ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Clustering Upper Limb Impairment in Stroke Patients Using Data from Inertial Measurement Units

Bachelor Thesis

J. Moser

January 19, 2025

Supervisor: Prof. Dr. Julia E. Vogt
Advisors: A. Ryser, J. Pohl, C. A. Easthope
Department of Computer Science, ETH Zürich

Abstract

A stroke often results in one-sided paresis, leading to persistent loss of function or muscle weakness. The ability of stroke survivors to use their upper limbs significantly affects their independence and quality of life. Patients seek out rehabilitation services to improve the experienced impairment in daily life. However, current clinical assessments of upper limb function, performed in a standardized setting, fail to capture the complexities of daily life. Therefore, these assessments often do not accurately reflect upper limb performance in everyday activities. Wearable movement sensors offer precise measurements of daily activity, providing opportunities for evaluations of upper limb performance. Yet, the practical implementation of these sensors remains challenging, possibly due to the complexity in accessing the information from them. This challenge forms the central motivation for this thesis.

In this work, we aim to derive a simplified categorization of real-life upper limb performance in stroke patients based on wrist sensor data. To achieve this objective, we propose two approaches: the first approach is based on performance features derived from sensor data, and the second approach learns suitable features directly from the raw time series using deep learning. The first approach aligns with existing methods, using five established upper limb performance features extracted from the sensors. These features capture various aspects of performance through different arm use patterns in one or both limbs. We then apply dimensionality reduction on this feature set to compute representations of upper limb performance. This is followed by clustering of the representations to arrive at our final performance assignments. However, selecting a suitable set of features is challenging, and the features themselves only capture simple relationships in the data. This motivates the use of the second approach, in which we directly learn relevant features from the raw sensor time series. We achieve this using a contrastive learning approach on slices of the time series. The embeddings learned by the model are then clustered to arrive at a classification of upper limb performance.

The results demonstrate that the feature-based approach provides a valid categorization of upper limb performance into low, medium, and high performance groups. The resulting clusters distinguish these groups across various aspects of upper limb performance. The deep learning approach, on the other hand, successfully captures upper limb impairment and even more complex patterns in the data. However, the final clustering solution does not achieve the desired quality, indicating the need for further refinement to establish the clinical applicability of this approach.

The proposed categories offer a practical framework for clinicians to evaluate therapy outcomes and monitor improvements in patients' functional abilities in daily life. This paves the way for a more efficient and patient-centered rehabilitation practice.

Contents

Chapter 1

Introduction

1.1 Introduction

Globally, stroke ranks as a primary cause of physical disability [1]. Affected people often suffer from one-sided paresis, which leads to impaired motor function and muscle weakness. Impairment in upper limbs (UL) pose a greater issue, as they are crucial for performing daily tasks. Physical damage also affects their mobility and social well-being [2]. Therefore, the degree to which stroke survivors regain use of their upper limbs is essential to improve their independence and quality of life.

To increase their functioning in daily living, stroke patients seek out rehabilitation programs [3]. These programs aim to improve motor function and promote independence in activities of daily living [4]. However, improvements in clinical assessments do not always translate into improvements in impairment experienced in daily living [5].

These types of impairment are distinguished in the World Health Organization's International Classification of Functioning, Disability, and Health model [6]. There, the concept of *performance*, representing what an individual does in their daily, unstructured environment, differs from *capacity*, which refers to their potential ability to perform tasks in controlled settings, such as clinical or laboratory environments.

Clinicians typically use standardized clinical assessments, such as the Fugl-Meyer Assessment, to quantify UL capacity in stroke patients [7]. Although these assessments are valuable in structured settings, they often fail to capture the complexities of movement in unstructured, real-world environments. This leads to a gap between measured capacity and actual performance impairment experienced in daily life by patients [8]. To capture activity performance, clinicians need to be able to observe patients in natural environments, either through measurement technologies or in-person observations [9].

The use of wearable sensors, particularly Inertial Measurement Units (IMUs), to track human UL movement offers the opportunity to bridge the gap between capacity and performance. These sensors enable efficient measurements of UL activity in daily life, which allow the evaluation and description of continuous activity profiles of patients [10]. The metrics derived from these measurements can accurately reflect UL performance in daily activities, making them essential for evaluating the real-life effectiveness of rehabilitation programs [9]. While these sensors have shown to be a promising tool in research, they have not yet found use in practice. This leads us to the central problem addressed in this work.

1.2 Problem Statement

A wide range of performance features can be calculated from IMU data. These features represent different aspects of UL performance, such as the duration and intensity of movement during daily activities [9]. While these features provide valuable insights, their extensive number and complexity make it difficult for clinicians to interpret and incorporate them into therapy [11]. As a result, it remains difficult to effectively assess and track changes in UL performance of stroke patients.

This is the central challenge addressed in this work. Specifically, we want **to develop a simplified and clinically accessible classification system, derived from IMU data, that effectively represents UL performance in daily life.**

1.3 Motivation

Leveraging IMUs offers a unique opportunity to capture UL performance in real-world settings, providing insights that structured clinical evaluations cannot. Simplifying access to the data is essential for enabling clinicians to incorporate these insights into routine rehabilitation practices. By providing an accessible classification of UL performance impairment based on sensor data, clinicians can improve their understanding of therapy outcomes, personalize treatment plans, and improve overall rehabilitation success.

This thesis seeks to provide a practical way for clinicians to access IMU data by developing a clinically accessible UL performance classification system. By addressing this need, we hope to encourage the broader adoption of IMUs in clinical practice, ultimately improving the efficiency and personalization of stroke recovery therapies.

1.4 Objectives

This work addresses the challenge of simplifying the interpretation of IMU measurements to better integrate them into clinical rehabilitation practices. To achieve this, we present two approaches. In the first approach, we build on a study from Barth et al. [12]. In their work, they suggested the formation of distinct UL performance categories based on prioritized performance features, calculated using sensor data.

Objective 1: To validate and reproduce the approach of Barth et al. by deriving a classification based on performance features extracted from IMU data.

However, this classification approach has two main limitations. First, it relies on the manual selection of a suitable set of performance features. This process is inherently challenging and may not capture all relevant aspects of UL performance. Second, the performance features themselves may lack a full representation of UL performance, as they only capture simple relationships in the data.

To address this limitation, we explore the use of unsupervised deep learning methods in our second approach. Specifically, we investigate whether contrastive learning can be applied directly to raw sensor data to derive meaningful representations of UL performance. This approach has the potential to capture more complex patterns and thus qualitatively expand the representation of impairments in daily living in stroke patients.

Objective 2: Design a classification system leveraging a deep learning architecture trained using contrastive learning techniques on raw IMU data.

Finally, the practical value of the classifications from both approaches will be evaluated by analyzing their internal structures and comparing them to clinical assessments of UL capacity. This comparison will investigate whether data-driven classifications better capture real-world performance, providing information on their potential to improve rehabilitation practices.

Objective 3: Evaluate the ability of data-driven classifications to capture real-world UL performance by analyzing their internal structures and comparing them to clinical assessments.

1.5 Thesis Structure

The following section provides an overview of the structure of this thesis.

- **Chapter 2: Related Work and Background:** This chapter reviews current clinical assessment methods, explores the use of Inertial Measurement Units (IMUs) in stroke rehabilitation, reviews related work in

the field, and discusses the potential of contrastive learning methods. In addition, it introduces the foundational techniques upon which our approaches are based.

- **Chapter 3: Data:** This chapter describes the data collection process, including the characteristics of the patient cohort and the demographic, clinical, and sensor data utilized in the study. It also explains the derivation of a clustering solution based on clinical assessments, which serves as a reference for evaluating data-driven clustering approaches.
- **Chapter 4: Methods:** This chapter outlines the methodology developed to create the final UL performance assessment based on sensor data. It provides a detailed explanation of the steps involved in the proposed approaches and describes the evaluation methods used to assess their performance.
- **Chapter 5: Experiments:** This chapter presents the experimental results of the two clustering approaches. It discusses the challenges encountered, the insights gained, and evaluates the alignment of the clustering solutions with clinical assessments.
- **Chapter 6: Conclusion:** The final chapter revisits the initial objectives and the problem statement to assess whether the findings have successfully addressed the research goals. It summarizes key insights, critically evaluates the limitations of the work, and provides recommendations for future research.

Chapter 2

Related Work and Background

In this chapter, we review related work and provide background that forms the basis for this study. First, we provide an overview of the Fugl-Meyer Assessment, a clinical capacity assessment commonly used in stroke rehabilitation, and discuss its strengths and limitations to accurately capture patient capacity and performance. Next, we explore the role of Inertial Measurement Units (IMUs) in representing patient performance and how they address some of the shortcomings of traditional clinical assessments. Then we will explain the two main studies upon which our work is based in detail, as well as the adaptations needed to fit it to our context. Finally, we provide a glossary of the machine learning techniques applied in our clustering approaches.

2.1 The Fugl-Meyer Assessment

Clinical assessments are used to assess UL capacity in stroke patients. These assessments provide standardized and quantifiable metrics to guide rehabilitation strategies and measure progress. The Fugl-Meyer Assessment (FMA) has become a gold standard for assessing capacity [7] [13]. In this work, the FMA serves as the clinical benchmark for comparing our data-driven approaches with clinical evaluations.

The FMA is one of the most widely used clinical tools to evaluate motor recovery and functional capacity in stroke patients. It is designed to assess impairments in motor function, sensory ability, balance, joint range of motion, and coordination [14]. The FMA is used frequently in both research and clinical practice.

Scoring System The assessment scores UL capacity on a scale ranging from 0 to 66. Higher scores indicate better motor capacity. The scoring is based on the patient's ability to perform specific tasks, such as reflex activity, voluntary movements, and coordination. Each task is graded on a three-point scale (0 =

cannot perform, 1 = partially performs, 2 = performs fully). This structured approach allows clinicians to quantify motor impairments and track recovery progress over time [15].

Strengths The FMA is widely regarded as the gold standard for assessing post-stroke motor impairments due to its strong psychometric properties, including high reliability and validity [16]. Its standardized scoring system ensures consistency across studies and clinical settings.

Limitations Despite its strengths, the FMA has notable limitations. It primarily assesses motor capacity in controlled clinical settings and does not account for how patient performance is affected in real-world settings [5]. In addition, FMA requires the administration by trained professionals, which can be time and resource intensive in clinical practice.

The complexity involved in capturing FMA scores can also be seen in other clinical assessments. This led to a trend towards simplifying the assessments to make them more accessible and less resource intensive [17]. Although this is not directly related to our objective of deriving a simplified representation of UL performance, it reflects a general effort to streamline assessments in clinical practice.

As mentioned earlier, traditional clinical assessments often do not accurately represent daily life performance. This limitation has pushed the use of Inertial Measurement Units (IMUs), in stroke rehabilitation research. The following section highlights related work that uses IMUs to assess and monitor UL performance.

2.2 IMUs in Stroke Rehabilitation

IMUs have become an increasingly popular tool due to their ability to provide objective, continuous, and real-world measures of movement. This section highlights two key approaches to analyzing IMU data, traditional feature-based methods and more advanced machine learning techniques, and discusses their applications and challenges.

Feature-Based Several studies have shown that IMU-derived features can provide a more accurate representation of UL performance compared to traditional clinical assessments. These feature-based approaches rely on carefully designed metrics, such as movement intensity, symmetry, and duration, to quantify UL performance. For example, Uswatte et al. [10] validated the use of IMU-based activity monitoring to assess upper limb use in stroke survivors, demonstrating that these measures offer an objective and real-world index of daily arm activity. Similarly, De Lucena et al. [18]

further showed that these kinematic measures encode novel information about patient UL performance, not captured by classical clinical assessments.

Machine Learning Recent advances in machine learning have opened new possibilities for analyzing IMU data. They enable the discovery of complex patterns and relationships that traditional feature-based methods may overlook. Boukhennoufa et al. [19] provide a comprehensive review of machine learning approaches applied to wearable sensor data in stroke rehabilitation, highlighting their potential to improve prediction accuracy and automate analysis. For example, Werner et al. [20] have focused on translating IMU data into clinically interpretable scores. They used IMU data to predict movement quality scores typically derived from clinical assessments, allowing less resource intensive clinical procedures.

Related to our work is the study by Felius et al. [21], which investigated unsupervised feature learning using variational autoencoders. While their approach differs in its specific objectives, it shares a similar goal of deriving meaningful representations of movement without relying on predefined labels or features.

However, as already mentioned, despite their potential, IMUs have not found much use in practice [11]. This encouraged the simplification of the IMU data, to make it more accessible to clinicians. That is the objective of the reference work upon which our first approach is based.

2.3 Feature-Based categorization by Barth et al.

Barth et al. proposed a categorization framework that serves as a simplified assessment of UL performance in daily life [12]. This section provides an overview of their approach. Furthermore, the differences between their methodology and the approach proposed in our study are discussed.

The Approach by Barth et al.

The primary objective of Barth et al. was to identify UL performance categories using sensor data from stroke-affected individuals and neurologically intact adult controls. Data was collected using bilateral wrist-worn accelerometers that provided continuous recordings of movement data. Using the data, they calculated UL performance features that served as proxies for different aspects of performance, such as duration, magnitude, or symmetry.

To form categories, they applied a Principal Component Analysis (PCA) to extract condensed embeddings of the calculated features. Subsequently, the embeddings were clustered using the K-Means algorithm to group the participants into performance categories.

Barth et al. evaluated a total of 12 clustering solutions, varying the number of clusters (3, 4, or 5) and the number of input features (12, 9, 7, or 5). They evaluated these solutions using quantitative metrics, considering both the explained variance and the simplicity of the feature calculation process. Their final proposed solution was a 5-cluster model using five selected UL performance features. The model was chosen because it balances interpretability, explained variance, and ease of implementation.

2.3.1 Comparison to our work

Our first approach aligns in large part with Barth et al. We apply PCA to a suitable feature set to calculate the performance representations and then group them using K-Means. In the following, we will provide an overview of the differences.

Differences

Several key differences distinguish our approach from the methodology of Barth et al.

- **Participant Population:** Our study exclusively involves participants with stroke, whereas Barth et al. included both individuals with stroke and neurologically intact adult controls. Furthermore, we use multiple measurements for each participant on different days to capture variations in rehabilitation states and environments, while Barth et al. used a single measurement per participant.
- **Feature Selection:** We started with a larger set of features than Barth et al. and therefore had to adjust our approach to find a suitable set of performance features. To identify such a set, we use the RelieF algorithm and evaluate the explained variance and the principal component weights derived from PCA. In contrast, Barth et al. filtered out features based on the difficulty of their calculation and their interpretability in a clinical context.
- **Number of Clusters:** While Barth et al. explored 3-, 4-, and 5-cluster solutions, we exclusively focus on a 3-cluster solution. This decision is motivated by the need for simplicity and clinical interpretability. Furthermore, preliminary analyzes indicated that increasing the number of groups does not provide significant benefits in capturing meaningful performance distinctions.
- **Clustering Evaluation:** We expand the evaluation of the clustering approach beyond the internal structure analysis by incorporating comparisons with clinical assessments of capacity. This additional validation emphasizes the potential advantages of our approach over traditional assessment methods.

2.3.2 Limitations

The work of Barth et al. provides a strong foundation for sensor-based categorization of UL performance, and our approach extends this foundation by incorporating additional dimensions of analysis. However, using features as the basis for this approach comes with inherent limitations.

Choosing a suitable set of features from the wide variety of established features is challenging. In addition, the selected features typically only capture simple relationships in the activity patterns of the patients but fail to represent the more complex temporal dynamics inherent in real-world movement data. To address these limitations, we propose to leverage deep learning techniques directly on the sensor data. In particular, we explore contrastive learning methods to learn representations of UL performance.

2.4 The SimCLR Framework

This section introduces the foundational work on contrastive learning on which our second approach is based. Specifically, we build on the SimCLR framework proposed by Chen et al. [22]. This framework has demonstrated remarkable success in unsupervised representation learning. In the following, we provide an overview of the SimCLR framework and describe the adjustments we made to adapt it to the context of this work.

2.4.1 Key Components

The primary objective of SimCLR is to learn meaningful measurement representations from unlabeled data using contrastive learning. Contrastive learning is implemented by defining positive and negative pairs of data samples. The model then maximizes similarity between representations of positive pairs, while minimizing it with negative samples using a contrastive loss function. In their work, Chen et al. defined positive pairs as different augmented views of the same data sample, while negative pairs consist of all other samples in the batch. The framework consists of the following four key components:

- **Data Augmentation:** Two random augmentations are applied on the same sample to create two distinct augmented views.
- **Neural Network Encoder:** A base encoder maps the augmented views to a representation space. These representations can be used for downstream tasks, such as clustering and classification.
- **Projection Head**: A secondary encoder maps the representations to a latent space where the loss function is applied. This step is removed for downstream tasks.

- **Contrastive Loss:** A specialized cross-entropy loss, commonly referred to as NT-Xent (Normalized Temperature-Scaled Cross-Entropy Loss), is applied to the latent representations. This loss encourages high similarity between positive pairs while separating negative pairs.

With this architecture, the framework achieves state-of-the-art results in unsupervised learning [23].

2.4.2 Adaptions to our context

Our objective with this framework is to learn meaningful representations of UL performance. Since the SimCLR framework was originally designed for computer vision tasks, several adaptions are needed to fit it to our context. Specifically, we adjust the data augmentations to suit the nature of time series data and the backbone encoder to process the data effectively. In addition, we adjust the sampling strategy to fit to the preprocessed data that provides the foundation to train our model.

The augmentations, the backbone model and the sampling strategy are specified in Section 4.2.2.

The next section includes a glossary of the machine learning techniques referenced in this work. This is followed by an overview of the dataset, presented in the next chapter.

2.5 Machine Learning Glossary

RelieF Algorithm is a feature selection method commonly used in machine learning to identify relevant features. It estimates their importance based on their ability to distinguish between data instances. The algorithm operates by iteratively sampling instances and evaluating the contribution of each feature to the difference between the nearest instances of the same and different classes (referred to as nearest hits and nearest misses). RelieF is particularly effective in handling noisy data and capturing feature interactions, making it a valuable tool in domains such as healthcare, where datasets often exhibit high dimensionality and complexity [24].

Principal Component Analysis (PCA) is a widely used, non-parametric method for dimensionality reduction that identifies simplified structures underlying complex datasets. By projecting data onto a lower-dimensional space, PCA captures the directions of maximum variance. This method helps uncover hidden patterns and reduce redundancy in high-dimensional data. This technique is particularly suited for preprocessing in clustering and machine learning tasks. It is also a valuable tool for visualizing higher-dimensional data [25].

K-Means Clustering is one of the most widely used unsupervised learning algorithms to partition data into k distinct clusters based on feature similarity. The algorithm operates iteratively by assigning each data point to the cluster with the nearest centroid and then recalculating the centroids as the mean of the points in each cluster. The limitations of K-Means lie in the sensitivity to the initial placements of the centroids, the choice of a suitable number of clusters k and handling with non-spherical or overlapping clusters [26].

Multilayer Perceptrons (MLPs) are a class of feedforward neural networks composed of an input layer, one or more hidden layers, and an output layer. Each layer consists of interconnected neurons (nodes) that use nonlinear activation functions to model complex patterns in the data. MLPs learn by minimizing a loss function through backpropagation and gradient descent. They are widely used because of their ability to approximate any continuous function. However, they require careful tuning of hyperparameters and can be prone to overfitting [27].

Long Short-Term Memory (LSTM) networks are a specialized type of recurrent neural network (RNN) designed to model sequential and time series data. LSTMs incorporate a memory cell and a set of gating mechanisms that regulate the flow of information and enable the network to retain long-term dependencies. This makes LSTMs particularly effective for tasks involving complex temporal patterns, which is a key advantage that we try to leverage to enhance the currently established features of IMU data, which only capture simple relationships [28].

Normalized Temperature-Scaled Cross-Entropy Loss (NT-Xent) is a contrastive loss function used in self-supervised learning to encourage similar embeddings for positive pairs while pushing apart embeddings for negative pairs. It uses cosine similarity to measure the closeness of the embeddings and incorporates a temperature parameter τ to scale the similarity scores.

For a positive pair of embeddings (i, j), the loss is defined as:

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)},$$

where $\text{sim}(u, v)$ is the cosine similarity of ℓ_2-normalized vectors, and the denominator sums over all other samples in the batch. NT-Xent is widely used in contrastive learning methods such as SimCLR [22], enabling models to learn meaningful representations without labeled data.

Chapter 3

Data

This chapter describes the data collection process, including details on the patients recruited for the study, the demographic and clinical data collected, and the IMUs used to collect the UL movement data. Data collection spanned multiple time points post-stroke to capture UL functionality in different environments and rehabilitation states. At the end of the chapter, we provide information on how we derive a clinical clustering, which is used to compare our data-driven approaches.

3.1 Patients

Clinical and movement-related information was collected from **93 stroke patients** as part of a prospective observational study conducted at the University Hospital Zurich. All patients provided written informed consent.

Inclusion Criteria

Participants were eligible for the study if they met the following criteria:

- Diagnosed with their **first stroke** (ischemic or hemorrhagic).
- **18 years or older**.
- **Motor impairment** affecting the upper or lower limb.
- **No history of neurological or other diagnoses** (e.g. Parkinson's disease, multiple sclerosis) that could have affected physical activity.

Exclusion Criteria

Participants were excluded from the study based on the following criteria:

- **Ethical contraindications** (e.g. vulnerable individuals).

- **Known substance abuse** (e.g. alcohol or drug dependency).
- **Non-compliance** with study procedures.

3.2 Demographic and Clinical Data

To contextualize the results, demographic data and clinical assessments were collected. Demographic data provides essential background information on the patient cohort, while the clinical assessments are used to evaluate UL motor capacity and recovery, and serve as a basis for comparison with our data-driven approaches.

Demographic Data

The collected demographic data included age, gender, time since stroke at each measurement point, handedness, and side of the paresis.

Clinical Data

Clinical data was available for approximately 84% of measurements, obtained using the Fugl-Meyer Assessment (FMA) to assess UL motor capacity and recovery (see Section 2.1 for more details on the FMA).

3.3 IMU Data

To quantify real-world UL activity, movement data was obtained from IMUs developed for research purposes by ZurichMOVE[1]. These devices measure motion through accelerometers and gyroscopes. Patients wore IMUs on both the affected and unaffected wrists to capture upper limb movement patterns during daily activities.

Devices and Placement

The IMUs include a 3-axis accelerometer, a 3-axis gyroscope, and other sensors such as a digital compass and an altimeter. However, in this study only accelerometer and gyroscope data are used. Each patient wore one device on the affected wrist and one on the unaffected wrist, secured with elastic straps. Some patients also wore additional sensors on their trunk and ankles. However, we did not use the data from these sensors since the focus of the study is on UL performance.

[1] https://zurichmove.com/

Measurement setup

Measurements were collected at five predefined time points post stroke. Specifically, days 3±2 (D3), 10±2 (D10), 28±4 (D28), 90±7 (D90), and 365±14 (D365) after stroke. Sometimes, measurements were performed over up to 3 consecutive days around each time point (for example, day 9-11 instead of only day 9). Assessments took place at the current location of stay, which was, depending on the patient's condition, either in a clinical environment or at home. Patients were instructed to wear the devices during their normal daily activities for the duration of each measurement day to capture representative movement data in natural environments.

Data provided

In this study, we use data from both the accelerometer and gyroscope of the IMU. The accelerometer measures acceleration (m/s^2) in three axes (x, y, z) to capture the intensity of movement. The gyroscope measures angular velocity ($^\circ/s$) in three axes (x, y, z) to capture rotational motion of the wrist. The acceleration and angular velocity data were originally recorded at 50 Hz but were downsampled to 1 Hz for clustering. The reason for this preprocessing step is explained in the Chapter 4. Figure 3.1 shows a 10-minute window of the downsampled data.

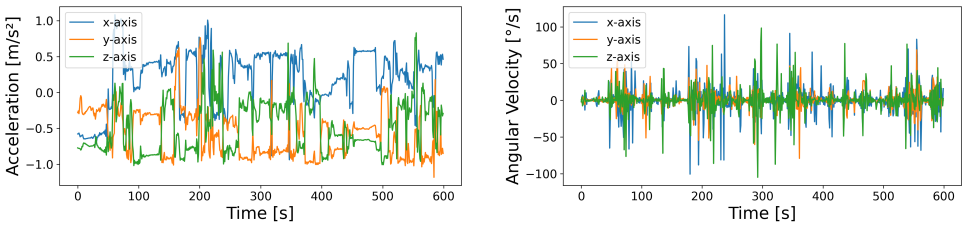

Figure 3.1: Visualization of the IMU sensor data over a 10-minute window: (a) acceleration data along the x, y, and z axes, and (b) angular velocity (gyroscope) data along the x, y, and z axes.

Each full day of IMU data is stored in a separate CSV file and processed offline. Further details on the respective preprocessing steps are provided in Section 4.1.1 and Section 4.2.1.

3.4 Data Summary

The dataset used in this study consists of IMU data from 93 stroke patients across five measurement points, resulting in a total of 340 recorded measurements. The demographic characteristics of the cohort are summarized in Table 3.1. Concordance indicates how many patients have their dominant side as the side affected by the stroke. UL capacity was assessed using the

Fugl-Meyer assessment (FMA). However, FMA scores are not available for all measurements. Specifically, 287 measurements include the corresponding FMA scores. Although this restriction exists, all 340 measurements are used in the derivation of the clustering solutions.

Table 3.1: Patient demographics across the five measurement days. The dataset includes 93 patients, each with one to five measurements. No significant differences in demographics are observed across the measurement days. Differences between measurement points are primarily attributed to the rehabilitation state and environment of the patients (e.g. clinical environment early after stroke, and home environment later).

Variable	Day 3	Day 9	Day 28	Day 90	Day 365
Count	87	83	50	81	39
Age Mean, years	68.18	68.05	65.88	67.58	68.95
Age Range, years	33-91	33-91	33-91	33-91	41-91
Sex, female	42.5%	42.2%	40.0%	45.7%	41.0%
Handedness, right	83.9%	83.1%	82.0%	86.4%	84.6%
Concordance	51.7%	49.4%	54.0%	53.1%	51.3%
Affected Side, right	47.1%	44.6%	48.0%	46.9%	41.0%

3.5 Clinical Clustering

To evaluate the data-driven clustering solutions, we analyze both their internal structures and their alignment with established clinical assessments. For this purpose, we derive a clinical comparison clustering based on FMA scores. Hoonhorst et al. have provided and validated a clustering based on FMA scores in five categories [29]. Although useful, this constraint limited their applicability for comparing it to alternative cluster solutions with a different number of groups.

To overcome this limitation and allow for greater flexibility, we derive a clinical cluster by applying the K-Means algorithm directly to the FMA scores. This approach enables us to generate a 3-cluster solution while still grounding the clustering on clinically meaningful data. The resulting comparison clustering provides a reference against which the data-driven clustering solutions are evaluated, offering a study of their agreement and differences.

The following chapter will detail the methodologies for both approaches, feature-based and using deep learning, and describe their implementation.

Chapter 4

Methods

The purpose of this chapter is to describe the methods used to group stroke patients according to UL performance using sensor data. In addition, we provide a clear and reproducible explanation of how we compare and evaluate clustering strategies, while highlighting the rationale behind each methodological choice. We explore two main methodologies: clustering based on performance features derived from IMU data and clustering based on learned embeddings from raw sensor data using deep learning techniques.

Both approaches share similar steps at a high level. First, we preprocess the data to prepare them for the following steps. Next, we pass the preprocessed data through an embedding pipeline designed to learn and extract suitable representations of UL performance. We then apply a clustering algorithm to the performance representations to compute accessible categories of UL performance. Finally, we evaluate the internal clustering structure and compare the clustering with clinical assessments. Fig. 4.1 visualizes these steps required to arrive at the final clustering.

We divide this chapter into two main sections, corresponding to the two clustering approaches. For each approach, we specify the preprocessing and clustering pipeline. At the end of the chapter, we describe the evaluation protocol used to assess the quality of the data-driven solutions.

4.1 Feature-Based Clustering

This section provides a detailed explanation of the first approach, in which we group patients based on features derived from IMU data. First, we describe how we calculate features from the raw data and outline the methods we use to construct a meaningful feature set that captures different qualities of UL performance. Following the approach of Barth et al., we apply PCA to

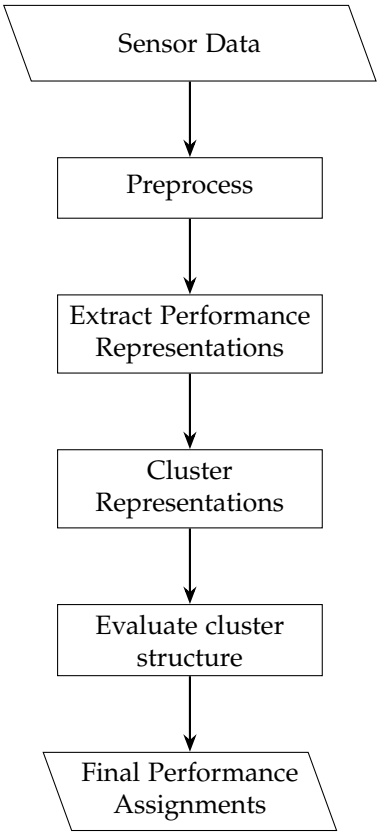

Figure 4.1: High-level overview of the two proposed approaches for clustering upper limb performance: the feature-based approach and the deep learning-based approach. The process begins with raw sensor data and involves five key steps: preprocessing (Sections 4.1.1 and 4.2.1), extracting performance representations (Sections 4.1.3 and 4.2.2), clustering the representations (Sections 4.1.4 and 4.2.4) and evaluating the cluster structure (Sections 4.3).

this feature set to compute representations of UL performance. Finally, we present the clustering algorithm that we use to identify patient groups.

4.1.1 Preprocessing

As described in Chapter 3, the sensor data from the IMUs consist of acceleration and gyroscope measurements along the three axes. In the literature, features derived from IMU data often prioritize acceleration measurements [9]. In line with the literature, we also focus exclusively on features based on acceleration data.

We derive two primary metrics from acceleration data: **activity counts** and **jerk**. We calculate activity counts using validated open source code provided by Brønd et al. [30]. Briefly, we apply a bandpass filter to the raw acceleration

data (0.25–2.5 Hz), downsample it to 10 Hz, and convert it to activity counts per 1-second epoch by summing the acceleration values within each epoch.

We compute jerk using a custom script. Jerk represents the rate of change in acceleration over time and is calculated as the norm of the derivative of the three acceleration vectors.

We process the derived metrics, activity counts and jerk, using custom software to compute UL performance features. These features measure various aspects of UL performance in everyday activities, including **duration**, **magnitude**, **variability**, **symmetry**, and **quality of movement** for either one or both upper limbs.

To account for differences in measurement scales (e.g. hours, counts, and ratios), we standardize the dataset of UL performance features using z-scores.

4.1.2 Feature Selection

Feature selection plays a critical role in ensuring that clustering produces meaningful results. Our objectives for feature selection are to derive a feature set that captures a diverse range of UL performance qualities, shows high explanatory power for the final clustering solution, and minimizes redundancy in the information provided.

The initial feature set consists of a large set of features that cover a wide range of qualities. For our baseline, we use the feature set proposed by Barth et al. We refine this feature set iteratively to better meet the stated objectives. First, we qualitatively evaluate the feature set proposed by Barth et al. to identify potential gaps in capturing key UL performance qualities or the presence of redundant information. We prioritize features that add to a balanced representation of UL performance qualities to ensure a comprehensive characterization of activity.

To quantitatively assess feature relevance for clustering, we apply the RelieF algorithm. This method evaluates the contribution of each feature to the data-driven clustering solution, helping us identify and prioritize the features that are most influential in distinguishing between clusters. This ensures that the clustering process is driven by meaningful and relevant features.

Lastly, we apply Principal Component Analysis (PCA) to the feature set to compute the explained variance and to analyze the feature weights associated with each principal component. We prioritize features that provide a large explained variance and have similar levels of importance in the principal components. This step also allows us to identify and address redundancy within the feature set.

4.1.3 Computing Performance Representations

In line with the approach proposed by Barth et al., we apply PCA to derive performance representations from the feature set. This ensures that only the most important patterns in the data contribute to subsequent clustering.

To obtain an unbiased estimation of the PCA, we ensure that the samples used to calculate the principal components are as uncorrelated as possible and reflect a diverse range of conditions. Specifically, we select one measurement per patient from different measurement days (e.g., Patient 1, Day 3; Patient 2, Day 90, etc.). This approach ensures that the principal components are computed on a balanced subset of the data, providing a consistent and unbiased basis for transforming all measurements.

After deriving the principal components, we use the first two components to compute performance representations for all measurements. These representations encapsulate most of the variance in the feature set and serve as proxies for UL performance.

4.1.4 Clustering

After computing the performance representations, we apply K-Means to group the representations into three clusters. The number of clusters can be specified, so a different number of clusters could also have been investigated. Barth et al. evaluated 3-, 4-, and 5-cluster solutions in their approach. However, we focus on a 3-cluster solution because it proves most promising based on two factors. First, interpretability, as the three clusters intuitively categorize the patients into low, medium, and high performance groups. Second, initial analysis did not show a clear advantage in clustering structure when using four or more clusters.

Ordering Clusters After running the clustering algorithm, we assign each measurement to a cluster. However, K-Means does not inherently assign an order to the clusters. For interpretability and evaluation, we impose an order on the clusters that aligns with the performance levels we aim to measure. To achieve this, we use the Hungarian Matching Algorithm, which solves the assignment problem by finding the optimal one-to-one mapping between two sets while minimizing the total cost of mismatches. In our case, we use this algorithm to align the data-driven clustering solution with the clinical clustering based on UL capacity.

Alternatively, we could order the clusters by ranking them based on the average feature values of each group. However, we choose to base the ordering on the clinical clustering, as we expect the two clustering solutions to largely agree.

4.2 Deep Learning Clustering

Barth et al. have demonstrated that clustering based on established features yield promising results [12]. However, this approach comes with notable limitations. Specifically, selecting appropriate features poses a significant challenge. In addition, features are simple statistical measures, such as means and medians, and therefore provide only a limited representation of overall UL performance. To address these limitations, we propose an alternative methodology that directly learns representations from the raw movement data.

This section provides a detailed explanation of our methodology for learning meaningful representations using deep learning. First, we describe the preprocessing steps required to prepare the data for training a deep learning model in a contrastive fashion. Then, we outline the training approach used to learn embeddings that encode performance characteristics. Finally, we explain how we cluster the learned embeddings to identify distinct patient groups.

4.2.1 Preprocessing

In this subsection, we outline the preprocessing pipeline that transforms raw IMU data into a format suitable for training the model and computing embeddings. Unlike the feature-based approach, this method incorporates both acceleration and gyroscope data from the IMU sensors.

Downsampling

We downsample the raw IMU data, originally sampled at 50 Hz, to 1 Hz. This step allows us to use activity counts as a filtering criterion in subsequent steps and offers efficient processing of longer activity windows. Although downsampling reduces granularity, we find this compromise acceptable because we focus on capturing global activity patterns over longer time ranges.

Relabeling

The raw IMU data are provided as time series originating from the left and right wrists. Because strokes can affect either side, and our objective is to learn side-independent performance representations, we relabel the data to distinguish between the affected and nonaffected sides. For example, if a patient's affected side is the right, we relabel the data from the right wrist as the affected side and the data from the left wrist as the nonaffected side. This relabeling strategy encourages that the model learns representations that reflect the characteristics of the affected and nonaffected sides.

Time Series Slicing

We divide the IMU time series into overlapping slices of 15-minute windows, with a 50% overlap between consecutive slices. This slicing step serves several purposes. First, it increases the effective size of the dataset, which is essential for training deep learning models. Second, it enables a more nuanced definition of positive and negative pairs for the contrastive loss function. Specifically, we define a positive pair as two slices of the same measurement, in contrast to the standard definition of a positive pair as two augmented views of the same slice, as used in vanilla SimCLR [22]. We provide a more detailed explanation of this approach and its underlying intuition in Section 4.2.2. Third, shorter time series improve gradient stability during training [31].

Slicing may disrupt certain long-term dependencies or remove critical information necessary to classify upper limb impairments. However, we design the window length to balance these risks with the advantages. Specifically, we assume that 15-minute windows are long enough to capture the relevant characteristics of upper limb performance while still providing the benefits of slicing. This trade-off between preserving temporal information and optimizing model performance is a key design decision, which we discuss further in Section 5.2.1.

After slicing the data, we filter slices with minimal activity (e.g., the slices recorded while the patients are asleep). This ensures that each slice in the final set of slices contains sufficient activity for the model to learn and capture UL performance. To identify and exclude low-activity slices, we calculate the proportion of active samples within each slice. Here, a sample refers to an individual data point in the downsampled 1 Hz time series. A sample is categorized as active if the activity count for either limb exceeds a threshold of 2, consistent with the activity threshold used in performance feature calculation in the literature [15]. In the final model, we classify a slice as active if at least 10% of its samples, equivalent to 90 seconds, are categorized as active. We determine this threshold by evaluating different percentages during model training and selecting a value that results in an equal distribution across clinical classifications. More details are provided in Section 5.2.1.

4.2.2 Contrastive Learning Approach

With the prepared slices, we now describe the training of a deep learning model to extract meaningful representations. Specifically, we use a contrastive learning approach, which aims to learn representations by bringing positive pairs closer together while pushing negative pairs apart. Our goal is to derive representations that focus on common activity patterns that reflect UL performance. To achieve this, we define slices from the same measurement as

positive pairs, encouraging the model to identify shared patterns of activity throughout the day.

The contrastive learning approach used in this work is based on SimCLR proposed by Chen et al. [22], taking into account the nature of our data and our objective. Figure 4.2 summarizes the four main components of the approach, and in the following sections we will provide detailed explanations for each of them.

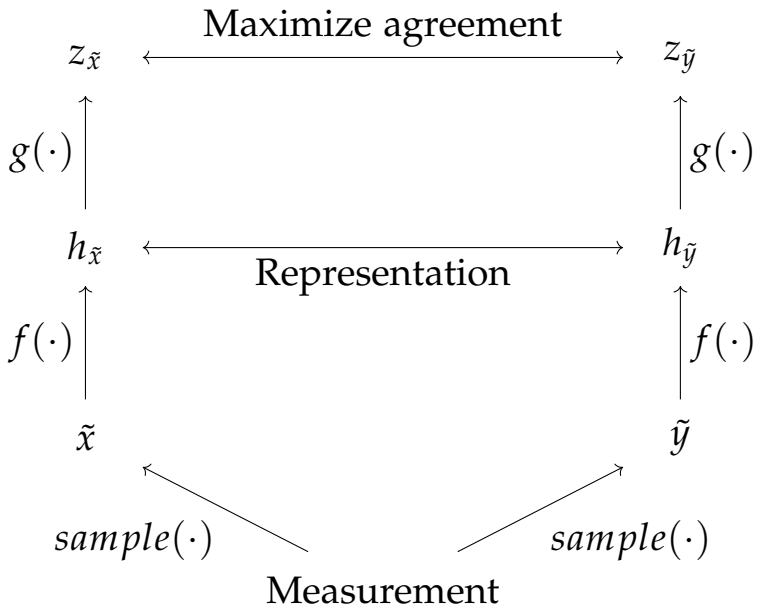

Figure 4.2: Overview of the contrastive learning approach. The process begins with sampling and augmentation, where two 15-minute activity slices are randomly sampled from a measurement and independently augmented to form a positive pair ($\tilde{x}$, $\tilde{y}$). These slices are passed through the backbone model $f(\cdot)$, a bidirectional LSTM, to extract performance representations ($h_{\tilde{x}}$, $h_{\tilde{y}}$). The representations are subsequently mapped to a latent space by the projection head $g(\cdot)$, producing embeddings ($z_{\tilde{x}}$, $z_{\tilde{y}}$). The NT-Xent loss is then computed on the embeddings, encouraging agreement between positive pairs while discouraging similarity with negative pairs. This framework is designed to learn invariant and discriminative UL performance representations for downstream tasks. This visualization is adapted from a figure by Chen et al., with modifications to include the measurement component.

Sample and Augment The first step in the process involves the sampling and augmentation module. The sampling works as follows: for a given measurement, we randomly sample two slices from the set of slices corresponding to that measurement. Each slice is then augmented using independent random augmentations, resulting in a pair of augmented views of the same measurement, denoted as $\tilde{x}$ and $\tilde{y}$. This pair is treated as a positive pair.

The intuition behind this approach is to encourage the model to learn both patient-specific and invariant representations of performance.

The augmentations used in this work include sequentially applying Gaussian noise, masking a continuous segment of the time series, and segmenting the time series into three mini-segments followed by shuffling their order. Augmentation is a commonly used technique for improving robustness in machine learning models, particularly in unsupervised learning, where it helps prevent overfitting and enhances generalizability. Examples of these augmentations are shown in Figure 4.3.

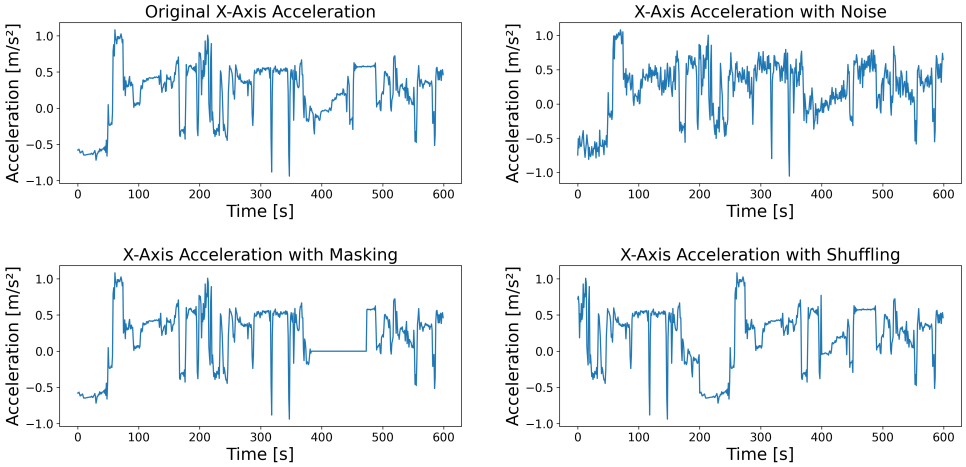

Figure 4.3: Visualization of augmentations applied: (a) Original data, (b) data with added noise, (c) data with a masked segment (set to 0), and (d) data after slicing and shuffling. The plots show only the x-axis acceleration, but these augmentations are applied to all axes of acceleration and angular velocity data during training.

During the computation of the final UL performance representations, all slices are passed through the model without sampling, and no augmentations are applied. This ensures that the embeddings reflect the raw characteristics of the input data and are consistent across slices.

Backbone Model The backbone neural network $f(\cdot)$ forms the core of this approach and is responsible for learning and extracting meaningful performance representations from input slices. In this work, we use a bidirectional Long Short-Term Memory (LSTM) network as the backbone model due to its ability to effectively handle sequential information and capture temporal dependencies. We compute the performance representation $h_{\tilde{x}}$ by concatenating the hidden states from both directions of the LSTM after processing the input slice $\tilde{x}$.

Once the model is trained, we use $f(\cdot)$ to extract performance representations for all unaugmented slices. These representations are subsequently used for clustering, while additional training steps, such as the projection head, are omitted.

Projection Head The projection head, denoted as $g(\cdot)$, maps the extracted performance representations to a latent space where the contrastive loss is computed. In this work, we utilize a small multilayer perceptron (MLP) with a single hidden layer as the projection head. As demonstrated in SimCLR [22], applying the loss to the mapped representations $z_{\tilde{x}}$, rather than directly to the original representations $h_{\tilde{x}}$, leads to improved representation quality.

Contrastive Loss Function The final step involves computing the contrastive loss on the projected embeddings. Specifically, we employ the NT-Xent loss (see Section 2.5), which is the same loss function used in vanilla SimCLR. The objective of this loss is to maximize the similarity between positive pairs while minimizing it for negative pairs within the latent space, thereby encouraging the model to learn discriminative representations.

4.2.3 Training Batch Design

In this section, we describe how we designed training batches to ensure that the backbone model $f(\cdot)$ learns meaningful performance representations.

Each batch contains N distinct measurements. To avoid duplicate measurements in the dataset, we choose N to be no larger than the number of distinct measurements available. This constraint is critical because of the way negative samples are defined. For each batch of N measurements, the sampling process results in a batch of $2N$ augmented slices. We do not explicitly sample negative examples. Instead, for each positive pair, we treat the remaining $2(N-1)$ augmented examples within the batch as negative examples. This approach is in line with the way Chen et al. defined negative samples [22].

It is important to note that while we make sure to not include the same measurement twice in a batch, there can be two distinct measurements from the same patient. This poses a risk of correlation and information leakage that may confuse the model. We judge this to be negligible because, on the one hand, we expect our model to learn representations that group patients with similar impairments together, even when treating them as negative examples. On the other hand, if we would have respected this, our maximum batch size would have been limited even further.

The number of batches can be either explicitly specified or, if unspecified, is determined by dividing the total number of slices in the dataset by the batch size. Due to the stochastic nature of random sampling, it is likely that not

all slices will be included in each epoch. However, with a sufficient number of epochs, this is unlikely to negatively impact the quality of the learned representations.

4.2.4 Extracting Embeddings and Clustering

In this section, we describe the pipeline used to derive the final cluster assignments for each measurement using the trained backbone model $f(\cdot)$ and the set of slices.

To extract the embeddings, all slices are passed through the trained backbone model $f(\cdot)$ without applying augmentations, resulting in a set of performance representations. These representations serve as the input for clustering, where we apply the K-Means algorithm to group the slices into clusters. We fix the number of clusters to three, as justified in Section 4.1.4.

After applying K-Means, each slice is assigned to one of the three clusters. However, since the goal is to derive a single cluster assignment per measurement, we have to aggregate the cluster assignments of all slices corresponding to a given measurement. The final cluster for a measurement is determined as the most frequent cluster assignment among its slices. To establish an order for the clusters, we apply the Hungarian Matching Algorithm, matching the data-driven clustering with the clinical clustering. This aligns with the method described in Section 4.1.4.

The decision to first cluster individual slice embeddings and then aggregate their assignments is intentional. By clustering on all slice representations, instead of a single representative embedding per measurement, we aim to preserve the diversity of performance patterns within a measurement. This approach allows for a broader assessment of variability in patient performance, which could provide valuable insight into their recovery state. For example, if a measurement does not exhibit a clear majority cluster assignment, this may indicate substantial variability in the patient's condition, potentially highlighting a recovery trajectory or an uncertain recovery state.

4.3 Evaluation Protocol

We evaluate the final clustering solution using a combination of quantitative and qualitative assessments. First, we examine the internal structure of the data-driven clusters by analyzing the means and ranges of selected UL performance features within each cluster. We choose these features to represent distinct aspects of UL performance. In addition, we create pairplot matrices to visualize the relationships between the input features and the cluster assignments.

Next, we compare the final clustering solution with the clinical clustering derived from FMA scores. We perform this comparison using several metrics, including normalized mutual information (NMI), adjusted rand index (ARI), accuracy, and confusion matrices. These metrics quantitatively assess the agreement between the two clustering solutions. We expect to observe a moderate level of agreement across all measurements.

To evaluate the structure of the clusters in the representation space, we generate scatterplots of the representations using the first pair of principal components (e.g., PC1 vs. PC2). We color the plots according to cluster assignments from both data-driven clustering and clinical clustering. This approach allows us to qualitatively evaluate the clustering structure and identify any visible patterns or separations in the data.

Finally, we investigate the discrepancies between the two clustering solutions by analyzing mismatches. Specifically, we calculate the means and ranges of UL performance features for patients whose cluster assignments differ between data-driven clustering and clinical comparison clustering. We further explore these discrepancies to determine which clustering solution more accurately reflects the underlying UL performance qualities.

Chapter 5

Experiments

This chapter presents the findings from applying the two clustering approaches introduced earlier to the movement dataset of stroke patients. To ensure a streamlined presentation, results and discussion are combined, allowing immediate interpretation of findings within their respective contexts. The chapter begins by discussing the outcomes of each individual approach, followed by a comparison of their performance and applicability.

5.1 Performance Feature Clustering

In this section, we present the results of clustering based on UL performance features. Recall that this approach involves three main steps. First, selecting a suitable set of features. Second, computing performance representations using the first two principal components obtained by applying PCA on the feature set. And third, applying K-Means to these performance representations to derive UL performance clusters.

We begin by outlining the process used to select the final set of features for clustering. Following this, we present the results of the clustering analysis. The results include an assessment of the internal structure of the clusters, their representation in the reduced feature space, and their alignment with clinical clusters. In addition, we explore mismatches between feature-based clusters and clinical clusters.

5.1.1 Feature Selection

This section describes the process of selecting and optimizing the set of features used to calculate the UL performance clusters. Starting with the feature set proposed by Barth et al., we evaluate its limitations and refine it to create a final set that better aligns with the objectives of this analysis. A

comprehensive list of all UL performance features considered in this study is provided in the Appendix Table A.1.

Table 5.1: Description of UL performance features used by Barth et al. We observe a bias towards features representing duration. In addition, only features based on activity counts are present.

UL Performance Feature	Quality	Description
Use Duration Affected	Duration	How long the affected limb was active, in minutes
Use Duration Nonaffected	Duration	How long the nonaffected limb was active, in minutes
Use Duration Ratio	Duration Symmetry	The ratio of minutes the affected limb was active to nonaffected
Acceleration Variability Affected	Variability	Standard deviation of the activity counts of the affected limb
Median AC affected	Magnitude	Median activity counts over the whole measurement in the affected limb, in activity counts

We begin the feature selection process with the feature set proposed by Barth et al. (see Table 5.1). A qualitative evaluation of this feature set reveals several observations. First, the set includes multiple measures of duration (affected, nonaffected, and ratio), which may introduce redundancy and a potential bias towards this quality in the clustering process. Second, all features are derived from acceleration counts, indicating the absence of a performance measure that captures the quality of movement based on jerk values. Barth et al. explicitly excluded jerk-related features due to their computational cost, as they computed jerk at 30 Hz.

Table 5.2: Feature importance scores computed using the RelieF algorithm applied to the clustering derived from Barth et al.'s UL performance features. The table highlights the relative importance of each feature in influencing the clustering. The results suggest that the use duration of the nonaffected side has minimal contribution to the clustering solution.

Feature Name	Importance Score	Rank
Use Duration Ratio	0.505	2
Median AC Affected	0.472	4
Use Duration Affected	0.445	7
Acceleration Variability Affected	0.268	14
Use Duration Nonaffected	0.049	35

To evaluate the contribution of these features to the resulting clustering, we applied the RelieF algorithm to the clustering derived using the features proposed by Barth et al. Table 5.2 shows the feature importance scores obtained from this analysis. The results indicate that the acceleration variability of the affected side has relatively low importance, while the use duration of

the nonaffected side contributes almost no significance to the final clustering solution. Notably, the minimal importance of the nonaffected side suggests that this feature has little to no correlation with UL performance.

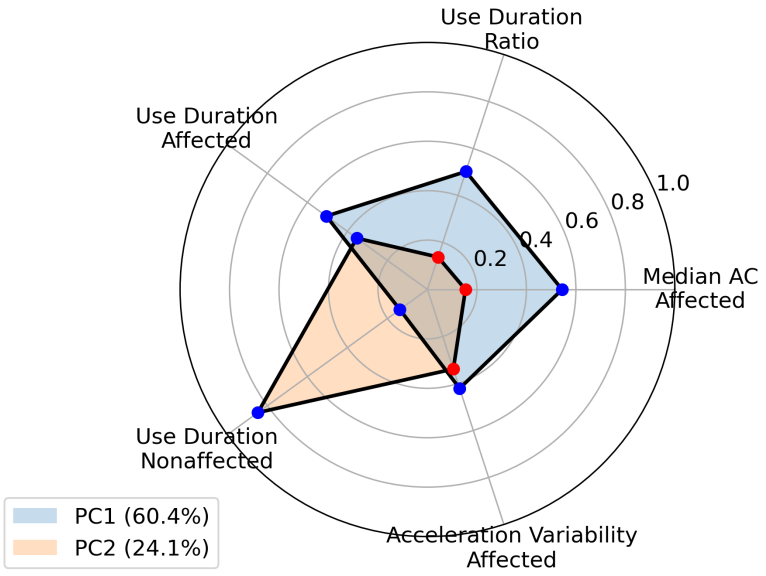

Figure 5.1: Explained variance and feature contributions to the principal components (PCs) after applying PCA to Barth et al.'s feature set. The radar plot visualizes the feature weights for the first two principal components (PC1 and PC2). The results show that the use duration of the nonaffected side contributes minimally to PC1 but significantly to PC2, suggesting that it encodes independent information. We view PC1 as the primary encoding of UL performance, as it contains moderate loadings of all other features.

Figure 5.1 illustrates the explained variance and feature contributions to the principal components after applying PCA to Barth et al.'s feature set. The analysis shows that the first principal component (PC1) captures the most variance (60.4%) but has unbalanced loadings, with the use duration of the nonaffected side contributing almost nothing. Additionally, the analysis reveals that use duration of the nonaffected side provides significant information in the second principal component (PC2). This supports the idea that the use duration of the nonaffected side may encode information independent of UL performance that are not directly relevant to the clustering objective. We hypothesize that PC1 primarily encodes UL performance, as it integrates moderate positive loadings from features that are representative of movement

quality, magnitude, and duration. The relatively low explained variance of PC1, indicates that a substantial portion of the variability in Barth et al.'s feature set may stem from irrelevant or redundant features. To address this, we refined the feature set to prioritize clinically significant features, with the aim of improving the representation of UL performance in the reduced feature space.

Table 5.3: Description of the final UL performance features. The final feature set was optimized based on insights from prior analyzes. The main change is the inclusion of "Mean Jerk Ratio" and "Median Magnitude Ratio."

UL Performance Feature	Quality	Description
Use Duration Affected	Duration	See Table 5.1
Use Duration Ratio	Duration Symmetry	See Table 5.1
Mean Jerk Ratio	Quality of Movement Symmetry	The mean ratio of jerk of the affected side to the nonaffected
Median AC Affected	Magnitude	See Table 5.1
Median Magnitude Ratio	Magnitude Symmetry	Median log of the ratio of activity counts of the affected to the nonaffected limb, capped at ± 7

After iteratively refining the feature set to better align with our objectives, which include capturing different qualities of UL performance, ensuring that all features are of high importance based on RelieF, and achieving a balanced PCA with a high explained variance, we arrive at the final set presented in Table 5.3. In this final set, we replaced the use duration of the nonaffected side and the variability of the affected side with the mean jerk ratio and median magnitude ratio. These replacements were made to introduce measures that capture movement smoothness based on jerk, a quality absent from the baseline feature set, and to include a highly important metric that offers a different assessment of magnitude.

Table 5.4: Feature importance scores derived from applying the RelieF algorithm to the clustering from the final feature set. The results demonstrate that all selected features rank relatively high in importance to the final clustering solution.

Feature Name	Importance Score	Rank
Median Magnitude Ratio	0.691	1
Use Duration Ratio	0.579	2
Median AC affected	0.424	6
Mean Jerk Ratio	0.420	7
Use Duration Affected	0.407	9

To evaluate the importance of the features in the final feature set, we applied the clustering pipeline to this feature set and then used the RelieF algorithm to analyze the resulting clusters. Table 5.4 presents the feature importance

scores derived from this analysis. The results indicate that all features rank relatively high in importance, providing confidence that each feature is represented in a meaningful way in the final clustering solution.

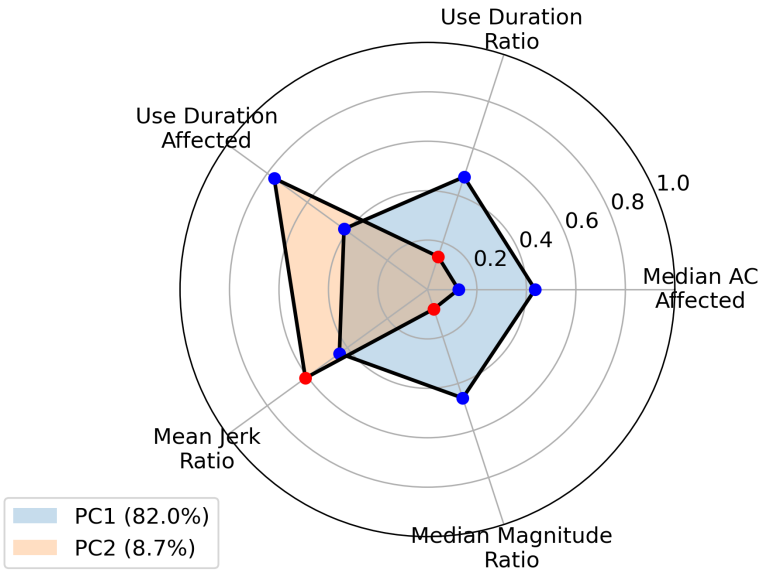

Figure 5.2: Weights of the first two principal components (PC1 and PC2) after applying PCA to the final feature set. The loadings for PC1 are distributed more evenly across all features. No single feature dominates the variance. This balanced distribution and the high explained variance highlights the improvements of the final feature set compared to the feature set proposed by Barth et al.

Fig. 5.2 illustrates the weights of the first two principal components after applying PCA to our final feature set. The first principal component (PC1) accounts for the largest proportion of the explained variance (82.0%) in the final feature set, suggesting that it captures the dominant patterns in the data. Additionally, PC1 exhibits balanced loadings across all features, indicating that it integrates multiple aspects of UL performance, such as movement magnitude, smoothness, and duration. This alignment supports the hypothesis that PC1 represents a general measure of UL performance, as opposed to being biased toward any single feature. As a result, we consider the feature set suitable for accurately capturing UL performance in the subsequent clustering analysis.

5.1.2 Clustering Results

This section presents the results of the clustering based on the final feature set. We begin by analyzing the internal structure of the clusters.

Internal Structure

Table 5.5: Summary of features representing different UL performance qualities across the three final clusters. For each cluster (Low, Medium, High), the table provides mean values along with ranges (in parentheses) for various performance metrics and for a capacity assessment. Patients with lower overall UL performance are grouped in the Low cluster, while those with progressively better performance are placed in the Medium and High clusters, respectively.

Feature name	Low (120)	Medium (96)	High (71)
Duration			
Affected (Min)	83.20 (9.5–278.9)	228.30 (14.5–417.4)	349.05 (18.2–596.8)
Ratio	0.28 (0.06–0.60)	0.73 (0.49–1.03)	1.04 (0.79–3.12)
Magnitude			
Affected (Median AC)	29.74 (15.5–48.5)	43.72 (28.7–61.7)	61.86 (18.3–86.0)
Ratio (Log, capped at ± 7)	-6.85 (-7.00—3.49)	-1.02 (-7.00–0.04)	0.00 (-2.33–3.36)
Variability			
Affected (SD)	48.67 (30.82–82.84)	54.40 (27.71–78.30)	68.89 (45.65–92.66)
Ratio	0.72 (0.38–1.27)	0.74 (0.41–1.08)	1.03 (0.64–8.18)
Movement Quality			
Affected (Jerk Mean)	0.25 (0.10–0.66)	0.45 (0.06–0.82)	0.76 (0.14–1.95)
Ratio	0.40 (0.18–0.70)	0.67 (0.44–1.14)	0.93 (0.58–2.11)
Clinical Assessment			
FMA Score	15.44 (1.0–61.0)	40.92 (12.0–63.0)	56.82 (8.0–66.0)

Table 5.5 summarizes a range of features that represent different aspects of UL performance. For each cluster and quality, we calculate averages and ranges of representative feature values for the affected side and the ratio of affected to nonaffected sides.

The mean values for each cluster highlight the differences in overall UL performance. Specifically, the low performance group consistently exhibits lower scores in all performance qualities, while the medium and high performance groups show progressively better performance. These findings suggest that the clusters effectively separate individuals based on their UL performance levels.

Fig. 5.3 illustrates the relationships between the final features and their pairwise distributions between the groups. The diagonal plots display the univariate distributions of each feature, allowing for a detailed analysis of how the feature values vary within the clusters. We expect some overlap between clusters, as individuals often lie on recovery trajectories where performance improvements occur gradually. This overlap reflects the continuous

nature of UL performance and recovery rather than discrete groupings. However, distinct separations are generally visible along all features, highlighting the clusters' ability to capture meaningful differences in the data.

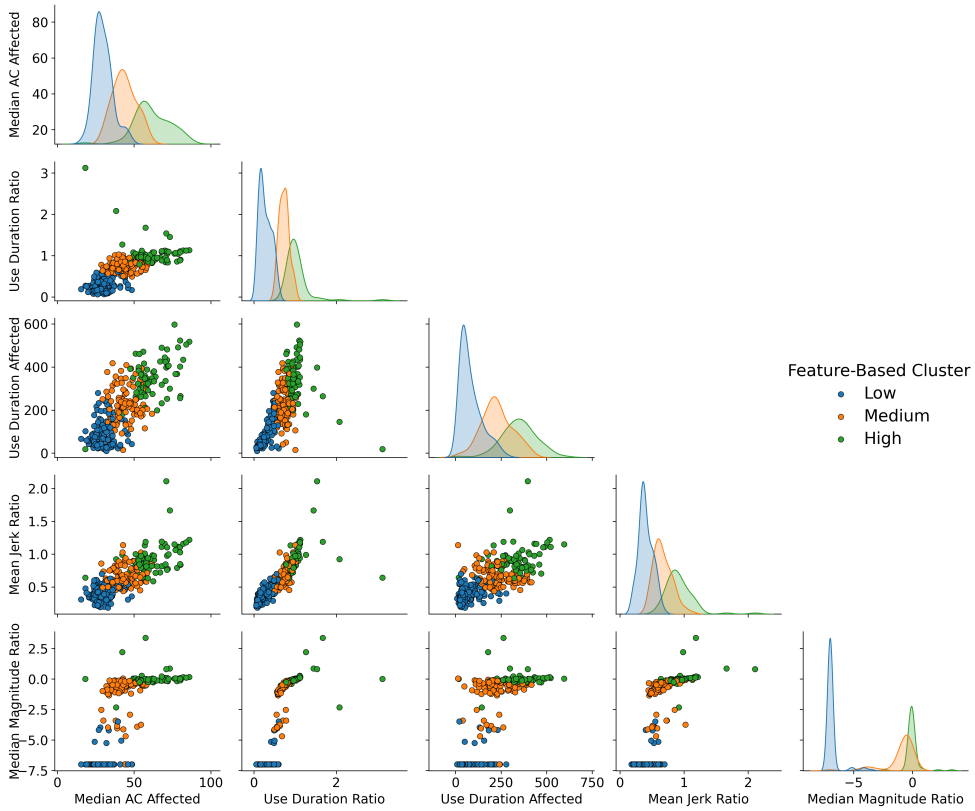

Figure 5.3: Pairplot of the final features across the three clusters. The diagonal plots show the univariate distributions for each feature and cluster. Pairwise scatterplots reveal the relationships between features. While some overlap exists, distinct separations are visible along several features, demonstrating the ability of the clustering to capture meaningful differences in the data and effectively group individuals based on their UL performance.

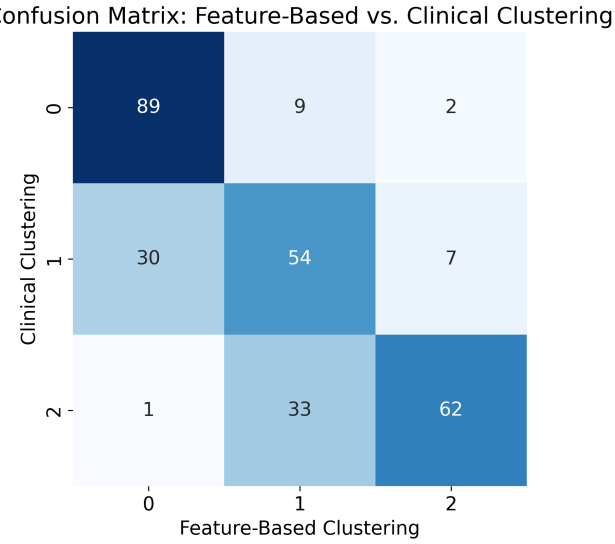

Figure 5.4: Confusion matrix comparing the feature-based clustering with the clinical clustering. The clustering achieves an accuracy of 71%, reflecting moderate agreement between the two cluster solutions. The data-driven clustering tends to cluster more conservative, as it assigns lower performance categories more frequently than the clinical clustering.

Clinical Comparison

Next, we evaluate how well the feature-based clustering aligns with the clinical clustering. The clustering achieves an accuracy of 71%, an adjusted rand index (ARI) of 0.387, and a normalized mutual information (NMI) of 0.393. Figure 5.4 presents the confusion matrix comparing the performance feature clustering with the clinical clustering.

Several observations can be made. First, the metrics and confusion matrix indicate moderate agreement between clinical clustering and data-driven clustering. Some divergence between the feature-based and clinical clustering is expected, as clinical capacity assessments do not perfectly reflect performance in real-world settings. However, we still expect them to largely agree, as capacity does relate to performance to a certain degree. We will investigate shortly whether our feature-based clustering better captures performance-related aspects. Second, data-driven clustering tends to be more conservative than clustering based on clinical assessments, as it tends to assign lower performance categories more frequently.

Two-dimensional performance representations enable cluster visualization, providing an opportunity to inspect clustering results and differences between data-driven and clinical clustering approaches. Fig. 5.5 illustrates these

representations, colored by the data-driven clustering on the left and the clinical clustering on the right.

Overall, we again observe a moderate degree of agreement between the two clustering approaches, as reflected in the visual alignment of the clusters. The first principal component appears to encode a significant amount of impairment-related information, as a clear hierarchy is evident along it in both feature-based and clinical clustering. This observation supports our hypothesis from the feature selection process, where we suggested that the first principal component captures overall UL performance.

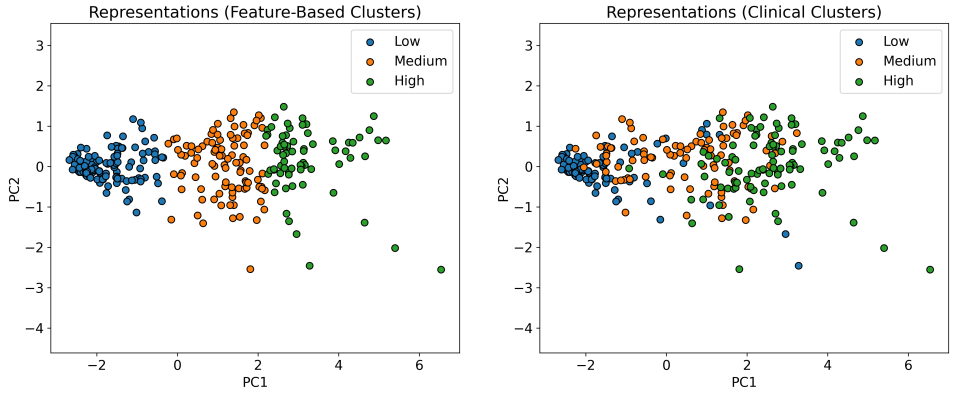

Figure 5.5: Feature-based performance representations colored by feature-based clustering (left) and clinical clustering (right). The first principal component encodes a significant amount of impairment-related information, supporting its role in capturing overall UL performance.

Mismatches

To compare and analyze mismatches between data-driven and clinical clustering, we calculate the averages of performance features across all cluster combinations. Fig. 5.6 presents two heatmaps: the left shows the average use duration of the affected side, and the right shows the average median magnitude of the affected side. Each cell contains the average value for a specific cluster combination, with the sample size indicated in parentheses.

The heatmaps highlight the ability of the feature-based clustering to capture performance-related differences. For use duration, clusters with a lower feature-based cluster assignment consistently exhibit shorter use durations compared to higher feature-based assignments, regardless of the clinical cluster. Similarly, for median magnitude, lower feature-based clusters consistently show smaller values than higher ones. These trends suggest that data-driven clustering provides a finer differentiation of UL performance levels, as it better distinguishes varying levels of performance than the clinical clustering.

Two notable outliers emerge: patients classified as high-performing by data-driven clustering but with low capacity according to clinical assessments. Further investigation reveals that these patients exhibit a high use duration ratio, indicating that while their overall limb use is limited, it is heavily biased toward the affected side. This imbalance likely influenced the model to assign them to a higher performance category.

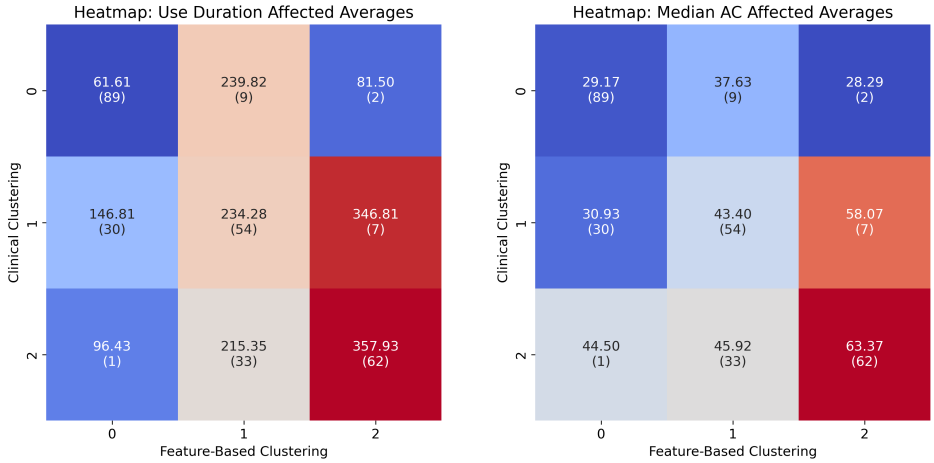

Figure 5.6: Heatmaps of average use duration and magnitude of the affected side across cluster combinations from data-driven and clinical clustering. The heatmaps show that, on average, the data-driven clustering approach better captures performance-related differences. Patients with a low capacity assignment, but high performance class (upper right) exhibit a high use duration ratio. This likely influenced their assignment to a higher performance category.

5.1.3 Reflection

In this section, we reflect on the performance of the feature-based clustering approach, highlighting its strengths, challenges, and potential directions for future work. This approach builds on the feature set proposed by Barth et al., addressing its limitations by reducing redundancy and incorporating additional features, such as jerk-related metrics, which provide unique insights into movement quality. By refining their feature set, we derive a meaningful clustering of UL performance that clearly distinguishes between low, medium, and high performance levels. In each category, individuals demonstrate similar values across the performance features analyzed. Furthermore, the analysis of mismatched clusters indicates that this approach outperforms clinical capacity assessments to capture performance-related differences. Therefore, this approach has achieved the desired initial goal of deriving an accessible categorization system to assess UL performance.

Despite these strengths, the approach has several challenges and areas for improvement. The feature selection process is not exhaustive and may overlook features that capture relevant relationships. For example, the RelieF algorithm used for feature selection relies on local feature importance, which may miss global interactions between features. Similarly, incorporating PCA in the feature selection process ignores potential nonlinear relationships between the features. These limitations highlight the need for future work to explore alternative feature selection strategies.

Additionally, the feature set primarily captures simple relationships in the data, such as means, medians, and cumulative measures like duration. While these features are interpretable and valuable, they may not fully capture more complex movement patterns, such as temporal dynamics, UL movement coordination, or variability over shorter timescales.

Future studies could also explore alternative dimensionality reduction techniques when computing performance representations. While this study experimented with additional principal components and clustering on raw features, these adjustments did not significantly impact clustering results (see Appendix Fig. A.1). Nonlinear techniques, such as t-SNE or UMAP, may provide additional insights by capturing relationships that PCA may overlook. Similarly, investigating alternative clustering algorithms, such as Gaussian Mixture Models (GMMs) or hierarchical clustering, may yield different perspectives on the data structure (see Appendix Fig. A.2).

The limitations of this feature-based approach motivate the development of a second technique leveraging deep learning. By eliminating the need for manual feature engineering and leveraging the ability to model temporal dynamics and nonlinear relationships, deep learning-based approaches offer a promising avenue to address the limitations of the current feature-based method. The experimental setup and results of this method are detailed and analyzed in the following section.

5.2 Deep Learning Clustering

In this section, we present the results of clustering based on embeddings computed using a contrastive learning approach. This method aims to capture complex movement patterns and identify meaningful clusters of UL performance, using deep learning. Recall that this approach involves three main steps. First, the measurements are split into 15-minute slices and filtered using a minimum activity threshold of 10%. Second, a contrastive learning approach is used to train a model to maximize the similarity of the embeddings for slices with similar movement characteristics. Third, the extracted embeddings are clustered using K-Means. The final cluster assignments of slices are aggregated at the measurement level, and the final assignment per measurement is derived by majority voting.

This section begins with a description of the experimental setup, including details on slice generation. We then provide an overview of the training process, including hyperparameter configurations, and evaluate whether the model effectively captures patient-specific behaviors. Lastly, we analyze the clustering performance of the resulting embeddings by examining their internal structure and comparing them to the clinical clustering.

5.2.1 Experiment setup

The experimental setup involves how we prepared the raw IMU measurements for training the deep learning model and how to configure the training process itself. This includes segmenting the data into manageable slices, applying activity filtering to ensure a minimum level of activity, and setting up the deep learning model for training. In addition, we describe the optimizer, hyperparameter tuning, and the key factors that influenced model performance.

Slicing

To prepare the IMU data for training a deep learning model, we segmented the continuous time series into slices of 15-minute windows. After slicing, we applied an inactivity threshold filter, such that at least 90 seconds of activity is present in the measurements. During our experiments, we tested different window lengths and activity thresholds. Table 5.6 summarizes the number of valid slices across different window sizes and threshold values. Selecting an appropriate window size and threshold is critical for training the model effectively, as these parameters determine the quality and quantity of data used to learn performance representations.

Another key consideration is the distribution of valid slices between patient performance levels. For example, a dataset heavily skewed toward high-performing patients would introduce bias and hinder meaningful clustering

Table 5.6: Number of valid IMU data slices across varying window sizes and activity thresholds. Valid slices are defined as having at least the specified proportion of activity within the window. Smaller window sizes and lower thresholds provide more segments, while larger windows and stricter thresholds reduce the number of valid slices. These parameters are critical for balancing data quality and quantity in training the deep learning model.

Window Size (min)	Threshold 10%	Threshold 25%	Threshold 50%
5	85,142	62,284	34,790
10	44,394	32,065	16,815
15	30,414	21,697	10,882
30	15,952	11,105	5,016
60	8,393	5,597	2,179

of the embeddings. To address this, we used categories from our clinical clustering as a proxy to evaluate the distribution of valid slices. Figure 5.7 illustrates the distribution of slices across clinical categories for a window size of 15 minutes. The results indicate that the lower thresholds (10% and 25%) maintain a relatively even distribution between categories, while the 50% threshold disproportionately reduces the number of valid slices from lower-performing patients.

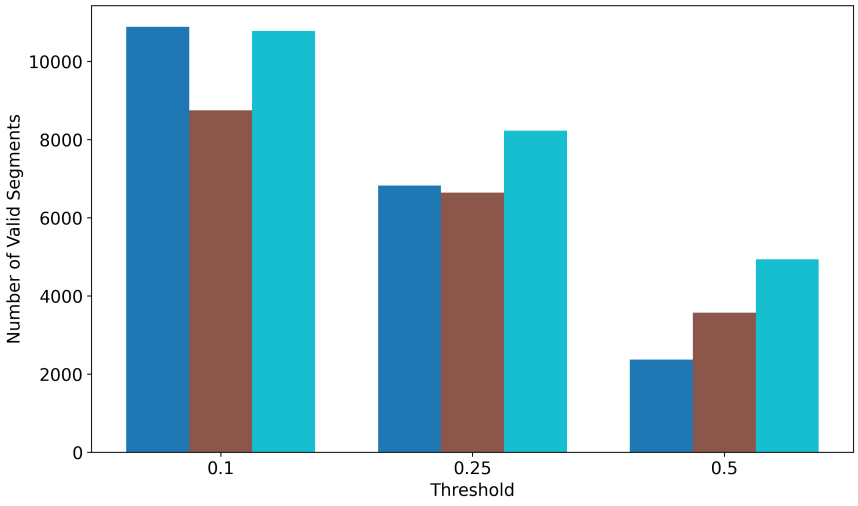

Figure 5.7: Distribution of valid slices across clinical capacity categories for a 15-minute window size at different activity thresholds. Lower thresholds (10% and 25%) maintain a relatively even distribution among capacity categories, while the 50% threshold disproportionately reduces valid slices from lower-performing patients, potentially introducing bias.

As shown in Table 5.6, increasing the window size reduces the total number of slices significantly. Similarly, applying stricter threshold filters drastically reduces the number of valid slices. However, as Figure 5.7 demonstrates, this reduction is not uniform across clinical categories. Higher thresholds

disproportionately impact patients with lower functional capacity, leading to a skewed dataset.

All combinations of window size and threshold were provided to the hyperparameter optimization framework during training. In cases of comparable model performance, we prioritized configurations with lower thresholds (10% or 25%) and medium-sized windows (15- or 30-minute slices). Training the model revealed that medium and longer windows worked best when paired with lower thresholds. Possibly, because they provide a sufficient number of slices while preserving meaningful activity patterns. Based on these findings, the final selection was determined to be 15-minute windows with a 10% threshold. This configuration strikes a balance between maintaining data quality and ensuring a representative distribution across patient categories.

Training

The training process focuses on developing a robust embedding encoder $f(\cdot)$ capable of capturing complex UL movement patterns. In the following, we provide additional information on training, including data split, optimization strategy, hyperparameter tuning, and insights gained during the training process. The models are trained using the procedure described in the pseudocode provided in Appendix Algorithm 1.

Train-Validation Split The training and validation datasets are split at the patient level, ensuring that 20% of the patients are assigned to the validation set. We take this approach to prevent information leakage between the training and validation sets, which can otherwise compromise the evaluation of the model's generalization performance.

Optimizer and Learning Rate Scheduler We employ the AdamW optimizer, which combines the benefits of adaptive learning rates with decoupled weight decay for better generalization [32]. A cosine learning rate scheduler is used to dynamically adjust the learning rate during training, promoting more effective convergence by gradually reducing the learning rate.

Hyperparameter Tuning and Monitoring We tuned the hyperparameters using Optuna, an efficient and flexible framework for automated hyperparameter optimization. Optuna is particularly advantageous in scenarios with a large number of hyperparameters, as it employs a principled approach to explore the search space and identify optimal configurations [33]. A comprehensive list of all tuned hyperparameters, along with their respective search spaces and selected values, is provided in Appendix A.2.

Validation Loss During the validation step, we fix the temperature parameter τ to a constant value, regardless of the τ used during training. This approach ensures comparability between models trained with different values of τ, since τ directly affects the loss calculation and can otherwise confound the comparison of parameters.

Training Insights We highlight key insights gained during training that guide our parameter choices and model design. The use of stronger augmentations, such as adding more Gaussian noise, effectively enhances the robustness of the embeddings to variations in movement patterns. Longer slice windows, combined with lower thresholds, help capture sufficient data while maintaining balanced patient representation. Additionally, medium-sized models, with appropriately tuned hidden dimensions and LSTM layers, demonstrate superior performance by effectively mitigating overfitting compared to larger models. These findings inform the model architecture, ensuring both stability during training and high-quality embeddings for clustering.

It is important to note that the ultimate goal of this approach is not simply to minimize the validation loss, but to enable the model to learn embeddings that accurately capture UL performance. While a low validation loss is an important indicator of the model's capability, it is not the sole criterion for selection. Instead, we prioritize models that achieve a balance of low loss and produce high-quality embeddings that effectively capture patient performance.

Final model We analyzed multiple models trained with different parameter setups, focusing on those with low validation losses, and evaluated the quality of their embeddings. Here, we present only our final model, which we select based on its ability to achieve both a reasonably low loss and embeddings suitable for downstream analysis. The exact parameters for this final model are provided in Appendix A.2.

After identifying the best configuration, we train the final model on the entire dataset to extract the embedding encoder $f(\cdot)$. Once trained, we analyze the learned embeddings to evaluate their ability to accurately represent UL performance. The next section provides a detailed analysis of these embeddings.

5.2.2 Learned Embeddings

As a first step to evaluate the quality of the learned embeddings, we examine whether the final model successfully groups the slice embeddings of the same measurement together. To do this, we apply PCA to reduce the dimensionality of the embeddings and plot the first two principal components.

Figure 5.8 visualizes the embeddings of ten different measurements, colored by measurement. The plot shows that slices from the same measurement form tight clusters. This behavior likely results from the variability introduced by sampling different slices, which forces the model to learn robust and invariant representations.

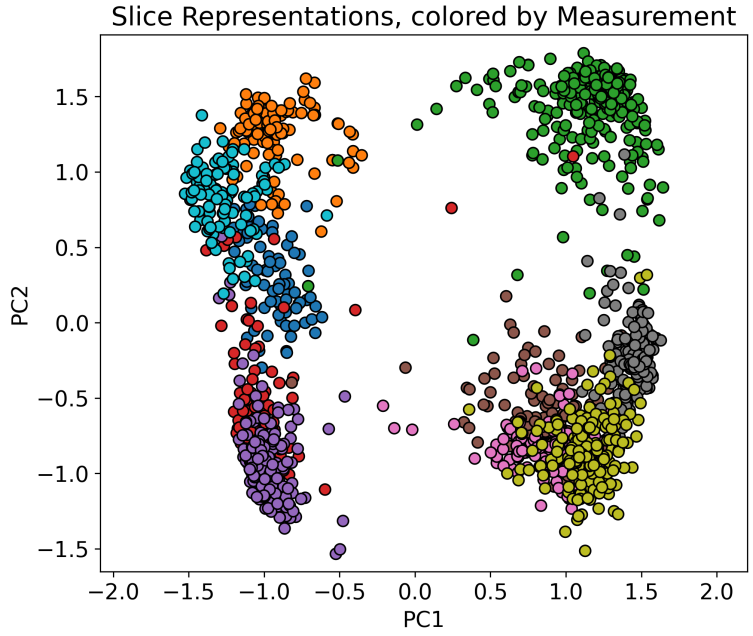

Figure 5.8: PCA projection of slice embeddings from ten different measurements, colored by measurement. The plot shows that slices from the same measurement form tight clusters, indicating that the model learns robust and invariant representations. A potential hierarchy is observed along the second principal component, and two distinct groups are evident along the first principal component.

Figure 5.8 also indicates a possible hierarchy between patients along the second principal component, as well as a clear clustering in two groups along the first principal component. This can be due to the random sampling of ten measurements rather than inherent properties of the embeddings. However, we will investigate whether such patterns are visible in the final assignments.

The results demonstrate that the model effectively distinguishes slices based on their measurement of origin. This observation suggests that the learned embeddings encode information that captures the variability introduced by different slices, potentially relevant to UL performance.

In the next section, we assess whether the learned embeddings capture clinically meaningful information and enable the grouping into performance categories.

5.2.3 Clustering Results

We analyze the final cluster assignments for each measurement to assess their internal structure and clinical relevance.

Internal Structure

Table 5.7: Means and ranges of UL performance features and capacity across the three deep learning clusters: low, medium, and high performance. The low and high performance clusters exhibit distinct characteristics, while the medium performance cluster shows less meaningful distinctions, suggesting limited utility in differentiating performance levels.

Feature name	Low (93)	Medium (123)	High (75)
Duration			
Affected (Min)	88.35 (9.5–376.51)	224.75 (18.17–596.81)	287.83 (49.12–517.04)
Ratio	0.29 (0.06–2.08)	0.71 (0.07–3.12)	0.85 (0.51–1.67)
Magnitude			
Affected (Median AC)	30.20 (15.5–48.67)	45.58 (18.25–84.00)	51.28 (28.67–86.00)
Ratio (Log, capped at 7)	-6.63 (-7.00—0.85)	-2.39 (-7.00–0.26)	-0.61 (-7.00–3.36)
Variability			
Affected (SD)	47.83 (27.71–82.84)	57.75 (34.45–92.66)	61.12 (39.22–86.40)
Ratio	0.71 (0.38–1.27)	0.83 (0.43–8.18)	0.85 (0.49–1.42)
Movement Quality			
Affected (Jerk Mean)	0.27 (0.10–0.73)	0.48 (0.09–1.46)	0.61 (0.15–1.95)
Ratio	0.40 (0.18–0.92)	0.67 (0.23–1.18)	0.79 (0.37–2.11)
Clinical Assessment			
FMA Score	15.08 (1.0–61.0)	37.62 (4.0–66.0)	51.05 (25.0–66.0)

Table 5.7 summarizes the means and ranges of the UL performance features and UL capacity for the three clusters in the final solution, labeled as low, medium, and high performance. The results show that the low and high performance clusters capture distinct performance characteristics. The medium performance group does not show meaningful distinctions in performance characteristics, raising questions about its utility in representing clinically relevant differences. From this analysis, it does not seem like the whole clustering solution yields a clinically meaningful grouping.

Clinical comparison

Next, we evaluate the alignment between the clustering results and the clinical groupings derived from FMA scores. The clustering achieves an accuracy of 52%, an adjusted rand index (ARI) of 0.179, and a normalized mutual information (NMI) of 0.245, indicating low agreement. Figure 5.9 shows the confusion matrix comparing the clustering based on performance features with the clinical clustering. In particular, the medium performance group (on the x axis) appears to cover a nearly equal distribution across all levels of clinical capacity, further suggesting that this category does not effectively distinguish between different levels of upper limb performance.

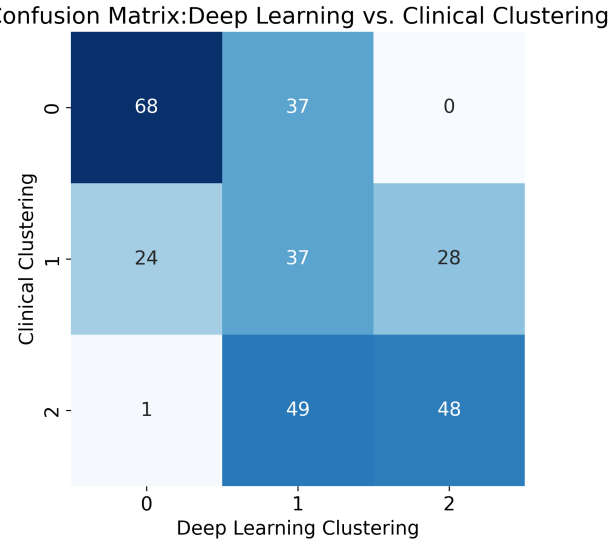

Figure 5.9: Confusion matrix comparing deep learning-based clustering with clinical clustering. The clustering achieves an accuracy of 52%, indicating low agreement. The medium performance group (x-axis: 1) distributes nearly equally across all clinical capacity levels, underscoring its limited ability to distinguish between different levels of upper limb performance.

To visualize the results, we averaged the principal components of the slice embeddings for each measurement and assigned colors based on the final cluster assignments. Figure 5.10 presents these visualizations, with the left panel showing the data-driven clustering and the right panel showing the clinical clustering.

The visualization suggests that the model has not primarily learned to distinguish UL performance but instead has captured another dominant property. However, the coloring based on the clinical clustering demonstrates that the model still encodes a hierarchy of impairments, visible along the second principal component. We observed similar patterns in the analysis of other models trained with slightly different parameters, indicating that this behavior is not unique to the specific model presented here.

The clustering algorithm appears to group slices based on the dominant binary characteristic, as evident in the separation along the first principal component. This structure conflicts with the objective of deriving clusters based on UL performance, which seems to be encoded along the second principal component. Due to the unsatisfactory clustering results caused by this dominant structure, we omit the analysis of mismatches and pairplots for this approach. Instead, in the next section, we investigate the nature of this dominant structure.

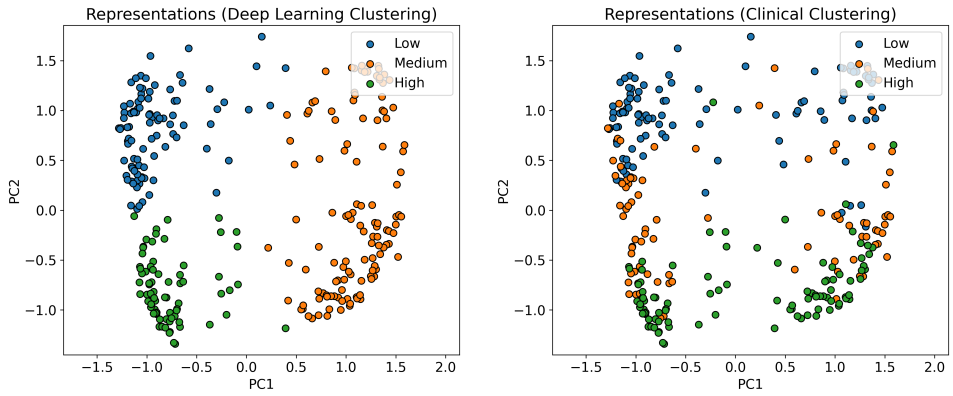

Figure 5.10: Visualization of final cluster assignments by applying PCA on the slice embeddings and averaging per measurement. The left panel shows clusters derived from the deep learning approach, while the right panel represents clinical clustering. The data-driven clustering highlights a dominant binary characteristic along the first principal component, which conflicts with the objective of distinguishing UL performance. However, the clinical clustering reveals that the model encodes a hierarchy of impairments along the second principal component.

Dominant Structure

To understand the structure learned by the model, we explored potential patterns in the embeddings. An intuitive hypothesis suggests that the impairment of the dominant side could be the dominant structure, as individuals often show different activity patterns if their dominant side is affected [34]. However, our visual analysis of the embeddings revealed that the model predominantly learned which side of the body was affected by the stroke as the distinguishing feature.

Figure 5.11 visualizes the measurement embeddings colored by the affected side, demonstrating that the clusters are clearly separated on the basis of this feature. This result is particularly surprising, as the model was not explicitly provided with side-related information (left or right) during training. More precisely, the data originally labeled as left and right were relabeled as affected and nonaffected, as described in Section 4.2.1.

To further investigate, we examined concordance between dominant side and affected side, as shown in Figure 5.11. Although there is some overlap between these features, it is visually evident that the affected side is the dominant structure captured by the model.

One plausible explanation for the model's ability to distinguish the affected side could be the presence of subtle side-related information embedded within the raw accelerometer and gyroscope data. For example, when reaching for an object, the movement direction differs between the left and right hands: the left hand moves to the right, and the right hand moves to the

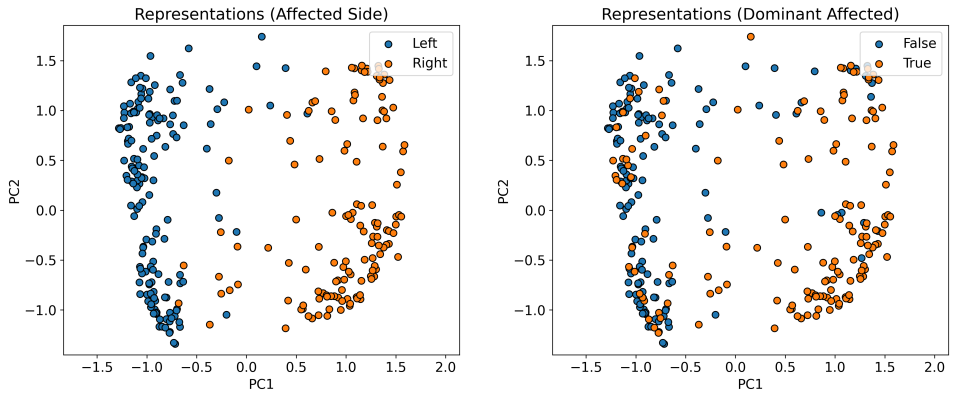

Figure 5.11: Measurement embeddings, colored by affected side (left) and by concordance of dominant and affected side (right). The clear separation in the left panel indicates that the model predominantly learned to distinguish based on the affected side, despite not being explicitly trained with lateralization information.

left. These directional patterns are reflected in the horizontal y-axis values of the accelerometer data, where the left hand shows positive values and the right hand shows negative values. Since the preprocessing steps do not account for this mirroring effect, the model may captures this information. While this reflects the model's sensitivity to subtle details in the data, it also highlights challenges in controlling what the model learns when working with complex signals.

5.2.4 Reflection

This section reflects on the results of the deep learning-based approach, outlining its strengths, technical challenges, and opportunities for future improvements. We concentrate on technical aspects specific to this method, deferring a broader reflection on the two approaches to Section 5.3.

The proposed approach leverages contrastive learning to train a deep learning model on 15-minute slices of activity, aiming to learn meaningful representations of UL impairment and performance. By carefully tuning the hyperparameters, the model demonstrates its ability to identify patterns within the data, such as a clearly reflected hierarchy of impairment in the visualized embeddings. Additionally, it captures patient-specific behaviors, shown by its ability to group slices from the same measurement together. Interestingly, the model also learns more complex characteristics, such as the affected side, despite not being explicitly trained on side-specific information. However, this dominant structure interferes with the primary objective of deriving an accessible and clinically relevant category system of UL performance.

This unintended focus on lateralization disrupts the clustering algorithms applied to the learned embeddings, reducing the clinical relevance of the resulting clusters. It underscores the need for more sophisticated preprocessing techniques to address such biases. Although measurements are relabeled to affected and nonaffected to remove explicit side information, subtle side-related biases remain. Addressing these biases can encourage the model to focus on clinically relevant aspects of UL performance.

Future research can explore the development of advanced preprocessing techniques to eliminate these side-related biases more effectively. Ensuring that mirrored movement patterns are accounted for during preprocessing helps the model achieve its intended focus. Another direction involves exploring higher-resolution data, such as the original 50 Hz accelerometer and gyroscope signals without resampling, as this provides the model with more granular information, potentially improving its ability to learn meaningful representations of UL performance.

Further investigation of alternative unsupervised learning frameworks is also beneficial. Techniques such as other contrastive learning frameworks or variational autoencoders provide complementary approaches to representation learning. Additionally, conducting in-depth analyses of the embeddings learned by the deep learning model yields valuable insights, similar to the methodology employed in the Autoencoder work [21].

In the next section, we compare the feature-based and deep learning-based approaches, discuss their broader implications, and highlight shared limitations and opportunities for improvement.

5.3 Comparison of the Two Approaches

In this section, we compare the two clustering approaches proposed in this thesis.

The first approach derives a clustering based on human-engineered UL performance features, producing a valid classification of UL performance into low, medium, and high performing groups. This method offers clear and interpretable clusters that reflect a broad range of UL performance qualities. Additionally, it outperforms clinical UL capacity assessments, providing clinicians with an accessible and practical tool for assessing daily life UL performance in Stroke Patients. However, this approach relies on simple statistical metrics, which may fail to capture more nuanced movement patterns, thereby limiting its ability to provide a complete representation of UL performance.

In contrast, the second approach applies a deep learning model trained directly on raw acceleration and angular velocity time series data. This

method demonstrates the ability to learn complex patterns in the data, capturing subtle movement qualities and patient-specific behaviors that extend beyond the scope of human-engineered features. Notably, it encodes impairment-related information and identifies the affected side of the stroke. However, the dominant structure learned by the model hinders its ability to produce a clinically meaningful classification system, as the clusters do not align with UL performance categories. This challenge highlights the need for further optimization to tailor the learned embeddings toward clinical relevance.

Currently, the first approach is more practical and readily applicable in clinical settings, as it already produces an interpretable and valid grouping of UL performance. In contrast, the second approach holds significant potential for uncovering richer, data-driven insights, particularly as preprocessing techniques and training strategies improve. This deep learning method could eventually extend the assessment of UL performance by capturing dimensions of movement quality and rehabilitation progress that traditional feature engineering cannot.

A promising avenue for future research lies in combining the strengths of both approaches. A hybrid system that integrates human-engineered features with the nuanced patterns captured by the deep learning model could offer a more comprehensive representation of UL performance. Such a system could enhance both the precision and interpretability of rehabilitation assessments, providing clinicians with clinically relevant insights while retaining the ability to capture subtle, complex patterns in the data.

In summary, the two approaches offer complementary advantages. The feature-based method delivers immediate clinical utility, while the deep learning approach provides a foundation for future innovations in unsupervised feature extraction. Together, they represent a step forward in developing robust and accessible tools for assessing UL performance in stroke rehabilitation.

Chapter 6

Conclusion

The primary objective of this thesis is to develop and validate a simplified categorization system for UL performance using IMU sensor data. By assigning stroke patients to one of three performance groups (low, medium, or high), this system provides clinicians with actionable insights into real-world patient functioning. The development and evaluation of two distinct clustering approaches represent a significant step toward achieving this goal.

The first approach uses human-engineered features extracted from raw sensor data. This method effectively produces valid UL performance groups with clear and interpretable distinctions between the three categories. The second approach applies a deep learning model trained directly on 15-minute windows of the raw sensor time series, capturing more complex patterns in the data. Although this method has shown the potential to enhance traditional UL performance assessments, the final classification results did not produce a satisfactory solution.

This work establishes the validity of the feature-based clustering approach as a robust tool for evaluating overall UL performance, while highlighting the potential of deep learning methods in rehabilitation research. It provides a foundation for future studies to refine and validate these methods and explore new applications.

6.1 Key Findings

Feature-Based The results of the feature-based approach demonstrate that this method produces valid and clinically interpretable groupings of UL performance. The feature set is refined to consist of five features that are both relevant to the final clustering and representative of different qualities of UL performance. When comparing this clustering solution to clinical

assessments, we show that on average it outperforms clinical assessments of capacity across different performance features.

Deep Learning-Based The deep learning-based approach also yields compelling results. On the one hand, the model successfully groups 15-minute activity windows from the same measurement together, highlighting its ability to learn global patient-specific behavior. Additionally, the model encodes a form of impairment within its embeddings, as observable from visual investigations where embeddings are colored based on clinical assessments. Finally, the model captures more complex patterns in the data, such as identifying the affected side of the stroke. These findings suggest that the features learned by the model may extend beyond the information captured by human-engineered features.

6.2 Significance and Contributions

The primary motivation for stroke rehabilitation is to help patients regain the ability to perform daily activities. To assess how well patients function in their everyday lives, clinicians require tools that evaluate UL performance in real-world environments. This thesis simplifies UL performance assessments by introducing a data-driven categorization system. The system provides clinicians with a practical and interpretable tool for evaluating therapy impact and collaboratively setting meaningful improvement goals with patients.

Unlike traditional clinical assessments, which are often limited to controlled environments, this work shifts the focus toward real-world performance, offering clinicians a more actionable and realistic understanding of a patient's rehabilitation progress. By categorizing UL performance into three interpretable groups, the proposed system bridges the gap between complex sensor data and the practical needs of clinicians, enabling more effective communication of progress and goals with patients.

This work also represents an important step toward improving and showing the potential of data-driven methods in stroke rehabilitation. The feature-based clustering approach offers an accessible and interpretable solution that can be easily integrated into existing clinical workflows, reducing reliance on time-intensive evaluations. At the same time, the deep learning-based approach demonstrates the potential to uncover subtle and complex movement patterns, laying the groundwork for future innovations in rehabilitation assessments.

Furthermore, the proposed methods could be generalized to assess motor function in other neurological conditions, as well as monitor motor impairments in aging populations or evaluate physical therapy outcomes. By

enabling continuous, objective, real-world performance tracking, this work lays the foundation for similar systems in diverse settings.

6.3 Limitations and Future Work

While this thesis makes significant contributions to the development of a simplified categorization system for UL performance, several limitations highlight areas for future improvement.

Dataset Size and Diversity One of the primary limitations lies in the dataset's size and homogeneity. The dataset used in this study is relatively small and exclusively includes stroke patients, which may limit the generalizability of the findings to other populations. A larger, more diverse dataset that includes individuals with other neurological conditions or healthy controls is essential for evaluating the broader applicability of the proposed methods. Additionally, while the dataset includes measurements from multiple time points, we did not explore changes over time. Future work could focus on leveraging longitudinal data to predict rehabilitation trajectories, offering valuable insights into recovery dynamics and enabling more personalized rehabilitation strategies.

Real-World Validation Another key limitation is the lack of validation in real-world clinical settings. Although promising results were achieved in this study, further studies are needed to evaluate how these tools integrate into clinical workflows. Ensuring that the categorization system provides actionable and meaningful insights for clinicians is critical to its adoption.

Feature Calculation and Selection The feature-based approach relies on human-engineered features derived from sensor data, which capture simple relationships. However, these features may overlook more complex aspects of UL performance. Additionally, the large number of available performance features increases the risk of focusing on irrelevant features. Selecting an optimal feature set remains challenging, and the proposed feature set should undergo critical testing and refinement to ensure its reliability and relevance.

Deep Learning Unsupervised learning captures more complex patterns in the data, but these patterns may not always align with the desired objectives. For example, our experiments show that the model's embeddings predominantly capture which side of the body is affected by the stroke, which interferes with the clustering algorithm's ability to represent UL performance. This highlights the need for advanced preprocessing techniques or other deep learning approaches to ensure the models primarily learn patterns related to UL performance. Additionally, while impairment-related patterns

are evident in the embeddings, the clinical applicability of these patterns requires validation. Future work could investigate and validate the use of deep learning for other related tasks in rehabilitation.

Combining the Two Approaches Finally, while this thesis explores feature-based and deep learning-based clustering approaches separately, combining the strengths of both methods could lead to a more robust and well-rounded system. A hybrid approach that integrates human-engineered features with deep learning embeddings has the potential to improve interpretability while capturing subtle, complex patterns in the data.

6.4 Closing Remarks

This thesis advances UL performance assessment in stroke rehabilitation by introducing a data-driven categorization system that simplifies real-world impairment experienced by stroke patients. By offering objective insights into patient performance and progress, these methods enhance clinical decision-making and enable more personalized rehabilitation strategies.

This thesis demonstrates the potential of traditional feature-based methods and modern machine learning approaches, offering a novel pathway for improving UL performance assessments. By integrating data-driven tools into clinical practice, this work lays the groundwork for more precise, patient-centered rehabilitation strategies that can transform how clinicians assess and monitor real-life recovery of stroke patients.

Beyond stroke recovery, the proposed methods could generalize to other neurological conditions, aging-related impairments, and physical therapy outcomes. By enabling continuous, real-world performance tracking, this work contributes to the development of precise, personalized rehabilitation strategies that improve patient outcomes. It underscores the transformative potential of data-driven tools in modern healthcare and sets the stage for future innovations at the intersection of data science and clinical care.

Acknowledgment I would like to thank the Medical Data Science group at ETH and the Lake Lucerne Institute for providing me with the opportunity and resources to conduct my bachelor thesis. I am also deeply grateful to my supervisors, Alain Ryser and Johannes Pohl, for their valuable feedback and guidance throughout this project. Their support has been essential in shaping this work. This project has been a highly rewarding experience, allowing me to contribute to the growing intersection of data science and clinical practice. By bridging the gap between data-driven innovations and patient-centered rehabilitation, I hope that this work inspires further advances in tools that improve the quality of care and outcomes for stroke patients.

Bibliography

[1] World Health Organization, *Global Health Estimates: Life expectancy and leading causes of death and disability*, 2024. [Online]. Available: https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates.

[2] I. J. de Rooij, M. M. Riemens, M. Punt, J.-W. G. Meijer, J. M. Visser-Meily, and I. G. van de Port, "To What Extent is Walking Ability Associated with Participation in People after Stroke?" *Journal of Stroke and Cerebrovascular Diseases*, vol. 30, no. 11, p. 106 081, Nov. 2021, ISSN: 10523057. DOI: 10.1016/j.jstrokecerebrovasdis.2021.106081.

[3] K. J. Waddell, R. L. Birkenmeier, M. D. Bland, and C. E. Lang, "An exploratory analysis of the self-reported goals of individuals with chronic upper-extremity paresis following stroke," *Disability and Rehabilitation*, vol. 38, no. 9, pp. 853–857, Apr. 2016, ISSN: 0963-8288. DOI: 10.3109/09638288.2015.1062926.

[4] G. Kwakkel *et al.*, "Motor rehabilitation after stroke: European Stroke Organisation (ESO) consensus-based definition and guiding framework," *European Stroke Journal*, vol. 8, no. 4, pp. 880–894, Dec. 2023, ISSN: 2396-9873. DOI: 10.1177/23969873231191304.

[5] K. J. Waddell *et al.*, "Does Task-Specific Training Improve Upper Limb Performance in Daily Life Poststroke?" *Neurorehabilitation and Neural Repair*, vol. 31, no. 3, pp. 290–300, Mar. 2017, ISSN: 1545-9683. DOI: 10.1177/1545968316680493.

[6] World Health Organization, *International Classification of Functioning, Disability, and Health: Children \& Youth Version: ICF-CY*. Geneva, Switzerland: World Health Organization, 2007.

[7] K. Baker, S. J. Cano, and E. D. Playford, "Outcome Measurement in Stroke," *Stroke*, vol. 42, no. 6, pp. 1787–1794, Jun. 2011, ISSN: 0039-2499. DOI: 10.1161/STROKEAHA.110.608505.

[8] C. E. Lang, K. J. Waddell, J. Barth, C. L. Holleran, M. J. Strube, and M. D. Bland, "Upper Limb Performance in Daily Life Approaches Plateau Around Three to Six Weeks Post-stroke," *Neurorehabilitation and Neural Repair*, vol. 35, no. 10, pp. 903–914, Oct. 2021, ISSN: 1545-9683. DOI: 10.1177/15459683211041302.

[9] A. David, T. Subash, S. K. M. Varadhan, A. Melendez-Calderon, and S. Balasubramanian, "A Framework for Sensor-Based Assessment of Upper-Limb Functioning in Hemiparesis," *Frontiers in Human Neuroscience*, vol. 15, Jul. 2021, ISSN: 1662-5161. DOI: 10.3389/fnhum.2021.667509.

[10] G. Uswatte, C. Giuliani, C. Winstein, A. Zeringue, L. Hobbs, and S. L. Wolf, "Validity of Accelerometry for Monitoring Real-World Arm Activity in Patients With Subacute Stroke: Evidence From the Extremity Constraint-Induced Therapy Evaluation Trial," *Archives of Physical Medicine and Rehabilitation*, vol. 87, no. 10, pp. 1340–1345, Oct. 2006, ISSN: 00039993. DOI: 10.1016/j.apmr.2006.06.006.

[11] C. E. Lang, J. Barth, C. L. Holleran, J. D. Konrad, and M. D. Bland, "Implementation of Wearable Sensing Technology for Movement: Pushing Forward into the Routine Physical Rehabilitation Care Field," *Sensors*, vol. 20, no. 20, p. 5744, Oct. 2020, ISSN: 1424-8220. DOI: 10.3390/s20205744.

[12] J. Barth, K. R. Lohse, J. D. Konrad, M. D. Bland, and C. E. Lang, "Sensor-Based Categorization of Upper Limb Performance in Daily Life of Persons With and Without Neurological Upper Limb Deficits," *Frontiers in Rehabilitation Sciences*, vol. 2, 2021, ISSN: 26736861. DOI: 10.3389/fresc.2021.741393.

[13] J. Pohl *et al.*, "Consensus-Based Core Set of Outcome Measures for Clinical Motor Rehabilitation After Stroke—A Delphi Study," *Frontiers in Neurology*, vol. 11, Sep. 2020, ISSN: 1664-2295. DOI: 10.3389/fneur.2020.00875.

[14] A. R. Fugl-Meyer, L. Jääskö, I. Leyman, S. Olsson, and S. Steglind, "The post-stroke hemiplegic patient. 1. a method for evaluation of physical performance.," *Scandinavian journal of rehabilitation medicine*, vol. 7, no. 1, pp. 13–31, 1975, ISSN: 0036-5505.

[15] J. See *et al.*, "A Standardized Approach to the Fugl-Meyer Assessment and Its Implications for Clinical Trials," *Neurorehabilitation and Neural Repair*, vol. 27, no. 8, pp. 732–741, Oct. 2013, ISSN: 1545-9683. DOI: 10.1177/1545968313491000.

[16] D. J. Gladstone, C. J. Danells, and S. E. Black, "The Fugl-Meyer Assessment of Motor Recovery after Stroke: A Critical Review of Its Measurement Properties," *Neurorehabilitation and Neural Repair*, vol. 16, no. 3, pp. 232–240, Sep. 2002, ISSN: 1545-9683. DOI: 10.1177/154596802401105171.

[17] H. T. Jordan, J. Che, W. D. Byblow, and C. M. Stinear, "Fast Outcome Categorization of the Upper Limb After Stroke," *Stroke*, vol. 53, no. 2, pp. 578–585, Feb. 2022, ISSN: 0039-2499. DOI: 10.1161/STROKEAHA.121.035170.

[18] D. S. de Lucena, O. Stoller, J. B. Rowe, V. Chan, and D. J. Reinkensmeyer, "Wearable sensing for rehabilitation after stroke: Bimanual jerk asymmetry encodes unique information about the variability of upper extremity recovery," in *2017 International Conference on Rehabilitation Robotics (ICORR)*, IEEE, Jul. 2017, pp. 1603–1608, ISBN: 978-1-5386-2296-4. DOI: 10.1109/ICORR.2017.8009477.

[19] I. Boukhennoufa, X. Zhai, V. Utti, J. Jackson, and K. D. McDonald-Maier, "Wearable sensors and machine learning in post-stroke rehabilitation assessment: A systematic review," *Biomedical Signal Processing and Control*, vol. 71, p. 103 197, Jan. 2022, ISSN: 17468094. DOI: 10.1016/j.bspc.2021.103197.

[20] C. Werner *et al.*, "Using Wearable Inertial Sensors to Estimate Clinical Scores of Upper Limb Movement Quality in Stroke," *Frontiers in Physiology*, vol. 13, May 2022, ISSN: 1664-042X. DOI: 10.3389/fphys.2022.877563.

[21] R. Felius *et al.*, "Exploring unsupervised feature extraction of IMU-based gait data in stroke rehabilitation using a variational autoencoder," *PLOS ONE*, vol. 19, no. 10, e0304558, Oct. 2024, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0304558.

[22] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," Feb. 2020.

[23] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A Survey on Contrastive Self-supervised Learning," Oct. 2020.

[24] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Machine Learning*, vol. 53, no. 1/2, pp. 23–69, 2003, ISSN: 08856125. DOI: 10.1023/A:1025667309714.

[25] J. Shlens, "A Tutorial on Principal Component Analysis," Apr. 2014.

[26] K. P. Murphy, "Machine learning - a probabilistic perspective," in *Adaptive computation and machine learning series*, 2012. [Online]. Available: https://api.semanticscholar.org/CorpusID:17793133.

[27] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org.

[28] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735.

[29] M. H. Hoonhorst, R. H. Nijland, J. S. van den Berg, C. H. Emmelot, B. J. Kollen, and G. Kwakkel, "How Do Fugl-Meyer Arm Motor Scores Relate to Dexterity According to the Action Research Arm Test at 6 Months Poststroke?" *Archives of Physical Medicine and Rehabilitation*, vol. 96, no. 10, pp. 1845–1849, Oct. 2015, ISSN: 00039993. DOI: 10.1016/j.apmr.2015.06.009.

[30] J. C. BRØND, L. B. ANDERSEN, and D. ARVIDSSON, "Generating ActiGraph Counts from Raw Acceleration Recorded by an Alternative Monitor," *Medicine & Science in Sports & Exercise*, vol. 49, no. 11, pp. 2351–2360, Nov. 2017, ISSN: 0195-9131. DOI: 10.1249/MSS.0000000000001344.

[31] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training Recurrent Neural Networks," Nov. 2012.

[32] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," Nov. 2017.

[33] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," Jul. 2019.

[34] J. K. Rinehart, R. D. Singleton, J. C. Adair, J. R. Sadek, and K. Y. Haaland, "Arm Use After Left or Right Hemiparesis Is Influenced by Hand Preference," *Stroke*, vol. 40, no. 2, pp. 545–550, Feb. 2009, ISSN: 0039-2499. DOI: 10.1161/STROKEAHA.108.528497.

Appendix A

Appendix

All code used in this project was written in Python, with deep learning modules implemented using PyTorch Lightning. The analyzes presented in this thesis were performed using custom code, except for the calculation of activity counts, which was performed using code provided by Brønd et al. [30]. and accessible in their Github repository https://github.com/jbrond/ActigraphCounts.

The full project code, including the deep learning models, preprocessing steps, and analysis scripts, is available in the GitHub repository associated with this project https://github.com/Jamoser123/StrokeImpairmentClustering.

Table A.1 shows all performance features used in our analysis.

Table A.1: Set of all UL performance features, available for the clustering

UL Performance Feature	Quality	Description
Use Duration (Non-)Affected	Duration	How long the respective limb was active, in minutes
Use Duration Unilateral (Non-)Affected	Duration	How long the respective limb was active and the other was inactive, in minutes
Use Duration Bilateral	Duration	How long both of the limbs are active at the same time, in minutes
Use Duration Ratio	Duration Symmetry	The ratio of minutes the affected limb was active to nonaffected
Use Duration Unilateral Ratio	Duration Symmetry	The ratio of time only the affected limb was active to the time only the nonaffected limb was active
Median AC (Non-)Affected	Magnitude	Median activity counts over the whole measurement in the respective limb, in activity counts
Mean AC (Non-)Affected	Magnitude	Mean activity counts over the whole measurement in the respective limb, in activity counts
Peak AC (Non-)Affected	Magnitude	Peak activity counts in the respective limb, in activity counts
Median AC Unilateral (Non-)Affected	Magnitude	Median activity counts over the whole measurement where only the respective limb was active, in activity counts
Mean AC Unilateral (Non-)Affected	Magnitude	Mean activity counts over the whole measurement where only the respective limb was active, in activity counts
Median AC Bilateral	Magnitude	Median activity counts summed up over both limbs, in activity counts
Mean AC Bilateral	Magnitude	Mean activity counts summed up over both limbs, in activity counts
Mean Magnitude Ratio	Magnitude Symmetry	Mean log of the ratio of activity counts of the affected to the nonaffected limb, capped at ± 7
Median Magnitude Ratio	Magnitude Symmetry	Median log of the ratio of activity counts of the affected to the nonaffected limb, capped at ± 7
Acceleration Variability (Non-)Affected	Variability	Standard deviation of activity counts in the respective limb, in activity counts
Acceleration Variability Ratio	Variability Symmetry	Standard deviation of activity counts in the affected limb to the nonaffected limb
Mean Jerk (Non-)Affected	Quality of Movement	Mean jerk in the respective limb, in jerk
Median Jerk (Non-)Affected	Quality of Movement	Median jerk in the respective limb, in jerk
Mean Jerk Ratio	Quality of Movement Symmetry	The mean ratio of jerk of the affected side to the nonaffected
Median Jerk Ratio	Quality of Movement Symmetry	The median ratio of jerk of the affected side to the nonaffected

All ratios have additionally been calculated and incorporated using an alternative method, namely calculating $\frac{val_{aff} - val_{nonaff}}{val_{aff} + val_{nonaff}}$.

Feature-Based

Clustering with Multiple Principal Components

This section presents visualizations of the clustering results using three principal components as embeddings or raw features. The aim is to provide an intuitive understanding of the impact of using multiple principal components (PCs) or no dimensionality reduction at all.

The clustering results, visible in Figure A.1, demonstrate that using three principal components changes only very few assignments. Clustering on the raw features directly introduces more changes, while still agreeing overall with the two principal component approach. For simplicity and the advantage of having condensed features as a base for clustering, we continued with two principal components.

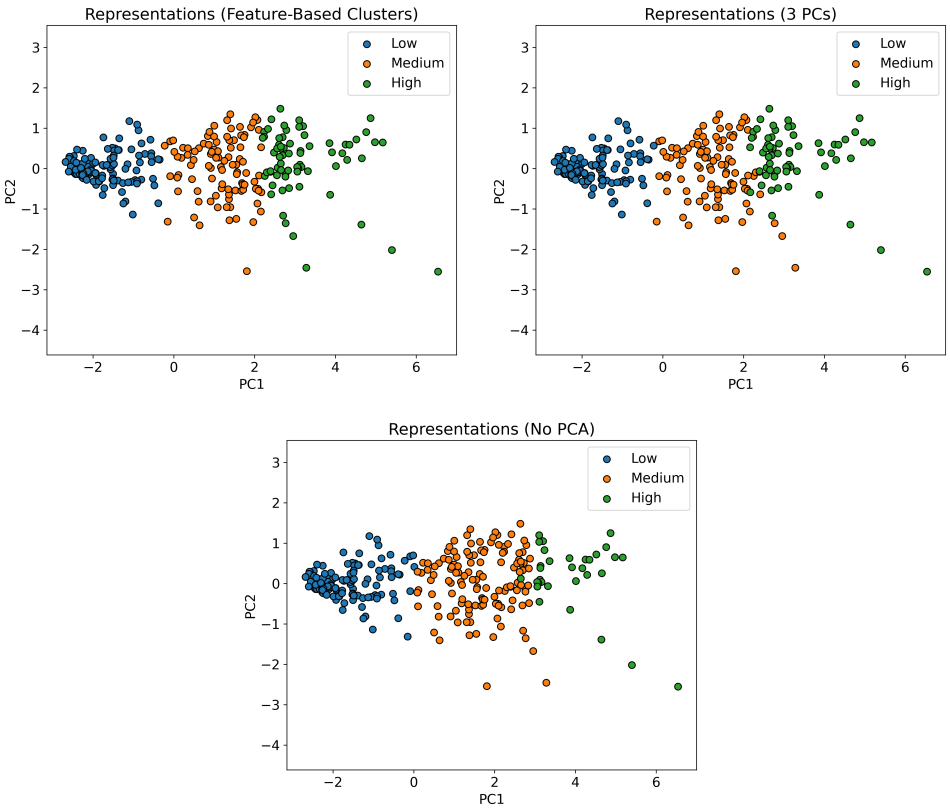

Figure A.1: Clustering results using 2 or 3 principal components or with clustering on the final feature set directly. The clusters agree overall, and the final representation choice does not drastically change the clustering.

Clustering using GMM's

This section compares clustering results using Gaussian Mixture Models (GMM) and K-Means. While K-Means has proven effective in capturing and distinguishing impairment levels, GMM often failed to provide the same level of robust separation. In certain feature sets, GMM was able to produce clusters of comparable quality to K-Means. However, in our final feature set, it struggled to achieve similar performance.

As illustrated in Figure A.2, the GMM clusters align poorly with the expected structure of the data. As shown earlier in the thesis, impairment is primarily encoded along the first principal component. In contrast, the K-Means approach maintains distinct clusters that correspond more closely to the expected levels of impairment. This pattern was consistent across other feature sets as well. While GMM, being more flexible, often overfit to alternative patterns in the embeddings, K-Means, with its more rigid structure, better captured the underlying data structure.

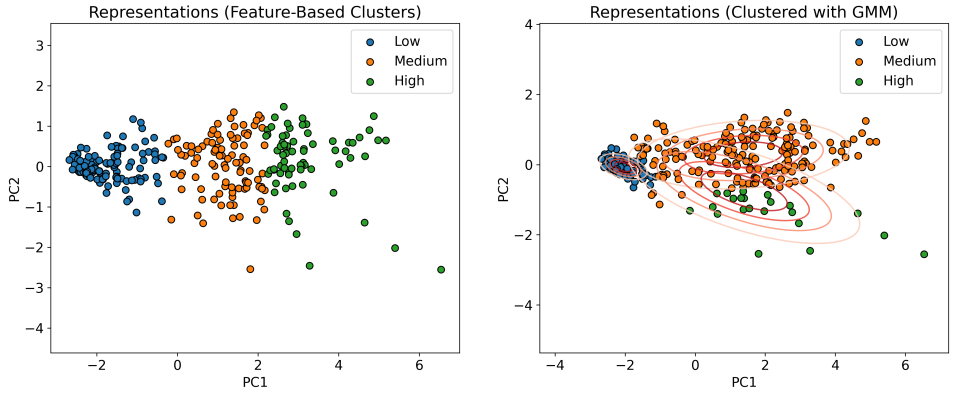

Figure A.2: Clustering using K-Means or GMM's on the performance representations (consisting of the first 2 PC's on the final feature set). It is visible that GMM's while more flexible, did not capture the wanted structure. K-Means on the other hand performed well, as shown in the thesis.

Deep Learning

Efficient Data Storage

Efficient data access is critical for both training and evaluation. Initially, the IMU data was stored in CSV format, with a single CSV file corresponding to each day of patient measurements. However, this format posed significant performance constraints due to the frequent opening and closing of files.

To address these limitations, the data was converted to the Hierarchical Data Format (HDF). HDF is a binary file format designed for the efficient storage

and retrieval of large datasets. By collecting all patient data into a single HDF file, we significantly improved data access performance. This optimization enabled faster training and evaluation of the model.

Ultimately, this leaves us with a single HDF file, that contains time series data for all patient measurements, as well as a list of indices for valid slices.

Preliminary Classifier for Clinical Clustering

Before employing a contrastive learning approach, we first aimed to determine whether meaningful information could be extracted from the raw IMU sensor data. To test this, we developed a classifier to predict clinical clusters based on activity slices. This approach served as a foundational step, helping us identify suitable hyperparameters and select an appropriate architecture for the backbone model, which we later used in the contrastive learning approach.

We experimented with two architectures: an LSTM and a 1-dimensional Convolutional Neural Network (CNN). Early experiments showed better performance with the LSTM, prompting us to proceed with this architecture. The idea was to train the LSTM such that a simple MLP applied to the hidden layer could accurately classify the clinical cluster of each activity slice. We used the same train/validation splitting strategy as in the final model.

The best classifiers achieve an average accuracy of around 75%, demonstrating that the models can approximate capacity from activity slices. To further investigate the model's robustness, we performed cross-validation by varying the train/validation splits. We observed significant variation in validation accuracy depending on the fold. Specifically, folds with fewer patients in the medium-capacity category achieved higher accuracy (up to 80%), while folds with more patients in this category performed worse. This is expected, as distinguishing low or high capacity is likely more straightforward, whereas medium capacity may require more contextual information than is available in a single activity slice.

Figures A.3, A.4, and A.5 provide insights into the model's performance. Figure A.3 shows the training and validation accuracy curves across folds using good hyperparameters. Figure A.4 presents the class-wise accuracy for clinical clusters and the confusion matrix for a fold with good accuracy, highlighting the model's strong performance in distinguishing between clinical categories. Figure A.5 shows similar plots for a fold with lower accuracy, where the model struggles to classify medium-capacity slices, underlining the challenges of this category.

Although this classifier was a preliminary step, it demonstrated that the raw IMU sensor data contains meaningful information, achieving a solid average

accuracy of 75%. This result highlights the model's ability to approximate capacity from activity slices and validates the feasibility of using deep learning techniques for this task.

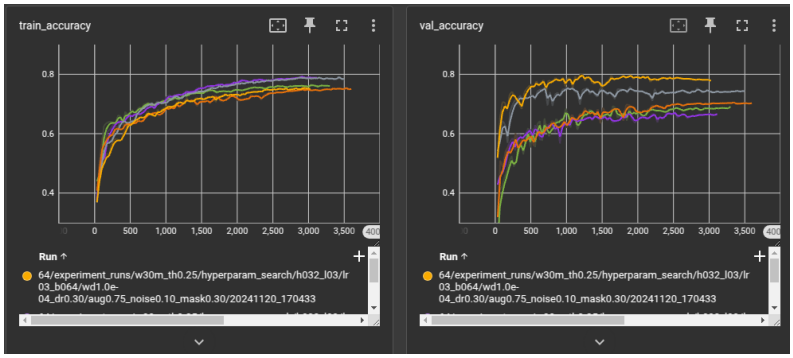

Figure A.3: Training and validation accuracy across folds using good hyperparameters.

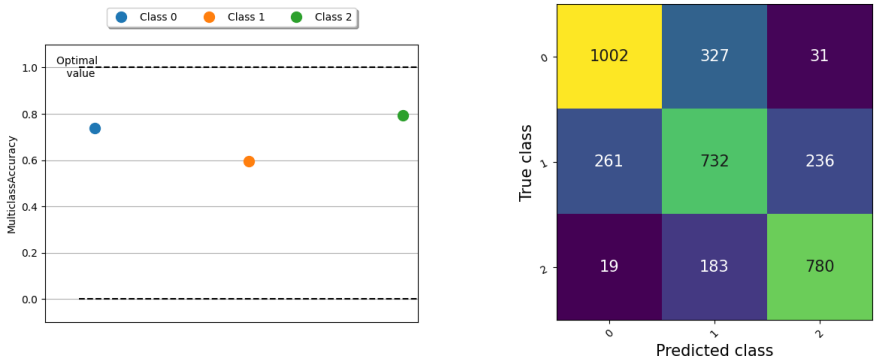

Figure A.4: (Left) Class-wise accuracy of the validation set for clinical clusters (Class 0: low capacity, Class 1: medium capacity, Class 2: high capacity). The model performs better for low and high capacities compared to the medium category. (Right) Confusion matrix for a fold with good accuracy. The model performs well at distinguishing between clinical clusters.

Training

Algorithm 1 shows the training algorithm used to train our encoder network $f(\cdot)$ in a contrastive fashion. Table A.2 shows all hyperparameters given to Optuna for hyperparameter optimization, and the respective final set of hyperparameters.

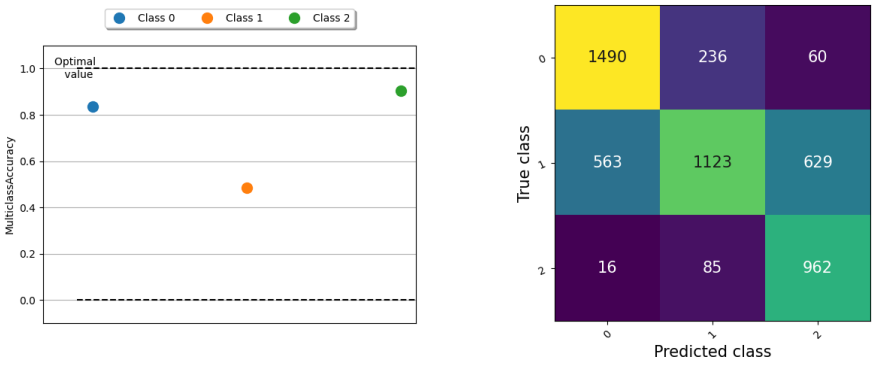

Figure A.5: (Left) Class-wise accuracy of the validation set for clinical clusters in a fold with low accuracy. (Right) Confusion matrix for the same fold. The model struggles with distinguishing medium-capacity slices. Note that the total number of slices can differ, since measurements provide a different number of valid slices.

Algorithm 1 Training Algorithm used to train the slice encoder $f(\cdot)$

Input: Batch with N Measurements, constant τ, Sampling Strategy $\mathcal{S}$ structure of f, g
for sampled minibatch $\{m_k\}_{k=1}^{N}$ **do**
  for $k \in \{1, \ldots, N\}$ **do**
    Sample two augmented slices $\tilde{x}, \tilde{y} \sim \mathcal{S}(m_k)$
    $h_{2k-1} = f(\tilde{x})$, $h_{2k} = f(\tilde{y})$
    $z_{2k-1} = g(h_{2k-1})$, $z_{2k} = g(h_{2k})$
  end for
  Compute similarities $s_{i,j} = z_i^\top z_j / (\|z_i\| \|z_j\|)$
  Define loss $\ell(i,j) = -\log \frac{\exp(s_{i,j}/\tau)}{\sum_{k \neq i} \exp(s_{i,k}/\tau)}$
  $\mathcal{L} = \frac{1}{2N} \sum_{k=1}^{N} \left[\ell(2k-1, 2k) + \ell(2k, 2k-1)\right]$
  Update f and g to minimize $\mathcal{L}$
end for
Return: Encoder network $f(\cdot)$

Table A.2: Hyperparameters Used for Optimization and Their Final Choices

Hyperparameter	Values	Final Choice	Description
Window size	{15m, 30m, 60m}	30m	The time window size for segmenting data. A larger window captures more temporal information but may smooth out fine-grained patterns.
Activity threshold	{10%, 25%, 50%}	10%	Minimum activity threshold to determine whether a segment contains enough information for analysis.
Batch size	{32, 64, 128}	64	Number of samples per batch during training. A larger batch size stabilizes gradient updates but may require more memory.
Dropout	{0.1, 0.3, 0.5}	0.3	Dropout rate applied to prevent overfitting by randomly setting a fraction of input units to zero during training.
Learning rate	{1e-4, 1e-3}	1e-4	Initial learning rate for the AdamW optimizer. Controls the step size for updating model weights.
Weight decay	{1e-6, 1e-4}	1e-6	Weight decay (L2 regularization) to prevent overfitting by penalizing large weights in the model.
Hidden dimension (LSTM)	{16, 32, 64, 128}	64	Number of hidden units in each LSTM layer. Determines the model's capacity to capture sequential dependencies.
Hidden dimension (MLP)	{15, 32, 64}	32	Number of hidden units in the multilayer perceptron (MLP) used after the LSTM layers.
Number of layers (LSTM)	{2, 3, 4}	3	Number of stacked LSTM layers to increase the model's representational power for temporal features.
Augmentation noise	{0.05, 0.1, 0.2}	0.1	Standard deviation of Gaussian noise added to the input as a form of data augmentation.
Shuffle segments	{3, 4, 5}	3	Number of segments to shuffle during augmentation to encourage robustness in sequential learning.
Mask ratio	{0.15, 0.3}	0.15	Fraction of input segments masked during training for self-supervised learning.
Temperature	{0.05, 0.07, 0.1, 0.5}	0.07	Temperature parameter τ controlling the sharpness of the softmax distribution in the NT-Xent loss.
Epochs	{50, 100, 200}	100	Total number of training epochs. Determines how many times the model sees the entire dataset during training.
Early stopping	{5, 10, 15}	10	Number of epochs without improvement in validation loss before halting training to prevent overfitting.

Vanilla SimCLR sampling strategy

The vanilla SimCLR approach defines a positive pair as two augmented views of the same data sample. We applied this approach to our dataset by sampling a single slice from a measurement and applying two random augmentations to generate the positive pair. In contrast, our approach samples two different slices from the same measurement as the positive pair. All other training steps remained the same.

Interestingly, the vanilla SimCLR sampling strategy resulted in a much lower validation loss for the model (approximately three times lower). However, upon investigating the embeddings, we observed significant differences in the learned representations.

Figure A.6 shows the slice embeddings from 10 measurements, colored by measurement. On the left, we have our proposed sampling strategy, and on the right, the vanilla SimCLR sampling approach. Our strategy results in embeddings where slices from the same measurement are grouped together. This is desirable, as we expect measurements from the same individual, taken over the course of a day, to reflect similar impairment levels. In contrast, the vanilla SimCLR sampling results in more scattered embeddings, indicating that the approach may not effectively capture this expected consistency within measurements.

Furthermore, as shown in Figure A.7, when we color the embeddings using clinical clustering (low, medium, high capacity), our approach demonstrates a much clearer hierarchical structure. This indicates that our sampling strategy better preserves the level of impairment in the learned representations, whereas the vanilla SimCLR approach fails to capture this pattern effectively. While the lower loss achieved by the vanilla SimCLR strategy might seem promising, it ultimately results in embeddings that perform worse in capturing impairment.

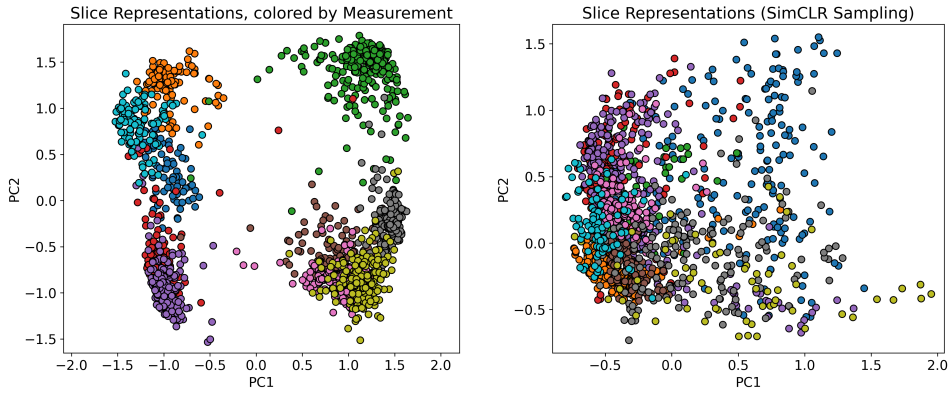

Figure A.6: Comparison of slice embeddings from 10 measurements using different sampling strategies. Left: Our sampling strategy (two different augmented slices from the same measurement). Right: Vanilla SimCLR sampling (two augmented views of the same slice). Our approach groups slices from the same measurement, reflecting expected consistency, while SimCLR results in scattered embeddings.

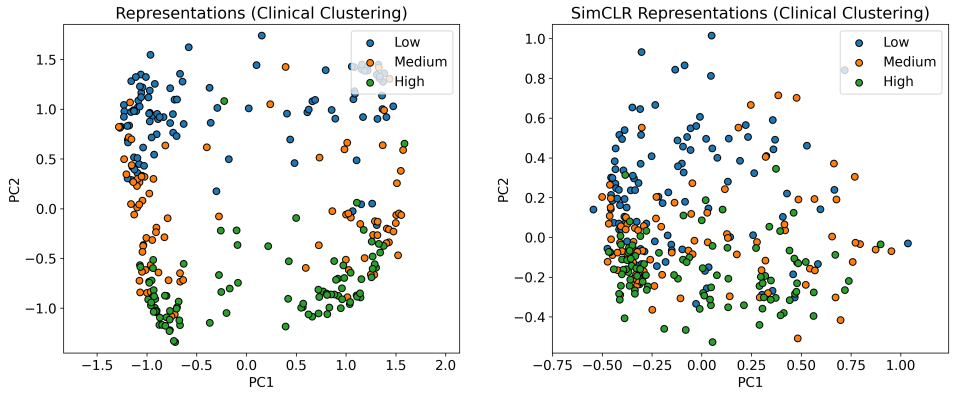

Figure A.7: Comparison of embeddings by clinical clusters (low, medium, high capacity) using different sampling strategies. Left: Our sampling strategy shows a clear clinical hierarchy. Right: Vanilla SimCLR sampling fails to preserve these clinical distinctions, resulting in less structured embeddings.

Breaking the Dominant Structure

During our exploration, we observed that the deep learning model primarily focused on distinguishing which side was affected rather than mainly capturing UL performance. To address this issue, we attempted several techniques to break the dominant structure in the data. These included:

- Randomly swapping the affected and nonaffected sides during training.
- Training the model using data from only one wrist, randomly choosing affected or nonaffected.
- Incorporating an adversarial approach during training, where an additional adversary was used to encourage the model to learn embeddings that did not allow for easy distinction between the affected and nonaffected sides.

While these methods helped reduce the dominance of the structure to some extent, none of them were sufficient to fully eliminate it. As a result, we chose not to include these techniques in the main work.

It is important to note that we initially focused on addressing this issue without considering the mirroring aspect of the data. The suggestion to explore mirroring came during the presentation of this project and was a significant revelation. This idea intuitively aligns well with the problem and will be the primary focus of our future work. I would like to express my gratitude for this insightful suggestion.

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Eigenständigkeitserklärung

Die unterzeichnete Eigenständigkeitserklärung ist Bestandteil jeder während des Studiums verfassten schriftlichen Arbeit. Eine der folgenden drei Optionen ist in Absprache mit der verantwortlichen Betreuungsperson verbindlich auszuwählen:

○ Ich bestätige, die vorliegende Arbeit selbständig und in eigenen Worten verfasst zu haben, namentlich, dass mir niemand beim Verfassen der Arbeit geholfen hat. Davon ausgenommen sind sprachliche und inhaltliche Korrekturvorschläge durch die Betreuungsperson. Es wurden keine Technologien der generativen künstlichen Intelligenz[1] verwendet.

○ Ich bestätige, die vorliegende Arbeit selbständig und in eigenen Worten verfasst zu haben, namentlich, dass mir niemand beim Verfassen der Arbeit geholfen hat. Davon ausgenommen sind sprachliche und inhaltliche Korrekturvorschläge durch die Betreuungsperson. Als Hilfsmittel wurden Technologien der generativen künstlichen Intelligenz[2] verwendet und gekennzeichnet.

◉ Ich bestätige, die vorliegende Arbeit selbständig und in eigenen Worten verfasst zu haben, namentlich, dass mir niemand beim Verfassen der Arbeit geholfen hat. Davon ausgenommen sind sprachliche und inhaltliche Korrekturvorschläge durch die Betreuungsperson. Als Hilfsmittel wurden Technologien der generativen künstlichen Intelligenz[3] verwendet. Der Einsatz wurde, in Absprache mit der Betreuungsperson, nicht gekennzeichnet.

Titel der Arbeit:

Clustering Upper Limb Impairment in Stroke Patients Using Data from Inertial Measurement Units

Verfasst von:
Bei Gruppenarbeiten sind die Namen aller Verfasserinnen und Verfasser erforderlich.

Name(n):	**Vorname(n):**
Moser	Janic

Ich bestätige mit meiner Unterschrift:
- Ich habe mich an die Regeln des «Zitierleitfadens» gehalten.
- Ich habe alle Methoden, Daten und Arbeitsabläufe wahrheitsgetreu und vollständig dokumentiert.
- Ich habe alle Personen erwähnt, welche die Arbeit wesentlich unterstützt haben.

Ich nehme zur Kenntnis, dass die Arbeit mit elektronischen Hilfsmitteln auf Eigenständigkeit überprüft werden kann.

Ort, Datum	**Unterschrift(en)**
Zürich, 19.01.2025	

Bei Gruppenarbeiten sind die Namen aller Verfasserinnen und Verfasser erforderlich. Durch die Unterschriften bürgen sie grundsätzlich gemeinsam für den gesamten Inhalt dieser schriftlichen Arbeit.

[1] z. B. ChatGPT, DALL E 2, Google Bard
[2] z. B. ChatGPT, DALL E 2, Google Bard
[3] z. B. ChatGPT, DALL E 2, Google Bard