown to be correlated with clinician ratings.

Although these devices represent significant advances in wearable sensor technology, they are inherently limited by their inability to correlate tremor data with spontaneous ADLs, which would provide crucial data for physicians [46]. Measuring the impact of tremor on ADLs is essential for understanding its real-world effects. The TETRAS scale [38], which includes both, performance and ADLs subscales, addresses this need. Gerbasi *et al.* demonstrated significant correlations between ADLs scores and quality-of-life measures, highlighting the importance of incorporating ADLs evaluations in tremor studies to assess treatment effectiveness and functional impact [46]. Current approaches, even those incorporating ADLs assessments like *Kinesia 360*™, often rely on structured tasks triggered by mobile applications. This method, though useful, imposes artificial constraints, limiting the capture of tremor's dynamic and spontaneous nature in real-world settings. Consequently, these systems may not fully reflect the true severity and impact of tremor on patients' daily functioning. In addition, many systems that aim to correlate the quantification of tremors with ADLs for a more precise evaluation continue to rely on supplementary tools such as video recordings or patient-reported diaries [20, 21, 22, 23]. This reliance on external input underscores the need for further innovation to achieve fully autonomous monitoring. Integrating ADLs classification capabilities directly into tremor quantification systems presents a potential way to improve their clinical utility and provide a more comprehensive assessment of tremor severity and impact on daily life.

2.3 Activity Classification prior to Tremor Quantification

The classification of ADLs is a critical step in studies that focus on patients with tremor or motor impairments, as it allows a contextualised evaluation of symptoms and their impact on functional ability. Various methods exist to extract activity information from raw sensor data, the process generally involves several key steps (see Figure 2.1). First, **pre-processing** is performed to clean the data, such as filtering noise or removing outliers. This is followed by **segmentation**, where the continuous time series data is divided into meaningful segments for analysis. Next, **feature extraction** is performed to compute relevant descriptors for each segment, often from the time or frequency domains. This step is particularly common in traditional machine learning (ML) methods, which rely on manually engineered features to repre-

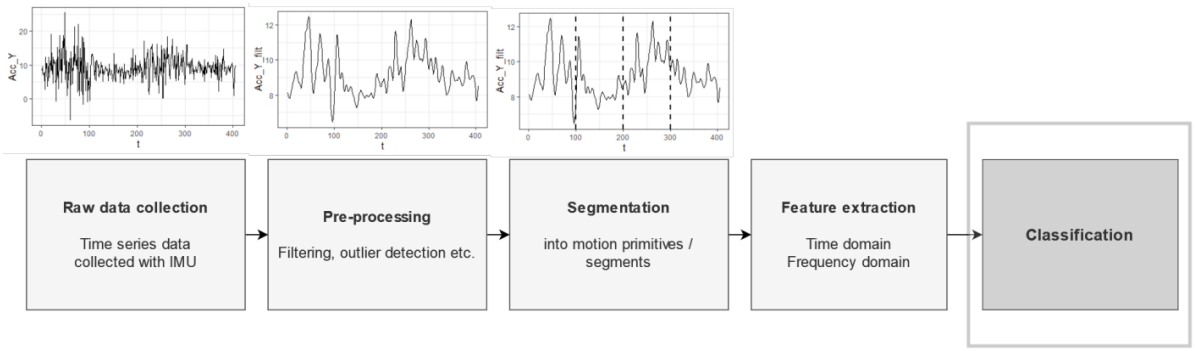


Figure 2.1: Typical Activity Recognition Chain to recognise activities from wearable sensors [1].

sent the data. In contrast, deep learning (DL) methods typically bypass this step by learning representations directly from raw sensor data, reducing the need for explicit feature engineering. In some cases, **dimensionality reduction** is applied to simplify the feature space while retaining critical information. Finally, **classification methods** are used, ranging from traditional ML algorithms to DL techniques, to identify and label specific activities [3, 1].

2.3.1 Data Collection

In the field of HAR, various sensors and ML methods are employed to accurately determine daily activities [47]. Among these, wearable sensors such as accelerometers and gyroscopes are widely used because of their accessibility and effectiveness in capturing relevant motion data. For example, the authors in [48] demonstrated that accelerometer data alone can achieve high classification accuracy, establishing it as a reliable primary sensor for activity recognition. Additionally, [49] an intelligent system was developed based on accelerometer data to classify movements, specifically for applications in ambient assisted living, highlighting the sensor's practical utility.

In some cases, the combination of multiple sensors can further improve the accuracy of task classification. In [50], the authors conducted a comparison of accelerometer and gyroscope data across various body positions—such as the hip, wrist, and ankle—and showed that sensor placement significantly impacts the recognition performance. Wrist-worn smartwatches can provide sensitive information on human activity as well as the information on whole body movement, as proven in [51]. Moreover, in [52] the authors integrated accelerometer and gyroscope data from a smartwatch to classify activities, demonstrating the value of sensor fusion in improving classification outcomes.

2.3.2 Preprocessing

Due to the inherent characteristics of inertial sensors, the acquired sensor data should first pass a pre-processing phase. This phase is crucial as it ensures that the signals maintain the relevant characteristics that carry information about the activities of interest. Pre-processing of acceleration and gyroscope signals may involve various steps such as calibration, unit conversion, normalization, resampling, synchronization, or signal-level fusion [53]. To clean and filter the signals, different

types of filter can be employed, including low-pass median filters, Laplacian filters, Gaussian filters, and median filters. Furthermore, high-pass filters can be applied to separate dynamic body acceleration from gravitational acceleration, allowing a clearer analysis of activity-related components [3]. When working with patients affected by tremors, signal processing becomes more complex due to the components introduced by the tremor itself, which must be carefully handled to avoid misclassification or data distortion.

2.3.3 Segmentation

Windowing is a crucial step in the signal segmentation process for activity recognition, though its application is often vague and lacks consensus on the ideal window size [26]. The main challenge lies in selecting an appropriate window length, as it directly impacts both the performance and accuracy of the models used for activity classification [54]. Windowing approaches are commonly divided into three types [26]:

- **Activity-defined windows** partition the sensor data stream based on activity changes, detecting the start and end points for each activity. These changes are often identified through variations in frequency characteristics.
- **Event-defined windows** focus on detecting specific events, such as heel strikes or toe-offs in gait analysis, to define data segments. While useful in controlled settings, this method may not generalise well to real-world conditions due to the non-uniform distribution of events, making it less suitable for at-home autonomous monitoring.
- **Sliding windows** are the most widely used method, particularly in real-time applications, due to their simplicity and minimal preprocessing requirements. Data is split into fixed-size windows with or without overlaps, depending on the application.

Several studies have explored the impact of window size on classification accuracy. For instance, smaller window sizes allow for quicker activity detection and reduced computational demands, but may fail to capture sufficient information for correct classification. On the other hand, larger window sizes may be more suitable for recognizing complex activities but come at the cost of increased computational load and the risk of mixing multiple activities into a single window. This would lead to biased classification outcomes in continuous recognition, as the activity that dominates the frame will be represented more compared with other activities [27, 54]. Similarly, the authors in [55] found that smaller windows improved shot-duration activity recognition but increased computational complexity, while larger windows reduced the load but decreased accuracy for detecting rapid activity changes. In DL approaches, Nunavath *et al.* [56] explored varying window lengths of 3, 5, and 10 seconds (with 50% overlap) and noted the trade-offs between window size and recognition performance. The window length is thus considered a selective tuning parameter that requires careful optimization based on the characteristics of the activity being studied.

2.3.4 Features

In traditional ML models, feature vectors are essential for the classification process. To obtain these vectors, feature extraction is used to capture the main characteristics

of a data segment that effectively represent the original signal [57]. This process helps to distinguish between different activities [58]. Once extracted, these characteristics are used as inputs to the classification algorithms.

Three main types of features are typically extracted from the IMU signal data [59]:

- **Time-domain** features, such as mean, variance, standard deviation, auto-correlation function, interquartile range, and maximum and minimum values.
- **Frequency-domain** features, which include Fourier transform, discrete cosine transform, and wavelet transform.
- **Dimensionality reduction** techniques, such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA).

Additionally, more advanced methods can be used for feature extraction from a sequence of sensor events, such as the baseline method and sensor window mutual information (SWMI) transform [54].

A summary of the most commonly used features for the classification of ADLs based on IMU is presented in Table 2.1.

Type	Features
Time-Domain	Mean
	Variance, Std. Dev., Mean Abs. Dev.
	RMS
	Cum. Histogram
	Zero or Mean Crossing Rate
	Derivative
	Peak Count & Amp.
	Sign
Frequency-Domain	Discrete FFT Coef.
	Spectral Centroid
	Spectral Energy
	Spectral Entropy
	Freq. Range Power
Time-Frequency Dom.	Wavelet Coef.
Domain-Specific	Time-Domain Gait Detection
	Vertical or Horizontal Acceleration

Table 2.1: The most widely used features in ADLs classification [3].

2.3.5 ADLs Classification

Activity classification for ADLs involves a complex problem with many degrees of freedom, as it is defined by several key elements within human activity recognition systems [53]. The classification and analysis of activities can be approached either offline or online, depending on system requirements. For non-interactive applications such as health monitoring, offline classification is commonly used. Furthermore, the system must be robust and user-independent, capable of generalizing well across various individuals, and resilient to temporal variations caused by factors such as sensor displacement or drifting sensor responses. Additionally, activity recognition

can be continuous, where the system automatically detects the occurrence of activities in real-time, or isolated/segmented, where the system assumes that input data is pre-segmented, classifying each segment as belonging to a single activity [53]. In this work, we will focus on both ML and DL approaches, some of which have been applied to patients, addressing both sequential and continuous activity identification methods.

2.3.5.1 Machine Learning based ADLs Classification Techniques

Human activity recognition is primarily approached as a supervised classification problem, where various ML algorithms are employed to build models that can accurately classify different activities based on sensor data. Among the traditional ML techniques used for task classification, Decision Trees (DT), k-Nearest Neighbors (kNN), Support Vector Machines (SVM), and Naïve Bayes (NB) are some of the most popular methods. These algorithms have been applied to process sensor data for various applications, from detecting daily activities to more complex behaviors like fine motor movements [3].

SVM, in particular, have gained attention due to their high accuracy when working with limited data and their suitability for offline tasks [27]. SVM was identified as the most frequently used classifier in a study by [1], where it appeared in 42 articles, followed closely by neural networks and kNN. Studies such as [60] and [61] demonstrated the successful application of SVM, C4.5 DT, and kNN classifiers to detect activities using data from smartphones and smartwatches. Additionally, in [62] the authors compared the performance of NB and kNN classifiers using accelerometer, further supporting their efficacy in HAR.

These traditional ML techniques have been widely used in domains ranging from movement classification to eating behavior detection and fine motor activity recognition [60, 61, 62, 50, 63, 52, 64, 49, 65]. However, factors such as sensor placement, data fusion, and classifier selection can significantly influence the overall accuracy and performance of these models, making the careful choice of method crucial to success.

Although there is a wide range of articles discussing ADLs classification, few studies focus on this problem in the context of tremor patients, aiming to combine it with tremor quantification. We will now review several studies that have explored ADLs classification in combination with tremor quantification, particularly in PD and ET patients.

Salarian *et al.* [66] developed a kinematic sensor-based system for continuous monitoring of PD patients during daily life. The system incorporated five sensors to detect and quantify tremor, bradykinesia, gait, and posture. By using a combination of statistical and fuzzy classifiers, the study achieved accurate classification of basic body postures and transitions, such as sitting and standing. Tremor quantification was performed based on spectral estimation, which demonstrated strong correlations with clinical scores. This study was one of the first to bridge laboratory-controlled studies with real-world, free-living scenarios, offering clinically relevant insights into motor function.

In a similar way, Zwartjes *et al.* [67] introduced an ambulatory monitoring system designed to analyze motor activity and symptom severity simultaneously. The system

used four inertial sensors and was validated with six PD patients and seven healthy controls. Activities such as sitting, standing, lying down, standing up, and walking were classified using a DT algorithm.

Serrano *et al.* [6] proposed a methodology for segmenting and recognizing ADLs using inertial sensor data, with a focus on upper-limb activities in tremor patients. The study involved 28 participants, including 15 patients with ET and 13 with PD. A NB classifier was used to identify activities and tasks, providing insight into the classification of activities in patients with tremors.

In a more recent study, Nguyen *et al.* [24] focused on the detection and segmentation of unstructured ADLs in patients with PD, using algorithms based on inertial sensor data. Nine older adults with PD performed tasks in a simulated free-living environment. The developed algorithms combined nonlinear transformations and adaptive thresholds to identify and segment activities such as sitting, standing, walking, reaching, and turning.

2.3.5.2 Deep Learning based ADLs Classification Techniques

Although ADLs classification has traditionally relied on conventional ML techniques, such as SVM, DT, and kNN, the fact that they require manual feature extraction is both time and resource consuming [68]. In contrast, DL models have become more popular due to their ability to automatially extract meaningful features from raw data, leading to superior performance, especially when dealing with large-scale datasets [68]. Moreover, DL approaches often provide better generalization across diverse scenarios, making them particularly suitable for handling the complexities inherent in human activity recognition tasks [55, 51]. Recent advances in DL architectures, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and hybrid models, have significantly enhanced the accuracy of ADLs classification. These models effectively address challenges such as overlapping feature spaces and unstructured activity patterns while reducing the dependence on hand-crafted feature engineering [69]. Hereafter, we review several studies that show the use of DL in ADLs classification, with a specific focus on methodologies, segmentation techniques, and patient populations.

Nunavath *et al.* [56] explored the use of DL for classifying physical movement activity patterns using datasets in which volunteers performed different ADLs using a wrist-worn IMU and sensors placed on their hips. The authors compared two architectures: a Deep Feedforward Neural Network (DNN) with five layers and a RNN using Gated Recurrent Units (GRUs). The study highlighted the advantages of RNNs, which showed better performance in distinguishing between closely related activities like climbing and descending stairs.

Ashry *et al.* [54] proposed a continuous ADLs classification framework using bidirectional LSTMs (Bi-LSTMs) to process IMU data from smartwatches. Their model utilised a feature descriptor combining autocorrelation, median, entropy, and instantaneous frequency, showcasing efficient activity recognition in continuous streams.

Gholamrezaii *et al.* [70] investigated the effectiveness of CNNs for HAR. They proposed an architecture that omitted pooling layers, replacing them with strided convolutions to reduce computational time without compromising accuracy. Their findings revealed that both 1D and 2D CNNs outperformed traditional hand-crafted feature

methods.

In [55], the authors analyzed the impact of sliding window segmentation on DL model performance using wrist-watch sensors. They evaluated CNN, LSTM, Bi-LSTM, and hybrid CNN-LSTM models on 11 ADLs from a HAR dataset. The CNN-LSTM hybrid achieved the highest accuracy with a 20-second overlapping sliding window.

Although DL has been widely applied to general ADLs classification, its integration with tremor quantification remains less explored. However, a few studies have targeted this intersection, addressing both ADLs recognition and tremor assessment in patients with PD and ET.

On one hand, Ni *et al.* [68] focused on ADLs classification for ET patients, combining it with tremor evaluation to support clinical diagnosis. Using smartwatch accelerometers, the authors collected data from 20 ET patients and 5 healthy controls performing activities such as extending arms, touching nose, writing, drawing spirals, pouring water, and simulating drinking. A Stacked Denoising Autoencoder (SDAE) was used to handle the small and noisy dataset. Resampling techniques were applied to address data imbalance.

On the other hand, Al-Majdi *et al.* [69] developed two models, an LSTM and a CNN-LSTM hybrid, to classify ADLs and detect falls using accelerometer data. The CNN-LSTM model, combining the feature extraction strengths of CNNs and the temporal sequence modeling capabilities of LSTMs, achieved higher accuracy than standalone LSTMs or traditional ML methods.

2.4 Publicly Available Datasets

Many of the studies reviewed in this work were validated using publicly available datasets, shown in Table 2.2. These datasets commonly use accelerometer sensors, some of which are wrist-worn, while others combine this information with gyroscopes and other additional sensors. Although these datasets provide a wide variety of activities, two major limitations persist. On the one hand, most existing studies focus on simple activities such as walking, running, and sitting, which involve whole-body movement. However, classifying more complex and varied daily activities such as cooking, eating, and working is necessary to enable broader applications [51].

On the other hand, none of these datasets involve patients, particularly those with neurological disorders such as PD or ET. It seems that there are few public datasets that focus on both ADLs classification and tremor analysis, and those who are aimed for tremor-related conditions often focus on falls, freezing, or postural activities such as "touching the nose" or "stretching the arms" [71, 72, 73]. Studies involving patients often fail to address complex task identification, focusing instead on postural and transitional activities [74]. Those that do tackle complex tasks usually require multiple IMUs, which can be burdensome for patients.

2.5 Limitations

Despite the advances in activity recognition, several challenges persist, particularly when dealing with patients with tremors. One challenge is intra-class variability,

Dataset	Sensors	Activities
DaLiAc [75]	A, G	Sitting, Lying, Standing, Walking, Stairs, Walking outside, Ascending stairs, Descending stairs, Treadmill running, Bicycling, Bicycling, Rope jumping
UCI HAR [76]	A, G	Walking, Walking upstairs, Walking downstairs, Sitting, Standing, Laying
PAMAP2 [77]	A, G, M, HR	Lying, Sitting, Standing, Walking, Running, Cycling, Nordic Walking, Watching TV, Computer Work, Car Driving, Ascending Stairs, Descending Stairs, Vacuum Cleaning, Ironing, Folding Laundry, House Cleaning, Playing Soccer, Rope Jumping, Other (Transient Activities)
WISDM [78]	A	Walking, Jogging, Upstairs, Downstairs, Sitting, Standing
ActiTracker [79]	A	Walking, Jogging, Stairs, Sitting, Standing, Lying
OPPORTUNITY [80]	A, G, M, O, AM	ADLs routines including: lying, grooming, walking, coffee preparation and drinking, sandwich preparation and eating, cleaning up, and a drill run (repetitive sequence of opening/closing furniture, toggling lights, cleaning, and drinking)
STISEN [81]	A, G	Biking, Sitting, Standing, Walking, Stair Up, Stair Down
RealWorld HAR [82]	A, G, GPS, L, MF, S	Climbing Upstairs, Jumping, Lying, Running, Sitting, Running, Jogging, And Walking
DHA [83]	A	Office working, reading, writing, eating, cooking, dish washing, walking, running, taking a transport

Table 2.2: Available datasets for sensor–based ADLs classification (A= Accelerometer, G= Gyroscope, M= Magnetometer, HR= Heart Rate, AM= Ambient Sensors, O= Object Sensors, L= Light Sensors, S= Sound Sensor, ECG= Electrocardiogram, EEG= Electroencephalogram, EOG= Electro–Oculogram, GPS= Global Positioning System, MF= Magnetic Field Sensor) [4].

where the same activity can be performed differently by the same individual or across different individuals. Factors like stress, fatigue, and emotional or environmental state can all influence how an activity is executed, impacting recognition accuracy. Inter-class similarity also presents difficulties. This situation occurs when activities themselves share similar movement patterns, making it hard to distinguish them based only on sensor data. Another common issue is class imbalance [53]. In real-world scenarios, some activities like sleeping or working occur frequently, while others like taking a sip of water are much less common. This imbalance can skew the performance of activity recognition models. Fortunately, techniques such as over-sampling (duplicating data points in the minority class) or generating synthetic data can help mitigate this problem [84].

Finally, the difficulties in data collection are made even more challenging by the problems faced when labelling the ground truth. In everyday life settings, accurately labelling activities becomes significantly more complex. Researchers have explored various methods to address this, including daily self-recall techniques where users report their activities later, experience sampling where users report activities prompted by random intervals, and even reinforcement learning or active learning approaches that all involve user interaction for supervised learning [85]. However, these methods are often time-consuming and prone to inaccuracies, further increasing the complexity of the problem [86, 87, 88]. While semi-supervised learning allows for part of the data to remain unlabeled, it still faces notable limitations due to the lack of standardised methods, with most systems implementing their own approaches [89, 90, 91] Semi-supervised learning in HAR, while allowing part of the data to remain unlabeled, faces notable limitations due to the lack of standardised algorithms or methods, with most systems implementing their own approaches [89, 90, 91].

Chapter 3

Methodology

In this chapter, the methodology used to design and develop a model for classifying complex Activities of Daily Living (ADLs) in patients with Parkinson's Disease (PD) and Essential Tremor (ET) is detailed. To achieve this primary objective, several sub-goals were addressed: firstly, the design and implementation of a battery of exercises tailored to evaluate tremor in home environments; secondly, the development of a machine learning classification model for short-duration, discrete sequences of ADLs as an initial approach to continuous identification; thirdly, the extension of the classification models to manage continuous, long-duration signals; and finally, the exploration of neural networks for the continuous classification of ADLs. The techniques used to address these objectives are outlined in this chapter.

3.1 Subject recruitment

The study included a total of 10 patients. Table 3.1 presents their demographic and disease-related characteristics. All patients were clinically diagnosed according to the consensus criteria established by the Movement Disorder Society Group [92]. Both ET and PD patients exhibited visible and persistent postural and kinetic tremors in their hands (either unilateral or bilateral), with some also showing tremors at rest. None of the patients had a history of neurological conditions other than PD or ET. The study was approved by the local ethics committee of the Hospital Universitario 12 de Octubre in Madrid, ensuring compliance with the Declaration of Helsinki. Furthermore, all patients provided written informed consent after being fully briefed about the study.

In addition, 45 healthy individuals participated as a control group. Their detailed demographic characteristics are also presented in Table 3.1.

3.2 Data collection

A smartwatch (Fitbit Sense, [93]) on the dominant wrist was used to collect all the data. The lightweight smartwatch is equipped with three orthogonal accelerometers for measuring linear acceleration (m/s^2) and three orthogonal gyroscopes for measuring angular velocity (rad/s). Motion data was sampled at 30 Hz. Then, all the data was transferred via Bluetooth to and Android device which sent the data via

Characteristic	ET patients	PD patients	Control participants
Total n	5	5	45
Age (years); mean ± SD	73 ± 7	71 ± 11	32 ± 17
Age (years); range	62-79	54-80	22-81
Disease duration (years)	3-33	3-15	N/A
Sex female; n (%)	2 (40)	1 (20)	19 (51.4)
Sex male; n (%)	3 (60)	4 (80)	18 (48.6)

Table 3.1: Characteristics of recruited participants for the study. The table provides demographic details including total sample size, mean age with standard deviation, age range, disease duration for patient groups, and distribution of sex among ET patients, PD patients, and control participants.

internet to a database for storage. During testing, each subject performed ten tasks of daily living (see Figure 3.1) at a self-chosen speed, in the same manner as they would normally execute them (see Table 3.2). These tasks can be grouped into four ADL groups:

- **Personal hygiene**: This group includes activities related to personal care and daily hygiene such as brushing teeth, combing hair or getting dressed which includes the buttoning task.
- **Feeding**: This group includes activities related to food intake and nutrition such as eating with a fork, cutting food with a knife, drinking, and opening a tupperware container.
- **Studying**: These tasks involve reading books or newspapers and taking handwritten notes.
- **Functional transfers**: Activities related to mobility in the environment such as opening and closing doors.

These tasks were selected according to previous studies [6] and available datasets [94, 95, 96, 97], as they encompass both fine (e.g., buttoning a shirt, turning pages, writing, and cutting food) and proximal movements (e.g., brushing teeth, combing hair, simulating drinking, and lifting the fork to the mouth), representing two levels of precision. Fine movements involve small muscles and precise control, whereas proximal movements engage larger muscle groups and require more general limb coordination. Handwriting, eating, dressing and self-care related activities are most prominently affected by tremor [15]. Also, these tasks represent activities that might interfere with tremor measurement at home, for example, handwriting and toothbrushing include repetitive motions that may resemble tremor. All tasks were performed with the dominant hand, and none were part of any standardized rating scale. Furthermore, the motion between repetitions and any periods of non-task activity were also recorded and labelled as 'no task' (NT).

3.2.1 Segmented setting

Part of this study was conducted to classify ADLs within the segmented or discrete setting, as outlined in Chapter 2. Data collection for patients was carried out at the hospital, while for control participants it took place in the laboratory. Before data

Figure 3.1: Matrix illustrating the 10 tasks performed by ET and PD patients during the ambulatory sessions.

collection, tasks were explained in detail to ensure that participants were familiar with them. The initial and final postures for each task were kept consistent. For seated tasks, participants began and finished with their hands placed flat on the table at shoulder width and elbows bent. For tasks performed while standing (such as buttoning a coat and turning a doorknob), participants started and ended in a relaxed upright stance with their arms naturally hanging at their sides. Each task was performed between 3 and 6 consecutive times, with 5 seconds rest between repeats. While the tasks were being performed, kinematic data from the sensor unit were gathered, and the subject's upper extremity was videotaped as a supplementary measure to evaluate the movements if necessary.

3.2.1.1 Additional Dataset

In this context, a complementary dataset [6] which contains records from four IMUs placed over the dominant arm was also used. From this dataset, the IMU placed at the third distal of the forearm was selected for analysis, focusing on the closest kinematic movements to the wrist. This dataset contains ET and PD diagnosed patients' data carrying out different ADLs, among them, the tasks shown in Table 3.2. The dataset is composed of acceleration and angular velocity in all three axes (x, y, z) from 16 patients, whose gender and age were not specified.

ADL type	Abbr.	Task	Task description
Personal hygiene	CH	Combing hair	Pick up comb from table - comb hair with dominant hand - put down comb
	BT	Brushing teeth	Pick up toothbrush - simulate tooth brushing covering the entire range - put down toothbrush
	BB	Buttoning	This task was performed standing up. Button the buttons of a lab coat in self-chosen order and unbutton.
Feeding	EF	Eating with fork	Pick up fork - select piece - bring to mouth - (repeat for 3 or 4 pieces) - put down fork
	CK	Cutting with knife	Pick up fork and knife (hold knife with dominant hand) - cut a piece from plasticine sheet - (repeat for 3 or 4 pieces) - put down fork and knife
	SD	Simulate Drinking	Pick up water bottle - bring to mouth - hold to mouth a few seconds (simulate drinking) - put bottle on the table
	OB	Open/close box	Pick up box container - open container - close container - put down container
Studying	WW	Writing	Pick up pen - write down name and surname - put down pen
	TP	Turning pages	Open book - turn 3 or 4 pages - close book
Functional transfers	TD	Open/close door	This task was performed standing up. Reach for doorknob - turn to open door - close door

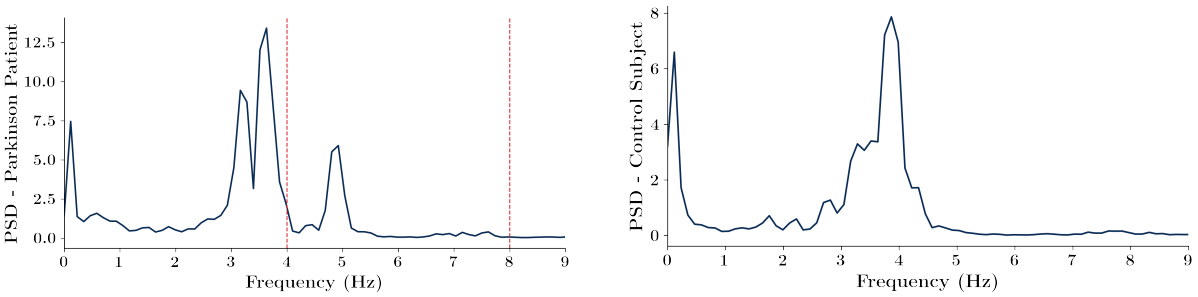
Table 3.2: Description of the tasks defined for the identification of contextual information. The table shows the code, and the description of the tasks selected to assess how patients and control subjects perform ADL's.

3.2.2 Continuous setting

In the second phase of the study, data collection was moved to a real world setting to capture continuous signals in a home environment. Participants were asked to wear a smartwatch for extended periods, allowing for natural and unstructured activity throughout the day. To annotate their actions, participants easily logged the start and end times of each task on a smartphone device, ensuring accurate alignment of sensor data with task execution. No specific sequence or order was imposed on the tasks, enabling a more realistic representation of daily activities. This setup aimed to mimic typical daily routines, providing valuable insights into the temporal variability and natural context of task performance, which are often absent in controlled laboratory environments. These signals were collected from eight control participants who had not participated in the data collection conducted in the segmented setting, further enriching the dataset with unbiased and independent contributions.

3.3 Data Processing

First, data preprocessing was carried out to address incomplete data. Signals that deviated significantly from the central time distribution of each task were excluded, as they likely represented faulty recordings that could result in misclassification.



(a) (b)

Figure 3.2: Spectrogram of a PD patient (a) and a control subject (b) performing the task 'brushing teeth', where the dashed red lines delimit the tremor component frequencies.

Then, all data were resampled to a uniform frequency of 30 Hz to ensure consistency with both the experimental dataset and the complementary dataset detailed in Section 3.2.1.1. To align with the pre-processing described in [6], a Butterworth filter was applied with a cutoff frequency of 4 Hz and a filter order of 6. This filtering approach is based on the distinction between voluntary movements and tremorous activity, which typically occurs around a 3.5 Hz threshold [98]. Even for tasks involving the highest frequency movements, such as brushing teeth, spectrogram peaks remain below 4 Hz [99]. Figure 3.2 illustrates this distinction, comparing the spectrograms of a control subject and a patient with PD during the brushing-tooth task. In the PD patient, voluntary movement peaks are observed below 4 Hz, while tremorous components appear around 5 Hz. Thus, the 4 Hz cutoff frequency was chosen to effectively filter out tremor while preserving voluntary movement (see Figure 3.2).

3.3.1 Window Design

Given the variability in signal duration across different tasks and patients, and the lack of consensus on an optimal window size [26, 28, 27], the signals were segmented using five types of overlapping windows of varying sizes: 2.5 seconds with a 1 second overlap, 5 seconds with a 2 second overlap, 10 seconds with a 4 second overlap, 15 seconds with a 6 second overlap, and 20 seconds with an 8 second overlap. This selection was guided by previous studies investigating the impact of window size on physical activity classification, detailed in Section 2.

After segmentation, statistical, frequency, and shape-related features were extracted from each window, following the methodologies in [97, 100, 6]. The extracted features included metrics such as mean, standard deviation, median, maximum, minimum, variance, RMS (Root Mean Square), entropy, centroid, bandwidth, flatness, energy, skewness, and kurtosis. In addition, the difference between the first and last values of each window was calculated. These features, representing various aspects of the signal, resulted in five distinct collections of 94 features corresponding to the five segmentation approaches.

To manage the complexity and size of the dataset and to mitigate potential overfitting due to high-dimensional feature spaces, various feature selection and dimensionality reduction techniques were applied. These methods aimed to identify the most relevant features for activity classification while preserving the essential characteristics

of the original data.

3.3.2 Feature Subset Selection

To optimise model performance by reducing dimensionality, we applied feature selection techniques to identify the most relevant variables within our dataset. Specifically, we used the *SelectKBest* method [101], a filter-based feature selection approach that evaluates and ranks features based on their statistical relationship with the target variable (i.e., the task label). Filter-based methods rely on statistical metrics to score features independently of the model, ensuring an efficient selection process. Given the characteristics of our data and the nature of the problem, an ANOVA test was used to measure the dependencies between the features and the target variable.

Using the importance scores generated by the ANOVA test, we selected the top 60 features with the highest scores to build the final feature subset. This threshold was determined graphically, observing the distribution of the importance of characteristics in identifying a cut-off point [102]. By selecting the most informative features, this approach ensured a balance between dimensionality reduction and the preservation of critical information for accurate task classification.

3.4 Classification of ADLs

3.4.1 Segmented Setting Classification

As described in Chapter 2, in this setting, the system assumes that input data are pre-segmented, classifying each segment as belonging to a single activity.

3.4.1.1 Algorithm Design

The algorithm used to classify the ADL segment followed a two-phase pipeline, as shown in Figure 3.3. In the first phase, the data was divided into training and testing sets, taking into account the five different windowing processes (2.5, 5, 10, 15, and 20 seconds). Consequently, five classifiers were trained, each corresponding to a distinct window segmentation, and their predictions were evaluated. Within each sequence, defined as a task trial for a specific patient, the segmented windows were grouped. A label for the entire sequence was determined using the mode of the predicted labels. If the mode was ambiguous, the label with the highest mean probability was assigned. This approach ensured that each sequence was labelled along with an associated prediction probability. The process was repeated independently for each classifier, corresponding to the various window sizes.

In the second phase, the predictions obtained from all classifiers for each sequence were aggregated. The predicted labels and probabilities of each classifier were compared. A similar procedure to the one used in the first phase was applied: if a clear mode existed among the predicted labels, the sequence was assigned to that label, accompanied by the highest probability among the mode predictions. In cases where no distinct mode was present, the label corresponding to the maximum probability was selected. At the end of the pipeline, each sequence was assigned a final predicted task label along with an associated classification probability, providing a robust and consistent output for activity recognition.

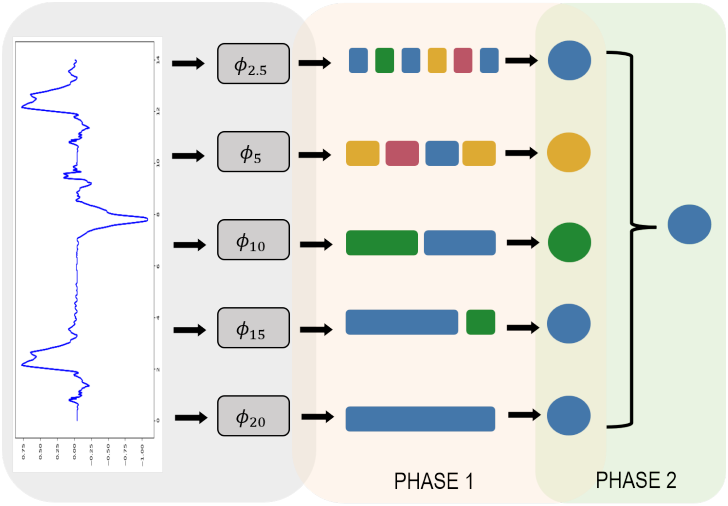


Figure 3.3: Representation of the design algorithm for different time length windows integration for the classification of ADLs. Where ϕ represents each classifier trained with the different window-sized segmented dataset.

3.4.1.2 ML CLassifiers

Three different types of classifiers were implemented, chosen based on the state-of-the-art reviewed in Chapter 2: Random Forest (RF) [103], Support Vector Machines (SVM) [104], and Extreme Gradient Boosting (XGB) [105]. These models were selected to perform multiclass classification, and each one was trained and tested for this purpose. Figure 3.4 illustrates the pipeline followed, from data collection to classification, showing how the different classifiers were applied for ADL identification.

In addition, each model was hyperparameter tuned to optimise performance. Grid-search cross-validation [101] was applied to assess the impact of modifying classifier-specific parameters on accuracy. After determining the optimal parameters for each model, a 7-fold cross-validation procedure was used for validation. To prevent data from the same patient from appearing in both the training and testing sets simultaneously, a group-stratified k-fold cross-validation method was used. This approach maintained a train-test split of 85-15%, ensuring that each fold contained at least one ET or PD patient, thus preserving the proportionality of the classes. This strategy helped to provide robust and generalisable results for activity classification.

3.4.1.2.1 Random Forest

RF is a powerful ML algorithm known for its ensemble approach, where multiple decision trees work together to improve predictive performance [103]. In this model, each decision tree independently learns from bootstrapped samples drawn from the training dataset. Every tree makes predictions by evaluating feature values, progressing through nodes, and splitting branches until reaching a final leaf node that represents the target class. The model then aggregates the predictions of all trees using majority voting to produce a single output.

The effectiveness of RF relies on fine-tuning its hyperparameters, especially the number of estimators and the maximum depth of the trees. Increasing the number of estimators typically improves the model's performance by reducing the bias or error

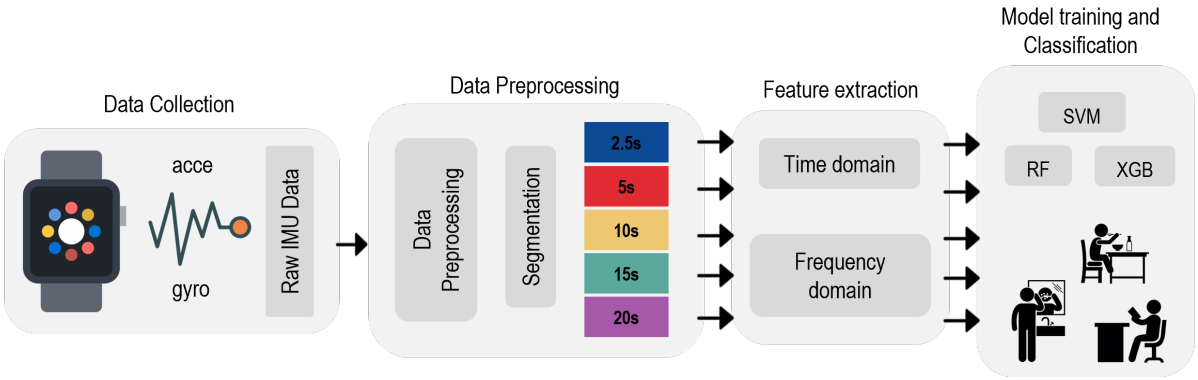


Figure 3.4: Framework of ADL identification using ML approaches. First, inertial data is collected through sensors. Then, feature extraction is performed using five different windowing processes. This results in segmented datasets of varying window sizes, which are then fed into ML classifiers for final activity classification.

of each individual tree and leveraging their combined knowledge. However, adjusting the maximum depth of the trees can influence the model's complexity, enabling it to capture more intricate patterns in the data. However, setting excessively high values for these hyperparameters can lead to overfitting, reducing the model's ability to generalise well to unseen data. To optimise these parameters, grid search was used to tune the number of estimators and the maximum depth. A hyperparameter search was performed using 5-fold stratified group cross-validation, testing five different values for the number of estimators (100, 200, 300, 400, 500) and six different values for the maximum depth (None, 2, 5, 10, 15, 20), with the goal of maximising the model's accuracy for each dataset.

3.4.1.2.2 Support Vector Machine

SVM is a classification algorithm that aims to find a hyperplane capable of separating different classes in the training dataset [104]. It does this by identifying support vectors and maximising the margin between these vectors and the hyperplane to improve generalisation. While SVM is initially designed for linear separability, it can also manage nonlinear classification through the use of kernel functions. Key hyperparameters, such as the regularisation term C and the type of kernel (e.g., linear, sigmoid, polynomial, and radial basis function (RBF)), are typically fine-tuned to identify the optimal model for the dataset.

To optimise these parameters, grid search cross-validation was employed to evaluate various values of C (ranging from 1 to 1000) and kernel types (RBF and polynomial). The process systematically tested different combinations of these parameters to identify the best fit to the data. By iterating over the hyperparameter values and assessing their performance in terms of accuracy, grid-search ensured that the final model was well tuned, providing the highest possible accuracy while minimising the risk of overfitting. This evaluation was repeated across different folds for patients and for each individual time-length classifier (2.5, 5, 10, 15, and 20 seconds), allowing for a comprehensive comparison of the model's performance at various time intervals. The best-performing SVM models, along with their respective C and kernel parameters, were then selected for final use.

3.4.1.2.3 Extreme Gradient Boosting

XGB is a ML model designed for classification tasks that combines decision trees with the boosting ensemble technique [105]. This approach works by sequentially refining predictions and correcting errors made by previous models in an attempt to minimise the loss function through gradient descent.

XGB offers two critical hyperparameters for optimisation: maximum depth and learning rate. To mitigate overfitting, particularly in smaller datasets, the maximum depth is typically constrained. Grid-search cross-validation was carried out across the five time window classifiers, testing various combinations of learning rates (ranging from 0.1 to 0.3) and tree depths (from 3 to 7). This procedure allows for fine-tuning the model, reducing the risk of overfitting, and ensuring improved performance.

3.4.2 Continuous Setting Classification

As described in Chapter 2, in this setting, the system automatically detects the occurrence of activities, assuming that each sequence may contain different task labels.

3.4.2.1 Machine Learning Approach

3.4.2.1.1 Segmentation

The feature extraction process is modified for continuous signals. Instead of relying on pre-segmented task sequences, we start with a single continuous signal containing multiple tasks. To segment the sequences, window sizes of 5s, 10s, 15s, and 20s were used, discarding the 2.5s windows, since they were found to provide unsatisfactory results in the segmented setting tests. By applying these window sizes with the previously defined overlaps, for each window, the majority label within the window is recorded, and its temporal position is stored within a start index. For the last window, both the start and end indexes are saved. This information is then combined with all the extracted features (both temporal and spectral), allowing to trace back what part of the original signal each window came from. Maintaining temporal information facilitates later analysis and allows to map the predictions back to the original signal.

3.4.2.1.2 Algorithm Design

The algorithm designed for the segmented setting (see Section 3.4.1.1) was slightly adapted. The algorithm now uses four classifiers instead of five, but follows a similar two-phase process. The 2.5-second classifier was discarded due to its poor performance in the discrete setting, which will be further explained in Section 4.

First, the four trained models for the segmented setting were loaded and the continuous signals recorded (detailed in Section 3.2.2) were used as the test group. Each of the four loaded window classifiers provides labels for each window. With the complementary information extracted during the segmentation process, the segmented signal is unrolled to return to the continuous signal. Now, the data point, instead of being a time window, is a specific time instance. To handle and manage multiple labels per data point caused by the overlap within the windows, the label is selected based on the mode or the maximum associated probability.

Finally, the predictions from the four classifiers are combined for each time instance, following the same scheme as in the segmented setting. The final task label and

prediction probability for each sequence were determined by repeating the same procedure as in the first phase, selecting the mode or the label with the maximum probability in case of ambiguity. Thus, the algorithm is very similar to the previous one, maintaining the same concept of combining classifiers with different window sizes based on the mode and prediction probability.

3.4.2.1.3 ML Classifiers

The SVM classifier was used due to its better performance in the previous settings, which will be discussed in Chapter 4. The classifier was trained with data from the segmented setting, and optimised through hyperparameter tuning with GridSearch and validated with 7-fold cross-validation, maintaining class proportionality. The train-test ratio was set to 85-15%. Once the model was trained, the preprocessed long-duration signals from the test group were introduced to identify and classify tasks in continuous sequences.

3.4.2.2 Deep Learning Approach

The continuous signals were then classified using various DL architectures. In this study, the signals were segmented into fixed 15-second windows with a 6-second overlap, ensuring sufficient temporal context while maintaining computational efficiency. The model was trained and validated on data collected in a controlled setting, which was split into training and validation subsets—75% for training and 15% for validation—both derived from the controlled setting data. A separate 10% of the data came from continuous setting data, which was used for testing. The model's generalization capabilities were then evaluated on continuous signals. The model's input was multivariate, consisting of accelerometer and gyroscope signals across all three axes. It outputs a predicted label for each data point, enabling offline classification of continuous signals.

Finally, a post-processing step was applied in which labelled sequences shorter than a specified threshold were merged with neighbouring sequences, given that they shared the same label and exceeded a minimum length. This step helped to reduce misclassification and improve consistency in the predicted labels.

3.4.2.2.1 DL Architectures

Three different neural network architectures were trained to classify the continuous data, each using different techniques to effectively capture temporal and spatial information, as shown in Figure 3.5.

Bidirectional LSTM

This architecture implements a Bidirectional Long Short-Term Memory (LSTM) layer to capture temporal dependencies in both forward and backward directions, making it well-suited for tasks where the temporal order is crucial. The input consists of sequential data with a shape of $(\text{TIME\_STEPS}, 6)$, representing six features derived from accelerometer and gyroscope signals. To prevent overfitting, dropout layers were included, and L2 regularization was applied to enhance generalization. The output layer uses a softmax activation function with the number of units corresponding to

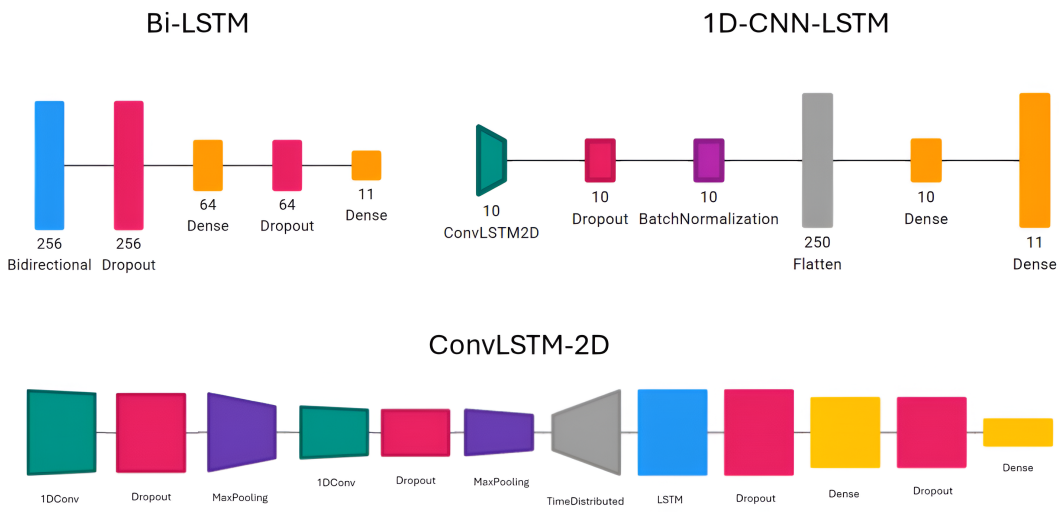


Figure 3.5: Different DL architectures implemented. Figures created with *Net2Vis* [2]

the target activities. The model is optimised using the Adam optimiser to ensure stable training, while the categorical cross-entropy loss function is used with accuracy as the evaluation metric.

1D-CNN-LSTM

This architecture combines Convolutional Neural Networks (CNNs) and LSTMs to leverage both spatial and temporal features. The model begins with a unidimensional convolutional layer with ReLU activation to extract spatial features from each time step independently. As before, dropout layers were included for regularisation, along with max-pooling layers to reduce dimensionality. After the convolutional blocks, a Flatten layer prepares the data for temporal processing. Then, an LSTM layer follows, which captures temporal dependencies in the data. Dense layers and L2 regularization are incorporated to further improve generalization. The output layer uses a dense layer with a softmax activation function, where the number of units corresponds to the unique task identifiers in the dataset. This architecture effectively integrates spatial and temporal information, thereby enhancing classification performance.

ConvLSTM-2D

This architecture combines the strengths of convolutional and LSTM layers in a single ConvLSTM2D operation, making it highly effective for spatiotemporal data. The model begins with a ConvLSTM2D layer, which is similar to an LSTM layer, but both input transformations and recurrent transformations are convolutional [106]. This layer extracts both spatial and temporal features simultaneously. Dropout layers with a rate of 70% are added to prevent overfitting, followed by BatchNormalization to stabilise and accelerate training, and an L1 activity regulariser to enforce sparsity. The final output layer uses a softmax activation function, with the number of units corresponding to the unique task identifiers in the dataset. This architecture is compact yet powerful, effectively combining convolutional, recurrent, and dense components.

In summary, the BiLSTM architecture is designed to focus solely on capturing tempo-

ral dependencies. In contrast, the 1D CNN-LSTM model combines the spatial feature extraction capabilities of Convolutional Neural Networks (CNNs) with the temporal modeling of Long Short-Term Memory (LSTM) networks. This allows the model to process spatial features at each time step independently through CNNs, followed by LSTMs to capture the temporal dependencies across the sequence. On the other hand, the ConvLSTM2D architecture integrates both spatial and temporal feature extraction into a single layer, where convolutional filters operate within the LSTM framework, capturing both dimensions, spatial and temporal, simultaneously.

Chapter 4

Results and Discussion

This chapter presents and discusses the main findings of this work.

4.1 Segmented Setting

As a reminder, in this setting, the signals were collected sequentially, with each signal corresponding to a single task. The classification problem was addressed using three types of classifiers: RF, SVM, and XGB (see Section 3.4.1.2). In addition, a combination approach with varying temporal window sizes was tested to evaluate its impact on performance (see Section 3.4.1.1).

The accuracy results are summarised in Table 4.1, showing that the SVM classifier outperformed the others in all metrics evaluated, closely followed by RF. In particular, all metrics surpassed the 80% threshold, demonstrating the robustness of the models. The task-specific precision and recall values for SVM are detailed in the table shown in Figure 4.1. The algorithm achieved high precision and recall ($\sim 85\%$) for most tasks. However, tasks such as cutting with a knife (CK), open/close a box (OB), and turning pages (TP) exhibited lower precision, falling below the 80% mark.

Model	**Accuracy**	**Precision**	**Recall**	**F1 score**
RF	83.06%	83.94%	82.13%	82.85%
SVM	**84.88%**	**86.00%**	**84.88%**	**84.98%**
XGB	82.39%	83.22%	82.47%	82.38%

Table 4.1: Average results obtained by the three tested classifiers (RF, SVM, and XGB)

An analysis of the performance of each different sized-windowed classifier individually (see Figure 4.2) reveals a trend in which larger windows generally result in better classification performance across the three tested classifiers. In particular, no tasks were best classified with the 2.5 second window, and only two tasks, EF and CH performed better with the 5-second window. The modal performance across the three classifiers (RF, SVM, and XGB) was observed with the 15-second window. There is no clear pattern observed across the three classifier types, other than larger windows tend to reach higher results. Tasks like EF and CH, which involve repetitive

Task	Precision	Recall	F1 score
BB	86.32%	92.02%	88.76%
BT	93.35%	91.99%	92.45%
CH	89.36%	86.01%	87.03%
CK	77.14%	84.46%	79.94%
EF	83.29%	85.07%	83.96%
OB	71.42%	70.28%	70.30%
SD	93.76%	88.22%	90.64%
WW	91.84%	87.36%	89.06%
TP	77.07%	71.77%	73.81%
TD	96.19%	92.35%	93.99%
M-Avg	85.97%	84.95%	85.01%
W-Avg	86.00%	84.88%	84.98%
Acc			**84.88%**

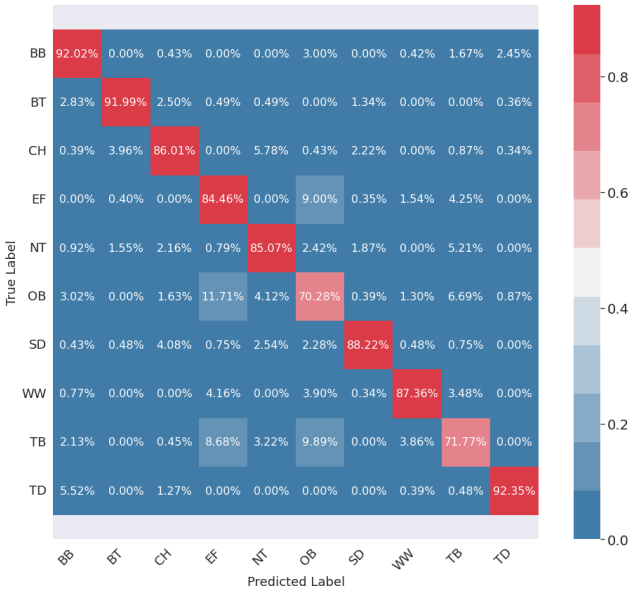


Figure 4.1: Results of the ensemble SVM classifier for ADL classification. On the left, performance metrics per task in terms of precision, recall, and F1-score, as well as the weighted and unweighted average metrics. On the right, the confusion matrix illustrating the classifier's performance on continuous signals.

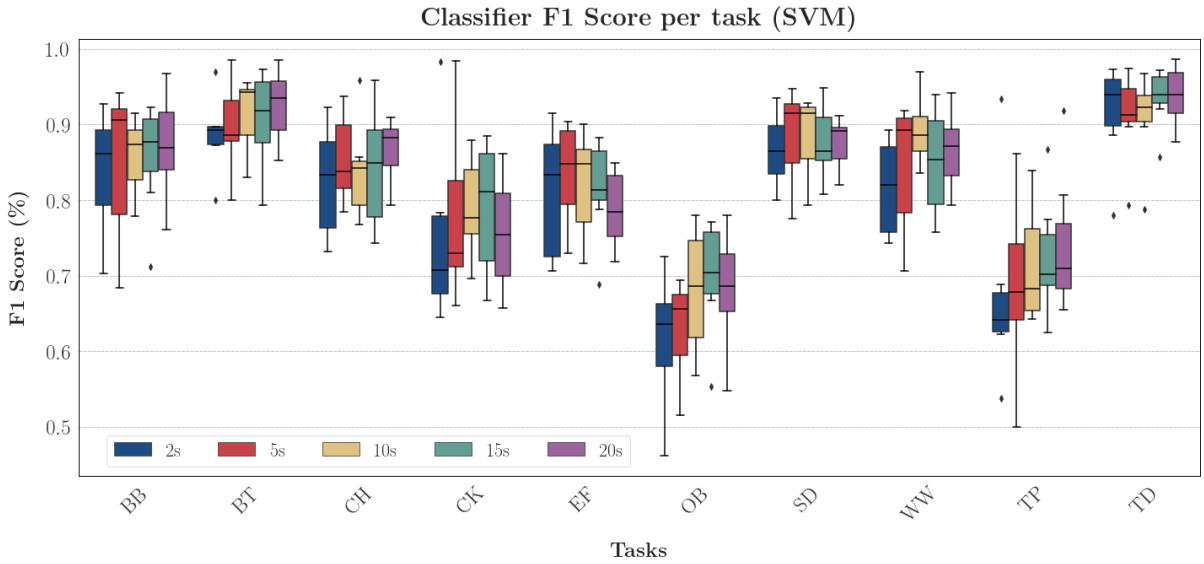


Figure 4.2: Bar graph showing the precision obtained in the classification per task using different time windows with SVM classifiers. The error bars represent the standard error of the mean (SEM).

movements encompassing the entire arm, seem to be better classified with smaller windows. Conversely, tasks like TP and TD that involve fine movements seem to perform better with larger windows. Other tasks like BB, CK, or OB show better performance with intermediate-sized windows of 10-15 seconds.

The tasks with the lowest classification performance were OB, and TP, with an approximately 69% and 73% F1-score respectively. After examining the distributions

of these tasks, we observed that they both involve bimanual object manipulation at a distal point from the body. Since our recordings are unilateral, capturing only the dominant wrist, classifiers struggle to differentiate between these tasks with sufficient accuracy. People often use both hands for these tasks, especially if the dominant hand is somewhat affected by tremor. Therefore, their execution varies significantly depending on individual preferences and motor habits. For instance, in CK, some participants may hold the knife with their dominant hand, while others use their non-dominant hand. Similarly, for OB, variability arises from differences in hand dominance and the specific techniques used to open a box-shaped recipient.

Despite these findings for individual windows, the ensemble models outperformed single-window approaches (see Figure 4.3), showing the potential efficacy of combining results from different window sizes.

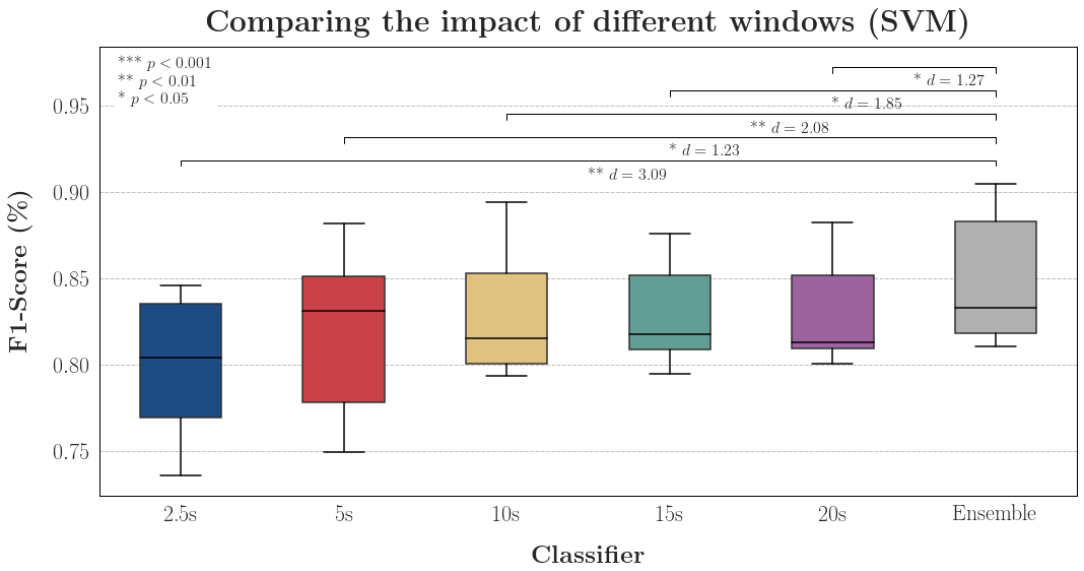


Figure 4.3: Boxplots illustrating the F1-Scores (%) of individual classifiers using different window sizes (2.5s, 5s, 10s, 15s, 20s) and the ensemble classifier. Significant differences between classifiers are indicated with asterisks (*, **, ***), and effect sizes (Cohen's d) are reported. The ensemble classifier demonstrates superior performance across most comparisons.

A statistical analysis evaluated the performance of the SVM ensemble classifier compared to individual (single window size) classifiers using the F1-Score as the primary evaluation metric. A paired *t-test* was conducted to compare the performance, as the scores for the ensemble and individual classifiers were paired and derived from the same datasets. The assumptions of the *t-test*, including the normality of the differences between the paired scores, were rigorously evaluated using the Shapiro-Wilk test, along with visual analysis like Q-Q plots and Kernel Density Estimation (KDE) plots (see Figure 4.4). Effect sizes were calculated using *Cohen's d* to quantify the magnitude of the observed differences. To address the issue of multiple comparisons (n=5), a *Holm-Bonferroni* correction was applied to adjust the significance threshold, minimizing the risk of Type I errors (false positives). The results of this analysis are summarised in Figure 4.3, which presents boxplots of the F1-scores for each classifier, highlighting the ensemble classifier's superior performance.

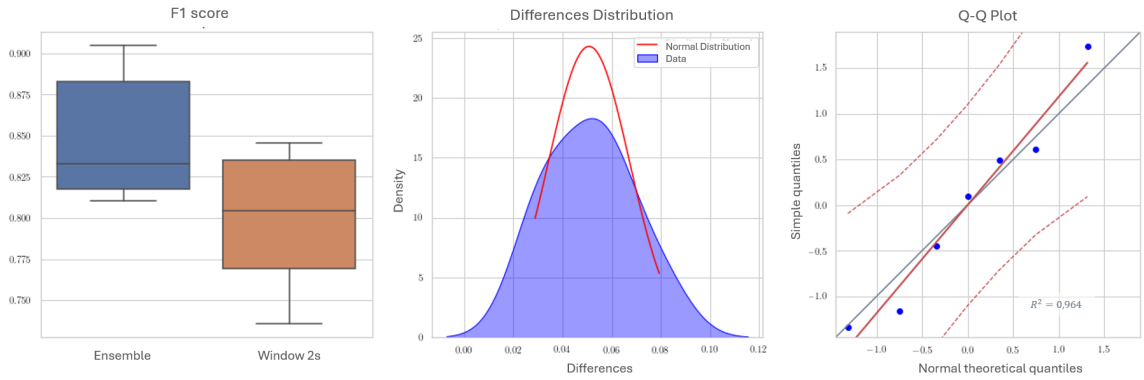


Figure 4.4: Comparison of F1-scores for ensemble and 2-second window approaches (left). The density plot (centre) shows the distribution of differences between both models predictions, while the Q-Q plot (right) illustrates the alignment of data quantiles with theoretical normal quantiles with a 95% confidence interval.

This comprehensive approach ensured the validity and robustness of the statistical analysis. However, it is important to note that the sample size (n=7 (k-folds)) limits the statistical power of the tests. Therefore, repeating the analysis with a larger dataset would be recommended to increase the confidence in the reliability and robustness of the results.

4.2 Continuous Setting

As a reminder, in this setting, the signals were collected in such a way that each signal corresponds to several different tasks. Specifically, the tasks CK and EF have been merged, as it made more sense to label the activity as 'eating' rather than distinguishing between the two. Furthermore, the task 'signing name' has been generalised to 'writing,' which aligns better with the context of daily life activities. An additional important classification is 'no-task (NT)', which refers to any activity that does not fall under one of the 10 target tasks. This classification helps to identify when a relevant task is being performed and when it is not.

4.2.1 ML Approach

The classification problem was addressed using the SVM classifier, as it was found to provide better results in the discrete setting (Figure 4.1). Once again, a combination approach with varying temporal window sizes was tested to evaluate its impact on performance. Therefore, the classifiers used were trained with all data collected in the discrete setting and then tested on continuous signals to evaluate their inference capabilities on longer and more complex signals (see Section 3.4.2.1). In this stage, four classifiers are used instead of five, as the 2.5-second classifier was discarded due to its poor performance in the previous setting.

The table in Figure 4.5 shows the metrics used to evaluate the performance of the ensemble model on continuous signals. The overall accuracy is 72.77%. However, there is a notable variability between tasks, following a similar pattern as in the segmented setting classification results, although with certain exceptions. The simple (M) and weighted (W) average metrics are also presented to provide a comprehensive

view of the model's performance. The simple averages give equal importance to each class, whereas the weighted averages account for class imbalance by considering the number of instances in each category, providing metrics that reflect the model's performance in real-world conditions, where certain classes naturally occur more frequently. It is worth noting the difference between the simple and weighted metrics, as they underscore the level of class imbalance in our test set, which will be discussed in further detail in Section 5.

Task	**Precision**	**Recall**	**F1 score**
BB	26.81%	79.74%	39.93%
BT	66.32%	88.31%	75.75%
CH	50.34%	85.68%	63.42%
EF	72.75%	58.84%	65.06%
NT	85.59%	77.65%	81.43%
OB	29.63%	54.36%	38.35%
SD	57.40%	48.57%	52.62%
WW	59.77%	58.05%	58.90%
TP	56.13%	76.23%	64.66%
TD	62.42%	21.06%	31.50%
M-Avg	56.70%	64.85%	57.16%
W-Avg	76.51%	72.77%	73.70%
Acc			**72.77%**

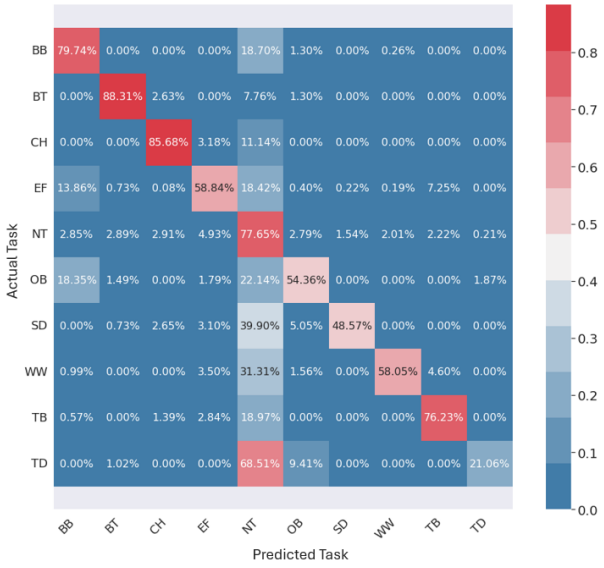


Figure 4.5: Results of the ensemble SVM classifier for ADL classification. On the left, performance metrics per task in terms of precision, recall, and F1-score, as well as the weighted and unweighted average metrics. On the right, the confusion matrix illustrating the classifier's performance on continuous signals.

The confusion matrix in Figure 4.5 further illustrates the model's performance across classes. Although the matrix demonstrates a strong diagonal, indicating accurate predictions for many tasks, it also reveals significant misclassifications, particularly false positives with the NT class. This suggests that, although the model is generally effective, there are specific areas where its ability to distinguish between tasks could be improved.

For tasks such as BB and OB, which are potentially more complex due to their bi-manual nature, precision is low (<30%), despite a high recall, indicating the presence of false positives, as shown in Figure 3. In contrast, tasks such as BT and CH exhibit balanced performance metrics, suggesting more reliable classification. The task with the highest metrics is NT, likely due to its greater representation in the dataset. The confusion matrix supports this observation, revealing that many tasks are misclassified as NT, probably due to class imbalance.

When comparing the table in Figure 4.1 with the table in Figure 4.5, it is evident that in these tests, the classification results for tasks TD, SD, and BB worsen compared to the discrete setting, whereas the results for EF and TP show partial improvement. This may be attributed to the representation and proportion of these tasks in the new test set. Tasks such as TD and BB had little representation in the test set of continuous signals, which contained task EF in a much higher proportion. The OB

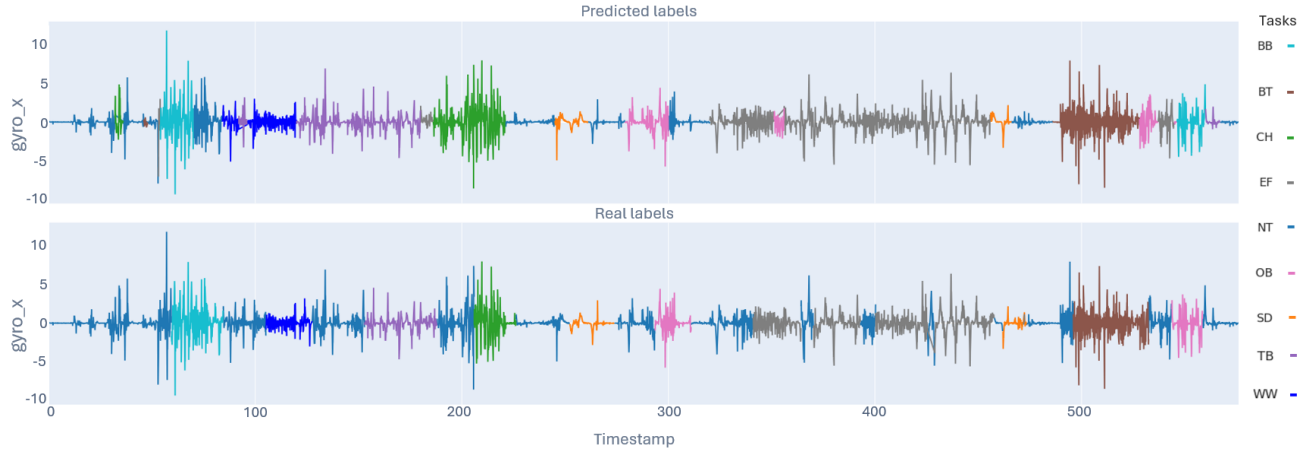


Figure 4.6: Graphs that compare the predicted labels by the ensemble-SVM model with the real ones in a continuous sequence of 10 minutes. Each colour represents a different task label.

task continues to be particularly challenging to classify.

To illustrate the performance of the ensemble model, a graph similar to the one shown in Figure 4.6 are plotted for each participant in the test group. In these graphs, different colours represent the various tasks performed over time, including the NT label, which indicates periods where no specific activity is carried out. The upper section of the graph displays the tasks performed by the participants, while the lower section shows the tasks classified by the model. Some misclassifications are evident in the graphs, such as instances where tasks of interest are detected during NT periods (e.g., in the 0–50s interval, a CH (green) sequence is falsely detected during an NT period. A similar example can be seen in the 550–600s interval, where an EF task (grey) appears as a false positive in Figure 4.6). Additionally, the model occasionally identifies very short, incoherent signals that do not align with the expected activity patterns (e.g., in the 0–50s interval, a very short BT (brown) sequence is falsely detected during an NT period, which is inconsistent with the expected duration of the task. A similar case occurs around the 550s point, where very short OB (pink) and TB (purple) sequences are detected between tasks in Figure 4.6). To address these issues, post-processing techniques were incorporated into the DL approach to correct potentially erroneous classifications using contextual information.

4.2.2 DL Approach

After testing the results of combining different window sizes with more traditional ML techniques, and after expanding our dataset with the continuous signals collected, we decided to experiment with DL, which has become widely used in recent state-of-the-art research. Therefore, after designing the three architectures outlined in Section 3.4.2.2.1, the models were trained using segmented signals from the first setting for both training and validation. Once the parameters were fine-tuned, the models were tested on the continuous signals that had been recorded.

Table 4.2 presents the comparative results obtained with the three different DL architectures. The BiLSTM model demonstrated superior performance across all metrics,

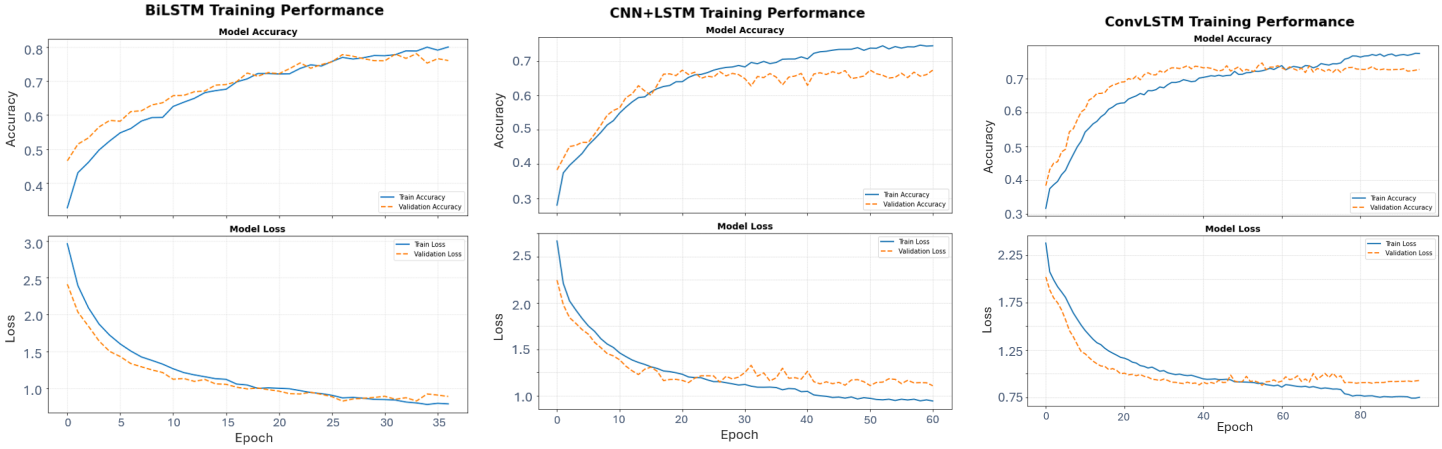


Figure 4.7: Training curves from the BiLSTM model (left), the CNN+LSTM model (centre), and the ConvLSTM model (right). Validation set's curves are shown in orange and train's in blue.

achieving the highest results. Despite its superior performance, the results do not represent a significant improvement over those previously obtained with traditional machine learning techniques. This highlights several areas for potential improvement. Firstly, increasing the training data volume, as DL architectures typically require large amounts of data to reach their full potential. Secondly, further optimizing the neural network architectures through more exhaustive hyperparameter tuning. Lastly, exploring data augmentation techniques specific to time series could enhance the model's generalisability and robustness.

Figure 4.7 presents the three sets of training performance plots for different models (CNN+LSTM, ConvLSTM, and BiLSTM). From this, it can be observed that, among the three models, the BiLSTM achieves the best performance, with high accuracy, low loss, and notable stability between training and validation metrics. The ConvLSTM follows, due to its stability and generalisation, though its performance is lower than the BiLSTM. In contrast, the CNN+LSTM model shows signs of overfitting and could benefit from further adjustments.

In Figure 4.8, it can be seen that the performance of the BiLSTM model for the classification of ADL varied between tasks, following a pattern similar to the one observed in the results of the SVM ensemble (Figure 4.5). As shown in the table, some tasks exhibited high recall but low precision (i.e. BB, CH, OB), suggesting that while the model is sensitive in identifying true positives it also produces a significant number of false positives in these tasks. This may be due to a greater class imbalance in this test group. High-performance activities with F1-score values above 70% included BT,

Model	**Accuracy**	**Precision**	**Recall**	**F1 score**
BiLSTM	**75.17%**	**78.21%**	**75.17%**	**75.89%**
CNN+LSTM	54.23%	78.83%	54.23%	60.40%
ConvLSTM	59.81%	70.61%	59.81%	62.45%

Table 4.2: Average results obtained by the three tested DL models.

Task	Precision	Recall	F1 score
BB	37.04%	75.52%	49.70%
BT	70.29%	90.71%	79.20%
CH	60.29%	83.36%	69.97%
EF	61.27%	81.11%	69.81%
NT	88.61%	75.86%	81.74%
OB	23.35%	30.10%	26.30%
SD	52.74%	42.89%	47.31%
WW	81.41%	69.85%	75.19%
TB	77.55%	78.76%	78.15%
TD	28.62%	27.51%	28.05%
M-Avg	58.12%	65.57%	60.54%
W-Avg	78.21%	75.17%	75.89%
Acc			**75.17%**

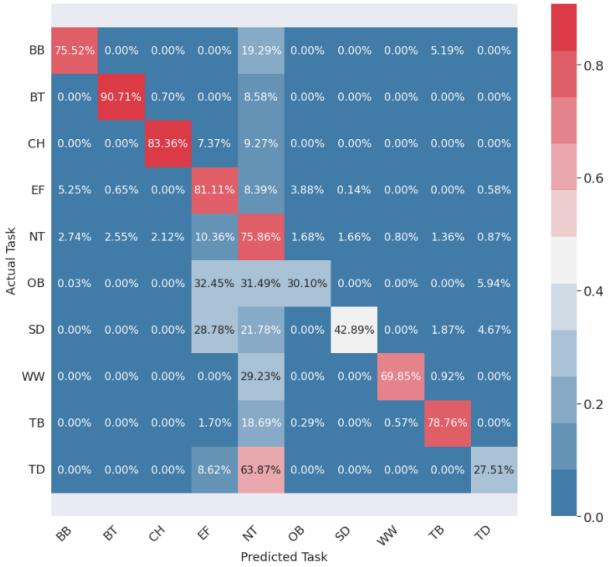


Figure 4.8: Results of the ensemble BiLSTM model for ADL classification. On the left, performance metrics per task in terms of precision, recall, and F1-score, as well as the weighted and unweighted average metrics. On the right, the confusion matrix illustrating the model's performance on continuous signals.

EF and TP. These tasks involve repetitive and distinct motion patterns, such as hand-to-mouth or cyclical arm movements, which are easily captured by both acceleration and gyroscope data.

Medium performance tasks, included BB and WW. These activities benefited from recognisable motion patterns, though individual variations, particularly in writing styles, likely impacted accuracy. Lower performance tasks, with recall below 50% and F1-score below 30%, included TD and OB. The brief duration, variability in execution, and similarities with other reaching movements made these actions more difficult to classify.

A comparison of the confusion matrix in Figure 4.8 with that in Figure 4.5 reveals a noticeable reduction in false positives for the NT class. Furthermore, the diagonal in the BiLSTM confusion matrix appears to be more defined, reflecting improved clarity and accuracy in task classification.

To illustrate the performance of the ensemble model, Figure 4.9 presents a graph similar to the one shown in Figure 4.6, representing the same signal. A comparison between the two reveals that, although discrepancies with the true labels remain, the classification has improved. Short, incoherent signals are no longer present, probably due to the implementation of post-processing techniques that homogenise segments shorter than a certain threshold based on their local context. However, some misclassifications persist, such as the OB (orange) task in the 550–600s segment, which is labeled as TB (brown), and certain false positives in the SD and OB regions (e.g., a false positive for SD (dark blue) around 50s, or NT (blue) misclassified as OB in the 190–200s interval). These observations suggest that further improvements are necessary, which will be discussed in the next chapter of this work.

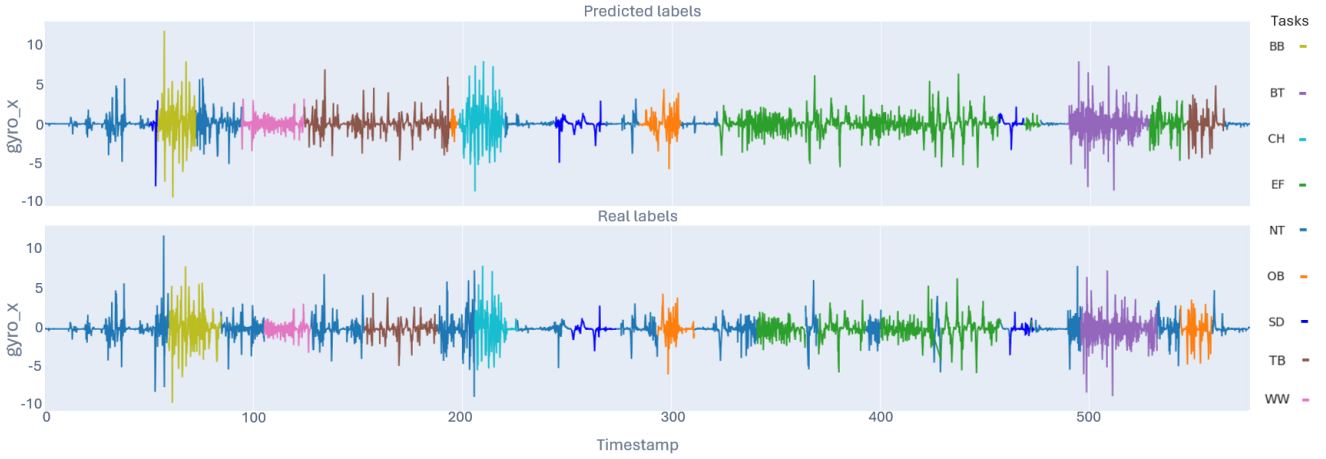


Figure 4.9: Graphs that compare the predicted labels by the BiLSTM model with the real ones in a continuous sequence of 10 minutes. Each colour represents a different task label.

4.3 Building a Dataset

Another key result of this work is the creation of a dataset consisting of signals from ET and PD patients, as well as healthy participants, recorded while performing complex daily living tasks. The dataset contains multivariate time series collected using a wrist-worn sensor, which includes gyroscope and accelerometer data from three axes. The recordings cover both discrete and continuous activities, ensuring a comprehensive representation of the patient's movement patterns. In total, the dataset contains 10 task classes (EF and CK are considered as one class), including activities such as reading, eating, and writing, along with a 'no-task' class to capture periods when the patient is not engaged in a specific activity.

From the distribution of proportions shown in Figure 4.10, we can observe that the majority class is NT, followed by the EF task. Apart from these two, the remaining tasks have a relatively similar representation in the dataset. Furthermore, it can be seen that the total proportion of data from subjects with and without tremor is approximately a 2:1 ratio, with a significantly larger amount of data from subjects without tremor.

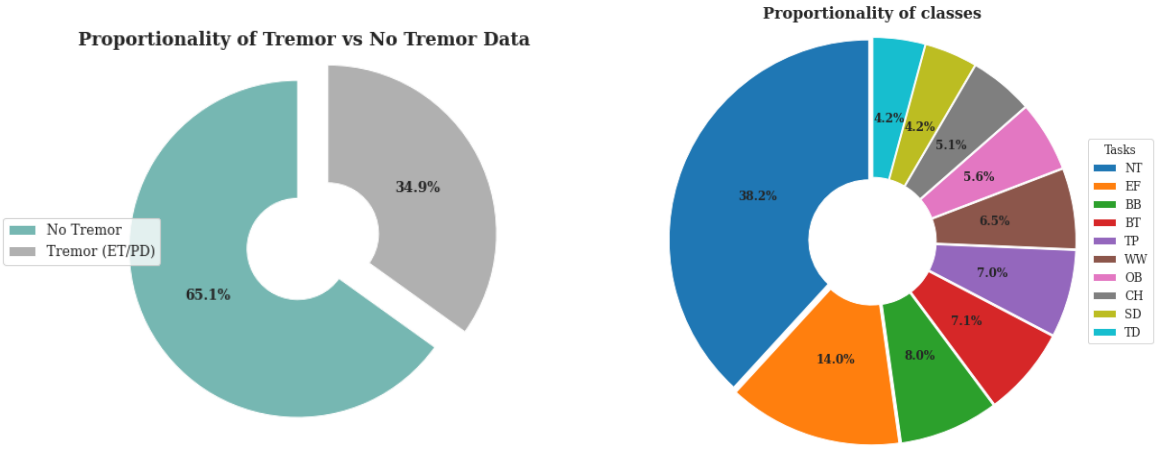


Figure 4.10: Pie chart showing the proportion of classes (ADLs) and the tremorous component in the collected data

Chapter 5

Conclusions and Future Work

This work explored the classification of ADLs using wearable inertial sensors in patients with tremor, specifically those diagnosed with PD and ET. The primary objective was to develop a model capable of accurately classifying complex ADLs in both controlled and real-life settings.

To address this objective, various approaches and methodologies were used. First, we described the use of inertial recordings as a minimally obtrusive technology for extracting movement patterns to identify task execution. A recording protocol was carefully designed and a specific battery of exercises was selected, considering tasks that were common in daily life and informative to compare patterns related to tremors. On the one hand, the problem of segmented classification was studied, where each segment or sequence is assumed to consist of a single task. This scenario was addressed using classical ML techniques. On the other hand, the continuous setting was also approached, where long signals are recorded, and each segment or sequence may contain multiple tasks with varying start and end times. This second problem was tackled using both classical ML techniques and more recent DL approaches.

For the first problem (segmented setting), an approach was tested that combines different window lengths, ranging from 2.5 to 20 seconds, to preprocess and segment the ADL signals. This was done with the aim of addressing the lack of consensus in choosing an appropriate window size for feature extraction from time-series signals in HAR and ADL classification problems. To explore this, five different classifiers were trained with the differently segmented datasets, and an ensemble model was built by combining the predictions of these classifiers to provide the most accurate class label.

To address the second problem, two distinct approaches were used. The first was to adapt the SVM ensemble to classify continuous signals by incorporating temporal information during the segmentation process prior to feature extraction. The second approach involved DL techniques. Three types of architectures were built: BiLSTM, which considered only temporal information; CNN+LSTM, which extracted features using one-dimensional convolutions before processing them temporally; and ConvLSTM, which integrates both spatial and temporal feature extraction into a single layer, where convolutional filters operate within the LSTM framework. Post-processing techniques were applied to reduce the number of misclassifications using local information.

In terms of the conclusions drawn from the classification of activities, several key insights have emerged regarding the performance of the classification models and the challenges faced in ADL recognition:

For the segmented classification, the SVM classifier demonstrated the best overall performance compared to RF and XGB, achieving accuracy metrics above 80%. Combining models trained with different window sizes further enhanced performance, particularly when compared to using individual windows. However, tasks that involved bimanual manipulation of objects, such as cutting with a knife (CK), opening/-closing boxes (OB), and turning pages (TP), showed lower classification accuracy. This could be mainly due to the unilateral nature of the recordings, which captured movement only from the dominant wrist, resulting in incomplete representations of these tasks. In terms of temporal windows, the analysis indicated that larger windows (15 seconds) typically resulted in better performance across the classifiers tested. However, tasks involving repetitive arm movements, such as eating with a fork (EF) and combing hair (CH), were classified more accurately with smaller windows, suggesting the need for an adaptable windowing approach depending on task characteristics.

For continuous classification with ML, where multiple tasks were recorded in a continuous stream, the SVM classifier achieved an overall accuracy of 72.77%. The performance varied notably across different tasks. Tasks like buttoning (BB) and opening/closing boxes (OB) had lower accuracy, while tasks such as brushing teeth (BT) and combing hair (CH) showed more balanced classification metrics. The 'no task' (NT) class performed particularly well, likely due to its higher representation in the dataset. When applying DL techniques, the BiLSTM model demonstrated the best performance for classifying continuous signals, possibly because it has a simpler architecture and the test data was limited. However, this improvement was not substantially better than the results achieved using the SVM classifier in ML. The BiLSTM model effectively reduced false positives for the NT class and improved classification clarity. Tasks like brushing teeth (BT), eating (EF), and turning pages (TP) performed well under this model. The application of post-processing techniques also improved the consistency and reliability of the classifications.

One of the limitations identified in the study is the reduced size of the dataset, particularly with regard to the continuous signals that made up the test group. Additionally, the dataset is highly imbalanced, showing a clear bias towards the majority class, 'no task' (NT), which is reflected in the classification results. Therefore, to improve the reliability and quality of the results, especially in the continuous setting, it would be necessary to increase the size of the dataset. Resampling techniques could be applied to improve model testing or to expand the dataset, achieving a better balance between classes, as well as between tremor and non-tremor data. Furthermore, data augmentation techniques could be explored to artificially increase the size and variability of the data set, improving the ability of the model to generalise. Additionally, undersampling or oversampling methods could be considered to address class imbalance, ensuring a more balanced representation of tasks across the dataset. Also, expanding the participant pool will support model generalisation and allow for meaningful comparisons with DL approaches, enhancing the system's robustness and real-world applicability. Additionally, to enhance the representativeness of the dataset, it would be beneficial to include a wider range of activity context, considering additional ADLs affected by tremor. Conversations with patients have highlighted tasks such as inserting a key into a lock or eating with a spoon, suggesting potential

directions for future expansions of the study.

In conclusion, the study highlighted the potential of wearable sensors and both ML and DL techniques for classifying ADLs, while also revealing challenges in handling complex tasks, particularly when data is unilateral or there is class imbalance. Further improvements are needed, especially in the continuous classification scenario, to enhance accuracy and consistency. These advancements could significantly aid healthcare professionals in evaluating prescribed treatments and medication dosages, potentially reducing reliance on trial-and-error approaches. Future work should focus on exploring optimal window sizes for classification and real-time activity analysis to assess the effectiveness of therapies for PD and ET, deepening the understanding of tremor-related patterns and improving classification algorithms.

Bibliography

[1] C. Kranzinger, S. Bernhart, W. Kremser, V. Venek, H. Rieser, S. Mayr, and S. Kranzinger, "Classification of Human Motion Data Based on Inertial Measurement Units in Sports: A Scoping Review," *Applied Sciences*, vol. 13, p. 8684, July 2023.

[2] A. Bäuerle, C. van Onzenoodt, and T. Ropinski, "Net2vis – a visual grammar for automatically generating publication-tailored cnn architecture visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 6, pp. 2980–2991, 2021.

[3] A. Avci, S. Bosch, M. Marin-Perianu, R. Marin-Perianu, and P. Havinga, "Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey," pp. 167–176, 01 2010.

[4] Z. Baloch, F. K. Shaikh, and M. A. Unar, "Deep Architectures for Human Activity Recognition using Sensors," *3C Tecnología_Glosas de innovación aplicadas a la pyme*, pp. 14–35, May 2019.

[5] A. Anouti, W. C. Koller, and K. City, "Articles tremor disorders diagnosis and management," 6 1995.

[6] J. I. Serrano, S. Lambrecht, M. D. del Castillo, J. P. Romero, J. Benito-León, and E. Rocon, "Identification of activities of daily living in tremorous patients using inertial sensors," *Expert Systems with Applications*, vol. 83, pp. 40–48, 10 2017.

[7] B. Thanvi, N. Lo, and T. Robinson, "Essential tremor - the most common movement disorder in older people," 7 2006.

[8] F. Demrozi, R. Bacchin, S. Tamburin, M. Cristani, and G. Pravadelli, "Toward a wearable system for predicting freezing of gait in people affected by parkinson's disease," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, pp. 2444–2451, 9 2020.

[9] M. A. Thenganatt and J. Jankovic, "The relationship between essential tremor and parkinson's disease," *Parkinsonism Related Disorders*, vol. 22, pp. S162–S165, 2016. Proceedings of XXI World Congress on Parkinson's Disease and Related Disorders, December 6-9, 2015, Milan, Italy.

[10] P. Crawford and E. E. Zimmerman, "Tremor: Sorting through the differential diagnosis," *American family physician*, vol. 97, pp. 180–186, 2018.

[11] G. Mostile, J. P. Giuffrida, O. R. Adam, A. Davidson, and J. Jankovic, "Correlation between kinesia system assessments and clinical tremor scores in patients with essential tremor," *Movement Disorders*, vol. 25, pp. 1938–1943, 9 2010.

[12] E. D. Louis, K. J. Wendt, S. M. Albert, S. L. Pullman, Q. Yu, and H. Andrews, "Validity of a performance-based test of function in essential tremor," *Archives of Neurology*, vol. 56, p. 841, 7 1999.

[13] E. D. Louis, L. Barnes, K. J. Wendt, B. Ford, M. Sangiorgio, S. Tabbal, L. Lewis, P. Kaufmann, C. Moskowitz, C. L. Comella, C. C. Goetz, and A. E. Lang, "A teaching videotape for the assessment of essential tremor," *Movement Disorders*, vol. 16, pp. 89–93, 1 2001.

[14] S. Fahn, "Classification of movement disorders," 5 2011.

[15] D. A. Heldman, J. Jankovic, D. E. Vaillancourt, J. Prodoehl, R. J. Elble, and J. P. Giuffrida, "Essential tremor quantification during activities of daily living," *Parkinsonism and Related Disorders*, vol. 17, pp. 537–542, 8 2011.

[16] K. Frank, M. Josefa, V. Nadales, P. Robertson, and M. Angermann, "Reliable real-time recognition of motion related human activities using mems inertial sensors."

[17] I. Vanmechelen, H. Haberfehlner, J. D. Vleeschhauwer, E. V. Wonterghem, H. Feys, K. Desloovere, J. M. Aerts, and E. Monbaliu, "Assessment of movement disorders using wearable sensors during upper limb tasks: A scoping review," 1 2023.

[18] L. Sigcha, L. Borzì, F. Amato, I. Rechichi, C. Ramos-Romero, A. Cárdenas, L. Gascó, and G. Olmo, "Deep learning and wearable sensors for the diagnosis and monitoring of parkinson's disease: A systematic review," *Expert Systems with Applications*, vol. 229, p. 120541, 11 2023.

[19] D. Rodríguez-Martín, J. Cabestany, C. Pérez-López, M. Pie, J. Calvet, A. Samà, C. Capra, A. Català, and A. Rodríguez-Molinero, "A New Paradigm in Parkinson's Disease Evaluation With Wearable Medical Devices: A Review of STAT-ONTM," *Frontiers in Neurology*, vol. 13, p. 912343, June 2022.

[20] C. L. Pulliam, M. A. Burack, D. A. Heldman, J. P. Giuffrida, and T. O. Mera, "Motion sensor dyskinesia assessment during activities of daily living," *Journal of Parkinson's Disease*, vol. 4, no. 4, p. 609–615, 2014.

[21] C. L. Pulliam, D. A. Heldman, E. B. Brokaw, T. O. Mera, Z. K. Mari, and M. A. Burack, "Continuous assessment of levodopa response in parkinson's disease using wearable motion sensors," *IEEE Transactions on Biomedical Engineering*, vol. 65, p. 159–164, Jan. 2018.

[22] M. D. Hssayeni, J. Jimenez-Shahed, M. A. Burack, and B. Ghoraani, "Dyskinesia estimation during activities of daily living using wearable motion sensors and deep recurrent networks," *Scientific Reports*, vol. 11, Apr. 2021.

[23] A. Salarian, H. Russmann, C. Wider, P. R. Burkhard, F. J. G. Vingerhoets, and K. Aminian, "Quantification of tremor and bradykinesia in parkinson's disease using a novel ambulatory monitoring system," *IEEE Transactions on Biomedical Engineering*, vol. 54, p. 313–322, Feb. 2007.

[24] H. Nguyen, K. Lebel, S. Bogard, E. Goubault, P. Boissy, and C. Duval, "Using inertial sensors to automatically detect and segment activities of daily living in people with parkinson's disease," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, pp. 197–204, 1 2018.

[25] B. Jiang, J. J. Han, and J. Kim, "A wearable in-home tremor assessment system via virtual reality environment for the activities in daily lives (adls)," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 1117–1120, IEEE, 7 2022.

[26] O. Banos, J.-M. Galvez, M. Damas, H. Pomares, and I. Rojas, "Window size impact in human activity recognition," *Sensors*, vol. 14, pp. 6474–6499, 4 2014.

[27] N. Yala, B. Fergani, and A. Fleury, "Towards improving feature extraction and classification for activity recognition on streaming data," *Journal of Ambient Intelligence and Humanized Computing*, vol. 8, pp. 177–189, 4 2017.

[28] T. Gu, Z. Wu, X. Tao, H. K. Pung, and J. Lu, "epsicar: An emerging patterns based approach to sequential, interleaved and concurrent activity recognition," in *2009 IEEE International Conference on Pervasive Computing and Communications*, pp. 1–9, IEEE, 3 2009.

[29] P. Urwyler, L. Rampa, R. Stucki, M. Büchler, R. Müri, U. P. Mosimann, and T. Nef, "Recognition of activities of daily living in healthy subjects using two ad-hoc classifiers," *BioMedical Engineering OnLine*, vol. 14, p. 54, Dec. 2015.

[30] V. R. Chifu, C. B. Pop, D. Demjen, R. Socaci, D. Todea, M. Antal, T. Cioara, I. Anghel, and C. Antal, "Identifying and Monitoring the Daily Routine of Seniors Living at Home," *Sensors*, vol. 22, p. 992, Jan. 2022.

[31] Leidy Tatiana Ordoñez Mora, Diana Patricia Sánchez, Jorge Enrique Daza Arana, Lina Johanna Álvarez Toro, Paola Teresa Penagos Gómez, Marysol Valencia Buitrago, Claudia Fernanda Giraldo Jiménez, Mónica Yamile Pinzón Bernal, María Mercedes Naranjo Aristizábal, Jennifer Jaramillo Losada, and Sandra Milena Carabalí Cachimbo, "Evaluación de autocuidado y actividades de la vida diaria," in *Evaluación de la función neuromuscular*, pp. 325–350, Colombia: Universidad Santiago de cali, 2020.

[32] Z. Hussain, M. Sheng, and W. E. Zhang, "Different Approaches for Human Activity Recognition: A Survey," *Journal of Network and Computer Applications*, vol. 167, p. 102738, Oct. 2020. arXiv:1906.05074 [cs].

[33] A. Avci, S. Bosch, M. Marin-Perianu, R. Marin-Perianu, and P. Havinga, "Activity Recognition Using Inertial Sensing for Healthcare, Wellbeing and Sports Applications: A Survey," in *23th International Conference on Architecture of Computing Systems 2010*, pp. 1–10, Feb. 2010. Journal Abbreviation: 23th International Conference on Architecture of Computing Systems 2010.

[34] A. Antonini, H. Reichmann, G. Gentile, M. Garon, C. Tedesco, A. Frank, B. Falkenburger, S. Konitsiotis, K. Tsamis, G. Rigas, N. Kostikis, A. Ntanis, and C. Pattichis, "Toward objective monitoring of Parkinson's disease motor symptoms using a wearable device: wearability and performance evaluation of PDMonitor®," *Frontiers in Neurology*, vol. 14, p. 1080752, May 2023.

[35] M. Knudson, T. H. Thomsen, and T. W. Kjaer, "Comparing Objective and Subjective Measures of Parkinson's Disease Using the Parkinson's KinetiGraph," *Frontiers in Neurology*, vol. 11, p. 570833, Nov. 2020.

[36] M. D. S. T. F. on Rating Scales for Parkinson's Disease, "The unified parkinson's disease rating scale (updrs): status and recommendations," *Movement Disorders*, vol. 18, no. 7, pp. 738–750, 2003.

[37] C. G. Goetz, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, G. T. Stebbins, M. B. Stern, B. C. Tilley, R. Dodel, B. Dubois, *et al.*, "Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): process, format, and clinimetric testing plan," *Movement disorders*, vol. 22, no. 1, pp. 41–47, 2007.

[38] C. Swartling, "The Essential Tremor Rating Assessment Scale," *Journal of Neurology and Neuromedicine*, vol. 1, pp. 34–38, July 2016.

[39] S. Sapienza, O. Tsurkalenko, M. Giraitis, A. C. Mejia, G. Zelimkhanov, I. Schwaninger, and J. Klucken, "Assessing the clinical utility of inertial sensors for home monitoring in Parkinson's disease: a comprehensive review," *npj Parkinson's Disease*, vol. 10, p. 161, Aug. 2024.

[40] M. H. Monje, G. Foffani, J. Obeso, and Sánchez-Ferro, "New sensor and wearable technologies to aid in the diagnosis and treatment monitoring of parkinson's disease," *Annual Review of Biomedical Engineering*, vol. 21, p. 111–143, June 2019.

[41] A. Channa, N. Popescu, and V. Ciobanu, "Wearable solutions for patients with parkinson's disease and neurocognitive disorder: A systematic review," *Sensors*, vol. 20, p. 2713, May 2020.

[42] A. Tzallas, M. Tsipouras, G. Rigas, D. Tsalikakis, E. Karvounis, M. Chondrogiorgi, F. Psomadellis, J. Cancela, M. Pastorino, M. Waldmeyer, S. Konitsiotis, and D. Fotiadis, "Perform: A system for monitoring, assessment and management of patients with parkinson's disease," *Sensors*, vol. 14, p. 21329–21357, Nov. 2014.

[43] C. Moreau, T. Rouaud, D. Grabli, I. Benatru, P. Remy, A.-R. Marques, S. Drapier, L.-L. Mariani, E. Roze, D. Devos, G. Dupont, M. Bereau, and M. Fabbri, "review imus en parkinson!!," *npj Parkinson's Disease*, vol. 9, p. 153, Nov. 2023.

[44] K. R. Chaudhuri, A. Hand, F. Obam, and J. Belsey, "Cost-effectiveness analysis of the Parkinson's KinetiGraph and clinical assessment in the management of Parkinson's disease," *Journal of Medical Economics*, vol. 25, pp. 774–782, Dec. 2022.

[45] J. P. Giuffrida, D. E. Riley, B. N. Maddux, and D. A. Heldman, "Clinically deployable Kinesia™ technology for automated tremor assessment," *Movement Disorders*, vol. 24, pp. 723–730, Apr. 2009.

[46] M. E. Gerbasi, R. J. Elble, E. Jones, A. Gillespie, J. Jarvis, E. Chertavian, Z. Smith, M. Nejati, and L. C. Shih, "Associations Among Tremor Amplitude, Activities of Daily Living, and Quality of Life in Patients with Essential Tremor," *Tremor and Other Hyperkinetic Movements*, vol. 14, p. 22, May 2024.

[47] S. Balli, E. A. Sağbaş, and M. Peker, "Human activity recognition from smart watch sensor data using a hybrid of principal component analysis and random forest algorithm," *Measurement and Control*, vol. 52, pp. 37–45, Jan. 2019.

[48] N. Ravi, N. Dandekar, P. Mysore, and M. Littman, "Activity recognition from accelerometer data.," vol. 3, pp. 1541–1546, 01 2005.

[49] F. G. da Silva and E. Galeazzo, "Accelerometer based intelligent system for human movement recognition," in *5th IEEE International Workshop on Advances in Sensors and Interfaces IWASI*, pp. 20–24, 2013.

[50] J. Parkka, M. Ermes, K. Antila, M. van Gils, A. Manttari, and H. Nieminen, "Estimating intensity of physical activity: A comparison of wearable accelerometer and gyro sensors and 3 sensor locations," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 1511–1514, 2007.

[51] M.-C. Kwon and S. Choi, "Recognition of Daily Human Activity Using an Artificial Neural Network and Smartwatch," *Wireless Communications and Mobile Computing*, vol. 2018, p. 2618045, Jan. 2018.

[52] H. Wang, T. T.-T. Lai, and R. Roy Choudhury, "Mole: Motion leaks through smartwatch sensors," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, MobiCom '15, (New York, NY, USA), p. 155–166, Association for Computing Machinery, 2015.

[53] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Computing Surveys*, vol. 46, pp. 1–33, Jan. 2014.

[54] S. Ashry, T. Ogawa, and W. Gomaa, "CHARM-Deep: Continuous Human Activity Recognition Model Based on Deep Neural Network Using IMU Sensors of Smartwatch," *IEEE Sensors Journal*, vol. 20, pp. 8757–8770, Aug. 2020.

[55] S. Mekruksavanich, P. Jantawong, and A. Jitpattanakul, "Effect of Sliding Window Sizes on Sensor-Based Human Activity Recognition Using Smartwatch Sensors and Deep Learning Approaches," in *2024 5th International Conference on Big Data Analytics and Practices (IBDAP)*, (Bangkok, Thailand), pp. 124–129, IEEE, Aug. 2024.

[56] V. Nunavath, S. Johansen, T. S. Johannessen, L. Jiao, B. H. Hansen, S. Berntsen, and M. Goodwin, "Deep Learning for Classifying Physical Activities from Accelerometer Data," *Sensors*, vol. 21, p. 5564, Aug. 2021.

[57] N. Krishnan, C. Juillard, D. Colbry, and S. Panchanathan, "Recognition of hand movements using wearable accelerometers," *JAISE*, vol. 1, pp. 143–155, 01 2009.

[58] J.-Y. Yang, J.-S. Wang, and Y.-P. Chen, "Using acceleration measurements for activity recognition: An effective learning algorithm for constructing neural classifiers," *Pattern Recognition Letters*, vol. 29, pp. 2213–2220, 12 2008.

[59] S. Preece, J. Goulermas, L. Kenney, D. Howard, K. Meijer, and R. Crompton, "Activity identification using body-mounted sensors — a review of classification techniques," *Physiological measurement*, vol. 30, pp. R1–33, 05 2009.

[60] M. Shoaib, S. Bosch, H. Scholten, P. J. M. Havinga, and O. D. Incel, "Towards detection of bad habits by fusing smartphone and smartwatch sensors," in *2015 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, pp. 591–596, 2015.

[61] N. Ravi, N. Dandekar, P. Mysore, and M. Littman, "Activity recognition from accelerometer data.," vol. 3, pp. 1541–1546, 01 2005.

[62] B. Banu and B. Mokhtiar Sheriff, "A study of different daily human activities from smart devices using ai," 01 2022.

[63] Y. Dong, J. Scisco, M. Wilson, E. Muth, and A. Hoover, "Detecting periods of eating during free-living by tracking wrist motion," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, p. 1253–1260, July 2014.

[64] D. Riboni and C. Bettini, "Cosar: Hybrid reasoning for context-aware activity recognition," *Personal and Ubiquitous Computing*, vol. 15, pp. 271–289, 03 2011.

[65] R. I. Ramos-Garcia and A. W. Hoover, "A study of temporal action sequencing during consumption of a meal," in *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, BCB'13, (New York, NY, USA), p. 68–75, Association for Computing Machinery, 2013.

[66] A. Salarian, *Ambulatory monitoring of motor functions in patients with Parkinson's disease using kinematic sensors*. PhD thesis, École Polytechnique Fédérale de Lausanne (EPFL), 2006.

[67] D. G. M. Zwartjes, T. Heida, J. P. P. van Vugt, J. A. G. Geelen, and P. H. Veltink, "Ambulatory monitoring of activities and motor symptoms in parkinson's disease," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 11, pp. 2778–2786, 2010.

[68] Q. Ni, Z. Fan, L. Zhang, B. Zhang, X. Zheng, and Y. Zhang, "Daily Activity Recognition and Tremor Quantification from Accelerometer Data for Patients with Essential Tremor Using Stacked Denoising Autoencoders," *International Journal of Computational Intelligence Systems*, vol. 15, p. 1, Dec. 2022.

[69] K. Al-Majdi, R. S. AL-Musawi, A. H. Ali, S. A. Abudlmoniem, and Y. S. Mezaal, "Real-Time classification of various types of falls and activities of daily livings based on CNN LSTM network," *Periodicals of Engineering and Natural Sciences (PEN)*, vol. 9, p. 958, Sept. 2021.

[70] M. Gholamrezaii and S. AlModarresi, "A time-efficient convolutional neural network model in human activity recognition," *Multimedia Tools and Applications*, vol. 80, pp. 19361–19376, May 2021.

[71] P. M. Roggen, Daniel and J. Hausdorff, "Daphnet Freezing of Gait." UCI Machine Learning Repository, 2010. DOI: https://doi.org/10.24432/C56K78.

[72] B. M. Bot, C. Suver, E. C. Neto, M. Kellen, A. Klein, C. Bare, M. Doerr, A. Pratap, J. Wilbanks, E. R. Dorsey, S. H. Friend, and A. D. Trister, "The mpower study, parkinson disease mobile data collected using researchkit," *Scientific Data*, vol. 3, Mar. 2016.

[73] Julian Varghese, Alexander Brenner, Lucas Plagwitz, Catharina van Alen, Michael Fujarski, and Tobias Warnecke, "PADS - Parkinsons Disease Smartwatch dataset," Mar. 2024.

[74] J. Russell, J. Inches, C. Carroll, and J. Bergmann, "A five-sensor imu-based parkinson's disease patient and control dataset including three activities of daily living," 2023.

[75] H. Leutheuser, D. Schuldhaus, and B. M. Eskofier, "Hierarchical, multi-sensor based classification of daily life activities: comparison with state-of-the-art algorithms using a benchmark dataset," *PLoS ONE*, vol. 8, no. 10, p. e75196, 2013.

[76] J. L. Reyes-Ortiz, D. Anguita, A. Ghio, L. Oneto, and X. Parra, "Human activity recognition using smartphones data set," 2013.

[77] D. Reiss, M. Stricker, G. Haring, D. H. Hütten, M. S. Böhm, and H. W. Schümmer, "Pamap2 physical activity monitoring for health care applications," 2012. Accessed: 2024-12-20.

[78] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," in *Proceedings of the 4th International Conference on Mobile Computing and Ubiquitous Networking (ICMU)*, pp. 1–8, 2011.

[79] M. A. Hamade, W. R. H. Kanaan, G. N. Yannakakis, and E. A. Petropoulos, "Antitracker: A dataset for human activity recognition using wearable sensors," 2016. Accessed: 2024-12-20.

[80] E. J. G. Real, E. R. D. Rojas, P. A. G. Canueto, and D. P. F. S. Gonzalez, "Opportunity: A new dataset for human activity recognition using wearable sensors," in *Proceedings of the 2014 International Joint Conference on Neural Networks (IJCNN)*, pp. 3889–3896, 2014.

[81] F. S., E. R., H. N., and R. M., "Stisen: A dataset for sensor-based temporal inertial signals for event recognition," in *Proceedings of the 2017 International Conference on Artificial Intelligence and Data Science*, pp. 235–240, IEEE, 2017.

[82] F. J. Reyes, J. A. D. L. Rosa, and H. R. H., "Real-world human activity recognition dataset," in *Proceedings of the 2016 International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous)*, pp. 1–8, ACM, 2016.

[83] M. U. Al, M. Javed, I. Razzak, and M. Guizani, "Dha: A new dataset for daily human activity recognition using wearable sensors," in *Proceedings of the 2013 International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, pp. 266–271, IEEE, 2013.

[84] R. Abdel-Salam, R. Mostafa, and M. Hadhood, "Human activity recognition using wearable sensors: Review, challenges, evaluation benchmark," 2021.

[85] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Communications Surveys amp; Tutorials*, vol. 15, no. 3, p. 1192–1209, 2013.

[86] M. Stikic, D. Larlus, and B. Schiele, "Multi-graph based semi-supervised learning for activity recognition," in *2009 International Symposium on Wearable Computers*, IEEE, Sept. 2009.

[87] M. Stikic, D. Larlus, S. Ebert, and B. Schiele, "Weakly supervised recognition of daily life activities with wearable sensors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, p. 2521–2537, Dec. 2011.

[88] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, (New York, NY, USA), p. 92–100, Association for Computing Machinery, 1998.

[89] A. Ali, R. C. King, and G.-Z. Yang, "Semi-supervised segmentation for activity recognition with multiple eigenspaces," in *2008 5th International Summer School and Symposium on Medical Devices and Biosensors*, pp. 314–317, 2008.

[90] D. Guan, W. Yuan, Y.-K. Lee, A. Gavrilov, and S. Lee, "Activity recognition based on semi-supervised learning," in *13th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA 2007)*, pp. 469–475, 2007.

[91] T. Huynh and B. Schiele, "Towards less supervision in activity recognition from wearable sensors," in *2006 10th IEEE International Symposium on Wearable Computers*, pp. 3–10, 2006.

[92] G. Deuschl, P. Bain, and M. Brin, "Consensus statement of the movement disorder society on tremor," *Movement Disorders*, vol. 13, pp. 2–23, 10 2008.

[93] Fitbit, Inc., "Fitbit sense." `https://www.fitbit.com/global/us/products/smartwatches/sense`, 2024. Accessed: 2024-06-26.

[94] A. Reiss, "Pamap2 physical activity monitoring," 2012.

[95] D. Roggen, A. Calatroni, L.-V. Nguyen-Dinh, R. Chavarriaga, and H. Sagha, "Opportunity activity recognition," 2012.

[96] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Computing Surveys*, vol. 46, pp. 1–33, 1 2014.

[97] M. G. Martin, *Contributions to Human Motion Modeling and Recognition Using Non-Intrusive Wearable Sensors*. PhD thesis, Universidad Politécnica de Madrid, 2022.

[98] J. Gallego, E. Rocon, J. O. Roa, J. Moreno, and J. L. Pons, "Real-time estimation of pathological tremor parameters from gyroscope data," *Sensors*, vol. 10, pp. 2129–2149, 3 2010.

[99] E. Inada, I. Saitoh, Y. Yu, D. Tomiyama, D. Murakami, Y. Takemoto, K. Morizono, T. Iwasaki, Y. Iwase, and Y. Yamasaki, "Quantitative evaluation of toothbrush and arm-joint motion during tooth brushing," *Clinical Oral Investigations*, vol. 19, pp. 1451–1462, 7 2015.

[100] A. Ruiz-Vitte, E. Carbone, B. Larraga, E. Rocon, and Álvaro Gutiérrez, "The importance of integral time length windows for the classification of activities of

daily living based on machine learning techniques," in *Proceedings of the XLI Congreso Anual de la Sociedad Española de Ingeniería Biomédica*, (Cartagena), 11 2023.

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

[102] D. Kavya, "Medium," 2 2023.

[103] K. Fawagreh, M. M. Gaber, and E. Elyan, "Random forests: from early developments to recent advancements," *Systems Science & Control Engineering*, vol. 2, pp. 602–609, 12 2014.

[104] W. S. Noble, "What is a support vector machine?," *Nature Biotechnology*, vol. 24, pp. 1565–1567, 12 2006.

[105] T. Chen and C. Guestrin, "Xgboost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, ACM, 8 2016.

[106] F. Chollet, "Keras," *GitHub*, 2015.

Appendix A: Legal and Ethical Assessment

This research project explores the classification of ADLs using wearable inertial sensors to enhance tremor assessment in patients with PD and ET. Given the involvement of biometric data collected through accelerometers and gyroscopes, compliance with data protection and AI regulations is crucial. Although the system is developed for research purposes and is not currently subject to commercial regulatory frameworks, a legal and ethical assessment is conducted in this appendix.

Legal Classification and Risk Categorisation

The AI system in this project could be classified as a high-risk system under the European Union AI Act if commercialized, due to its potential impact on health and well-being. High-risk AI systems are subject to strict regulatory obligations, including algorithmic transparency, comprehensive risk management, data traceability, and technical documentation. Additionally, under the EU Medical Device Regulation (MDR), the system could be considered a medical device, requiring certification and post-market surveillance.

The classification of this system also depends on its intended use. If employed solely as a research tool without clinical decision-making implications, it could be considered a limited-risk AI system, with transparency obligations but fewer regulatory constraints. In any case, compliance with relevant legal frameworks would be essential to ensure ethical deployment and avoid potential liabilities.

Data Protection and Privacy Compliance

Since the system relies on personal and biometric data, adherence to the General Data Protection Regulation (GDPR) is paramount. Before data collection, explicit informed consent was obtained from all participants. They were clearly informed about the nature of the collected data, its purpose, and their rights under GDPR, including access, rectification, erasure, restriction, objection, and portability.

If commercialized, additional privacy measures would be required. These would include implementing robust anonymization or pseudonymization techniques, defining security protocols for data breaches, and appointing a Data Protection Officer (DPO). A Data Protection Impact Assessment (DPIA) should also be conducted to evaluate potential risks and mitigation strategies.

Bias, Fairness, and System Limitations

The dataset used in this study originates from real-world inertial sensor recordings, along with a small complementary dataset. To mitigate biases, an analysis of the dataset distribution was performed before model development. However, certain demographic groups may be underrepresented, potentially impacting model generalizability. For example, there is a significant bias toward younger and healthier participants compared to elderly tremor patients. Additionally, there is a gender imbalance, with more male than female participants. These underrepresented groups might exhibit different movement patterns that are not adequately captured in the dataset.

Beyond population biases, technical limitations should be considered. The models are trained using specific movement patterns, which may not generalize to unobserved activities. Additionally, errors in classification could lead to misinterpretations in tremor severity assessment. If the system were integrated into healthcare workflows, human oversight would be necessary to prevent automation bias and ensure that clinical decisions remain under human supervision.

Trustworthiness and Compliance with Ethical AI Guidelines

To ensure a reliable AI system, this project aligns with the European Commission's ALTAI (Assessment List for Trustworthy Artificial Intelligence) framework. The system adheres to seven key principles: human oversight, ensuring that clinical professionals remain in control of decision-making; technical robustness, with model validation to minimise errors and ensure reliability; privacy protection, safeguarding personal data integrity in compliance with GDPR; transparency, documenting and disclosing the system's functionality and limitations; fairness, conducting bias analysis to ensure equitable model performance; social and environmental impact, evaluating the potential consequences of the system on healthcare; and accountability, maintaining records of system behaviour for auditability and compliance.

Final Considerations

Overall, this research project integrates legal and ethical principles to foster responsible AI development in healthcare. While the system is currently designed for research purposes, potential commercialization would require compliance with strict legal standards, particularly under the EU AI Act, GDPR, and MDR. Addressing these considerations ensures not only regulatory compliance but also the development of a transparent, reliable, and ethically responsible AI-driven tremor assessment system.