FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Unsupervised and Supervised Strategies for Identification of Activity Profiles in Prosthesis Users

Érica Ferreira Gomes

MASTER'S THESIS

Master in Bioengineering - Biomedical Engineering

Supervisor: Hélder Oliveira, PhD, FCUP | INESC TEC

Co-supervisor: Sofia Assis, Adapttech

July 24, 2024

Unsupervised and Supervised Strategies for Identification of Activity Profiles in Prosthesis Users

Érica Ferreira Gomes

Master in Bioengineering - Biomedical Engineering

July 24, 2024

Abstract

Every year, tens of thousands of amputations are performed in the United States, with a growing number of individuals living without a lower limb. To identify the activity level of these individuals, a classification system called K-level system was created. The attribution of this level of activity is made by the clinician and people assessed with a higher K-level are eligible for reimbursement of more technologically advanced prosthetic components, impacting their quality of life. The attribution of K-levels is done taking into account self-reports, the experience of clinicians and tests carried out in a clinical environment, which is subjective and does not accurately reflect the level of activity of individuals during their daily lives. Therefore, this work aims to address these problems, using machine learning to identify K-levels based on data collected by Adapttech's Motio StepWatch System during the daily lives of amputees.

This project includes the analysis of dataset characteristics, preprocessing, and data augmentation techniques, exploration of clustering and supervised classification methods and testing of various feature reduction and selection methods. The results were obtained using two datasets: only activity metrics or the combination of activity metrics with personal characteristics.

In unsupervised analysis, the results obtained for K-means and GMM were similar, with 3 being the best number of clusters obtained for all approaches. In none of the cases did the clusters correspond exactly to the clinical K-levels. However, it is possible to verify a separation of the data regarding the values of activity metrics that are associated with each level of activity.

In the supervised analysis, for multiclass classification the best result was obtained for RF with PCA using the dataset with only activity metrics (test F1-score of 0.58, accuracy of 0.70, and balanced accuracy of 0.67). For binary classification, the best results were obtained for SVM with the combined dataset (test F1-scores, accuracy, and balanced accuracy all with value 1.0), with a much higher performance than best performance obtained for multiclass task.

This project highlights the potential of machine learning techniques using activity metrics collected during the daily lives of amputees to support and improve the effectiveness of K-levels assessment, which has proven to be particularly useful in differentiating between K2 and higher levels of activity. The explored approaches could, in the future, be incorporated into a decision support system for clinicians and feedback systems for patients.

Keywords: Amputation, K-level, Prosthetic Components, Machine Learning

Resumo

Anualmente, são feitas dezenas de milhares de amputações nos Estados Unidos, havendo cada vez mais indivíduos que vivem sem um membro inferior. Para identificação do nível de atividade desses indivíduos, foi criado um sistema de classificação denominado K-level system. A atribuição desse nível de atividade é feita pelo clínico, sendo que pessoas avaliadas com um K-level mais elevado são elegíveis para um financiamento de componentes de próteses mais complexos tecnologicamente, tendo impacto na sua qualidade de vida. A atribuição de K-levels faz-se tendo em conta self-reports, a experiência dos clínicos e testes feitos em ambiente clínico, o que é subjetivo e não reflete de forma precisa o nível de atividade dos indivíduos no seu dia-a-dia. Assim, este trabalho surgiu com o intuito de resolver estes problemas, tendo como objetivo utilizar técnicas de machine learning para identificação de K-levels usando dados recolhidos pelo Adapttech's Motio StepWatch System durante o dia-a-dia de pessoas amputadas.

Este projeto inclui a análise das características do dataset, técnicas de pré-processamento e data augmentation, exploração de métodos de clustering e de classificação supervisionada e o teste de vários métodos de redução e seleção de features. Os resultados foram obtidos usando dois datasets: só métricas de atividade ou a combinação de métricas de atividade com características pessoais.

Relativamente à análise não supervisionada, os resultados obtidos para K-means e GMM foram semelhantes, sendo que o melhor número de clusters obtido foi 3 para todas as abordagens testadas. Para nenhum dos casos os clusters corresponderam exatamente aos K-levels clínicos. No entanto, é possível verificar uma separação dos dados tendo em conta os valores de métricas de atividade que estão associadas a cada nível de atividade.

Na análise supervisionada, para classificação multiclasse o melhor resultado foi obtido para RF com PCA usando o dataset só com métricas de atividades (test F1-score de 0.58, accuracy de 0.70 e balanced accuracy de 0.67). Já para classificação binária, os melhores resultados foram obtidos para SVM com o dataset combinado (test F1-scores, accuracy, e balanced accuracy todos com valor 1.0), com uma performance muito superior à obtida para a tarefa multiclasse.

Este projeto evidencia o potencial de técnicas machine learning usando métricas de atividade recolhidas durante o dia-a-dia de pessoas amputadas para apoio e melhoria da eficácia da avaliação do K-levels, tendo-se revelado particularmente útil na diferenciação entre K2 e níveis de atividade mais elevados. As abordagens exploradas podem no futuro ser incorporadas num sistema de apoio de decisão para os clínicos e sistemas de feedback para os pacientes.

Keywords: Amputação, K-level, Componentes Protésicos, Machine Learning

Acknowledgements

Ao longo destes 5 anos, varios foram os momentos e pessoas que me marcaram e tornaram este percurso memorável. Sem qualquer ordem de importância:

Ao orientador Prof. Hélder Oliveira, pela disponibilidade para orientar esta tese e por todos os conselhos, sempre com a sua perspectiva mais científica.

À orientadora Sofia Assis por ter sido um apoio enorme durante todo o processo de desenvolvimento da tese. Estarei para sempre grata pelo incrível acompanhamento e por tudo o que aprendi com a Sofia. Às restantes pessoas que fazem parte da Adapttech, por me terem recebido tão bem e com um ambiente tão agradável.

Aos meus pais, por sempre me terem dado as melhores condições possíveis para eu ter um bom percurso, por confiarem em mim e por sempre terem compreendido quando não podia ir a casa aos fins-de-semana ou quando ía mas passava o tempo todo a estudar. À minha irmã Bárbara por acreditar tanto em mim e nas minhas capacidades, muito mais do que eu própria acredito. Ao Carlos por ser a companhia das cervejas e pelas longas conversas. Ao Mike e à Kiara, os meus dois parceiros de 4 patas, que sempre que chego a casa me recebem com o maior entusiasmo e animam os meus dias com lambidelas, brincadeiras e alguma destruição.

À Paula, que foi o meu braço direito e me acompanhou em todas as fases de desenvolvimento desta tese. Desde os almoços no jardim ate aos momentos de stress e preocupação que levaram a horários de sono muito estranhos (amiga, acordar às 4h30 da manhã não é saudável).

Aos Bitchus por terem tornado estes 5 anos muito mais divertidos. Por mais jantares organizados à última da hora e mais jogos de futebol.

Às French Belgian Fries, amizades que se solidificaram com o ERASMUS e que me deram muito apoio e boa disposição ao longo deste último ano caótico.

A Metal&Bio por me ter proporcionado tantos bons momentos, por muitas vezes me ter tirado da minha zona de conforto e ter tornado o meu percurso académico ainda mais especial.

To Professor Peter Janssen, PhD student Ophelie Saussus, PhD student Sofie De Schrijver, and everyone at the Laboratory for Neuro- & Psychophysiology, KU Leuven, where I spent one semester doing an internship. It was an amazing experience that taught me a lot and made me grow so much.

À FEUP por me ter feito crescer e ensinado que depois da tempestade vem sempre a bonança.

Érica Gomes

"The human spirit, you know what they say about the human spirit?
Is is harder than a rock and more delicate than a flower petal."

Nadia Hashimi, The Pearl that Broke Its Shell

Contents

List of Figures

List of Tables

Abbreviations

2MWT	Two Minute Walk Test
6MWT	Six Minute Walking Test
10MWT	10 Meter Walk Test
AK	Above-Knee
AMP	Amputee Mobility Predictor
AMPPRO	Amputee Mobility Predictor With Prosthesis
AMPnoPRO	Amputee Mobility Predictor Without Prosthesis
BIC	Bayesian Information Criterion
BK	Below-Knee
CNNs	Convolutional Neural Networks
DTW	Dynamic Time Warping
FN	False Negative
FP	False Positive
GAN	Generative Adversarial Network
GMM	Gaussian Mixture Model
ICR	Instantaneous Center of Rotation
KBest	Select K-Best
KNN	K-Nearest Neighbors
LOGO CV	Leave-One-Group-Out Cross-Validation
MFCL	Medicare Functional Classification Level
MPK	Microprocessor-Controlled Knee
NC	Not Conclusive
NND	Non-Normal Distribution
PCA	Principal Component Analysis
PEQ	Prosthetic Evaluation Questionnaire
PLUS-M	Prosthetic Limb Users Survey of Mobility
PVD	Peripheral Vascular Disease
RF	Random Forest
RFE	Recursive Feature Elimination
RPE	Rating of Perceived Exertion
RMSE	Root Mean Squared Error
SACH	Solid Ankle Cushion Heel
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
TF	Transfemoral
TN	True Negative
TP	True Positive
TT	Transtibial
TUG	Timed Up and Go Test
WSS	Within-Cluster-Sum of Squared Errors

Chapter 1

Introduction

Lower-limb amputation is one of the oldest surgically performed procedures, with its origins tracing back to prehistoric times [1]. Today it is mainly a clinical practice, but that was not always the case. Archaeological findings and ancient drawings provide glimpses into the early practices of limb amputation, where Neolithic humans survived not only traumatic amputations, but also ritualistic and punitive limb-loss. Additionally, mummies from different cultures with cosmetic replacements for amputated extremities have been unearthed, indicating that mobility was important even in ancient cultures.

Amputation of limbs evolved during periods of war, where battlefield injuries required surgical interventions. Prior to the development of gunpowder in the 14th century, the majority of injuries were caused by weapons used for stabbing, clubbing, and cutting [2]. However, these early amputation procedures had significant risks, such as hemorrhage, shock, and sepsis. Since anesthesia was yet to be discovered, patients had to undergo the procedure while being conscious and aware. Often, patients were physically restrained by assistants and offered alcohol to relieve their pain.

Despite these challenges, the safety and effectiveness of amputation treatments were progressively enhanced by advancements in surgical techniques, such as hemostasis and improved perioperative conditions. A crucial development was the discovery of antisepsis, which dramatically reduced the mortality rate related to limb amputation. Further advancements in antibiotics, anesthesia, and pain control during the second half of the 20th century made amputation surgery a safer and more controlled procedure [2].

In the current days, amputation continues to be a significant medical intervention, with an estimation of 30000 to 40000 amputations being performed annually in the United States. The prevalence of individuals that live with limb loss is on the rise, with approximately 1.6 million in 2005 and expecting that number to escalate to 3.6 million until 2050 [3].

Having delved into the historical evolution of lower-limb amputation, it is clear that this surgical intervention has changed significantly over time. While in ancient times amputations were often a result of battlefield injuries and other traumatic events, the contemporary landscape of amputations reflects different causes. Nowadays, lower-extremity amputations are typically undertaken for the following reasons [1]:

- **Peripheral Vascular Disease (PVD):** The primary cause of limb amputation in the United States is ischemic disease, predominantly peripheral vascular disease. This condition often affects elderly individuals with diabetes mellitus, who are prone to developing peripheral neuropathy leading to trophic ulcers, gangrene, and osteomyelitis. Individuals with diabetes mellitus account for approximately 50% of PVD cases, with an estimated 65,000 lower-extremity amputations performed annually for this group.

- **Trauma:** Despite advancements in safety equipment and limb-salvage surgery techniques, traumatic limb loss still occurs due to industrial and motor vehicle accidents. These accidents often result in high-grade open fractures along with soft-tissue loss, nerve damage, among other complications that cannot be adequately reconstructed.

- **Tumors:** The primary goal in treating malignant bone tumors is to remove the lesion with minimal risk of recurrence. Even though limb-salvage surgery has become the preferred approach, amputation remains necessary in cases where limb salvage is not feasible or presents an increased risk of local recurrence compared to amputation.

- **Infections:** Under certain conditions, the treatment of sepsis with vasoconstrictor agents may inadvertently lead to vessel occlusion and subsequent extremity necrosis. When this happens, it may be necessary to amputate the limb.

- **Congenital limb deficiency:** Congenital limb deficiencies resulting from absence or malformation of limbs account for a small percentage of amputations, even though they are less common. These procedures are primarily performed in the pediatric population due to failure of partial or complete formation of a limb segment.

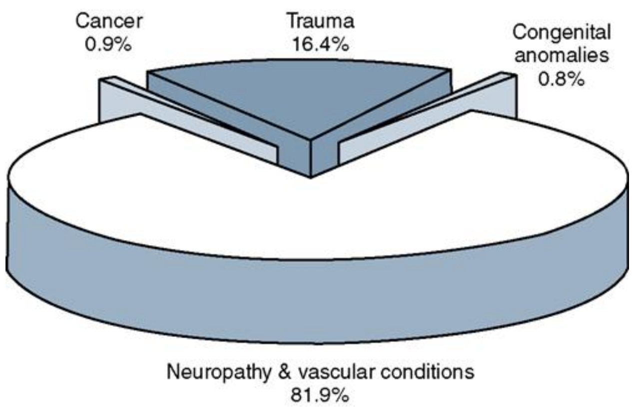


Figure 1.1: Causes of amputation by percentage. Extracted from Etiology of Amputation.

Diabetes in conjunction with dysvascular disease constitutes the predominant cause of limb losses in the United States, representing approximately 81.9% of cases [1], followed by trauma at

[1]https://musculoskeletalkey.com/etiology-of-amputation/. Accessed on 2024-04-01.

16.4% (Figure 1.1). While the specific percentages of each cause may vary across different regions, the overall order remains consistent worldwide. Exceptions exist in conflict-heavy regions, where trauma predominates. However, globally, the primary causes of lower-limb amputations are consistent.

Lower-limb amputations are categorized into various levels according to the location of the amputation site and the extent of limb loss [2]. It expresses the degree of amputation which influences the type of prosthetic device and guides treatment decisions. The most common levels of lower-limb amputation are represented in Figure 1.2.

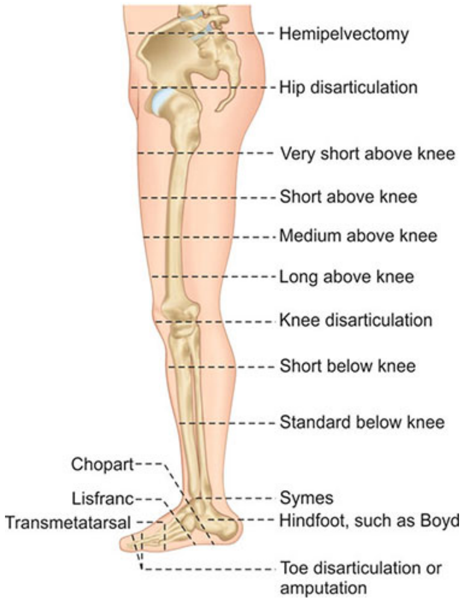


Figure 1.2: Levels of lower-limb amputation. Extracted from Lower-Extremity Amputations: Background, Indications, Contraindications.

The level of lower-limb amputation represents a delicate balance between ambulatory capacity and tissue healing ability. The higher the level of amputation, from below-knee (BK) to above-knee (AK), the greater the energy expenditure required for walking. This progression leads to a reduction in walking speed and a rise in oxygen consumption. Moreover, the location of the amputation must ensure proper circulation to promote the healing process. Proper circulation is vital for the formation of a skin envelope capable of withstanding the axial and shear stresses imposed by the prosthetic device [2]. This trade-off is important to take into account while deciding on the best level of amputation, since it directly affects the individual's mobility and functional abilities.

After an amputation, the patient does not receive a permanent prosthesis immediately. There is a period of time when the amputee is without a prosthesis or with a temporary prosthesis, to allow the remaining limb to fully shrink. A few weeks or months after amputation, a permanent prosthesis is built. The components of a lower limb prosthesis are:

[2] https://emedicine.medscape.com/article/1232102-overview. Accessed on 2024-04-01.

- **Socket:** is the part of the prosthesis that encases the residual limb. Initially, a temporary socket (preparatory socket) may be used while the surgical incision heals and the swelling decreases. The prosthetic socket protects the residual limb while enabling weight-bearing and load distribution.

- **Suspension mechanism:** this component attaches the prosthesis to the residual limb. It can be done by using belts, wedges, locks or suction to achieve secure attachment.

- **Knee Joint (if needed):** an articulating knee joint may be incorporated into the prosthesis if necessary, offering options ranging from a simple single-axis hinge joint to more complex polycentric axis designs. These mechanisms provide flexibility, allowing knee flexion while maintaining stability during weight-bearing.

- **Pylon:** also known as the shell, the pylon connects the socket to the terminal device. Recent advancements have introduced dynamic pylons capable of axial rotation and energy absorption. They can be in endoskeletal or exoskeletal, based on functional and aesthetic considerations.

- **Terminal Device / Foot:** this component is the final part of the prosthesis. Typically, it resembles a traditional foot. However, there are some specialized devices available for athletes and individuals with specific needs. Usually, ankle function is integrated into the terminal device. Yet, separate ankle joints may be beneficial in some cases, even though they may add weight and require increased energy expenditure and limb strength to control.

(a) Transtibial amputee

(b) Transfemoral amputee

Figure 1.3: The two most common types of lower-limb prostheses. Extracted from Below knee prostheses (a) and iFIT Transfemoral Prosthesis (b).

The type of prosthetic legs recommended for lower extremity patients depends on whether they require devices for above or below knee amputations. The two predominant types of lower-limb prostheses are transtibial (TT) and transfemoral (TF) prosthesis (Figure 1.3).

Every component within a prosthesis serves a distinct purpose, from offering support and stability to facilitating movement and functionality. By carefully selecting the appropriate components based on the individual's needs and activity level, prosthetists can enhance the effectiveness and comfort of the prosthesis. This also contributes to the improvement of the overall rehabilitation process and long-term outcomes for amputees.

The United States' Medicare Functional Classification Level (MFCL) or K-level system was adopted by Medicare in 1995 to describe the mobility and functional abilities of individuals with lower-limb amputations. It classifies patients into five activity levels, from K0 to K4 [4]. Beyond analyzing the functional abilities of individuals, it also helps to determine the appropriate prosthetic components for each person. The MFCL levels definitions are:

- **K0:** Does not have the ability or potential to ambulate or transfer safely with or without assistance and a prosthesis does not enhance quality of life or mobility.
- **K1:** Has the ability or potential to use a prosthesis for transfers or ambulation on level surfaces at fixed cadence. Typical of the limited and unlimited household ambulator.
- **K2:** Has the ability or potential for ambulation with the ability to traverse low-level environmental barriers such as curbs, stairs, or uneven surfaces. Typical of the limited community ambulator.
- **K3:** Has the ability or potential for ambulation with variable cadence. Typical of the community ambulator who has the ability to traverse most environmental barriers and may have vocational, therapeutic, or exercise activity that demands prosthetic utilization beyond simple locomotion.
- **K4:** Has the ability or potential for prosthetic ambulation that exceeds the basic ambulation skills, exhibiting high impact, stress, or energy levels, typical of the prosthetic demands of the child, active adult, or athlete.

This classification system was developed to assist clinics in selecting the most suitable prosthetic components based on the individual's potential to be successful with the prosthesis. The main purpose was to ensure that the recommended components align with the individual's functional abilities and rehabilitation goals. The lower limb prosthesis is pivotal to the fulfillment of its user needs, may they be functional, psycho-cognitive, ergonomic or others like cost and durability. Therefore, the prosthetist needs to ensure that the prosthetic device is comfortable, lightweight design, durability, aesthetic appeal, low maintenance requirements, and appropriate degree of mechanical function for the individual's K-level [5].

1.1 Motivation

As will be detailed in Section 2, prosthetic components have characteristics which target specific K-level (or K-level range). Typically, individuals rated with higher K-level by the clinical team will be eligible for reimbursement of more technologically advanced and more functional components [4]. For example, a prosthetic foot with energy return usually target prosthetic user with moderate activity levels, rather than a home ambulator. Such component allows for a more energy efficient gait, however, does not improve patient outcomes while being more expensive than other options.

On the other hand, inappropriate prescription of prosthetic components can have negative effects on amputees. It may cause discomfort, limit mobility and raise the risk of getting hurt. Continuing the above example, the keel of a prosthetic foot with energy return might be too stiff for patients with low activity levels which can be counterproductive. Furthermore, inappropriate prescriptions not only compromise the physical well-being of users but also have financial implications for payers.

There is no standard methodology for assessing patients' K-levels. Typically, prosthetists routinely evaluate amputees' functional level based on a combination of factors gathered through patient interview, such as patient history, comorbidities, residual limb condition, motivation to ambulate, and the K-level guideline descriptors (i.e., variable cadence, capability to clear environmental barriers) [6]. However, this assessment process is susceptible to variability and subjectivity, since each prosthetist may have a slightly different interpretation of the functional level characteristics.

One major limitation is the reliance on patients' self-reports to determine activity levels. Studies have shown that patients are inaccurate at reporting their activity and self-reports have a poor correlation with objective measures of activity, being less reliable than standardized outcome measures [7]. The paper by Godfrey et al. describes a study with 86 lower limb prosthetic users in which they were looking into how walking metrics (like cadence and daily number of steps) correlated with self-perceived changes in walking ability throughout an extended period of time (6 months). They found that there was not a clear correlation between people's self perception and the actual impact on the measure walking ability, nor was there any particular bias towards over- or under-reporting. Therefore, supporting that free-living walking can yield additional insights into the person's potential [8].

There is another relevant question about the difference between an individual's ability to perform at a high functional level during clinical tests and their functional behavior in the community. The functional activity capacity conducted in clinical settings may not accurately represent amputees' actual performance in daily life. Some research indicated that more than half of older people classified with a high clinical K-level exhibit low functional level behavior in the community [9]. The opposite might also be true, individuals who exhibit low functional capacity in clinical tests may engage in higher levels of activity in community environments. This emphasizes the complexity of evaluating functional levels in individuals with limb loss and highlights how crucial it is to consider daily-life functionality when assessing prosthesis users.

Due to the importance of measuring functional activity during daily life to complement traditional clinical assessments in prosthetic care, there is a growing number of studies measuring functional walking performance during daily life [10] [11] [12] [13] [14]. Objective outcomes measures obtained from wearable devices provide valuable information about individuals' actual activity levels, which improves the accuracy of prosthetic prescription. The incorporation of outcome measures into routine practice protocols aims to ensure that each patient is evaluated based on standardized quantitative data, while decreasing reliance on patient self-report and individual prosthetist judgments when assigning K-level. This would improve the overall health care cost-effectiveness and enable patients to receive prosthetic interventions adapted to their unique needs.

The present work is developed at Adapttech (adtechnologies, s. a.) [3]. Adapttech is a biomedical innovation company developing solutions to improve the quality of life of people with physical limitations (lower-limb amputations). Motio StepWatch System, one of its products, aims to revolutionize the evaluation of lower limb prosthetic users. By gathering real-world data reflecting each individual's daily activity, it enables the Rehabilitation Team to make more informed decisions on each individual's prosthetic rehabilitation. This system has two main components:

- **Motio StepWatch device:** remote monitoring device (Figure 1.4a) specifically designed for lower-limb prosthetic users, that gathers patients' real-world daily activities.
- **Motio App:** web-based application used to control the system (Figure 1.4b).

(a) Motio StepWatch device

(b) Motio App

Figure 1.4: Motio StepWatch System. Extracted from www.adapttech.eu

This monitoring device tracks patients' daily activities during a minimum of 7 days and provides clinicians with a report with step data such as cadence, distance, speed and other metrics that allow the clinicians to accurately classify their patients' K-level. The Motio's hardware is built upon the clinically approved and FDA-listed Modus Health StepWatch 4, a robust wearable device that has undergone extensive validation across various research activities. Apart from analyzing the patient's community ambulation data, it facilitates clinicians in acquiring standard outcome measures and performing clinical tests. The Motio StepWatch System empowers clinicians,

[3]https://www.adapttech.eu/. Accessed on 2024-04-01.

insurance companies and everyone involved in the patient's rehabilitation to look at objective and real-world data. This ensures that the k-level assessments are as accurate as possible and that the patient is getting adequate care. Through the utilization of this system, stakeholders can make informed decisions based on quantifiable data, ultimately enhancing the quality and effectiveness of prosthetic interventions.

Despite the advancements in prosthetic care, the subjective nature of K-level assessment and the differences between clinical tests and real-world functionality highlight the need for novel approaches, to enhance prosthetic intervention accuracy and efficacy. One promising direction is the implementation of machine learning techniques into prosthetic care. By analyzing vast amounts of data, machine learning algorithms can provide insights into patterns that may not be immediately identifiable using more conventional methods. It is already used in a variety of areas, such as bioinformatics and image analysis and it has shown advancements by overcoming existing limitations. Machine learning has drawn attention for its ability to efficiently interpret and predict large amounts of data, since it can automatically learn from it.

This presents an opportunity to revolutionize prosthetic design and user experience by leveraging advanced data-driven strategies to identify unique activity profiles among prosthesis users.

1.2 Goal

The main goal of this work is to use machine learning techniques (both supervised and unsupervised learning) for the identification of activity profiles in prosthetics users. By utilizing data collected from individuals' daily life, the aim is to classify the K-level for amputees using machine learning approaches.

While bridging the gap between technology and patient needs, these approaches would support and complement the decision-making process of healthcare professionals in assessing the activity levels among prosthetic users. Through this integration of data analysis techniques, the goal is to provide healthcare professionals with valuable tools to improve the individualized care and rehabilitation of individuals with limb loss.

1.3 Document Structure

This document is structured as follows. In Chapter 2, it is done an explanation of prosthetic components, the reimbursement process, the tools used to support K-level assessments and the data stages associated with Motio StepWatch System. Chapter 3 presents state of the art and techniques for machine learning in different areas of orthotics and prosthetics, sports and medical area. Chapter 4 analyzes the dataset and explains the preprocessing and data augmentation. Chapter 5 approaches unsupervised learning and presents the results associated. In Chapter 6, supervised learning and feature selection methods are explored, showing the results obtained. Lastly, Chapter 7 highlights the conclusions and reflects on possible paths to explore in the future.

Chapter 2

Background

As introduced in the previous chapter, the goal of this project is to investigate the utilization of activity metrics collected in prosthetic users to classify prosthetic users K-levels. In this section, our goal is to better understand the importance and impact of such assessment, by analysing:

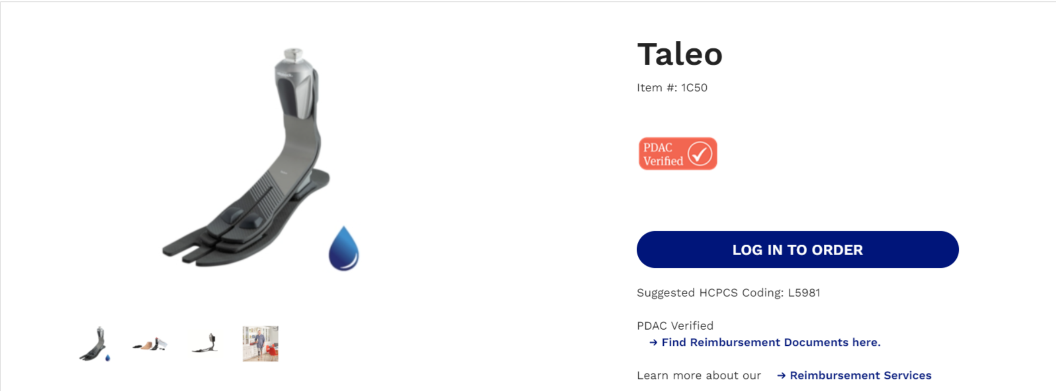
- the categories in each segment of prosthetic component and how they target low or higher activity levels of patients;
- the reimbursement process (particularly in the United States);
- the tools clinicians use to support their assessment with the payers;
- the data stages that the current Motio StepWatch system goes through.

The MFCL (or K-level system) is a widely scale used by clinicians to assess the mobility for individuals with lower-limb loss, also evaluating the person's rehabilitation potential. It is composed of five activity levels (K0 to K4), each with a very specific definition outlining the individual's functional activity requirements.

The K-level system significantly influences the selection of prosthetic components for each individual. Higher k-levels are associated with more advanced and complex prosthetic components [4].

2.1 Lower-limb Prosthetic Components

Lower-limb prosthetic components (including knees, feet and liners) have a wide variety of options. Each option has different features and functionalities to address specific needs of individuals with different levels of functional ability (Figure 2.1). Healthcare professionals explore the multiple options to customize prescriptions according to each individual's needs, preferences and lifestyle. Some factors that are taken into consideration are activity level, terrain preference, comfort and durability.



(a) Mechanical foot

Specifications

Activity Level	K3
Maximum Body Weight	330 lbs (150 kg)
Heel Height	Normal footshell (N) 10 +/- 5mm, Slim footshell (S) 15 +/- 5 mm
Footshell Shape	Normal (N), Slim (S)
~ Min. Clearance Height (Normal, Size 26 cm)	158 mm (6 1/4")
Product weight (w/ footshell) (Size 26 cm)	24.34 oz / 690 g
Sizes	22 cm - 30 cm
Footshell Color	Beige (4), Brown (15)
Side	Right (R), Left (L)
Standard Warranty	3 years

(b) Specifications

Figure 2.1: Specifications of a mechanical foot. Extracted from Ottobock.

2.1.1 Prosthetic Knees

There is a wide variety of prosthetic knees in the market, but they can be divided into two main categories: mechanical and microprocessor knees [1] [2].

2.1.1.1 Mechanical Knees

The least expensive knees are the mechanical ones, which can be single-axis or multiaxis/polycentric. For all prosthetic knees, it is necessary to have some kind of stability mechanism, which can either be manual or a weight-activated locking system. Additionally, they require a system to control the flexion and extension movement [3]. This can be accomplished with friction or a hydraulic/pneumatic control.

[1] https://www.amputee-coalition.org/resources/prosthetic-knee-systems/. Accessed on 2024-04-25.

[2] https://www.waramps.ca/pdf/english-site/ways-we-help/artificial-limbs-and-devices/lower-limb/knees-for-adults.pdf. Accessed on 2024-04-25.

[3] https://www.physio-pedia.com/Prosthetic_Knees. Accessed on 2024-04-25.

Single-axis knees (Figure 2.2a), or simple hinge type knees, execute a simple rotation around the knee axis during flexion or extension. They can feature either manual or automatic blocking of the flexion, which is useful for users with poor muscle strength [1]. To increase stability while standing, a manual lock can be incorporated. Additionally, constant friction control can also be added to prevent the leg from swinging too quickly. For regular prosthetic fitting of amputees with adequate muscle control and/or in situations with constrained financial resources, a suitable option is the use of knees without blocking. Single-axis knees have the advantage of being very simplistic, durable, light, and economical. However, there are some disadvantages, as users rely on their own muscle power in the limb to keep the knee stable during heel contact and standing.

Polycentric knees (Figure 2.2c) have multiple axes of rotation, typically 4 or 7 axes. Its Instantaneous Center of Rotation (ICR) is positioned considerably higher and posterior compared to the mechanical axes when the knee is in extension (Figure 2.2b). This characteristic is an advantage by significantly enhancing knee stability against involuntary flexion during heel strike and transitioning into stance phase. It also allows an easy swing and sitting down. The ICR also reduces the prosthesis length at the start of toe-off and enables foot clearance. This prosthetic knee is suitable for patients with potential for independent mobility in their homes, community and during extra activity [15]. Additionally, pneumatic or hydraulic elements can be incorporated for a possible variation in walking speed. However, polycentric knees have some disadvantages. Firstly, in comparison to single axis knees, these are heavier and have more parts that need servicing. Furthermore, most polycentric knees lack stance flexion resistance, so cannot yield during sitting, ramps, or stairs. Consequently, users need to actively generate a knee extension moment during the stance phase to prevent potential knee buckling and consequent falls.

Manual locking knees (Figure 2.2d) allow automatic locking of the knee joint with weight bearing, but users have the option to manually lock them [1]. Such prosthetic knees are indicated for K1 patients or debilitated individuals with difficulties in controlling the prosthetic knee. However, a disadvantage is that the hip joint must be raised to originate clearance between the foot and the ground during walking.

Safety knees (Figure 2.2a), with weight-activated stance control, have incorporated a constant friction system in the knee. This mechanism applies braking force upon weight-bearing, preventing knee buckling [2]. The rest of the time, the knee freely swings until pressure is applied. This kind of knee offers remarkable stability, which makes it suitable for first-time users, individuals that are quickly fatigued, older or less active people [1]. Safety knees can compensate for users who may inadvertently put weight on a partially bent knee, as the knee will brake if that happens. During sitting, users must take the weight of the leg to allow it to bend, which means that they will not be able to use the prosthetic side during sitting motion. Consequently, the natural knee flexion during toe-off does not happen. Due to friction, the patient movement will be slower and with smaller steps.

Pneumatic or hydraulic knees (Figure 2.2e) have a pneumatic or hydraulic component integrated into a single-axis or polycentric knee. These components control the swinging motion of the knee and enable walking speed variation [2]. As speed increases, the valve in the cylinder closes gradually, restricting the air or fluid flow to control knee flexion, enabling faster walking. Conversely, in slower walking, the fluid or air flow is less restricted, allowing for more knee flexion and slower gait. Pneumatic control compresses air as the knee flexes, storing energy to release it during knee extension. Spring coils can be added to enhance control during gait. Pneumatics or hydraulics are also incorporated into computerized prosthetic knees, with hydraulics being more suitable for highly active individuals. These systems enable weight to be maintained on the prosthetic leg while sitting, with the knee assisting in this movement. The resistance in the knee allows climbing down step-over-step maintaining weight on the leg before and during the motion. In comparison to pneumatics, hydraulic knees are heavier, more expensive, and require more maintenance [1].

2.1.1.2 Microprocessor-controlled Knees

More technologically advanced knees, known as microprocessor-controlled knees (MPKs) (Figure 2.2f), have computerized control systems and receive feedback from sensors in the knee joint and/or foot. This feedback allows the knee to adjust the range and speed of flexion in real-time to suit individual needs. These systems work similarly to an hydraulic system, where there is a computer regulating the opening and closing of valves, changing the flow of hydraulic fluid [2]. Typically, MPK incorporates a hydraulic actuator, while a powered knee has a motor actuator. In a motor-powered knee, knee extension is powered to help standing up from a seated position, with controlled resistance being provided when sitting down. Some components also integrate stumble recovery mechanisms to prevent falls.

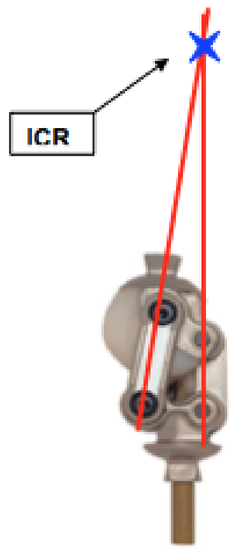
These knees have adjustable settings that may be modified using a computer or mobile device, to accommodate different walking/activity modes. These knees allow step-over-step descent on stairs, enhancing user mobility and safety. Once users become used to the control dynamics of these knees, they require less cognitive effort. However, they have some disadvantages. Learning how to use it entails an initial steep learning curve and a commitment to gait re-education. Despite their advanced features, MPKs are the most expensive option available. Their battery needs to be charged periodically. Additionally, these knees are the heavier ones and are compatible only with a specific foot selection. Lastly, environmental conditions, such as exposure to water, heat, or cold, as well as kneeling, can potentially damage these prosthetic devices.

2.1.2 Prosthetic Feet

Regarding prosthetic feet, they are expected to provide passive plantar flexion in early stance, maintain a neutral position in mid stance and enable toe hyperextension in late stance. However, this optimal function is not achieved in all cases, existing a wide variety of possible ankle-foot combinations. To categorize this, we can distinguish between two main types of feet: non-articulated and articulated feet [4].

[4]https://www.physio-pedia.com/Prosthetic_Feet. Accessed on 2024-04-25.

(a) Ottobock single axis knee joint, brake titanium, 3R49



(b) Instant Centre of Rotation location

(c) Ottobock Modular Polycentric EBS Knee Joint 3R60

(d) Ottobock Prosedo 3R31

(e) Ottobock's knee joint, monocentric, with rotation hydraulics, 3R80

(f) Ottobock Genium X3

Figure 2.2: Examples of prosthetic knees in the market. Extracted from Ottobock and Prosthetic Knees.

2.1.2.1 Non-articulated Feet

Solid Ankle Cushion Heel (SACH) (Figure 2.3a) is the less dynamic foot and it has no ankle articulation. In this foot, the shock absorption primarily occurs in the heel, while the forefoot stimulates foot dorsal flexion. There are SACH feet with different degrees of heel amortization. Although no longer the standard of care, SACH foot is now used in initial fittings/preparatory prostheses or for individuals with low activity levels [5].

There can be dynamic SACH feet, that incorporate a flexible keel within the foot to allow forefoot flexion when under load. When the load is released, the keel returns to its original shape and gives the user a push [5].

[5]https://www.pw.co.nz/products/categories/feet-and-ankles. Accessed on 2024-04-25.

Their keel is shorter, measuring approximately three quarters of the length of many higher activity carbon feet, which results in a poorer or faster transition from toe-off to sound foot. Additionally, the heel height is fixed. This foot is easily worn out by active users because of its simple one bolt attachment failure.

2.1.2.2 Articulated Feet

Articulated feet have a specific range of motion in the sagittal plane to mimic both plantar and dorsiflexion, while also combining movement in the sagittal and frontal planes to add inversion and eversion. This attempt to replicate anatomical ankle movements serves both cosmetic and functional purposes: after heel contact, immediate forefoot contact occurs and promotes ground reaction forces that stabilize the prosthetic knee in extension.

The **single-axis foot** (Figure 2.3b) allows plantar and dorsiflexion [5], by ensuring that the entire sole of the foot makes ground contact quickly, which enhances prosthetic knee stability [6]. This foot is suitable mostly for prosthetic users that need stability around the prosthetic knee (transfemoral). However, it adds weight to the prosthesis and demands more frequent maintenance compared to a non-articulated foot.

The **multi-axis foot** (Figure 2.3c) enables plantar and dorsiflexion, as well as side-to-side movements [6]. It absorbs a portion of the stress of walking, reducing pressure in the socket and protecting the skin and prosthesis from wear. However, it is heavier and more expensive compared to SACH or single-axis feet. Additionally, it exhibits poor shock absorption and energy return. Small bearing surfaces and bumpers may result in wear and noise and, as they wear, they may become looser, compromising user control.

Feet with energy return (Figure 2.3d), also known as dynamic response feet, are designed with a keel designed to function like a spring, typically made from molded carbon fiber plates [5]. This design enhances energy response during the toe-off phase, replicating the natural push-off of a human foot. These feet store and release energy during gait. The resulting "push off" promotes a more symmetrical gait pattern. Some models include a split-toe design for mimicking inversion and eversion movements. Due to their dynamic response, these feet typically encourage individuals to increase their activity levels from moderate to high. This type of prosthetic feet can also reduce impact and stress on the sound leg during walking [6]. Additional features such as a shock absorber or torque adapter can be incorporated into the prosthetic foot, to decrease stress on the residual limb by absorbing the impact and rotation. There are also versions with adjustable ankles, allowing users to wear shoes with different heel heights. Feet with energy return are generally more expensive than previous models and the keel may be stiff for individuals with low activity levels.

Microprocessor feet (Figure 2.3e) are controlled by a computer and sensors. The computer processes information coming from the sensors regarding the prosthesis, limb and environment, adjusting the ankle's speed and range of motion as needed [6]. Actions such as ankle positioning,

[6]https://www.amputee-coalition.org/resources/prosthetic-feet/. Accessed on 2024-04-25.

setting damping resistance and driving the ankle motor during the stance phase can be controlled. Sensors may include goniometers, accelerometers, gyroscopes, and torque sensors. These feet can adapt to different environments or situations by changing the speed or range of motion. Ankle alignment can also adjust for different situations, such as wearing heeled shoes, carrying a backpack, or walking uphill. They offer active push-off and dorsiflexion. To make adjustments to specific foot actions in different situations, these feet can be connected to a mobile device or a computer to customize settings. These prosthetic feet have some disadvantages: are battery-powered, so they require charging; electronic components may fail or break; their additional features add weight; these prosthetic feet are more expensive than the others; are not waterproof and are sensitive to environmental conditions

There are also **Motor Powered Ankles**, which, in addition to the microprocessor, have a motor for active push and dorsiflexion.

(a) SACH foot

(b) Ottobock single axis foot 1H40

(c) Ottobock 1M10 multiaxial foot

(d) Ottobock Taleo without shell

(e) Ottobock Meridium

Figure 2.3: Examples of prosthetic feet in the market. Extracted from Ottobock.

2.1.3 Prosthetic Liners

Prosthetic liners are a layer between the residual limb and the prosthetic socket, having a crucial role in amputee safety and comfort [7]. This component is crucial since the soft tissue of the residual

[7] https://mcopro.com/blog/an-overview-of-prosthetic-liner-types-benefits-more/. Accessed on 2024-04-25.

limb is not used to bear loads. Additionally, liners also enhance the prosthesis's functionality by contributing to the suspension of the prosthesis.

Despite the wide variety of prosthetic liners available, there is no standard practice to determine the appropriate liner for each patient, so clinicians have to rely on their personal experience.

Even though the material properties have been extensively studied, their impact on functionality is not well understood. Liners help distribute pressures over the residual limb, but pain tolerance varies with the type of soft tissue.

Typically, prosthetic liners need to be replaced every 6 months. To reduce the stress on the product and extend its lifespan, it is common practice to provide the patients with two liners simultaneously. This practice also facilitates cleaning and maintenance, since liners are not breathable[7].

2.1.3.1 Materials

Various types of liners can be differentiated based on their materials, offering patients a range of options according to their specific needs.

The first attempts of liners were **foam liners**. Typically made from open and closed cell foams formed around the residual limb. Even though they are still used in some practices, these liners are now considered outdated due to their inferior suspension, durability, and cushioning compared to modern alternatives. Pelite is a closed-cell foam used as a prosthetic liner. However, studies showed weaknesses in its cell structure when vacuum-formed, indicating possible limitations in its effectiveness.

Thermoplastic elastomer liners stand out for their softness and forgiveness. They are particularly well-suited for individuals who are sensitive or that were amputated recently, as well as those with low to moderate activity levels that are looking for comfort [7]. However, their softness also represents a negative point, since the gels that hold the liner together usually have a shorter lifespan compared to other alternatives. One example of this type of liner is the Össur Iceform (Figure 2.4a), whose price is between €218.95 and €233.95.

Silicone liners are a reasonable option for prosthetic liners because they have a well-balanced combination of softness and durability [7]. This type of liner not only retains some of the softness of thermoplastic elastomer liners, but also provides increased durability similar to urethane liners. These characteristics make silicone liners a better option for users who are moderately active and need supplies that last longer. Even though some developments have been made to introduce new silicone formulas with enhanced cushioning properties, silicone does not offer as much cushioning for individuals with sensitive limbs as other materials. However, this is not necessarily a drawback. For some active amputees, the sensation of firmness is more valuable than softness during high-impact sports. An example of a silicone liner is the Össur Iceross Comfort liner (Figure 2.4b), designed for low to moderate activity users, and it is marked with a price of €466.95.

Urethane liners are often used in prosthetics with vacuum sockets. They are the least soft liners. However, they provide flow properties, allowing the gel within the liner to flow away from high pressure regions. This process results in more balanced distribution of pressure and therefore a more even fit [7]. The flow properties of urethane liners make them excellent for high-activity wearers who require a combination of durability and adaptability. An example of a urethane liner is the Ottobock Anatomic 3D PUR liner (Figure 2.4c).

Mineral oil gel liners are highlighted for their remarkable softness and forgiving properties [8]. These liners provide skin-nourishing vitamins to the wearer and a close and contoured fit. They are particularly useful for people with bony residual limb or hypersensitive areas. However, it is important to keep in mind that this type of liners is the least durable and requires more frequency replacement. An example of a mineral oil gel liner is the Silipos Duragel Pin liner (Figure 2.4d), marked with a price of €185.95. Specifically designed for low to moderate activity levels, this liner is especially suitable for people with sensitive skin or diabetes.

(a) Össur Iceform

(b) Össur Iceross

(c) Ottobock Anatomic 3D PUR

(d) Silipos Duragel Pin liner (coin for reference)

Figure 2.4: Examples of prosthetic liners in the market. Extracted from https://amputeestore.com/.

2.1.3.2 Shape

Prosthetic liners are available in two different shapes: cylindrical (uniform) and conical forms. These shapes suit the majority of residual limb shapes and sizes. However, for individuals with residual limbs that do not fit these two shapes, it is possible to obtain tailor-made or custom prosthetic liners [8]. These custom liners are specifically made to fit the unique contours of the user's residual limb and to provide the best comfort and functioning possible.

[8]https://amputeestore.com/blogs/amputee-life/prosthetic-liners-silicone-gel-differences. Accessed on 2024-04-25.

2.1.4 Suspensions

The prosthetic liner selection is influenced by the suspension used. The suspension is responsible for keeping the socket secured to the residual limb during movement [9]. An efficient suspension improves energy transfer and control of the prosthesis, while decreasing discomfort and abrasions. There are various suspension methods that suit the different needs and activity levels:

- **Cuffs, straps and belts (Figure 2.5):** are the traditional method of auxiliary suspension [10], keeping the socket securely attached during movement. This method can vary from a single soft strap to a rigid band and it is valued for being an inexpensive, reliable and durable method.

Figure 2.5: Ottobock TF Suspension Belt. Extracted from Ottobock

- **Lanyard (Figure 2.6):** involves attaching a cord or strap to the lower end of the liner [10]. When the individual inserts the residual limb with the liner, the cord or strap passes through a small hole in the bottom of the socket. Then, the cord is tightened to draw the residual limb into the socket and the lanyard is secured using a mechanism placed outside the prosthesis. This suspension method is mostly used in K1 and K2 patients with above knee amputations.

Figure 2.6: Lanyard suspension mounted in socket. Extracted from Amputee Guide.

[9] https://www.physio-pedia.com/Lower_Limb_Prosthetic_Sockets_and_Suspension_Systems. Accessed on 2024-04-25.

[10] https://www.pw.co.nz/products/categories/suspension-systems. Accessed on 2024-04-25.

- **Self-suspending socket:** this suspension is created by the socket brim (top parte) on the residual limb. This is obtained by extending and narrowing the brim over the joint [10]. The suspension in this method is not very strong.

- **External sleeves (Figure 2.7):** are snugly fitting covers that extend from the socket to the thigh. They intend to seal the air in the socket to create suspension and can be made of different materials such as neoprene, silicone or copolymer gel [10]. External sleeves are a reliable suspension and are not very expensive, being mostly used with transtibial sockets.

Figure 2.7: Sealing sleeve suspensions. Extracted from Ottobock.

- **Pin lock:** metal components that connect to a lock mechanism within the socket. To secure the prosthesis upon the insertion of the limb, the pin must insert into the locking mechanism at the bottom of the socket [10]. The pin is unlocked by pressing a release mechanism. Pin locks are indicated for K1 and K2 activity levels, for both above and below knee amputees.

- **RevoLock Lanyard (Figure 2.8):** developed by ClickMedical, it integrates the BOA system or RevoFit tightening mechanism with a snap or threaded insert that locks into the socket lock. It allows adjustment of tension within the socket to maintain proper suspension. Additionally, it automatically aligns the limb within the socket. This system has the advantage of being waterproof.

Figure 2.8: RevoLock Lanyard. Extracted from Ottobock.

- **Magnetic lock:** as an alternative to traditional pin locks, this system was designed to eliminate the need for pins when donning a prosthesis. Its purpose was to reduce the frustration experienced by users with moderate to high activity levels.

- **Suction without a liner:** old method of suspension, where the socket has an expulsion valve. This valve lets the air out, which creates a seal once the residual limb is inserted and the valve is closed, creating suction between the bare skin and the interior of the socket [10]. To remove the prosthesis, users have to open or push the suction valve, so that the air returns and the seal is broken. Since new interfaces have been developed, this approach has become obsolete.

- **Suction with a liner/ seal-in suspension (Figure 2.9):** eliminates the need for skin suction by wearing a liner on the residual limb to create suction within the socket. These liners are typically made of silicone or gel and they can have a membrane or silicone rings or ribs to create a seal. Air can escape through the valve as the socket is donned, maintaining suction when the valve is closed. To remove the prosthesis, users have to open or push the suction valve, allowing air to return and breaking the seal [10]. This system offers powerful suspension without the need for an external sleeve, except in cases of significant volume fluctuations. Suction suspensions are particularly beneficial for high activity level patients and they are used in both above-knee and below-knee patients. However, these systems tend to be relatively expensive.

Figure 2.9: Seal-in liner. Extracted from Seal In Liner

- **Vacuum-assisted suspension:** also known as Negative Pressure, Elevated Vacuum or Dynamic Vacuum. This suction suspension involves creating suspension through direct contact between the liner (or its membranes) and the socket wall. In this system, a pump mechanism removes air between the liner and the socket, generating a negative pressure across the entire surface [10]. To seal the system, an external sleeve or seal is used at the top of the socket. Vacuum-assisted suspension systems tend to be quite expensive.

- **Osseointegration:** involves implanting a device within the remaining bone [10]. Once healed, it allows the direct attachment of a prosthetic component to the implant. With this system there is no need for a socket or a liner. This surgical procedure can be done in patients with short residual limb or with above knee amputation (Figure 2.10). With osseointegrated systems, there is better mobility. However, the surgery and the specific components are expensive and there is a risk of infection and other complications. Extensive rehabilitation is also required.

(a) Osseointegration in BK amputation

(b) Osseointegration in bilateral, AK, amputation

Figure 2.10: Examples of osseointegration. Extracted from [16] (a) and HPC (b).

2.1.5 Reimbursement for a Prosthetic Device

When a patient receives a prosthetic device, the expenses with the prosthetic clinic and device might be eligible to be covered by an insurance. In the United States, that health insurance program could be Medicare, which provides insurance for Americans aged 65 or older and younger people with some disability status as determined by the Social Security Administration. Medicare provides health insurance for over 57 million individuals [11].

For a prosthetic component to be covered by Medicare, it has to meet specific criteria [12]:

- be eligible within a defined Medicare benefit category;
- be reasonable and necessary for diagnosis, treatment or functional improvement;
- meet all other relevant Medicare statutory and regulatory requirements.

Additionally, it has to be proven that the beneficiary [12]:

- has the ability to maintain the defined functional state or as the potential to reach a higher level within a reasonable period of time;
- has motivation to ambulate.

If a prosthesis is considered not reasonable and unnecessary, related additions will also be denied due to the same reasons. Having in consideration the definitions of each K-level, prosthesis will obviously be denied if the beneficiary's activity level is K0 [12].

[11] https://www.statista.com/topics/1167/medicare/#topicOverview. Accessed on 2024-04-25.

[12] https://www.cms.gov/medicare-coverage-database/view/lcd.aspx?LCDId=33787. Accessed on 2024-04-25.

In the following sections, we will be going through each prosthetic component and examining which type of prosthetic components are covered for which functional levels.

2.1.5.1 Knee

The selection of the type of knee for a lower limb prosthesis is made by the clinician based on the functional potential and needs.

The basis knee is the single axis, constant friction knee. However, other knee options may be considered according to the beneficiary's functional ability.

Fluid or pneumatic units, control additions, fluid, or electronic/microprocessor are covered for beneficiaries with a functional level minimum of 3 [12].

Addition to lower extremity prostheses, endoskeletal knee-shin system with powered and programmable flexion/extension assist control including any type of motor(s) is only eligible for coverage when the beneficiary mets the following conditions [12]:

- has a MPK of swing and stance phase type;
- MFCL 3
- has a documented comorbidity affecting hip extension and/or quadriceps function that impairs K3 level function when using a MPK alone
- is able to use a product that requires daily charging
- is able to understand and react to error alerts and alarms, that may be related to the function of the unit

There are other knee systems that are covered for beneficiaries with a functional level of 1 or above [12].

Coverage may be extended in the presence of adequate clinical documentation revealing the functional necessity of specific features of a type of knee.

2.1.5.2 Feet

The determination of the appropriate type of foot for the prosthesis is based on the functional requirements of the beneficiary.

SACH is the basic lower extremity prostheses, but other prosthetic feet must be considered according to the beneficiary's functional level.

For beneficiaries with a functional level of 1 or above, coverage is extended to include an external keel SACH foot or a single axis foot [12].

Flexible-keel foot or multiaxial foot are coverage for individuals with K2 or above [12].

Beneficiaries with a functional level of 3 or above may be qualified for more advanced technologies, such as a microprocessor controlled ankle foot system, dynamic response foot with multi-axial ankle, energy storing foot, shank foot system with vertical loading pylon, flex foot system, flex-walk system or equal [12].

It is important to keep in mind that microprocessor foot or ankle systems with power assist are not covered due to insufficient evidence that it is reasonable and necessary. Additionally, an user-adjustable heel height feature will also be denied for the same reasons [12].

Once again, coverage may be extended if there is sufficient clinical documentation indicating the functional necessity of specific features.

2.1.5.3 Ankles

An axial rotation unit is eligible for coverage for beneficiaries with functional level of 2 or above[12].

2.1.5.4 Hips

A pneumatic or hydraulic polycentic hip joint is eligible for coverage for beneficiaries with functional level of 3 or above [12].

2.1.5.5 Sockets

Regarding coverage of sockets, more than two diagnostic (test) sockets for a single prosthesis are considered unreasonable and unnecessary unless there is a document in the medical record justifying the need [12]. No more than two of the same socket inserts at the same time are permitted per individual prosthesis.

When it comes to socket replacement, it may be considered reasonable and necessary with adequate documentation of functional and physiological needs. Some reasons are: changes in the residual limb, changes in functional needs, irreparable damage due to excessive beneficiary weight or the prosthetic demands of highly active amputees [12].

2.1.5.6 Proposed Coverage Criteria Modifications

Recently, a modification in the coverage criteria for microprocessor-controlled prosthetic knees was proposed. It argues that the extension of this component to MFCL 2 has a significant impact on the quality of life of those beneficiaries [13].

Latest advancements in research indicate that MPKs are associated with improved stability when standing and walking, which results in significant improvements in mobility [17]. It has also shown to reduce the risk of fall due to improved technology that detects stumble and adapts knee resistance to provide support [18] [19]. Consequently, a MPKs reduces the fear of falling [20] [21], a feeling reported by 50% of individuals with lower limb amputation [22] and that results in activity limitation and reduced quality of life.

This scientific evidence indicates that MPKs would have relevant outcomes if used as a therapeutic option for MFCL 2, having a significant impact on the quality of life of those beneficiaries

[13] https://www.cms.gov/medicare-coverage-database/view/lcd.aspx?lcdid=39777. Accessed on 2024-04-25.

and preventing the incidence of falls, common in this population. On top of that, no studies suggest that the benefits of MPKs are exclusive to individuals with MFCL 3 or above, but the current research is not sufficient to extend MPKs coverage to MFCL 1 people.

With these alterations, the criteria for fluid, pneumatic, or electronic/microprocessor control additions in prosthetic knees would be expanded to allow the coverage for beneficiaries with K2[13]. The first demand would be the performance of a clinical evaluation, in order to evaluate the beneficiary's functional level. It would also be required supporting documentation in the medical record detailing, considering the beneficiary's overall medical health and the reasons behind selecting the correspondent knee. This document has to at least contain information on how the chosen knee will improve the beneficiary's functional health outcomes, such as reducing the risk of falls, preventing injuries and reducing the energy expenditure and help the beneficiary to perform their daily life activities. It is also mandatory to register that all lower-level knee systems have been considered and discarded due to the beneficiary's specific functional needs. So, these more complex knees would be covered for beneficiaries with a functional level minimum of 3 or 2 if all the previous requirements are met.

In addition, for beneficiaries with functional level K2, coverage of an electronic/microprocessor-controlled knee requires the integration of technology to detect when the user trips or stumbles and automatically stabilize the knee, potentially preventing a fall. The individual must also be capable of using a product that requires daily charging and also be able to understand and respond to error alerts and alarms (related to the function of the component) that may appear [13].

As a consequence, the criteria for higher-level foot systems would also be expanded to include beneficiaries with functional level of 2. So the criteria for the coverage of a microprocessor-controlled ankle foot system, energy storing foot, dynamic response foot with multi-axial ankle, flex foot system, flex-walk system or equal, or shank foot system with vertical loading pylon would be beneficiaries with MFCL 3 or above, or with MFCL 2 under the two following conditions [13]:

- meets the functional level 2 coverage criteria for a fluid, pneumatic, or electronic/microprocessor control addition for a prosthetic knee;
- a higher-level foot is needed to use the prescribed knee system in a safe and effective way.

2.2 Clinical K-level Assessment

To assign the patient's potential K-level, the clinician gathers information such as patient's medical history, comorbidities, motivation to ambulate and K-level guideline descriptors [6]. Considering only this kind of information, the K-level attribution is subjective, relying only on self-report and the clinical experience. Therefore, it became important to develop standardized methods and objective outcome measures to complement the K-level assessment. Nowadays, there are standardized methods available to provide a more objective K-level evaluation in clinical settings. Those tools allow collecting Patient Reported Outcome Measures and Performance Based Outcome Measure. We will introduce some of both types.

2.2.1 Prosthesis Evaluation Questionnaire

The Prosthetic Evaluation Questionnaire (PEQ) is an instrument developed to assess aspects related to the quality of life and function of individuals using prostheses. It is used to examine the perceived difficulty in performing prosthetic functions and mobility, providing information on the user's experiences and challenges.

PEQ contains 82 items organized into nine functional domain scales that are not dependent on each other, allowing their individual use for evaluating specific domains of interest [14]. Each scale contains multiple questions and some additional individual questions. These scales have been validated for internal consistency and temporal stability.

It is important to note that PEQ provides valuable insights into prosthetic-related quality of life and function, but it does not include standard demographic questions, such as age, level of amputation or years since amputation. Therefore, clinicians may supplement the PEQ with additional demographic questions if they want to obtain information about the individual's background and context.

2.2.2 Amputee Mobility Predictor

Amputee Mobility Predictor (AMP) is an assessment tool used to evaluate functional ambulatory potential of lower-limb amputees, with (AMPPRO) or without (AMPnoPRO) a prosthesis [6]. This is a quick and easily administered test, taking between 10 to 15 minutes to complete, and it requires little equipment. The test consists of various items that analyze a variety of mobility-related tasks, organized with a increasing level of difficulty, to assess the individual's overall functional ability [6]:

- **Items 1 and 2:** test the ability to maintain sitting balance. If the amputee is not able to perform these tasks, it indicates limited potential for prosthetic use and the individual is categorized as a level K0.
- **Items 3 to 7:** evaluate the ability to maintain balance while transferring from chair to chair and standing. These skills are essential for individuals classified as K1, who need a prosthesis for transfers and basic standing activities.
- **Items 8 to 13:** include challenging activities related to standing balance. Success in these tests indicates that the individual has the potential to be a safe household ambulator, corresponding to level K2.
- **Items 14 to 20:** assess the quality of gait and the ability to navigate overcome specific obstacles. These skills indicate that the individual can perform most tasks more easily, which corresponds to a level K3 or K4.
- **Item 21:** considers the use of specific assistive devices [6].

[14] https://orthocareinnovations.com/wp-content/uploads/2021/11/PEQ.pdf. Accessed on 2024-04-27.

The scoring system of AMPRO goes from 0 to 42 points. For individuals without a prosthesis, the AMPnoPRO has a maximum score of 38 points. This change is due to the elimination of item 8, single-limb standing, since standing on the prosthetic side is impossible [6]. The AMP scores range for each MFCL level are presented in Table 2.1.

Table 2.1: AMP Scores Range [6].

	K0	**K1**	**K2**	**K3**	**K4**
AMPnoPRO	0-8	9-20	21-28	29-36	37-43
AMPPRO	N/A	15-26	27-36	37-42	43-27

The use of assistive devices during testing can increase the potential total score possibilities by 5 points (to 43 and 45 for AMPnoPRO and AMPRO, respectively) depending on the type of assistive device used.

The majority of the items have 3 possible scores: 0 when the individual is not able to perform the task, 1 for a minimal level of achievement or when some help was needed, and 2 for a total independence or task mastery [6].

The AMP is useful for assessing individuals with unilateral or bilateral lower-limb amputations. For bilateral amputees, adjustments to the scoring system have been made to accommodate their specific needs (AMP-Bilat test), ensuring a fair evaluation of their mobility potential [23].

2.2.3 Two Minute Walk Test

The Two Minute Walk Test (2MWT) is a measure of self-paced walking ability and functional capacity, particularly useful for people who are unable to complete the longer Six Minute Walk Test. This test has been used as an outcome measure for multiple populations and for patients with a variety of conditions, including those with lower-limb amputation [24].

During this test, the individual is instructed to walk as fast as possible for a duration of two minutes, walking the maximum distance they can, safely, without any physical assistance. Rest breaks are allowed if needed, but the clock continues to run. To measure accurately the distance walked, a measuring wheel is recommended. The 2MWT reference values for each MFCL level are presented in Table 2.2

If the result of 2MWT changes by at least 34.3 meters, it is indicative of a clinically significant change of the patient's condition (unilateral transtibial or transfemoral prosthetic user) [25].

Table 2.2: 2MWT Reference Values [26].

MCFL K-Level	**2MWT Distance (m)**	**2MWT Gait Speed (m/min)**
K0-K1	-	-
K2 (n=7)	81.7 ± 26.9	41.4 ± 13.4
K3 (n=70)	138.4 ± 28.5	69.4 ± 14.3
K4 (n=24)	177.9 ± 31.1	89.0 ± 15.7

2.2.4 Six Minute Walk Test

The Six Minute Walking Test (6MWT) is a similar test, but in this case it considers the distance walked during a period of six minutes. It is used to evaluate aerobic capacity and endurance [15].

This test was originally developed to assess patients with cardiopulmonary problems, but nowadays it is used for other conditions such as lower-limb amputation. In the context of lower-limb amputation rehabilitation, the 6MWT is a widely used clinical tool to evaluate progress and rehabilitation needs. It can also assist in the process of assessing the functional level to a patient. It works similarly to 2MWT. During this test, patients are instructed to walk as fast as they can during 6 minutes without any physical help [15]. They can rest if necessary but the clock continues to run. The 6MWT reference values for each MFCL level are presented in Table 2.3

For 6MWT, an increase in the distance walked during the six minutes indicates improvement in basic mobility. So, in order to consider a significant change in user's condition, a difference of at least 45 meters has to be identified in unilateral transtibial and transfemoral prosthetic users [25].

Table 2.3: 6MWT Reference Values [15].

MFCL K-Level	**K0-K1 (n=18)**	**K2 (n=43)**	**K3 (n=67)**	**K4 (n=39)**
Reference Values (m)	50 ± 30	190 ± 111	299 ± 102	419 ± 86

2.2.5 Ten Meters Walk Test

The 10 Meter Walk Test (10MWT) is a measure used to evaluate walking speed in meters per second over a short distance [16]. This test allows the assessment of functional mobility, gait and vestibular function.

10MWT requires a clear pathway with at least 10 meters, with lines indicating the start, finish and 2-meter intervals from either end. During this test, individuals are instructed to walk at two different speeds: self-selected speed and fast speed. In the first case, participants walk at their normal comfortable pace until the finish line. For the fast speed, they are instructed to walk as fast and safely as possible to the finish line. Clinicians should start the time when the patient crosses the 2-meter mark (to allow some distance of acceleration at the beginning) and stop timing when the patient passes the 8-meter mark (to allow some distance of deceleration at the end) [16]. It is recommended to record three trials for each speed condition and then obtain the average to have a more accurate assessment. The 10MWT reference values for each MFCL level are presented in Table 2.4.

[15] https://cpb-us-w2.wpmucdn.com/sites.udel.edu/dist/2/3447/files/2019/01/6-Minute-Walk-Test_Updated-January-2019-1rb8aq5.pdf. Accessed on 2024-04-27.

[16] https://neuropt.org/practice-resources/anpt-clinical-practice-guidelines/core-outcome-measures-cpg. Accessed on 2024-04-27.

Table 2.4: 10MWT Reference Values [27].

MFCL K-Level	K1	K2	K3	K4
Self-selected speed (m/s)	0.17	0.38	0.63	1.06

2.2.6 Timed Up and Go Test

The Time Up and Go (TUG) is a widely used test used to evaluate the fall risk and the progress of balance, sit-to-stand and walking. This test was initially designed for elderly people, but is now used in other populations such as amputees.

In the TUG test, patients wear their regular footwear and may use a walking aid if needed. The patient starts in a seated position. Upon the clinician's command, the patient stands up, walks a distance of 3 meters, turns around, returns to the initial point, and sits down [28]. The time stops once the patient is seated again.

The cut off scores indicate the risk of falls depend on the population. Amputees have a cut off score of 19 seconds [28], which means that an amputee who takes longer than 19 seconds to perform this test, may be at a higher risk of falls.

2.2.7 Prosthetic Limb Users Survey of Mobility

The Prosthetic Limb Users Survey of Mobility (PLUS-M) is a self-report tool developed with rigorous psychometric methods, used to evaluate the mobility of adults with lower limb amputation who have experience using a prosthesis. PLUS-M was designed to be used in clinical practice and research [17].

PLUS-M focuses on measuring the mobility of prosthetic users, evaluating respondents' perceived ability to perform actions involving the use of both lower limbs (from ambulating around the house to outdoor recreational activities) [17]. These activities are related to locomotion (continuous and repeatable movement patterns) and/or postural transitions (movement between different positions or activities), which are the two primary forms of movement. Responses are evaluated in terms of the difficulty experienced by the respondent in performing each activity, with options including "without any difficulty", "with much difficulty", "with some difficulty", "with a little difficulty", and "unable to do". It is important to note that PLUS-M™ focuses on reported ability rather than actual performance, excluding unintentional movements or activities performed with the assistance of devices or another person.

PLUS-M T-scores range from 17.5 to 76.6, with a mean of 50 and a standard deviation of 10. A T-score of 50 represents the average mobility reported by the development sample, with higher T-scores indicating greater mobility. For example, a respondent who receives a T-score of 60 has reported a level of mobility approximately 1 standard deviation above the mean reported by prosthetic users in the development sample (Figure 2.11). Therefore, compared to the people

[17] https://plus-m.org/about.html. Accessed on 2024-04-27.

in the PLUS-M development sample, this respondent reported higher mobility than almost 84% of the people.

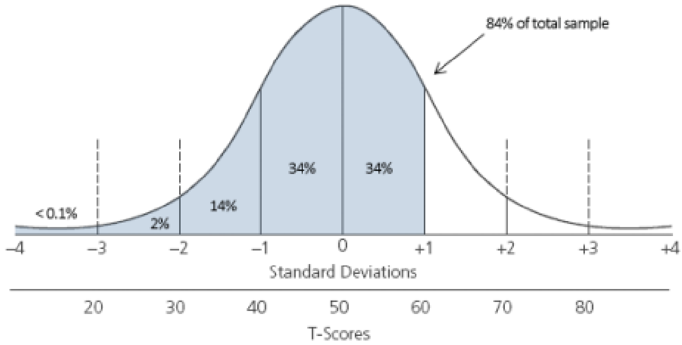


Figure 2.11: PLUS-M T-score of 60. Extracted from PLUS-M Users Guide.

2.2.8 Shortcomings of In-clinic Performance Tests

Even though these in-clinic performance tests are convenient to use, there is evidence that they may not accurately reflect what happens in the community during patients' daily lives.

In a study conducted by Kim et al., the practical applicability of a wearable system incorporating IMU and GPS technology in assessing the functional ability of individuals with LLA was evaluated [29]. The study assessed various parameters, including cadence, walking speed, their distributions, and geographical variations. The research explored the potential advantages of using performance measures from daily life compared to conventional in clinical assessments. For this research, seventeen participants with LLA were recruited (including two K2, twelve K3, and two K4 classifications). A control group of fourteen non-disabled individuals was also included for comparison purposes. Group differences and disparities between home and community environments were only evident when considering walking speed. Measurements of cadence and walking speed taken in clinical settings indicated higher values compared to those recorded during participants' daily activities. These findings indicate that there might be a difference between the functional activity capacity when tests are made in clinical settings and the amputees' actual performance in walking activities in the community.

2.3 Motio StepWatch Data

To develop this project, it is crucial to analyze and understand the data stages through which the data passes in the current Motio StepWatch system [18]:

1. **Series of Timestamps:** a list of timestamps in which strides are taken or an aliveness indicator is recorded.

[18] https://www.adapttech.eu. Accessed on 2024-04-27.

2. **Time Series of Step Count:** the magnitude of the signal represents the number of steps taken within the chosen resolution, with a minimum resolution of 1 second.

3. **Time Series of Daily Activity Metrics:** a time series where the magnitude of the signal represents each of the activity metrics per day.

4. **Tabular Data:** encompasses global activity metrics derived from the above time series. The global activity metrics are presented in Table 2.5.

Table 2.5: Global Activity Metrics Descriptions.

Global Activity Metric	**Description**
Daily Average Steps	Average of daily number of steps taken
Daily best steps	Maximum number of steps taken in a day
Daily average top cadence	Average of daily maximum cadence. The top cadence in a day is the single fastest minute
Daily best top cadence	Maximum of daily maximum cadence
Daily average top speed	Average of daily maximum speed. The top speed is the fastest single minute in the day
Daily best top speed	Maximum of daily maximum speed across all days
Daily average energy expenditure	Average of daily energy expended during ambulation
Daily best energy expenditure	Maximum of daily energy expended during ambulation
Daily average maximum continuous walking	Average of daily maximum continuous walking. A period of continuous walking starts whenever strides are detected in a 15 second window and ends when no strides are detected in a 10 second window
Daily best maximum continuous walking	Maximum of daily maximum continuous walking
Ambulation energy score	By considering weight and walking intensity, it provides insights into the patient's overall energy usage during walking activities
Peak performance score	Incorporates the person's average 30 fastest 1-min walking spurts each day and captures the best metrics recorded
Cadence variability score	Incorporates the proportion of time at high, moderate, and low walking cadences
Active time	Average of daily time active (minutes with at least 1 step)
Active time at low cadence	Average of daily time active at low cadence (minutes at cadence 1-30 steps/min)
Active time at medium cadence	Average of daily time active at medium cadence (minutes at cadence 31-80 steps/min)
Active time at high cadence	Average of daily time active at high cadence (minutes at cadence more than 80 steps/min)
Two minutes walking test	Meters walked during two minutes, showing the average and standard deviation
Six minutes walking test	Meters walked during six minutes, showing the average and standard deviation
Ten meters walking test	Gait speed in meters per second during ten meters, showing the average and standard deviation

2.4 Summary

In this chapter, it became clear the subjectivity of assessing an individual's K-level and the insufficient efficiency of the currently available standardized methods. This K-level attribution is crucial for the selection of prosthetic components, among the huge variety of types available for each component. Medicare coverage plays a crucial role in ensuring access to essential prosthetic components for an enhanced quality of life for the beneficiary. Medicare coverage criteria are contingent upon an individual's K-level. So, it is extremely important to make a precise classification of the functional activity level, not only in the clinical environment, but also during daily life.

Chapter 3

Literature Review

This chapter explores the application of unsupervised and supervised strategies for the identification of activity profiles in prosthesis users, drawing insights from recent research in the prosthetic field and in other fields that have a similar purpose to this project. Additionally, due to the diverse data formats that we have access to (explained in Chapter 2), this chapter also describes approaches of data augmentation, including both time series and tabular data augmentation.

3.1 Machine Learning

3.1.1 Unsupervised Learning

Unsupervised learning techniques have garnered significant attention across various fields, due to their ability to extract hidden patterns, structures, and correlations within data. Unsupervised learning can be divided into association (when we want to discover rules that describe large portions of the data) or clustering (when we want to discover the inherent grouping in the data).

K-means clustering is a technique that belongs to the clustering category. It is used to find the optimal number of centers (K) for a dataset. It assigns data points to one of the K clusters according to the distance between the data points and the centers of the clusters, in a way that the distance between the centers and the data points is minimized [30]. This leads to objects being similar within a group and distinct between groups.

Finite mixture models are a principled statistical approach for clustering, in which each component probability corresponds to a cluster [31]. Gaussian mixture model (GMM) is a clustering technique that is used to obtain the probability of a particular point belonging to a cluster. This method is made by several Gaussians, donated by $k \in \{1, \dots, K\}$, where K is the number of clusters in a dataset. Mixture model clustering has the advantage of allowing the observations to have partial participation in each group, not fully belonging to one exclusive group. Therefore, each observation has indexed membership to all clusters.

3.1.2 Supervised Learning

Unlike unsupervised learning, the data used in supervised learning is labeled and known. The input data is divided into training and testing sets [32]. In supervised learning, there are two main categories of approaches: classification and regression. In classification, the model must classify samples according into two or more classes. If there is only two classes, it is a binary classification, while if we have three or more classes, it is a multiclass classification. On the other hand, in regression the model predicts a continuous variable.

Linear methods are the most simple supervised learning approaches. They consider a linear relationship between the features and the label.

Another group of very simple supervised learning algorithms is the **Neighbourhood-based methods**. They learn how to label new points based on how close they are to the previously labeled points.

In **Tree-based methods**, the model builds a decision tree in which each instance is stratified into multiple regions based on the existing labels. After this division is complete, each region is associated with a certain class of label or a range of expected values. **Random Forest** (RF) is one example of a model that belongs to this category. RF uses a number of decision trees during its training phase. A random subset of features is used to build each tree, introducing variability among individual trees and reducing the risk of overfitting.

It is also possible to use algorithms to design hyperplanes in space that split the data based on their labels. This methodology is known as **Support Vector Machines** (SVMs). This method has some tolerance since not every point in a hyperspace region need to have the same label. However, the distance between a label's boundary-defining points and the points that define the boundary of another label should be maximized.

A **Neural Network** is another machine learning method that makes decisions in a similar way to the human brain, mimicking the way biological neurons work together. Their structure is composed of an input layer, several hidden layers, and an output layer. The input layer incorporates the features and the output layer attempts to match the response label. This method is based on a hierarchical concept since, in the hidden layers, each layer tries to understand how the previous layer is related to the output layer. With this, the neural network may learn complex concepts by associating them with the combination of simple ones.

3.2 Data Augmentation

To evaluate machine learning algorithms' performance, researchers usually resort on predictive accuracy. However, this method is not suitable for imbalanced datasets, where one class has significantly more instances than the other. Two primary approaches have been taken by the machine learning community to address this problem: assigning different costs to training examples or re-sampling the original dataset, either by oversampling the minority class or under-sampling the majority class.

3.2.1 Data Augmentation for Tabular Data

Data augmentation is a technique commonly associated with image and text data, but it is also highly relevant for tabular data.

One approach used for data augmentation of tabular data is the **Generative Adversarial Network** (GAN) [33]. It has two components: a discriminator model and a generator model, that compete with each other. The generator creates samples by attempting to reproduce the distribution without replicating it, while the discriminator tries to identify the fake samples from the real ones.

Variational Autoencoders are neural generative models buit upon the Autoenconders architeture [33]. Traditional Autoenconders transform typical artificial intelligence problems into domain-shifting tasks by modifying inputs to produce slightly altered outputs. Variational Autoencoders enhance this process through probabilistic modeling. The Autoenconder structure consists of two primary components: an encoder (reduces the input data into a lower-dimensional latent space) and a decoder (reconstructs the input data from this latent representation). Synthetic data is created by interpolating within the latent space and decoding the interpolated values. By regularizing its training, Variational Autoencoder reduces the risk of overfitting commonly observed in Autoenconders, producing more diversed samples. While Autoenconders encodes the input into a singles point, Variational Autoencoders encode it into a probability distribution. To generate new data, Variational Autoencoders samples a point from this distribution and decodes it.

Chawla et al. proposed a method that uses the **Synthetic Minority Oversampling Technique** (SMOTE) algorithm combined with classifiers to improve the classification of imbalanced datasets [34]. The SMOTE algorithm oversamples the minority class (Figure 3.1), balancing the dataset and enhancing the classifier's performance. It does this by randomly selecting some center samples from the minority class and using the k-nearest neighbors (KNN) algorithm based on Euclidean distance to find K neighbors for each "central" sample. Then, it applies linear interpolation to blend each central sample with its corresponding neighbors to create new synthetic samples.

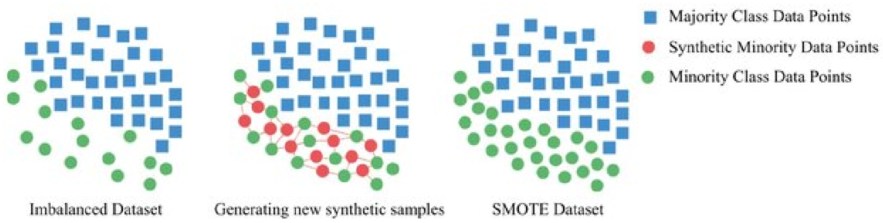


Figure 3.1: The visual representation of SMOTE oversampling for an imbalanced dataset. Extracted from [35].

To assess the SMOTE's effectiveness, nine datasets were used in this study, including numerical and categorical data. The results revealed that the SMOTE algorithm significantly improved the classification of minority class samples in imbalanced datasets [34].

SMOTE is beneficial for the following types of data: categorical data, numerical data, and mixed data types. However, it is noted that the classical SMOTE method does not effectively account for the characteristics of time series data, limiting its application in time series datasets [36].

3.2.2 Data Augmentation for Time Series

Data augmentation involves the creation of new data samples by transforming the existing ones. There is a major challenge associated with the generation of new data that maintains the correct label. To do so, deep domain knowledge is required. However, it is not always clear how to perform label-preserving augmentation in certain domains, such as wearable sensor data.

All techniques that take data input samples and synthesize new samples by directly transforming the data are considered traditional algorithms [37]. These algorithms include:

- **Rotation:** simulates different sensor placements by applying arbitrary rotations, introducing label-invariant variability.
- **Permutation:** perturbs the data location within a window by slicing and randomly permuting segments, since the fixed size window segmentation is arbitrary.
- **Time-warping:** changes temporal locations of samples by smoothly distorting time intervals.
- **Scaling:** changes data magnitude by multiplying by a random value.
- **Magnitude-warping:** changes sample magnitudes by convolving the data window with a smooth curve varying around one.
- **Jittering:** simulates additive sensor noise to increase robustness against noise and improve performance.
- **Cropping:** similar to image cropping, reduces dependency on event locations.

VAE and GAN can also be used to synthesise new time series [37]. The main difference between GAN and the traditional methods is that the first one creates new samples of synthetic data, while the second transformers the information from the real data to create new data.

To overcome the limitations of SMOTE associated with time series, more recently a new approach was proposed: Synthetic Minority Oversampling Data Augmentation Method Based on Dynamic Time Warping (DTW-SMOTE) [36]. Unlike traditional SMOTE that uses Euclidean distance to measure similarity between instances, DTW-SMOTE uses Dynamic Time Warping (DTW) distance. DTW uses an algorithm of dynamic programming to enhance the conventional one-to-one Euclidean distance method to one-to-many. To compare sample time series A and B, it creates a *mxn* matrix. The warp path from element (1,1) to (m,n) is defined by this matrix. The warp path with the minimal cost is the DTW path, representing the DTW distance between A and

B. Therefore, the goal of DTW-SMOTE is to select the k nearest neighbor of the central samples by applying KNN based on DTW. Extrapolation from the center is done for every neighbor, k times in total. Extrapolation is used in place of interpolation to overcome the lack of unusual and underrepresented sample availability.

The study conducted by Um et al. addresses the challenge of data augmentation for wearable sensor data to improve the classification of Parkinson's disease motor states using convolutional neural networks (CNNs) [38]. The original dataset included 30 patients' motor states. They took into consideration two frequent motor states associated with Parkinson's disease: dyskinesia, which is defined by involuntary movements of the extremities, and bradykinesia, which is characterized by decreased movement speed and may be accompanied by tremor. Firstly, they applied CNNs to the original dataset, then developed the data augmentation approaches to be tested, and finally made an experimental comparison of the proposed data augmentation methods. Those data augmentation methods included tradicional algorithms (Figure 3.2).

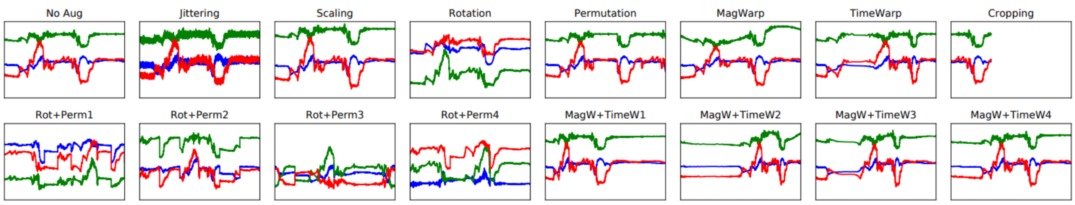


Figure 3.2: Data augmentation methods used in the experiments: jittering, scaling, rotating, permutating, magnitude- warping, time-warping methods. They also tested combinations of various data augmentation methods. Extracted from [38].

The proposed rotation, permutation, and time-warping methods effectively improved accuracy by 3.6-5.1%. Among the single data augmentation techniques, rotation was the one with the best results. Combining multiple data augmentation methods further enhanced performance, with the highest accuracy of 86.88% achieved by integrating rotation, permutation, and time-warping [38].

3.3 Feature Reduction and Selection

In machine learning, there is a common belief that having more features will result in better model performance. However, this is not always true. If we continue to increase the number of features beyond a certain point, the algorithm's performance tends to decline [39]. One way of reducing the model's complexity is to apply dimensionality reduction. It includes two main approaches: Feature Extraction (creation of a new feature subspace by extracting information from the original feature set) and Feature Selection (choosing a subset of the original feature set).

Principal Component Analysis (PCA) belongs to the first category. It is a linear dimensionality reduction method that transforms a set of correlated features in a high dimensional space into uncorrelated features in a low dimensional space. The purpose of PCA is to perform data reduction by replacing a large number of possible correlated variables with a smaller set of uncorrelated variables, while keeping as much information as possible [1]. These new variables are known as principal components and are linear combinations of the initial variables. These principal components are ordered in decreasing order of variance. The number of principal components created is equal to the number of original features and the total variance captured by all of the principal components is equal to the total variance in the original dataset. By selecting a subset of these components based on a threshold of explainability, the dimensionality of the original dataset is reduced. This threshold determines the number of principal components needed to retain as much of the original data's variance, thereby reducing the data's dimensionality while preserving essential information.

Linear Discriminant Analysis is another method of dimensionality reduction. Unlike PCA, Linear Discriminant Analysis is a supervised learning method that takes class labels into consideration. It aims to maximize the distance between the mean of each class and minimize the spreading within the class itself [1]. This is done by finding a new feature space that best discriminates between the different classes. The result is a set of new features, known as linear discriminants. In this method, it is assumed that the input data follows a Gaussian distribution, so if the apply it to not Gaussian data it may result in poor classification results.

Another dimensionality reduction technique is **t-Distributed Stochastic Neighbor Embedding** (t-SNE). It is primarily used for data visualization and captures complex, non-linear relationships in the data. It tries to find a two or three dimensional representation of the data that preserves the distances between data points as best as possible. It measures the similarity between data points in a high-dimensional space and represents this similarity as probabilities. Then, it builds a probability distribution that is similar in the lower-dimensional space and it uses a method known as gradient descent to minimize the difference between the two distributions [1]. This process enables t-SNE to capture the structure of the data, which makes it very helpful in identifying meaningful patterns and visualizing complex datasets.

There are several types of Feature Selection methods. **Filter** methods are approaches that do not use learning models, but use only features characteristics [40]. This type of feature selection involves two steps: choosing a criteria to rank and identifying the highest ranked features. **Correlation-based Feature Selection** [41] is a filter method that analysis the correlation between each pair of features and may also have in consideration the correlation between each feature and the target. It selects subsets of features that do not have a high correlation with each other, but are highly correlated with the target variable. Another filter-based aproach is **Select K Best** (KBest) [42]. It uses the statistical tests like chi-squared test, ANOVA F-test, or mutual information score to rank features based on their correlation. It has the limitation of only looking for linear correlations between features and the target.

Wrapper method is another category of selection techniques that identifies the feature subset that improves the model's performance. Therefore, these methods use a specific learning model to

[1] https://medium.com/analytics-vidhya/pca-vs-lda-vs-t-sne-lets-understand-the-difference-between-them-22fa6b9be9d0. Accessed on 2024-05-03.

select the subset of features [43] and are composed of three main stages: feature set selection, feature set evaluation and induction algorithm. **Recursive Feature Elimination** (RFE) is a wrapper method with the purpose of finding the best feature subset, by generating a new model recursively.

Embedded methods are a group of feature selection approaches that combine the best aspects of wrapper and filter-based techniques. They incorporate the feature selection procedure within the learning model [44].

The study made by Paszynski et al. is an example of the application of feature selection techniques in medical tabular data [45]. Their goal was to provide an analysis of different feature selection methods used in machine learning for multiclass classification tasks, focusing on medical tabular data. The evaluation was made based on three main steps: accuracy, stability and interpretability. According to the experimental pipeline, the authors compared the performance of classifiers using various feature selection methods: Variance threshold, Bayesian selection, KBest, RFE. The highest F1-score was obtained when using RFE, for all classifiers. The second-best performance was obtained for Bayesian Selection for Logistic Regression and KBest for Random Forest and XGBoost. According to the mean validation F1-score, the most stable classification was achieved with RFE for gradient boosting, Bayesian Selection for logistic regression, and K-Best selection for Random Forest.

3.4 Machine Learning in O&P

Machine learning is an important analytical tool that can be applied to healthcare, allowing the improvement of patient care [46].

Machine learning techniques have revealed promising results in different areas of orthotics and prosthetics (O&P). Choo et al. made a narrative review on the integration of machine learning with prosthetics and orthotics, referring to existing studies in the field [47].

3.4.1 Prosthesis Alignment

Proper alignment of prosthetics is crucial. If the prosthetics are not correctly aligned, the body's center of gravity can be incorrectly distributed over the prosthesis during standing or walking. It may lead to improper pressure and weight application, which can result in various complications. Currently, alignment adjustments are made based on clinician observation, which is highly subjective.

Zhang et al. investigated whether a SVM could detect transtibial prosthesis misalignment during walking using data from the ground reaction force [48]. They measure differente features from gait cicle, including peak vertical forces, timing of these peaks, braking and propulsion forces, and lateral and medial forces. The data was recorded from nine unilateral transtibial amputees who walked 50 meters with both normally aligned and misaligned prostheses. SVM acheived a detection accuracy of 96,67%, which reveals the potential of SVM in accurately identifying prosthetic misalignment based on gait data.

3.4.2 Prosthesis Selection

Selecting the appropriate prosthesis can maximize a patient's physical ability while minimizing excessive costs. The quality of a prosthesis is often associated with the pressure it exerts on the body surface. However, prosthesis attribution is traditionally determined by clinician observation, which is subjective.

Sewell et al. developed a method for improving tibial socket fitting by analyzing pressure data at the residual limb/socket interface during the different stages of gait cycle [49]. As input, they used strain data collected from a socket surface of one subject, (namely thirty pressure strain readings). An artificial neural network algorithm analyzed the data and generated as output interfacial pressures at the residual limb/socket interface. The artificial neural network predicted interfacial pressure with a high accuracy, with 8.7% difference from actual pressure values.

LeMoyne et al, develop a machine learning algorithm for classifying types of prostheses based on gait analysis data [50]. The gait of one transtibial amputee was analyzed while using both passive and powered prostheses. Gait characteristics were measured using a force plate, focusing on temporal (rise time to the first maximum symbolizing brake, time for a decrease from a push-off to the terminal aspect of the stance phase) and kinematic parameters (maximum force after breaking, the maximum force after a push-off and the average force profile of the stance). They used SVM to classify between passive and electric-powered prosthesis using gait characteristics of the patient as input. This algorithm revealed an accuracy of 100% when using temporal gait parameters and 80% with kinematic gait parameters.

3.4.3 Post-prosthesis Training and Education

After amputation, patients require proper training to use their prosthesis effectively. Some studies have focused on developing adaptive training systems to support this task, including systems that enable autonomous training.

Huang et al. develop a prosthesis training system for amputees to independently train with real-time feedback [51]. There is not a clear explanation of the number of participants with lower-extremity amputation. Radio frequency identification readers on prostheses track limb movements, providing immediate feedback for gait correction. They used Decision Tree, fast-learning neural networks, and SVM to classify gait patterns with high accuracy. As input they used initial angle, the length of perpendicular, time consumption per step. The model classify in normal gait pattern or abnormal gait pattern. SVM was the model with the highest accuracy, with 89,25%.

3.4.4 Fall Detection Systems

Amputees have a higher risk of falling compared to nonamputees, due to muscle weakness and reduced sensitivity. It is important for an individual with lower limb loss to perform their daily activities safely, since falls can have severe consequences. It may result in not only physical injuries but also psychological issues such as anxiety, depression, and loss of confidence.

Shawen et al. developed models to detect falls in amputees using data from nonamputees [52]. The study involved six amputees and ten nonamputees who simulated falls while carrying a cellphone equipped with sensors. To have data from the common daily activities, they were asked to perform actions such as sitting, standing, and walking after falling. The recorded data (speed when changing postures) was used to train four models: SVM, RF, gradient boosting, and extreme gradient boosting, with the purpose of distinguish between falling motions and nonfalling motions. The developed models had a accuracy of 90% or higher in classifying these two conditions.

3.4.5 Temperature Control in Prosthetics

Excessive heat and humidity inside a prosthesis can accelerate skin erosion and increase the risk of bacterial infections. So, automatic temperature detection systems within prosthetics would be highly beneficial.

Mathur et al. developed a model to predict skin temperature of residual limbs by measuring the socket liner internal temperature [53]. It was done by using Gaussian processes for machine learning. While two patients with traumatic transtibial amputation walked on a treadmill, the temperatures of their stumps and liners were recorded. The goal was to analyse how the type of liner material affects changes in skin temperature. The inputs were the liner temperatures and the outputs the skin temperatures. They tested a variety of materials, such as Iceross Comfort (silicone), OttoBock Technogel (polyurethane), and carbon fiber lay-up for sockets. This model was able to predict skin temperature with an error margin of ±0.5 degrees Celsius, closely matching the actual skin temperature.

3.4.6 Motion Control During Orthotics Use

Motion recognition systems can be effectively used to monitor processes and outcomes, providing feedback. These systems automatically detect human movement through sensor signals.

The studies by Lonini et al. [54] and Farah et al. [55] both demonstrate the application of machine learning algorithms in the context of gait analysis and training for individuals using orthotic devices.

Lonini et al. created a algorithm for classification of gait patterns that can be applied when patients are training their gait while wearing their orthotic devices [54]. In this study participated eleven healthy individuals and ten patients using knee-ankle-foot orthoses due to unilateral leg weakness. While their gait data was collected, they have to perform five different motions: sitting, standing, walking, going up stairs and going down stairs. The input data included a variety of parameters such as mean, moments, histograms, Pearson's correlation coefficient, power spectra and mean power in 0.5 Hz bins within 0–10 Hz. They explored three model types: personal model (trained on data from all patients and testes individually), global-patients model (trained on data from 9 patients and testes on the excluded one) and global-healthy model (trained on healthy subjects and tested on patient data). The supervised learning model used was RF with 100 trees for individual models and 50 trees for global models. All model types had the best accuracy regarding the sitting position (personal model, 92.0%; global-patients model, 86.7%; and global-healthy model, 88.1%). All the models had good performances in classifying the different motions.

Farah et al focuses on developing a machine learning algorithm for real-time gait phase recognition, specifically for controlling the gait of orthotic device users in real time [55]. In this study participated thirty healthy subjects included in the training set and twelve healthy subjects for validation set. While data on thigh and knee joint trajectories and body segments was collected, subjects had to walk on five types of ground: level ground, 7° declining slope, 7° inclining slope, 5° right crossing slope, and 5° left crossing slope. The algorithm developed included two main steps: first Logistic Model Tree was used to train the model using the collected gait data, then Transition Sequence Verification and Correction was implemented to improve gait phase recognition by solving the complexity of discontinuous classes. The results show that the models accurately identified different gait phases across various ground types and gait velocities.

3.4.7 Predicting Orthotics Use in the Future

Prognostic technology can be effectively used to prevent unnecessary prescriptions of orthotics for medical treatment. In 2021, a deep learning algorithm was developed by Choo et al. to predict the need for orthotics following a stroke [56]. They recorded data from 474 patients with stroke and the input variables included age, sex, stroke type, and muscular strength scores for various muscles measured within 30 days post-stroke. They used deep neural networks, logistic regression and RF. The output was the score for the ankle dorsiflexor muscle strength measured 6 months after the stroke. This score was used as the criterion for assessing the necessity of ankle–foot orthosis. It was devided that the ankle–foot orthosis was not necessary if the score was three or higher and ankle–foot orthosis was necessary if the score was less than three. The study utilized deep neural networks, logistic regression and RF. All models demonstrated excellent performance. This research indicated that machine learning could prevent unnecessary orthotic prescriptions for post-stroke patients.

This review made by Choo et al. [47] revealed that SVM was the most frequently used model in this field and emphasized the need for large datasets with diverse participants' characteristics to improve the generalizability and applicability of machine learning algorithms in prosthetics.

3.4.8 Identification of Functional Level Using Real-world Sensor Data

Performed a literature research to look for papers which used supervised learning to estimate or predict the functional level of a patient given data obtained from the usage of wearable sensors in the prosthetic users and patient information. Even though machine learning is already used in prosthesis for diverse purposes, there is not much work done related to identifying K-levels using real-world sensor data. To take a step further in this topic, Orendurff et al. incorporate wearable sensor data into the assessment of functional levels by evaluating the relationship between the clinical K-level and a K-level determined by a custom algorithm [4]. This custom algorithm was based on real-world functional activity levels from wearable sensor data during daily life. Twelve individuals with limb loss participated in this study. Each participant's K-level was assessed by a single treating prosthetist using gait observation and patient questionnaires in a clinical environment. Then, each participant was fitted with a StepWatch Activity Monitor on their prosthesis to record ambulatory activity over seven days. StepWatch has been found to be effective in counting steps for a wide variety of gait pathologies, such as poststroke, multiple sclerosis, and Parkinson's disease. The StepWatch data was processed using a custom algorithm that used four components, incorporating both community activity data (contribution of 75%) and clinical K-level assessments (contribution of 25%) to estimate the functional level. The three parameters related to community activity data included in the algorithm were: the average step rate from the most active minutes, the ratio of different step rates, and total daily steps to estimate energy expenditure. To evaluate the correlation between the prosthetist's clinical judgment of K-level and the K-level determined by the algorithm, they used a simple linear regression. It found a significant correlation between the prosthetist's K-level and the calculated K-level ($R^2 = 0.775$, $p < 0.001$). The results also indicate that real-world data could offer greater resolution in functional level assessment, potentially identifying high K2 ambulators who might benefit from more advanced prosthetic components. This study suggests that integrating real-world ambulatory data with clinical assessments can enhance the objectivity and accuracy of functional level determination in prosthesis users.

3.5 Machine learning: Sports and Medical Application

3.5.1 Unsupervised Learning

Within the domains of sports science and healthcare, researchers have increasingly turned to unsupervised learning methods to unlock novel insights and drive advancements in areas such as performance optimization, injury rehabilitation, and clinical diagnosis.

For instance, Shelly et al. proposes an approach to use data collected from wearable technology to better inform training decisions [30]. It involves the application of K-means clustering technique to group athletes from two seasons worth of data from an National Collegiate Athletic Association Division 1 American football team. In this study, different data was used, such as maximum velocity, IMA (count of intense movements that occur throughout each activity), various distance intervals ran (examples: distance ran 5 to 8 mph, distance ran 8 to 12 mph), and total distance ran for every game and practice. This data was used to obtain average game demands of each athlete, then used to create training groups based on each individual game demands. The resultant clustering groups showed similar results to traditional groupings used for training in American football. One major limitation was the inability to associate specific roles and physical attributes to individual athletes, which limited the analysis of the clusters created. In practice, coaches using this data could evaluate student-athletes and make informed decisions on appropriate training protocols. For example, if the reason for different groupings is the linebacker's defensive role, then the linebacker must practice in a group that closely matches the demands of their position in games.

The study developed by Soto-Valero et al. proposes a model-based cluster approach (Figure 3.3) to explore and categorize football players based on their performance data from FIFA video games [57]. The performance data was constituted by forty attributes belonging to eight main groups. Some examples of performance groups and features are: Physical (age, height, weight, body mass index), Movement (acceleration, sprint speed, agility, reactions, balance), and Defensive (marking, standing tackle, sliding tackle).

They use PCA to reduce data dimension and a GMM to cluster players into roles such as defenders, midfielders, forwards, and goalkeepers [57]. To determine the best clustering model, they used Bayesian Information Criterion (BIC) with different parameterizations and numbers of clusters. Then, players were labeled according to their attributed roles. With this cluster-labeled dataset, a gradient tree boosting model was trained to perform feature selection and rank attributes regarding their importance. The results of the cluster process showed clear differentiation between the different performance indicators that represent the four main groups of players. These groups represent association football main positions, indicating that physical and technical demands vary based on the player's specific role on the field. They also found that, among the profiles of the clustered players, the dribbling skill was the most discriminant feature.

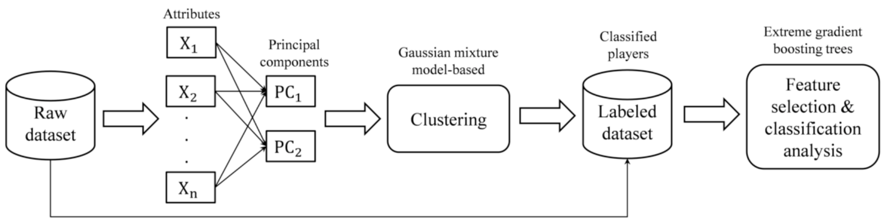


Figure 3.3: Graphical representation of the methodology followed. Extracted from [57].

In the field of stroke rehabilitation, unsupervised approaches also have a relevant contribution. Dolatabadiet al. intends to automatically categorize the gait pattern of mobility impaired individuals, to help to numerically analyze and study trends in the recovery of gait and track gait improvement or deterioration [58]. To do so, they apply mixture model clustering to analyze spatiotemporal gait parameters in individuals with pathological conditions and obtain groups of individuals with common attributes. The spatiotemporal gait features included were: gait symmetry ratio of swing time, stance time and step length, gait speed, and variability in step length and step time. This study uses data from sixty-eight stroke patients and twenty healthy adults to identify distinct gait patterns and generate a composite measure indicative of overall gait performance. A model-based Gaussian mixture clustering approach, along with the Bayesian Information Criterion (BIC), is applied to determine the optimal number of clusters. The selected clustering model effectively categorized participants' spatiotemporal gait data into three clinically meaningful groups: group 1 consisted of healthy participants, group 2 included individuals with intermediate-level gait parameters, and group 3 comprised the worst condition.

3.5.2 Supervised Learning

In sports, there is a Rating of Perceived Exertion (RPE) scale that evaluates how exhausting an exercise was for a specific person. RPE can be measure in two scales: the Borg RPE scale and a modified Borg category-ratio 10 RPE scale [2]. The first one goes from 6 (no exertion) to 20 (highest effort). The second one ranges from 0 (no exertion or resting) to 10 (pushing yourself to the limit). The quantification of subjective exertion during practice is important to adapt the exercises and avoid injuries resulting from overtraining. Since sometimes subjects do not accurately indicate their exertion during training, there is a necessity for complex external instrumentation to provide accurate workload quantification.

Carey et al. investigated the ability of machine learning methods to predict Borg category-ratio 10 RPE scores in professional Australian football players [59]. Forty-five professional Australian football players participated in this study. During a full season, data was recorded: global positioning system data, heart rate data, accelerometer data, and wellness questionnaires. The study used various machine learning algorithms (classification and regression), including neural networks, SVM, RF, and linear regression, to predict RPE values. In the regression approach, RPE response was considered a continuous real-valued number, while in classification RPE response was a discrete categorical variable. Three data preparation protocols were considered, and seven combinations of predictor variables were tested for each modeling approach. The predictor variables were chosen based on previous studies that found heart rate, running distances and speeds, accelerations, and wellness ratings impact athlete RPE. To evaluate the performance of each model, they used cross-validation and Root Mean Squared Error (RMSE). A RF model trained with only the three best variables achieved similar results to one trained with access to all variables, with a RMSE of 0.96 ± 0.08 and 1.09 ± 0.05, respectively.

Albert et al. presented a method to predict Borg RPE quantification of exertion during resistance training [60]. This study involved sixteen male participants that were equipped with six IMU sensors and electrocardiography sensors and data was recorded while they were performing squats on a flywheel training machine. From electrocardiography data, Heart Rate Variability parameters were calculated for windows of 30 seconds from the raw electrocardiography signals. From IMU data, eight statistical features were extracted for each window, including minimum, maximum, mean, median, root mean square, kurtosis, skewness, and standard deviation. To remove features that provide little information, feature selection algorithms were applied: Variance Inflation Factor thresholds, Univariate Selection, PCA and RFE. Four regression models were used to predict the subjective rating of perceived exertion based on the calculated features: Support Vector Regression with Radial Basis Function and linear kernels, RF Regression, and Gradient Boosted Regression Tree models. They used Root Mean Squared Percentage Error, Mean Absolute Percentage Error scores, Coefficient of determination (R2), and the Pearson correlation coefficient to evaluate the performance of the models. The authors found that the Gradient Boosted Regression Tree models model performed best, with the lowest Mean Absolute Percentage Error scores and Root Mean

[2]https://my.clevelandclinic.org/health/articles/17450-rated-perceived-exertion-rpe-scale. Accessed on 2024-05-03.

Squared Percentage Error values. In this study they considered that players were not restricted to integer responses when reporting their RPE. As a result, the RPE response is used as a continuous real-valued number, so it makes sense that in this case they only tested regression models.

In healthcare, supervised learning methods also have an important role. The biomedical area has been greatly impacted by machine learning, since it allows improved disease detection and prediction of biomedical interest [61].

An example is the study conducted by Alauthman et al., which assesses the performance of various machine learning algorithms at predicting the diagnosis of chronic liver disease while reducing its cost [33]. The liver is the largest organ in the body and is essential for food digestion and for processing the body's toxic substances. Liver diseases are a significant global health problem, with high incidence and mortality rates. However, the diagnosis of liver disease is complex and expensive. The dataset used in this study contains information regarding 583 Indian liver patients, such as quantity of total bilirubin, protein content, and amount of albumin in the patient. In this study, five machine learning algorithms were explored: Logistic Regression, KNN, Decision Tree, SVM, and Artificial Neural Network. The models' performance was assessed using accuracy, precision, recall, and f-1 score. To address the problem of unbalanced target classes, they applied GAN and SMOTE to guarantee that all data is distributed fairly and examined their effects on performance. In Figure 3.4, it is possible to see a flowchart with the main steps of the process of creating a model. The experimental results show that, with the used classifiers, SMOTE surpasses Generative Adversarial Networks in terms of accuracy and that KNN outperforms others with an average accuracy of 99%.

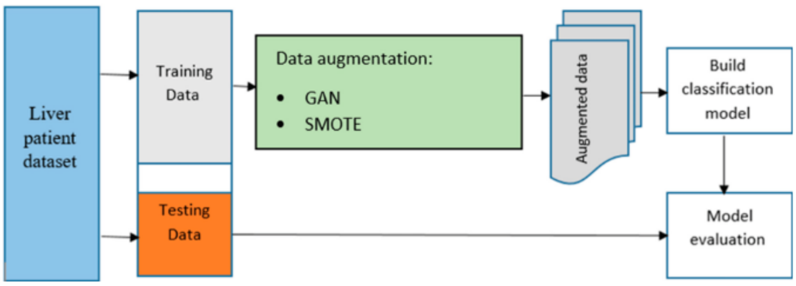


Figure 3.4: The architecture of the proposed approach. Extracted from [33].

3.6 Summary

In this chapter, a diverse range of potential advancements in patient rehabilitation and care associated with prosthetics are presented. From aligning prostheses to detecting falls, machine learning techniques provide adaptable solutions to various challenges within the field, with SVM as a prominent model. Due to the importance of measuring functional activity during daily life to complement traditional clinical assessments in prosthetic care, recent studies highlight the integration of real-world sensor data for assessing functional levels. From the exploration of machine learning techniques applied in sports and healthcare, it's possible to conclude that:

- Data Augmentation: When applied to time series, techniques like rotation, permutation, and time-warping have shown improvements in classification accuracy. On the other hand, for tabular data in unbalanced datasets, SMOTE proved to be a strong choice.
- Feature Reduction and Selection: Techniques such as PCA for feature extraction and methods like KBest and RFE for feature selection show potential in reducing model complexity and improving performance. These techniques have been particularly effective when applied to medical tabular data, enhancing model accuracy and stability.
- Unsupervised Learning: Methods such as K-means clustering and GMM have provided valuable insights into athlete performance and rehabilitation strategies.
- Supervised Learning: Algorithms such as SVM and Random Forests have been effective in predicting outcomes like RPE and diagnosing conditions such as chronic liver disease.

Having explored the application of machine learning in sports and healthcare, it is now important to reflect on how the principles and methods used in these fields can be adapted and applied to the field of prosthetics, particularly in identifying activity profiles for the assessment of K-levels. As referred in Chapter 1 and Chapter 2, the current assessment of K-levels is susceptible to variability and subjectivity, since it depends on the interpretation of the clinician. Since unsupervised learning techniques such as clustering do not depend on pre-attributed labels, they can be particularly useful in identifying activity profiles. By applying clustering methods such as K-means and GMM to metrics obtained from data collected from wearable sensors, it would be possible to group prosthetic users with similar activity patterns without the need to provide clinical K-levels. By doing so, it would allow a more objective analysis of amputees functional abilities. On the other hand, supervised learning techniques offer another view and could be utilized for the task of identifying K-levels based on labeled data (clinical K-levels).

To sum up, machine learning offers powerful tools for understanding human performance and health outcomes, opening the door for more personalized treatments and improved outcomes.

Chapter 4

Dataset

Before applying any machine learning approach, it is crucial to understand and analyze the available dataset. In this chapter, the dataset is presented, by detailing the patient population, exploring the activity metrics and explaining the preprocessing steps and the data augmentation process.

4.1 Data Acquisition Workflow

The data acquisition process using the Motio StepWatch includes eight steps to ensure correct usage. Initially, the Motio StepWatch has to be installed on the patient's prosthetic device. Then, the patient is subjected to some functional tests supported by the app, such as TUG, PLUS-M and AMPPRO. After this, it is necessary to set up the device to record activity with the Motio App. At this point, everything is ready and the patient carries out their life for a minimum of 7 days. After this period, the patient returns for a follow-up appointment, where the recorded data is downloaded from the device into the Motio App. Clinicians analyze the collected activity data, and a detailed report is generated with information from both the acquisition period data and the results from the functional tests.

4.2 Patient Populations and Acquisition Characteristics

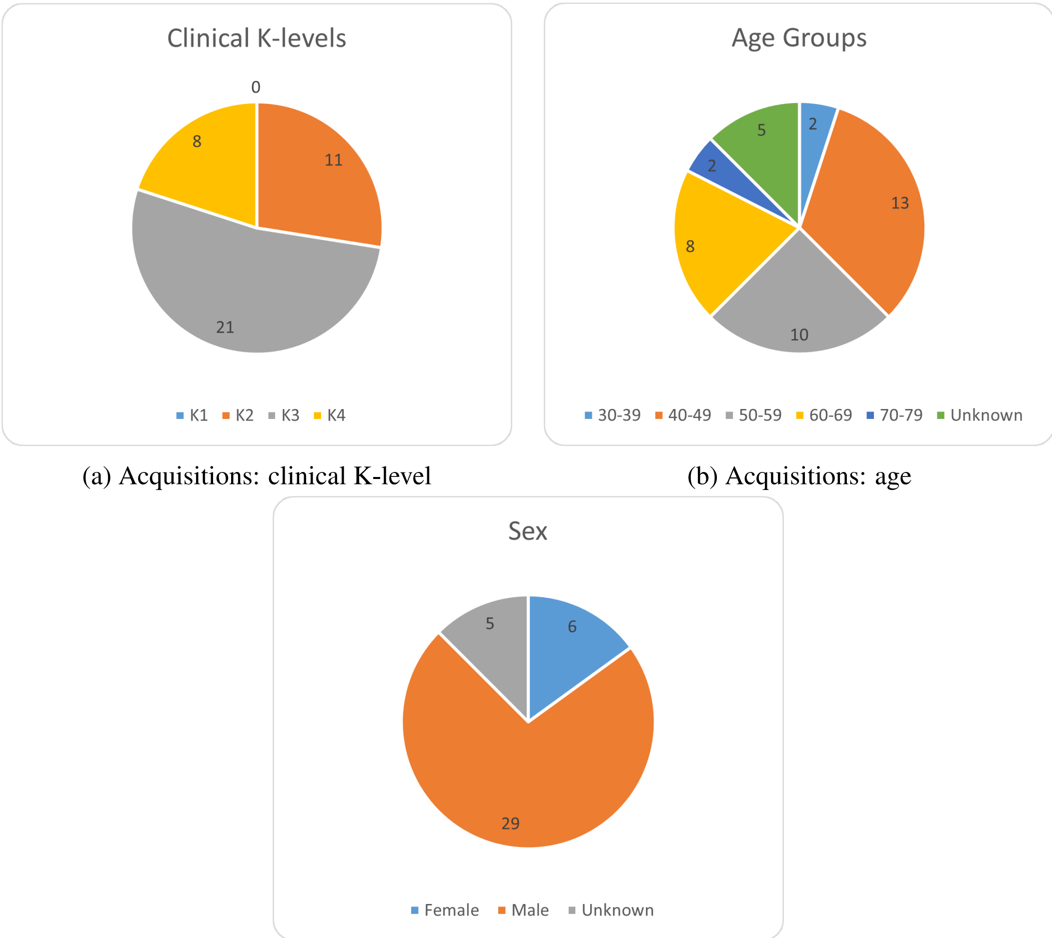
Before sending the patient out for their acquision period, the clinician manually records several data on the Motio App. This information is used as reference to accurately measure steps and cadence during acquisition period. The collected information include:

- Personal information: Patient ID, Age, Gender
- Prosthetic details: Years since amputation, years of prosthetic use, type of amputation, prosthetic components
- Physical measurements: weight, height, stride length

To differentiate patients, an unique alphanumeric identifier was created in order to maintain anonymity. Additionally, multiple acquisitions could be collected in the same patient at different timepoints therefore the start date was recorded as well. In the dataset, we have data from 35 distinct patients.

Additionally, it is also available a clinical K-level for each patient, assessed by a clinician based on the analysis of the data provided by Motio App and their professional recommendation.

Until June 2024, a total of 40 samples have been collected. The distribution of these acquisitions with respect to clinical K-level, age, and sex is shown in Figure 4.1.



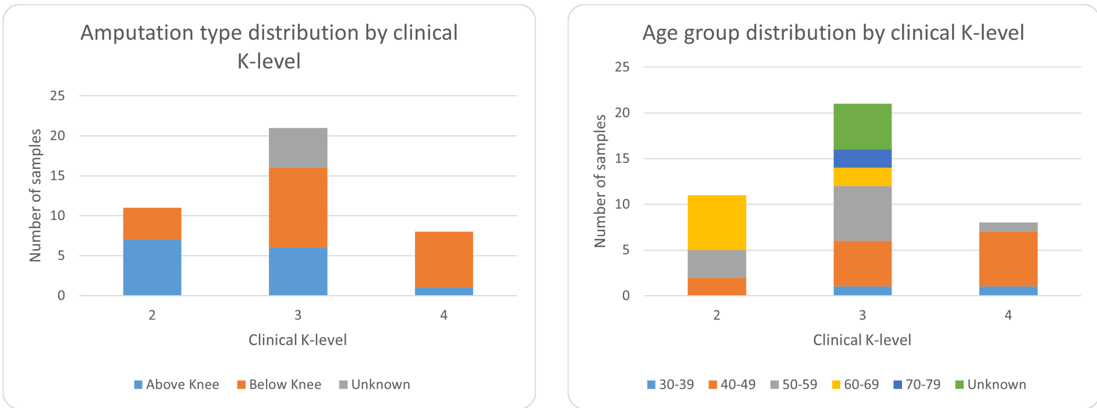
(a) Acquisitions: clinical K-level

(b) Acquisitions: age

(c) Acquisitions: sex

Figure 4.1: Dataset distribution of K-level, age, and sex.

Figure 4.1 shows that the dataset is not balanced regarding the number of samples for each K-level, with K3 having the highest amount of data. It also shows that there is no balance in the number of samples regarding sex, revealing that most of the patients are male. In the United States of America between 2016 and 2021, a total of 5 687 420 cases of limb loss or limb difference were registered. Of these cases, approximately 59% were men (3 342 746) and 41% were women (2 344 674). In the literature, there is evidence that sex influences the outcome of rehabilitation after amputation, with women having a poorer outcome than men [62]. Regarding the age groups, it is possible to see that the dataset covers a wide range of ages, but the majority of the patients are 40 or older, with few representations of younger people.



(a) Distribution: amputation type (b) Distribution: age group

Figure 4.2: Amputation type distribution and age group distribution, both for each clinical K-level.

In Figure 4.2, we have the distribution of type of amputation and age group. It is important to keep in mind that transtibial amputations are considered below knee amputations and transfemoral and knee disarticulation are considered above knee amputations. In Figure 4.2a, it is possible to observe that there are representations of both types of amputation in all clinical K-level.

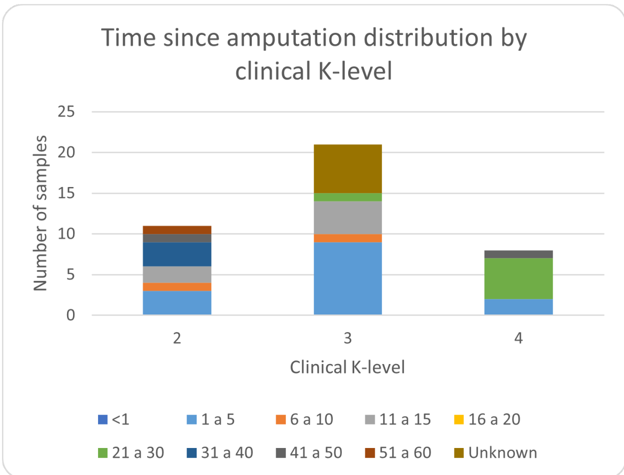


Figure 4.3: Time since amputation distribution for each clinical K-level.

In Figure 4.3, we have the distribution of time since amputation for each clinical K-level. Regarding time since amputation, the dataset covers a wide range of values, guaranteeing diversity of population. In literature, studies suggest that mobility can improve with time since amputation, but may begin to decline after some time [63]. Several factors have high importance in the quality of life of amputees, such as use of prostheses, comorbidities and phantom/residual limb pain [64]. Phantom pain decreases with time since amputation. Although there is no evidence of a direct association between phantom pain and mobility level, it is argued that factors such as phantom/residual limb pain and prosthesis use can indirectly influence both mobility and functional classification level, since pain leads to activity restrictions. Thus, mobility may improve with time since amputation, up to a certain limit.

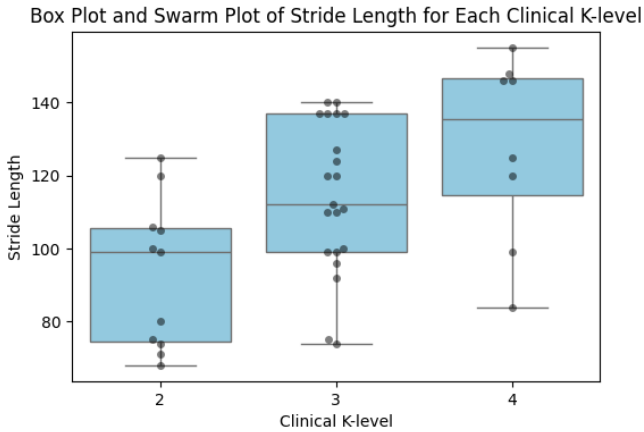


Figure 4.4: Stride length distribution for each clinical K-level.

Figure 4.4 shows the distribution of stride lengths for each clinical K-level. There is evidence that the stride length may be associated with functional mobility. According to the literature, the ability to adapt stride length to facilitate an appropriate speed for normal community activities is an indicator of functional mobility [29]. Thus, if the patient can increase their stride length when necessary, they demonstrate higher functional mobility. Additionally, participants with lower limb amputation walk faster and with longer stride lengths outside the home than inside.

In Figure 4.5, there is the distribution of duration of data acquisition. The number of days was obtained by selecting the first and last day of recorded activity for each patient and counting the number of days within this period.

It is important to highlight that no samples with 1 to 6 days were recorded, showing that all patients followed the indication of recording a minimum of 7 days. Additionally, 14 samples had an acquisition period of 14 to 20 days or more than 20 days, indicating that a significant number of patients collected data for more than twice the minimum required duration. The metrics below were calculated based on the average of 7 to 23 acquisition days.

4.3 Activity Metrics

The recorded data from 40 patients is downloaded into the Motio App, to generate the Motio Report with the calculated activity metrics. In this section, we present the analysis of those activity metrics (described in Section 2.3).

4.3.1 Mean and Standard Deviation

Table 4.1 shows the mean and standard deviation of each metric, splitting the dataset by clinical K-level (K2, K3, K4).

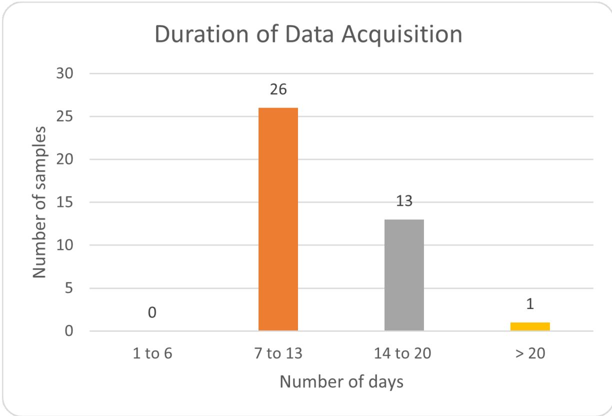


Figure 4.5: Distribution of duration of data acquisition.

Table 4.1: Mean and standard deviation of activity metrics for each K-level.

Activity Metric	K2 (Mean ± STD)	K3 (Mean ± STD)	K4 (Mean ± STD)
StepsDailyAverage	4634.36 ± 2489.01	4264.10 ± 2560.19	5645.00 ± 3597.58
StepsMax	7427.82 ± 3464.16	8176.57 ± 3661.99	10508.25 ± 6073.24
TopCadenceDailyAverage	84.46 ± 14.72	84.54 ± 25.48	86.57 ± 21.70
TopCadenceMax	100.36 ± 11.40	110.29 ± 16.23	112.25 ± 17.10
TopSpeedDailyAverage	0.66 ± 0.20	0.78 ± 0.21	0.92 ± 0.29
TopSpeedMax	0.78 ± 0.19	1.03 ± 0.15	1.20 ± 0.32
EnergyExpenditureDailyAverage	74.45 ± 37.77	69.17 ± 38.95	89.16 ± 55.36
EnergyExpenditureMax	117.28 ± 52.75	129.16 ± 55.87	163.73 ± 93.44
MaxContinuousWalkingDailyAverage	177.85 ± 108.05	180.38 ± 81.96	203.31 ± 126.48
MaxContinuousWalkingMax	356.15 ± 155.79	485.53 ± 258.93	533.84 ± 394.62
AmbulationEnergyScore	3.81 ± 1.13	3.98 ± 0.92	4.12 ± 0.81
PeakPerformanceScore	3.20 ± 0.80	3.02 ± 0.76	3.08 ± 0.76
CadenceVariabilityScore	3.05 ± 0.88	2.88 ± 0.70	2.99 ± 0.92
ActiveTimeAverage	1.46 ± 0.67	1.32 ± 0.78	1.72 ± 0.93
ActiveTimeLowCadenceAverage	0.53 ± 0.27	0.49 ± 0.32	0.63 ± 0.30
ActiveTimeMediumCadenceAverage	0.59 ± 0.28	0.51 ± 0.28	0.65 ± 0.34
ActiveTimeHighCadenceAverage	0.34 ± 0.32	0.31 ± 0.23	0.44 ± 0.38
TwoMinuteWalkTestsAverage	67.73 ± 17.62	85.96 ± 16.87	96.34 ± 24.85
TwoMinuteWalkTestsAverage_std	13.00 ± 4.69	18.60 ± 3.71	22.98 ± 7.99
SixMinuteWalkTestsAverage	231.24 ± 50.35	281.01 ± 52.10	368.62 ± 84.17
SixMinuteWalkTestsAverage_std	13.55 ± 5.96	22.93 ± 18.93	28.68 ± 13.80
TenMeterWalkTestsAverage	0.52 ± 0.13	0.65 ± 0.10	0.69 ± 0.12
TenMeterWalkTestsAverage_std	0.14 ± 0.05	0.19 ± 0.03	0.21 ± 0.04

For StepsMax, the mean number of steps increases with higher clinical K-levels, which suggests that patients with higher K-levels are capable of giving more steps.

For TopCadenceDailyAverage and TopCadenceMax, the mean values show relatively small changes across different clinical K-levels, with slight increases observed from K2 to K4. This suggests that there may be some variations in daily average cadence and higher maximum cadence values among patients with different K-levels, but the differences are not substantial.

Both TopSpeedDailyAverage and TopSpeedMax exhibit significant increases from K2 to K4, which indicate that patients with higher K-levels tend to walk faster.

EnergyExpenditureMax shows noticeable increases in mean values as the clinical K-level increases from K2 to K4. This suggests that patients with higher K-levels tend to expend more energy during physical activities, possibly due to increased mobility and more demanding activities.

Means from both MaxContinuousWalkingDailyAverage and MaxContinuousWalkingMax increase with higher clinical K-levels, which suggests that patients with higher K-levels can sustain continuous walking for longer periods.

AmbulationEnergyScore also demonstrates a slight increase in mean values from K2 to K4. This suggests a trend towards improved ambulatory energy among patients with higher K-levels, which may indicate better physical function and endurance during ambulation tasks.

TwoMinuteWalkTestsAverage and SixMinuteWalkTestsAverage show a significant increase in performance from K2 to K4, which suggests that patients with higher K-levels are able to walk longer periods with improved efficiency. Mean values of TenMeterWalkTestsAverage showed an increase from K2 to K4, indicating faster short-distance walking speeds at higher K-levels. It is also interesting to look at the mean values of the standard deviations of TwoMinuteWalkTestsAverage, SixMinuteWalkTestsAverage and TenMeterWalkTestsAverage, that increase with the higher clinical K-level. They represent the variability in cadence, indicating that higher K-levels correspond to people who are more able to change cadence.

The remaining metrics do not show a clear trend of increase or decrease as the clinical K-level rises.

To confirm these observations, we perform some methods to quantify them.

4.3.2 Spearman's Correlation Between Activity Metrics and K-levels

The strength and direction of a monotonic connection between two variables are measured using Spearman's correlation. The Spearman's correlation analysis between activity metrics and clinical K-level is important to reveal some significant relationships. The correlation between two variables is quantified by a number from -1 to +1. The 0 value means that there is no correlation and 1 means a perfect correlation, while the sign indicates the direction of the correlation (negative indicates an inverse correlation) [65]. In general, our expectation regarding these activity metrics is that when there is a correlation, it will be in a positive direction. We can classify correlation magnitudes into the following levels [66]:

- <0.20 - very weak
- 0.20 to 0.39 - weak
- 0.40 to 0.59 - moderate
- 0.60 to 0.79 - strong

- >0.80 - very strong

Table 4.2: Spearman's correlation coefficients and p-values, between each activity metric and clinical K-level.

Activity metric	Spearman's correlation with K-level	p-value
StepsDailyAverage	0.05	0.75
StepsMax	0.14	0.40
TopCadenceDailyAverage	0.09	0.58
TopCadenceMax	0.30	0.06
TopSpeedDailyAverage	0.26	0.10
TopSpeedMax	**0.54**	**<0.005**
EnergyExpenditureDailyAverage	0.03	0.86
EnergyExpenditureMax	0.11	0.50
MaxContinuousWalkingDailyAverage	0.05	0.75
MaxContinuousWalkingMax	0.17	0.30
AmbulationEnergyScore	-0.01	0.96
PeakPerformanceScore	-0.07	0.66
CadenceVariabilityScore	-0.03	0.86
ActiveTimeAverage	0.03	0.87
ActiveTimeLowCadenceAverage	0.05	0.75
ActiveTimeMediumCadenceAverage	-0.02	0.92
ActiveTimeHighCadenceAverage	0.14	0.40
TwoMinuteWalkTestsAverage	**0.42**	**0.01**
TwoMinuteWalkTestsAverage_std	**0.47**	**<0.005**
SixMinuteWalkTestsAverage	**0.58**	**<0.005**
SixMinuteWalkTestsAverage_std	0.35	0.11
TenMeterWalkTestsAverage	**0.46**	**<0.005**
TenMeterWalkTestsAverage_std	**0.52**	**<0.005**

Table 4.2 summarizes the Spearman's correlation coefficients and p-values for each activity metrics. It is possible to observe that the clinical K-levels have a significant positive correlation with some activity metrics, such as TopSpeedMax, TwoMinuteWalkTestsAverage, SixMinuteWalkTestsAverage and TenMeterWalkTestsAverage. In all these cases, the coefficients are in range 0.4-0.59 and p-values are less than 0.05. The remaining metrics show weak correlations with clinical K-levels, suggesting that these metrics do not seem to have a meaningful relationship with clinical K-levels.

4.3.3 Spearman Correlation Matrix

The Spearman correlation matrix is used to assess the relationship between two variables in the dataset. Figure 4.6 shows that various activity metrics have very strong relationships between each other.

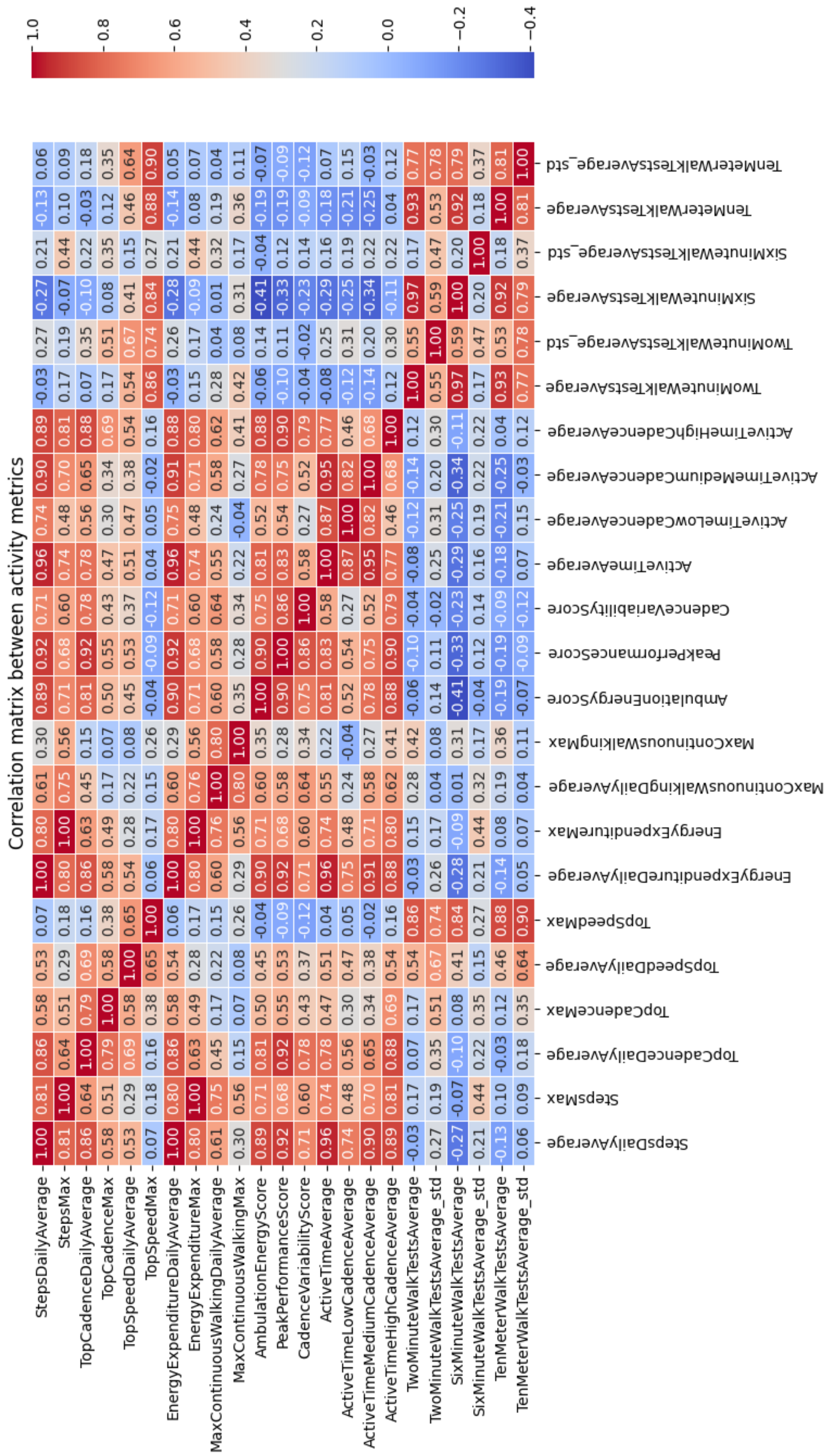


Figure 4.6: Spearman correlation matrix between activity metrics.

A strong positive correlation is observed between StepsDailyAverage and EnergyExpenditureDailyAverage (1.00), PeakPerformanceScore (0.916), and ActiveTimeAverage (0.96). Similarly, StepsMax shows a very high correlation with EnergyExpenditureMax (1.00).

Metrics related to cadence also show strong correlations. TopCadenceDailyAverage has a high correlation with PeakPerformanceScore (0.92).

EnergyExpenditureDailyAverage and PeakPerformanceScore have a strong positive correlation (0.92), which indicates that higher daily energy expenditure is closely associated with a higher peak performance score.

Several metrics related to activity duration, such as ActiveTimeAverage, are also highly correlated with other activity metrics. For instance, ActiveTimeAverage has strong correlations with EnergyExpenditureDailyAverage (0.96).

Performance metrics from walk tests also display significant correlations. Specifically, TwoMinuteWalkTestsAverage and SixMinuteWalkTestsAverage have a very high correlation (0.97), similarly to TwoMinuteWalkTestsAverage and SixMinuteWalkTestsAverage (0.93). Additionally, TwoMinuteWalkTestsAverage and TenMeterWalkTestsAverage also show a strong positive correlation (0.92).

These results indicate that a lot of information is being repeated, which reveals the need to apply feature reduction and selection methodologies.

4.3.4 Statistical Comparison using T-test

The t-test is one of the most commonly used statistical tests. Its purpose is to determine if there is a significant difference between the means of two groups, so we applied a two tailed t-test for independent samples. For the t-test to be valid, the data must follow a normal distribution. The t-test calculates a t-statistic and a p-value. If the p-value is less than 0.05, the difference between groups is considered statistically significant [67].

Table 4.3 summarizes the results of the t-test comparisons across different K-levels for various activity metrics. If the metric data did not follow a normal distribution, we recorded the result as "NND" (non-normal distribution). If the t-test did not yield a p-value less than 0.05, the result was considered "NC" (not conclusive). For cases where the p-value was less than 0.05, the t-statistic and p-value are provided.

Interestingly, most of the statistically significant differences between groups are found when comparing K2 and K3 or K2 and K4, with only one significant difference observed between K3 and K4. This indicates that, in the metrics that followed a normal distribution, the largest changes in activity metrics occur when moving from the lower to higher K-levels, with less pronounced differences between the higher K-levels.

4.4 Preprocessing of Time Series

To make a significant increase in the number of samples, while ensuring we are using the same standardized time window for metrics computations, we apply the sliding window method. Since

Table 4.3: T-test statistical comparison of activity metrics across K-levels. NND: non-normal distribution; NC: not conclusive.

Activity metric	**K2 vs K3**	**K2 vs K4**	**K3 vs K4**
StepsDailyAverage	NND	NND	NND
StepsMax	NND	NC	NND
TopCadenceDailyAverage	NC	NC	NC
TopCadenceMax	NC	NC	NC
TopSpeedDailyAverage	NC	t-statistic: -2.12, p-value: 0.05	NC
TopSpeedMax	t-statistic: -3.93, p-value: <0.005	t-statistic: -3.41, p-value: <0.005	NC
EnergyExpenditureDailyAverage	NND	NND	NND
EnergyExpenditureMax	NND	NC	NND
MaxContinuousWalkingDailyAverage	NC	NC	NC
MaxContinuousWalkingMax	NC	NC	NC
AmbulationEnergyScore	NND	NND	NND
PeakPerformanceScore	NND	NND	NC
CadenceVariabilityScore	NC	NC	NC
ActiveTimeAverage	NND	NND	NND
ActiveTimeLowCadenceAverage	NND	NC	NND
ActiveTimeMediumCadenceAverage	NND	NND	NND
ActiveTimeHighCadenceAverage	NND	NND	NND
TwoMinuteWalkTestsAverageStd	t-statistic: -2.77, p-value: 0.01	t-statistic: -2.78, p-value: 0.01	NC
TwoMinuteWalkTestsAverage_std	t-statistic: -3.58, p-value: <0.005	t-statistic: -3.23, p-value: <0.005	NC
SixMinuteWalkTestsAverageStd	t-statistic: -2.14, p-value: 0.04	t-statistic: -3.39, p-value: 0.01	t-statistic: -2.65, p-value: 0.02
SixMinuteWalkTestsAverage_std	NND	t-statistic: -2.28, p-value: 0.05	NND
TenMeterWalkTestsAverageStd	t-statistic: -3.05, p-value: <0.005	t-statistic: -2.83, p-value: 0.01	NC
TenMeterWalkTestsAverage_std	t-statistic: -3.85, p-value: <0.005	t-statistic: -3.36, p-value: <0.005	NC

we have samples with more than the minimum duration indicated (7 days), the sliding window method is used to extract 7-days samples from each original sample. To apply this technique, we used a window with a length of 7 days to extract samples from the original samples (since in literature 7 is the number of days recorded [4]). Then, the window moves forward the time step defined, in this case 1 day, to extract the next sample. This process is illustrated in Figure 4.7, with an example of an original sample of 9 days.

Table 4.4 summarizes the number of samples for each clinical K-level and in total before and after applying the sliding window. It is evident that the number of samples increases significantly by applying the sliding window technique. It is important to highlight that we kept track of which acquisition each new sample is originated from.

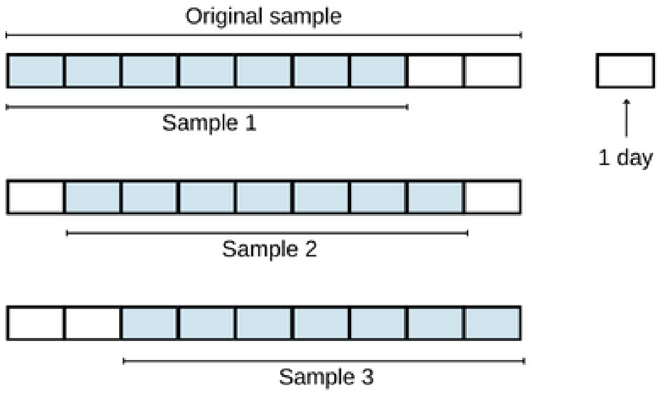


Figure 4.7: Sliding window technique exemplified with a sample of 9 days.

Table 4.4: Number of samples before and after applying the sliding window.

	Raw data	**Sliding window**
K2	11	87
K3	21	179
K4	8	50
Total	40	316

4.5 Preprocessing of Tabular Data

4.5.1 Patient Information

Considering the information explored in Section 4.2, we found that the following patient information might be interesting: age group, time since amputation, type of amputation, and stride length. However, as we can see in Table 4.5, there are some missing values.

Table 4.5: Number of missing data in appointment data.

Appointment data	**Number of missing data**
Age	33
Years since amputation	55
Type of amputation	33
Stride length	0

To address this problem, for the type of amputation, missing values were filled by considering the most frequent category (mode). For age, after converting the age from groups to integers (considering the minimum integer of each group), we handle missing age values by replacing them with the median age of the dataset. For time since amputation , we also filled missing values with the median value of the available data.

4.5.2 Activity Metrics

Table 4.6: Number of missing data in activity metrics.

Activity metric	Number of missing data
AmbulationEnergyScore	5
PeakPerformanceScore	5
CadenceVariabilityScore	5
TwoMinuteWalkTestsAverage_std	1
SixMinuteWalkTestsAverage	129
SixMinuteWalkTestsAverage_std	167

In Table 4.6, it is possible to observe the activity metrics with missing values and the respective quantities of missing data. For the activity metrics, missing values were filled using the minimum observed value for each metric.

4.6 Data Augmentation

As we have seen in this Chapter, the dataset is quite small and includes samples highly interdependent (same patient, in consecutive time windows). Apart from that, the amount of data for each K-level is not balanced. Therefore, to address these problems, data augmentation techniques are needed. This section focuses on explaining the data augmentation process used in this project (Figure 4.8), including two data augmentation methods: noise addition and SMOTE. The noise addition is applied to the series of timestamps, while SMOTE is designed to work on tabular data (not being suitable for time series) and specifically targets class imbalance by generating synthetic samples for the minority class (as explained in Chapter 3).

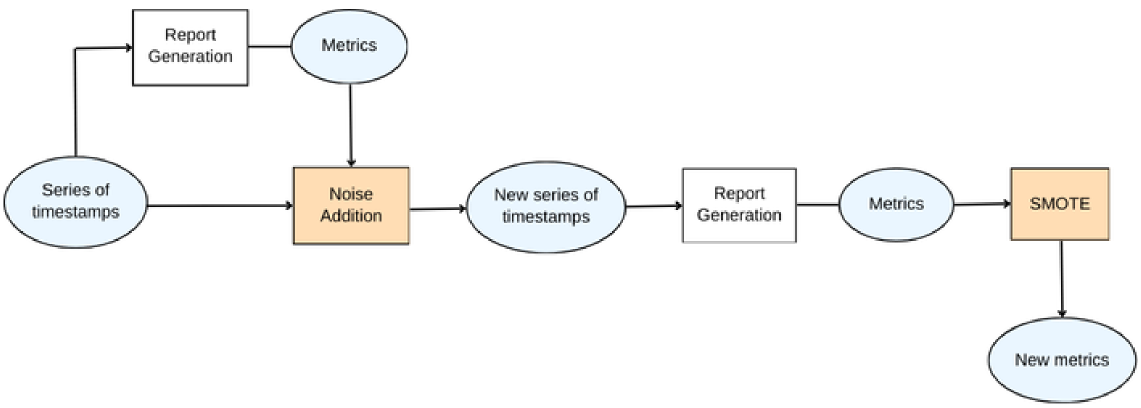


Figure 4.8: Representation of the data augmentation process used in this work.

4.6.1 Noise Addition

The StepWatch device has approximately 90% accuracy in recording daily steps [68]. The noise addition method is based on this accuracy to create new raw data by incorporating the inherent uncertainty present in the device used for data recording.

This process takes as input the StepsDailyAverage and the series of timestamps during 7 days. Knowing that the device has an uncertainty of 10% in the number of steps per day, it is possible to determine the potential range of the number of steps. The minimum adjustment value is calculated by subtracting 10% from the steps daily average, and the maximum adjustment value is determined by adding 10% to the steps daily average. These values are then multiplied by the number of days to obtain the total number of steps for both cases. Next, a random value within the adjustment range is generated. If the total steps are greater than this random adjustment, excess data points are removed until the desired adjustment is achieved. Conversely, if the total steps are less than the random adjustment, additional data points are added to the time series. The output of this process is one new data for each original data, along with a file containing appointment information.

This process was applied to all samples with the aim of increasing the dataset size, doubling the number of data samples (Table 4.7). The generated samples were subjected to the same pre-processing of tabular data, referred in Section 4.5.

Table 4.7: Number of samples before and after applying the noise addition method.

	Sliding window	**Noise addition**
K2	87	174
K3	179	358
K4	50	100
Total	316	632

4.6.2 SMOTE

The purpose of the SMOTE method is to address class imbalance by generating synthetic data for the minority class. This method is applied to the tabular data (activity metrics). The raw data is first prepared for SMOTE augmentation by normalizing metrics using the standard normalization, which transforms the distribution of each metric to have a mean of zero and a standard deviation of one. It can be expressed mathematically as $z = (x - \mu)/\sigma$, where μ is the mean and σ is the standard deviation. Then, missing values are filled as explained in Section 4.5.

SMOTE's first steps are the definition of the minority class and the number of neighbors (k) to consider. Usually, k is set to 5, which is the value that we used. This way, it is possible to calculate the number of samples required to expand the minority class by subtracting the number of samples from the minority class to the number of samples from the majority class. A subset of samples from the minority class is randomly selected for interpolation, referred to as central samples. A KNN model is fitter on the minority class data to find the nearest neighbors for those central samples. For each central sample and its nearest neighbors, new synthetic data is generated by interpolating between the central sample and its neighbors.

The output of the SMOTE method is a set of new synthetic metrics and the appointment data for each new sample generated, which corresponds to the appointment data of the central sample used to generate the new data.

4.7 Summary

With the analysis made in this chapter, it became clear that the dataset does not have many samples, with some belonging to the same patients, and it is not balanced. This motivates the application of data augmentation techniques, including noise addition and SMOTE.

Regarding activity metrics, it was possible to conclude two things. Firstly, it was found that a great amount of information is being repeated, which indicates that feature reduction and selection would be beneficial. Secondly, the majority of the statistically significant differences were between K2 and K3 or K2 and K4, not having much evidence on K3 and K4.

Taking into account the knowledge acquired about the dataset, the next chapters explore the application of this data in unsupervised and supervised techniques for the identification of activity levels in prosthetic users.

Chapter 5

Unsupervised Learning

After exploring the work done in P&O and other fields (Chapter 3) and analyzing the dataset (Chapter 4), it is time to explore clustering methods to group prosthetic users with similar activity patterns. Thus, two data approaches are explored: using activity metrics or appointment data with activity metrics (combined). Before applying the models, we augment data using noise addition and SMOTE. For clustering, two algorithm are tested: K-means and GMM. In this chapter, the methodology used is explained and the results are discussed.

5.1 Methods

The methodology used in this project for clustering activity patterns involves several steps, shown in Figure 5.1 and explained below.

5.1.1 Dataset

After applying the sliding window (presented in Section 4.4), the dataset (Dataset 1) contains 316 samples. However, this dataset still does not have many samples and is not balanced. Due to this, we apply the data augmentation process explained in Section 4.6. After noise addition, we obtain 632 samples. Then, since K3 is the majority class, we apply SMOTE to classes K2 and K4 to balance the dataset (Table 5.1). The dataset obtained after SMOTE is referred as Dataset 2. With SMOTE, new activity metrics are generated and, for each new sample, the appointment data corresponds to the appointment data of the central sample used to generate it.

Table 5.1: Number of samples in the dataset before and after applying SMOTE.

	Noise addition	**SMOTE**
K2	174	354
K3	358	358
K4	100	355
Total	632	1067

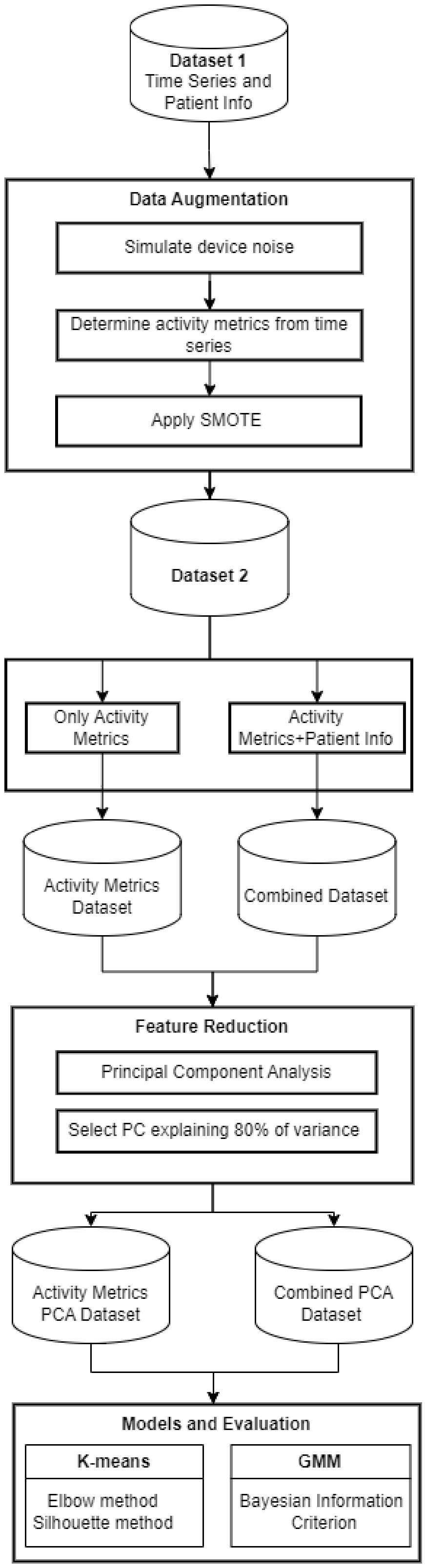


Figure 5.1: Methodology used for clustering prosthetic users.

Two data approaches are considered for clustering: only using activity metrics (total of 23 features) or using appointment data, including age, stride length, amputation type, and time since amputation, combined with activity metrics (total of 27 features). In both cases, it is necessary to fill the missing values as explained in Section 4.5. Then, numerical values need to be normalized. It is done by applying standard normalization so that data has a mean of zero and a standard deviation of one.

5.1.2 Feature Reduction

To reduce the number of features, we use PCA. Firstly, it is necessary to define how many components we want to keep. For that, we decided that a variance of 80% should be preserved. Then, PCA is performed with the number of principal components that meet that condition. The obtained PCA scores are incorporated into the clustering algorithms.

5.1.3 Clustering Approaches

5.1.3.1 K-means Clustering Algorithm

One of the most challenging issues in clustering is choosing the ideal number of clusters. For K-means model, the most commonly used method to determine the best number of clusters is the Elbow method. Using this process, we tested different numbers of clusters (k) from 2 to 20. The Elbow method is done by performing K-means clustering for each number of clusters and computing the Within-Cluster-Sum of Squared Errors (WSS), which is the sum of the squared difference between each data point and the centroid of each cluster, using the following equation:

$$\text{WCSS} = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$

where x_i is the i-th data point, μ_k is the centroid of the k-th cluster, and C_k is the k-th cluster.

Then, we plot these values to identify the point where the variation slows down, forming an elbow in the curve. The elbow point is considered the optimal number of clusters for K-means. However, the elbow point may not always be easy to identify, making it necessary to complement this assessment with other methods.

The silhouette method is an alternative way to assess the optimal number of clusters. It measures how similar a data point is to its own cluster, compared to other clusters.

The equation for calculating the silhouette coefficient for a particular data point i is:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where $a(i)$ is the average Euclidean distance between point i and all the other data points in the same cluster, and $b(i)$ is the average distance from point i to all clusters to which point i does not belong.

The silhouette coefficient ranges from -1 to 1. A coefficient of 1 means that the data point fits very well within the cluster and is far away from the neighboring clusters, a value of -1 means that the data may have been assigned to the wrong cluster and 0 indicates overlapping clusters.

We computed the silhouette coefficient for all data points for each k, having tested a range of clusters from 2 to 5.

To choose the correct k using the Silhouette plot, it is important to consider the following conditions:

1. **Maximization of Silhouette Scores:** High silhouette scores indicate strong separation between clusters, suggesting a better fit of data points to their respective clusters.

2. **Cluster Homogeneity:** All clusters must have higher silhouette scores than the average score of the entire dataset, represented by the red-dotted line in the plots.

3. **Balanced Cluster Sizes:** In datasets with balanced distributions, the silhouette method considers the width of clusters, representing the number of data points assigned to each cluster. It is recommended maintaining relatively consistent cluster sizes.

Once the optimal number of clusters is determined, K-means clustering can be applied. The K-means model is fitted with the PCA transformed data and the cluster labels are defined.

5.1.3.2 GMM Clustering Algorithm

Gaussian mixture model (GMM) clustering is a probabilistic model. It assumes the data is obtained from a mixture of many Gaussian distributions with unknown characteristics and tries to illustrate those normally distributed subpopulations in a large population.

To apply this model, it is necessary to first determine the best number of components. In this case, it was done using the Bayesian Information Criterion (BIC). The BIC is a good choice to be used with GMM, since it establishes an appropriate balance between model complexity (number of components) and model fit (likelihood), being designed to work with probabilistic models.

In this process, GMM models with different numbers of components (from 1 to 10) and covariance types (full, diagonal and spherical) are fitted with the data. Then, for each model, the BIC score is computed using the following equation:

$$\text{BIC} = -2\ln(L) + p\ln(N)$$

where L is the likelihood of the model, p is the number of parameters in the model and N is the number of data points.

Then, the BIC scores are plotted to help with the visualization. The model's ability to cluster the available data is higher when the BIC is smaller. This method also penalizes models with a large number of clusters to prevent overfitting. So, lower BIC values indicate a better model fit.

To help determine the best number of clusters, the gradient of the BIC scores is also obtained to analyze the point at which the BIC score varies the most and when that variation decreases and becomes almost constant.

After this analysis, the GMM model with the best number of clusters is selected.

5.1.4 Cluster Analysis

To visually evaluate the separation and overlap of the clusters, the data is projected into a 2D plot using the first and second principal components, since they are the ones that represent the most variance.

We also obtain plots with the distribution of clinical K-levels (K2, K3 and K4) within each cluster, to assess the composition of clusters in terms of patient activity levels attributed in clinical environment.

To compare the mean values of each feature across clusters, we use a radar chart. This plot allows the visualization of the main differences in activity between individuals within each cluster.

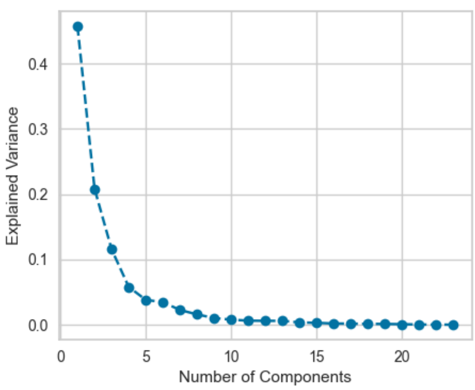
5.2 Results and Discussion

5.2.1 Approach 1: Activity Metrics Dataset 2

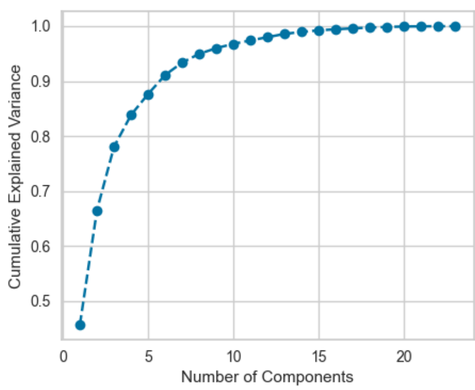
5.2.1.1 PCA

Figure 5.2a shows the individual explained variance for each principal component, for Dataset 2. The first principal component accounts for more than 40% of the variance, indicating that this component is the most important component. With an explained variance ratio of approximately 20% of the variance in the original dataset, the second component is less significant than the first one.

Figure 5.2b shows the cumulative explained variance, for Dataset 2. As the number of principal components increases, the cumulative explained variance approaches a saturation point, which indicates that additional components contribute progressively less to the overall variance. Since we wanted to preserve a minimum of 80% of the variance, four principal components were kept.



(a) Individual explained variance



(b) Cumulative explained variance

Figure 5.2: PCA using activity metrics Dataset 2.

5.2.1.2 K-means

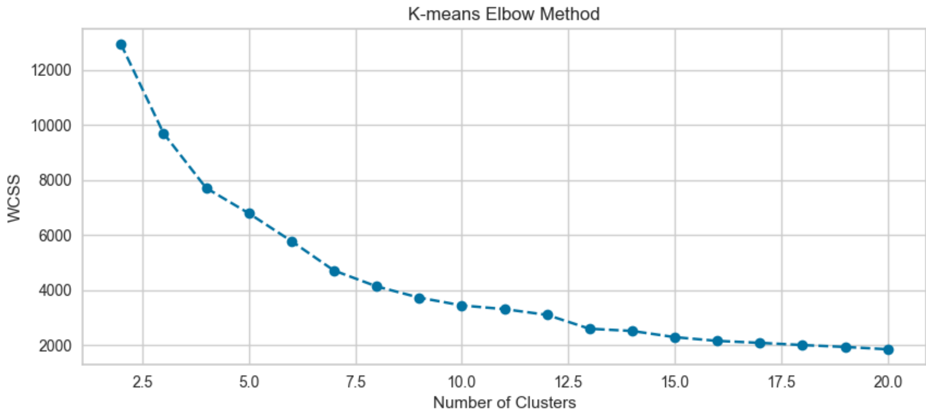


Figure 5.3: Elbow method applied to K-means using activity metrics Dataset 2.

The Elbow Method plot (Figure 5.3) shows the relationship between the number of clusters and the within-cluster sum of squares. From this plot, it is not easy to find the first elbow point, leaving some doubts between k=3 and k=4. This suggests that three or four clusters may be an appropriate choice for clustering the data.

As a complement to this assessment, it is important to analyze the plots obtained for the Silhouette method (Figure 5.4). Evaluating the condition 1 indicated in Section 5.1.3.1, it is possible to see in the plots that the maximum score is obtained for k=3. Although k=3 has the highest individual silhouette score, it is important to analyze the other factors referred in Section 5.1.3.1. All the clusters exhibit silhouette scores higher than the average score, so meet the criteria for condition 2. Regarding the width of the clusters, none of the k values have balanced cluster widths, so in this case we only consider the other criterions.

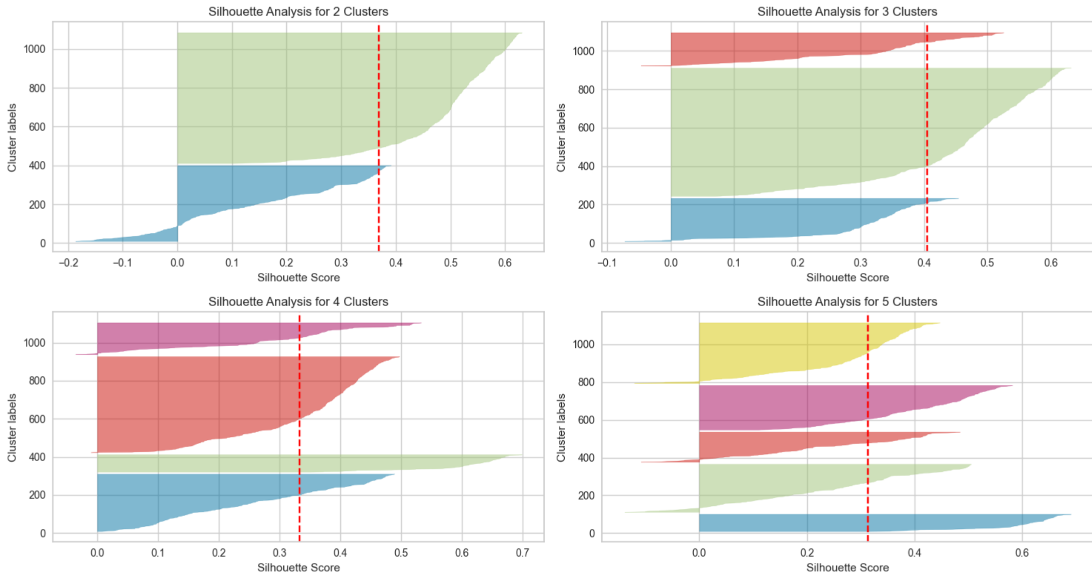
While the Elbow Method suggests k=3 or k=4, the Silhouette method indicates k=3. Therefore, considering both methods, k=3 seems to be the best choice for clustering this data using K-means.

Figure 5.5a represents the projection of the clusters obtained by the model, with different colors to allow the observation of the separate clusters. Since it is a 2D visualization, it was mandatory to select only 2 dimensions. So the first and second components were selected (representing the most variance).

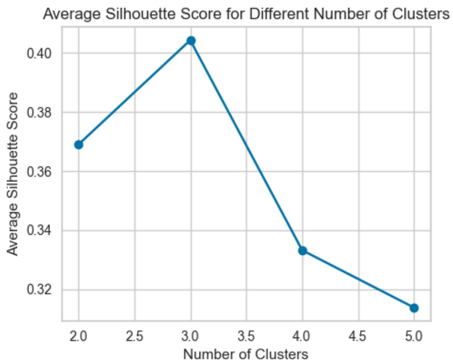
It is possible to observe that the clusters are well-defined, with minimal overlap between them, which indicates that the K-means algorithm successfully identified distinct groups within the data. Only by looking at this plot it is visible that, even though the dataset is balanced, the density of points varies between clusters.

In Figure 5.5b, we have the distribution of clinical K-levels across each cluster. This analysis helps to explore the composition of each cluster concerning the clinical K-levels.

- Cluster 0: contains predominantly samples with a clinical K-level of 4, with fewer samples of K3 and no instances of K2.

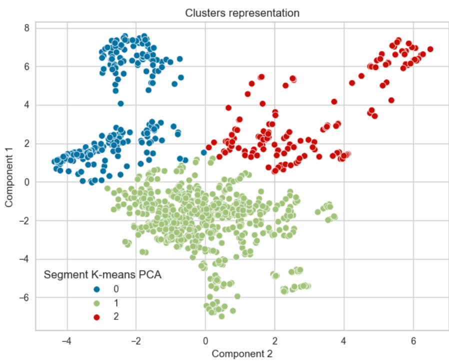


(a) Silhouette analysis plot for k=2 to k=5

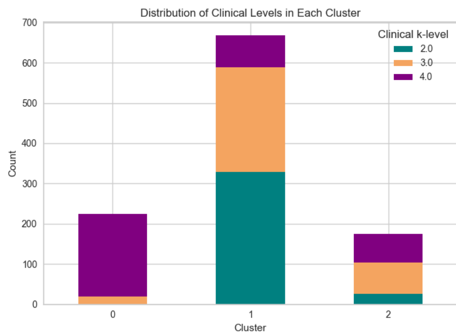


(b) Average silhouette score distribution

Figure 5.4: Silhouette method applied to K-means using activity metrics Dataset 2.



(a) Cluster representation



(b) K-level distribution across clusters

Figure 5.5: K-means: cluster analysis (k=3) using activity metrics Dataset 2.

- Cluster 1: contains the majority of the samples of K2 and some K3, having fewer instances of K4.

- Cluster 2: has a distribution with fewer instances at K2 and an equivalent number of K3 and K4.

The variation in the distribution of K-levels across the clusters suggests that the clusters represent different segments within the data. In short, Cluster 0 is highly concentrated with higher clinical K-levels, whereas Cluster 1 shows a distribution with medium to low clinical K-levels and Cluster 2 has a balanced distribution across medium to high clinical K-levels.

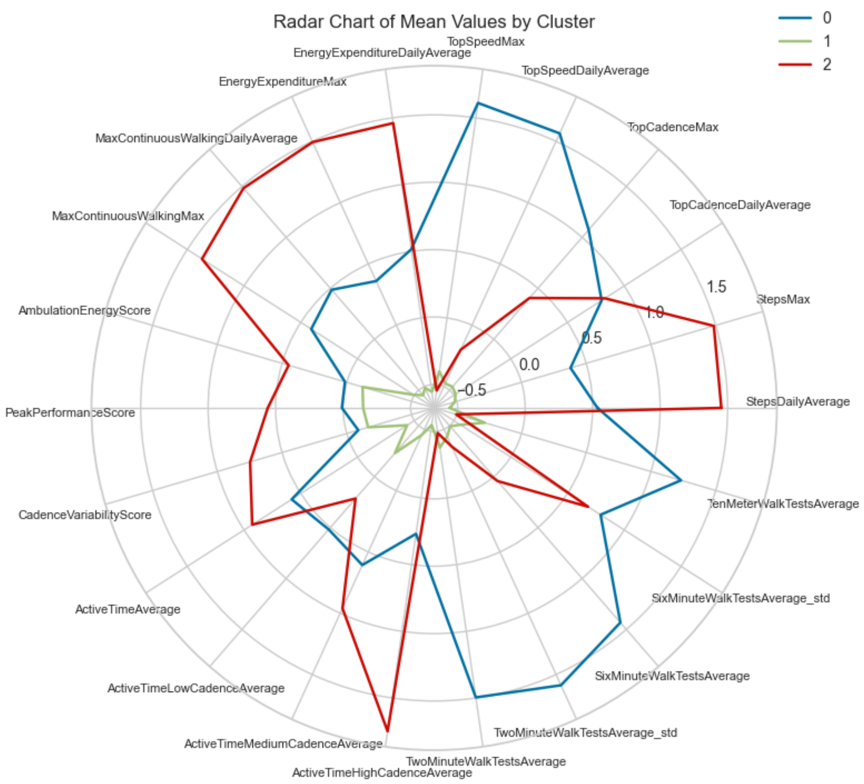


Figure 5.6: K-means: radar chart of mean values for each activity metric from Dataset 2, by cluster.

The radar chart in Figure 5.6 shows the mean values of each activity metrics in their standardized form for the clusters identified by the K-means on Dataset 2. Cluster 0 shows high values across most attributes, especially in TopSpeedMax, TopSpeedDailyAverage, TwoMinuteWalkTestsAverage, SixMinuteWalkTestsAverage and TenMeterWalkTestsAverage. There is evidence in literature that individuals classified as K3 are able to walk shorter distances when doing the 6MWT, compared to K4 [69]. There are also studies that show that lower activity levels are associated with slower gait speed and vice-versa [69][27]. Having in consideration the information in the literature and analysing the values of each activity metric, it is possible to infer that individuals that belong to Cluster 0 have high activity levels, which matches the distribution seen in Figure 5.5b.

In contrast, Cluster 1 consistently shows lower activity values across almost all metrics. Has we saw in Figure 5.5b, even though this cluster has some samples from high activity level, this cluster contains mainly samples with clinical K-level of 2 and 3, selecting almost all existing K2 samples. So, it makes sense that this cluster represents a group with lower overall activity level since it is made mainly by samples with medium to low clinical K-levels.

Cluster 2 shows particularly high values in some activity metrics, such as StepsMax, StepsDailyAverage, EnergyExpenditureMax, EnergyExpenditureDailyAverage, MaxContinuousWalkingDailyAverage and ActiveTimeHighCadenceAverage. Cluster 0 seems to identify samples in which the patient has a high amount of daily walking (MaxContinuousWalkingDailyAverage, StepsDailyAverage) but moves with a low cadence/speed, in contrast to Cluster 2, which despite the lower relative amount of activity, exhibits higher cadence, speed, and variation (TopSpeedMax, TopCadenceMax, TwoMinuteWalkTestsAverage, SixMinuteWalkTestsAverage and TenMeterWalkTestsAverage_std). In Figure 5.5b, we saw that this cluster has few K2 samples and a balance between K3 and K4. The radar chart suggests that Cluster 2 represents a group with relatively high activity levels, which matches the K-level distribution.

5.2.1.3 GMM

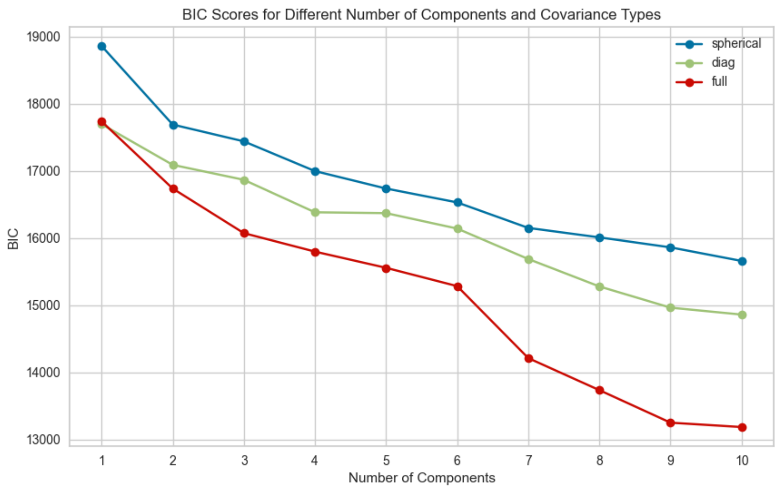
The BIC values (Figure 5.7a) exhibit a clear trend where the variation slows down as the number of clusters increases. The point where this variation becomes minimal corresponds to k=3. Additionally, the covariance type that has the lowest scores is the full covariance. Analyzing the gradient of the BIC curve (Figure 5.7b), it is possible to observe that from a cluster size of four beyond, the gradient becomes almost constant, which indicates that from this point the additional clusters would not benefit the model's performance. Concluding, this technique revealed that k=3 is the best choice for the GMM.

Figure 5.8a shows the 2D representation of the clusters obtained by the GMM model with k=3. The clusters are well-defined, but the density of points varies between clusters. They also have more overlap than in the previous model. This overlap is expected, due to the fact that GMM is a probabilistic model.

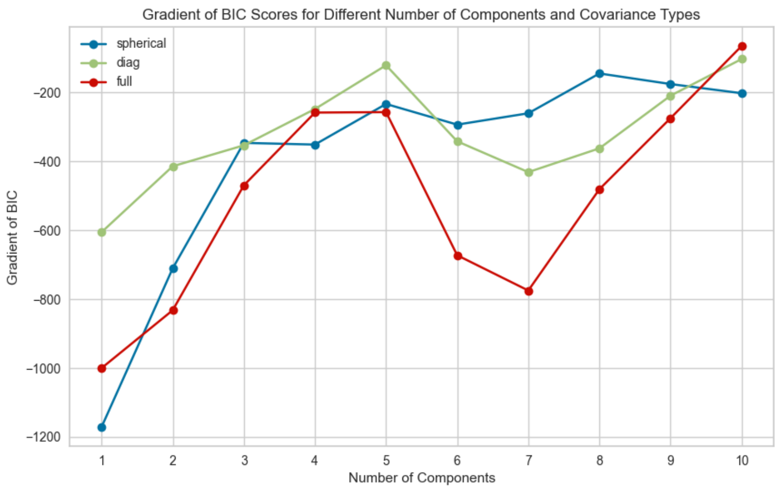
The distribution of clinical K-levels for each cluster is displayed in Figure 5.8b.

- Cluster 0: the majority of the samples have a clinical K-level of 4, with fewer samples of K3 and no K2.

- Cluster 1: contains the majority of samples with a clinical K-level of K2 and K3, with fewer instances of K4.



(a) BIC values



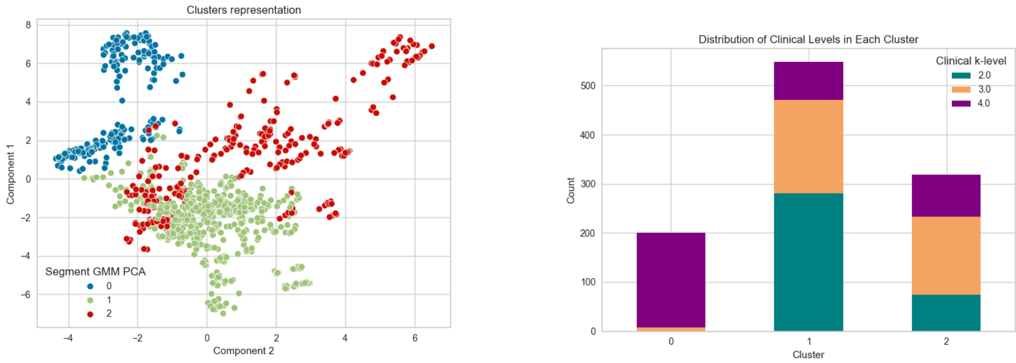
(b) Gradient of BIC

Figure 5.7: BIC applied to GMM using activity metrics Dataset 2.

- Cluster 2: has instances of all three clinical K-levels, with slightly more K3 than the other two K-levels.

Once again, it is possible to conclude that Cluster 0 contains mostly higher clinical K-levels, whereas Cluster 1 shows a distribution with medium to lower clinical K-levels and Cluster 2 has a more balanced distribution across the three clinical K-levels.

From the radar chart in Figure 5.9, it is possible to see that Cluster 0 has the highest mean values for most activity metrics (including TopSpeedMax, TopSpeedDailyAverage, TopCadenceMax, TopCadenceDailyAverage, TwoMinuteWalkTestsAverage, SixMinuteWalkTestsAverage and TenMeterWalkTestsAverage), which suggests that individuals in Cluster 0 have higher levels of functional ability. As referred before, higher values for some activity metrics such as 6MWT and speed



(a) Cluster representation (b) K-level distribution across clusters

Figure 5.8: GMM: cluster analysis (k=3) using activity metrics Dataset 2.

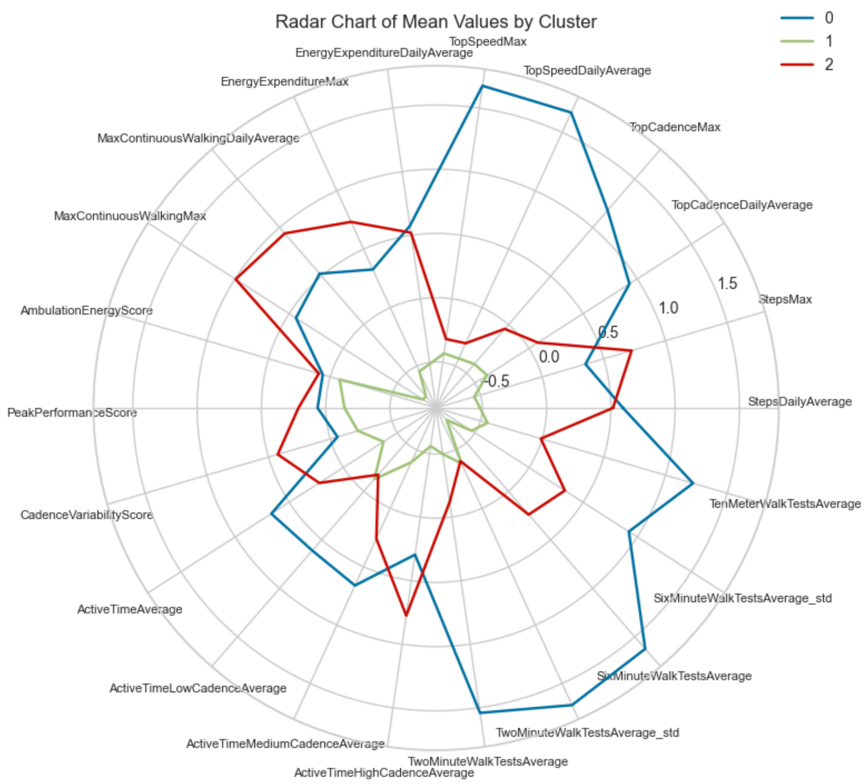


Figure 5.9: GMM: radar chart of mean values for each activity metric from Dataset 2, by cluster.

gait are associated with higher activity levels [69][27]. In fact, in the analysis of K-level distribution plot (Figure 5.8b), we saw that this cluster was mainly composed of K4 samples, which corresponds to the results obtain for activity metrics' values and the literature.

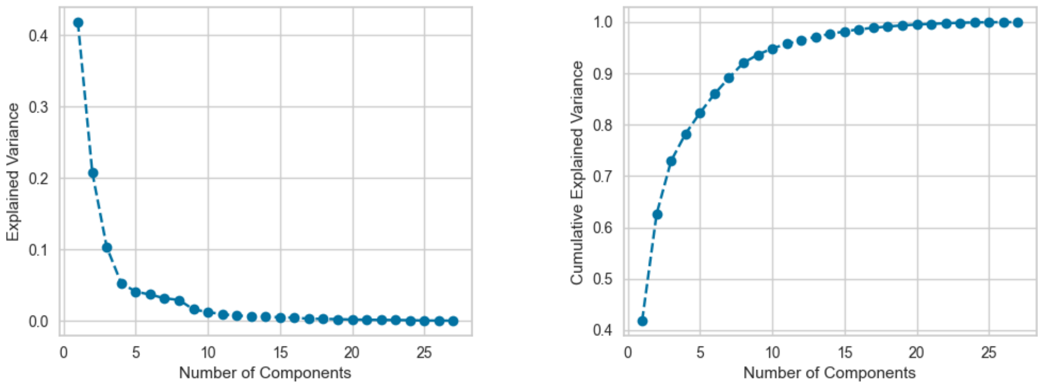
Almost all activity measures in Cluster 1 have negative mean values, indicating lower performance and activity levels than in the other clusters [69][27]. Has seen in K-level distribution plot (Figure 5.8b), this cluster has few samples of K4 and its mainly composed by individuals assessed as K2 and K3, which matches the activity metrics' values obtained.

Cluster 2 demonstrates moderate values for most metrics, with some metrics having relatively high values (MaxContinuousWalkingMax and MaxContinuousWalkingDailyAverage). The activity level in this cluster is balanced, being neither as high as in Cluster 0 nor as low as in Cluster 1. In Figure 5.8b, we saw that the Cluster 2 has samples from the three clinical K-levels, with slightly more K3 but still having a balanced distribution. This is related with the values seen for the activity metrics, that have medium values, not as high as Cluster 1 for most metrics.

5.2.2 Approach 2: Combined Dataset 2

5.2.2.1 PCA

In this approach, PCA was performed using both appointment data and activity metrics from Dataset 2. Similar to Approach 1, a significant amount of the variation is captured by the first principal component generated by PCA, representing more than 40% of the total variance (Figure 5.10a). However, five components must remain in order to meet the established threshold of 80% explained variance, as shown in Figure 5.10b.



(a) Individual explained variance (b) Cumulative explained variance

Figure 5.10: PCA using combined Dataset 2.

5.2.2.2 K-means

Figure 5.11 shows that the first elbow appears at k=3, which indicates that three clusters might be an appropriate choice. To validate this choice, we analyze the Silhouette Method (Figure 5.12). The maximum silhouette score is obtained for k=3 (Figure 5.12a). Furthermore, every cluster with k=3 satisfies requirement 2 since its silhouette scores exceed the dataset's average value (Figure 5.12b). Regarding the width of the clusters, it is not possible to use this criterion to help determine the best number of clusters, since none of the k values exhibit balanced cluster widths. Considering all this analysis, k=3 is seems to be the best option to use in K-means for this approach.

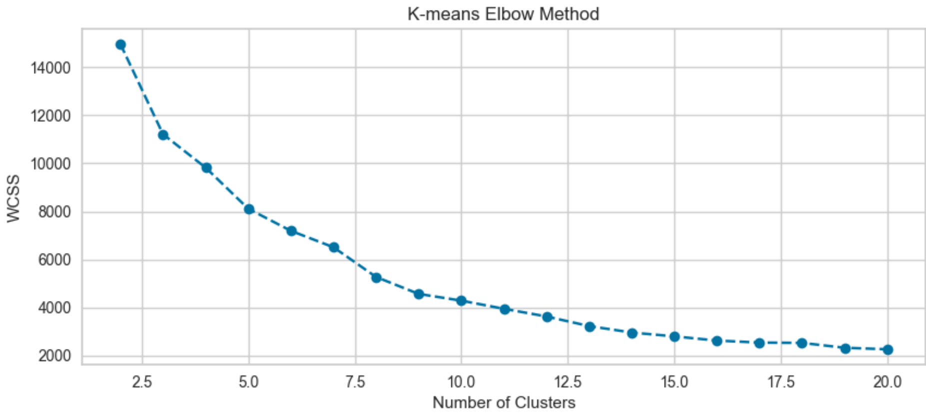


Figure 5.11: Elbow method applied to K-means using combined Dataset 2.

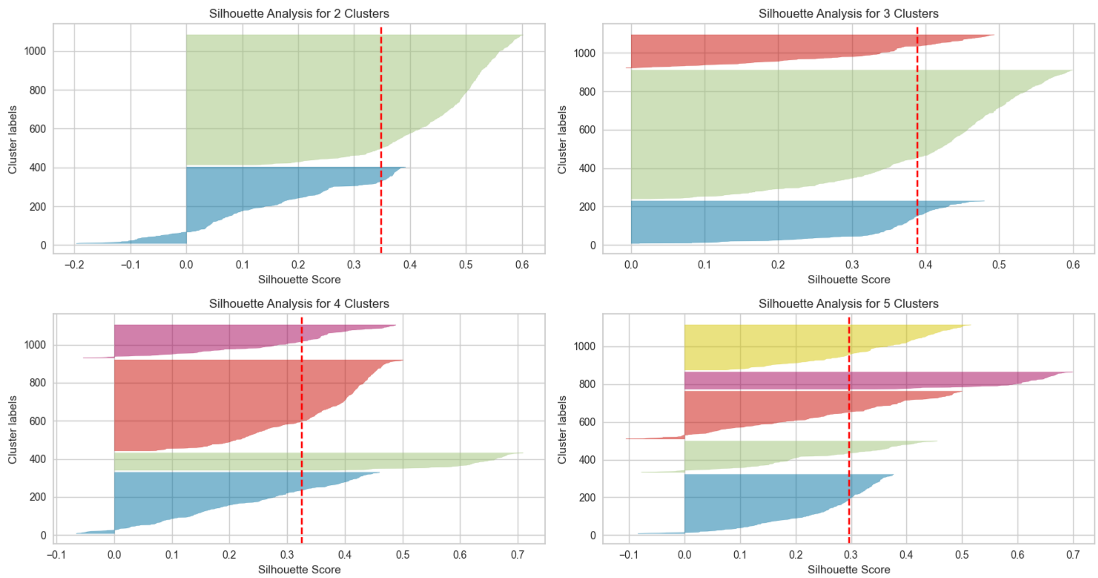
Based on the analysis of Figure 5.13, the results for the cluster representation and K-level distribution using K-means in this approach (using both appointment data and activity metrics) are similar to those obtained in Approach 1 (only using activity metrics).

The radar chart in Figure 5.14 shows the mean values (or the mode for categorical data) of appointment data and activity metrics for each of the three clusters identified by the K-means algorithm. The behavior of the activity metrics is similar to the obtained in Approach 1 for K-means.

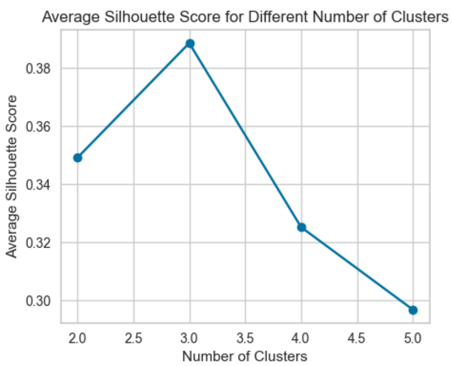
Regarding the appointment data, the average age of Cluster 0 is negative, while Clusters 1 and 2 have positive ages. This means that people from Cluster 0 are generally younger than those in the other two clusters.

Concerning time since amputation, Cluster 0 has a positive average for years since amputation, indicating that individuals in this cluster have more time since amputation on average than the individuals from the other two clusters. This aligns with the idea discussed in the Section 4.2 that mobility may improve with time since amputation, up to a certain limit [63]. The longer time since amputation in Cluster 0 may indicate that these individuals have had more time to adapt to the condition and improve mobility, which is related with the distribution seen in Figure 5.13b that shows that Cluster 0 contains mainly individuals with higher clinical K-levels.

When considering stride length, Cluster 0 shows the highest average, meaning that in general individuals in this cluster have longer strides compared to individuals in Clusters 1 and 2. This observation matches with the literature [29], also mentioned in Section 4.2, which suggests that the ability to adapt stride length is an indicator of functional mobility. The longer stride length in Cluster 0 may indicate better functional mobility, reflecting their ability to adapt to different walking conditions. Again, this is in agreement with the distribution seen in Figure 5.13b.



(a) Silhouette analysis plot for k=2 to k=5



(b) Average silhouette score distribution

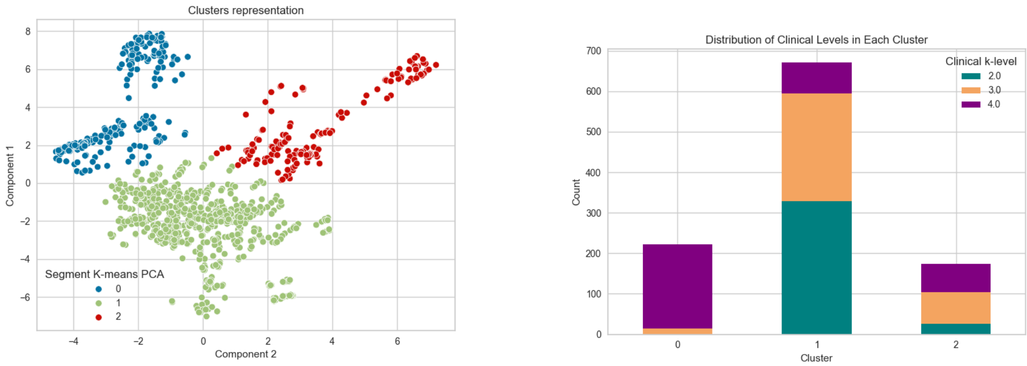
Figure 5.12: Silhouette method applied to K-means using combined Dataset 2.

5.2.2.3 GMM

For this approach, the BIC values (Figure 5.15a) decrease as the number of clusters increase. The first minimal variation point is observed at k=3 and the covariance type that has the lowest scores is the full covariance . Moreover, examining the BIC gradient (Figure 5.15b), when reaching a cluster size of four the gradient levels stagnate. Thus, this analysis supports that k=3 is the optimal number of clusters for the GMM in this approach.

The cluster representation (Figure 5.16a) shows the distribution of samples among the three clusters identified by the GMM. Similar to the results obtained in Approach 1, the clusters are well-separated. However, for this approach there is less overlap.

The distribution of clinical K-levels across each cluster is illustrated in Figure 5.16b. Cluster 0 consists of samples with clinical K-level 4 and Cluster 1 contains mainly samples from levels K2 and K3. This suggest that Cluster 0 has a higher functional ability and Cluster 1 a medium to



(a) Cluster representation

(b) K-level distribution across clusters

Figure 5.13: K-means: cluster analysis (k=3) using combined Dataset 2.

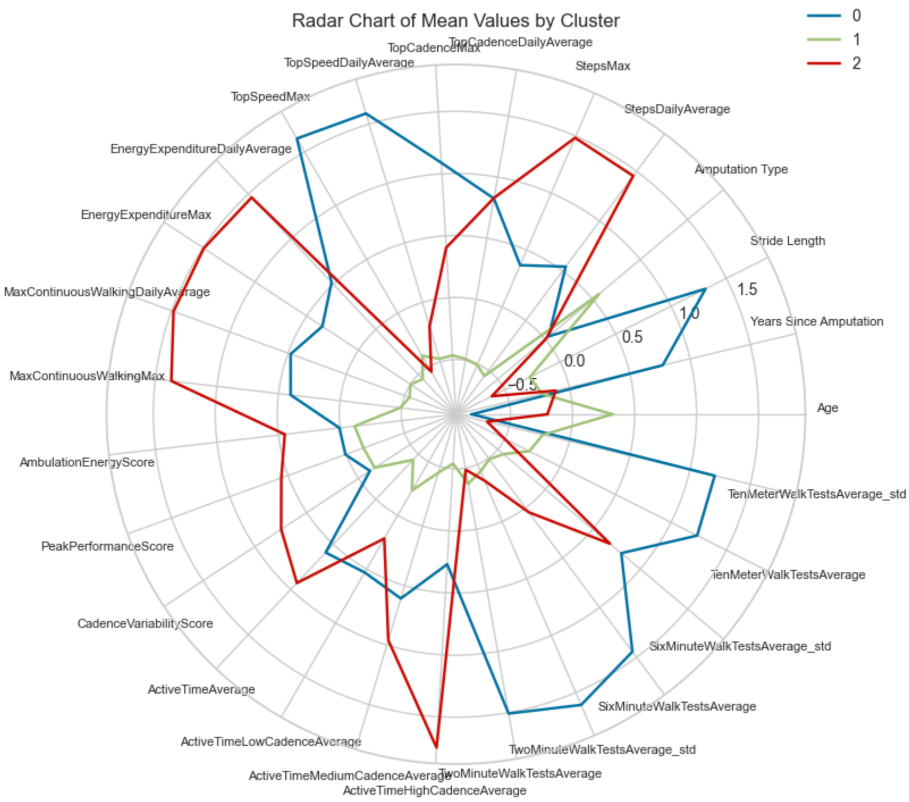
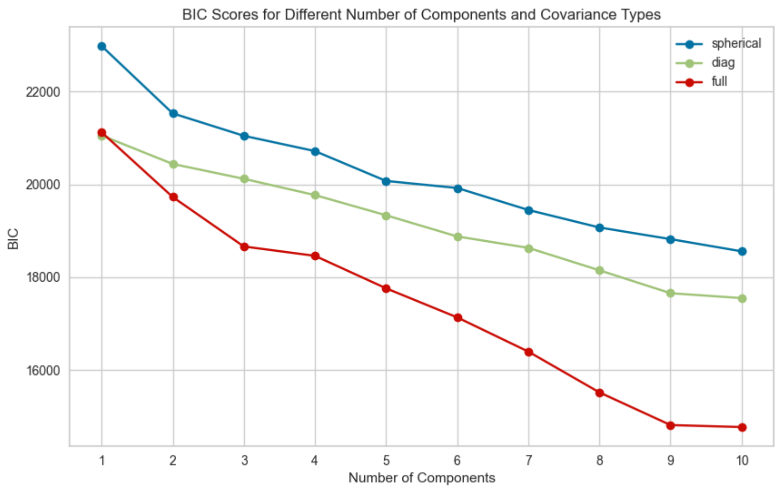
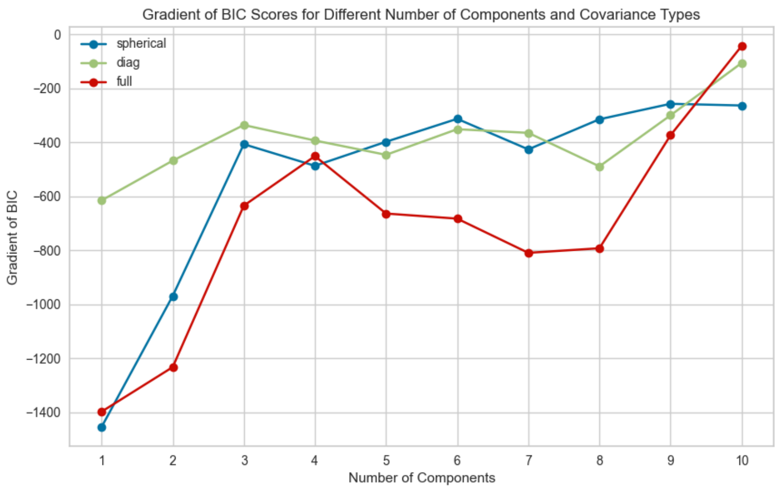


Figure 5.14: K-means: radar chart of mean values for each appointment data and activity metric from combined Dataset 2, by cluster.

low functional ability. Cluster 2 shows a balanced distribution across K3 and K4 levels with few K2 samples, including medium to high functional mobility levels.



(a) BIC values

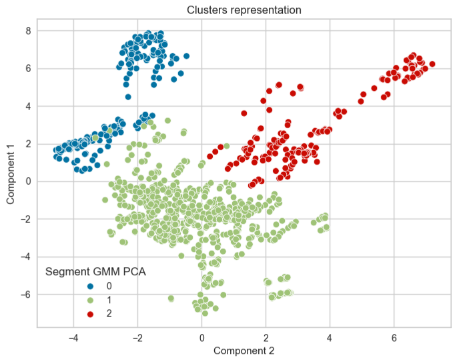


(b) Gradient of BIC

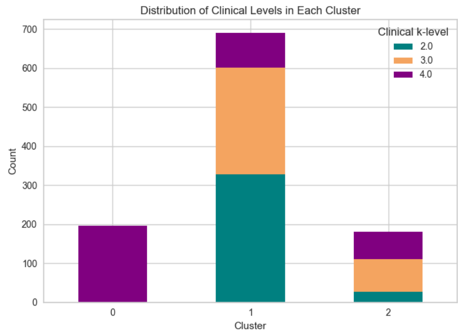
Figure 5.15: BIC applied to GMM using appointment data and activity metrics Dataset 2.

The appointment data and activity metrics mean values (or the mode for categorical data) for each of the three clusters are displayed in Figure 5.17.

As seen in K-means, the average age for Cluster 0 is negative and for Clusters 1 and 2 is positive, suggesting that individuals from Cluster 0 are typically younger than those in the other two clusters. Cluster 0 has a positive average time since amputation, which means that individuals that belong to this cluster have, on average, more time since amputation than people in the other two clusters. Cluster 0 also has the highest stride length average when compared with Cluster 1 and Cluster 2. As explained before, these observations for Cluster 0 are associated with more functional ability and higher activity levels, which corresponds to the distribution of clinical K-level seen in Figure 5.16b.



(a) Cluster representation



(b) K-level distribution across clusters

Figure 5.16: GMM: cluster analysis (k=3) using appointment data and activity metrics Dataset 2.

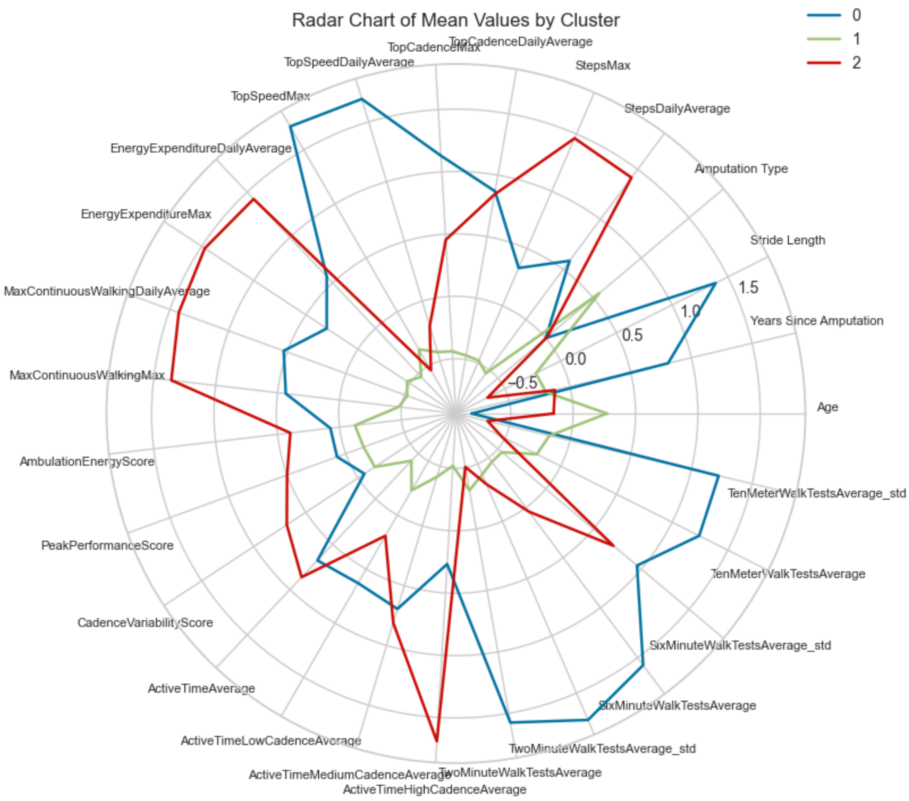


Figure 5.17: GMM: radar chart of mean values for each appointment data and activity metric from combined Dataset 2, by cluster.

5.2.3 Comparison: Results for Dataset 1

The results obtained for unsupervised learning using Dataset 1 are presented in Appendix A.

It is possible to see that the results are not very different from the ones obtained for Dataset 2. However, there seems to be a better separation between activity levels when using the balanced dataset (Dataset 2), with a cluster having higher functional ability, another low functional ability and the last with a balanced activity level.

5.3 Summary

The application of unsupervised algorithms revealed interesting results in grouping amputees with similar activity patterns.

The best number of clusters was k=3, in all models and approaches. In general, clusters were well-separated with minimal overlap and the distribution analysis showed that one cluster had higher functional ability, another low functional ability and the last had a balanced activity level.

The results for cluster distribution and constitution using both appointment data and activity metrics are similar to those obtained in using only activity metrics. However, they provide interesting information that Approach 1 do not. Results from Approach 2 suggest that individuals from Cluster 0 (higher functional ability) are typically younger than those in the other two clusters and have longer stride lengths compared to individuals in Cluster 1 and Cluster 2. Additionally, individuals from Cluster 0 appear to have, on average, more time since amputation than people in the other two clusters.

The results for K-means and GMM were similar, clustering data in a similar way, particularly when adding patient information to the activity metrics. One tiebreaker factor could be computational efficiency, since K-means is computationally faster than GMM [70]. Speed of clustering is a relevant factor for clinical application of models, so K-means model is more advantageous in this topic.

For none of the models and approaches the clusters corresponded perfectly to the clinical K-levels. However, in all cases, the data seemed to be clustered by activity metrics' values that are associated with activity levels. These results suggests that the assessment carried out in a clinical environment to assign K-level may, in fact, not be completely faithful to the individual's day-to-day ambulatory ability.

This type of analysis may be useful in the following applications:

- When the user is in the low activity cluster, it sends alerts to the patient to motivate and advise them.
- Helping the clinician in understanding the patient's profile and frame activity metrics to support their decisions regarding rehabilitation and prosthetic components.

Chapter 6

Supervised Learning

In parallel to the unsupervised analysis of activity profiles, explored in Chapter 5, it is also interesting to investigate the use of supervised learning to classify prosthetic users K-levels. Thus, using data augmentation, two datasets are obtained (activity metrics dataset and combined dataset) and various methods of feature reduction and selection are explored. The models used for classification are SVM and RF, performing binary and multiclass tasks. In this chapter, we explain the methodology used and discuss the results.

6.1 Methods

The methodology used in this project to classify activity levels of amputees consists in several steps and approaches (Figure 6.1) explained in this Section.

6.1.1 Dataset

As referred to in Chapter 3, in supervised learning it is used data with labels. In this project, the labels are the clinical K-levels. To train models and evaluate their performance on new data, it is necessary to split the dataset into training and test sets. Usually, the test set is made of approximately 20% of the total dataset. In this project, we select 6 patients from the original data to be used as a test set, corresponding to 7 raw samples and 27 after applying the sliding window (Table 6.1). This selection has in consideration two aspects: ensuring that all classes have representation and that samples from the same patient are in the same set.

Table 6.1: Number of samples in the test set before and after applying the sliding window.

	Raw data	**Sliding window**
K2	2	8
K3	3	11
K4	2	8
Total	7	27

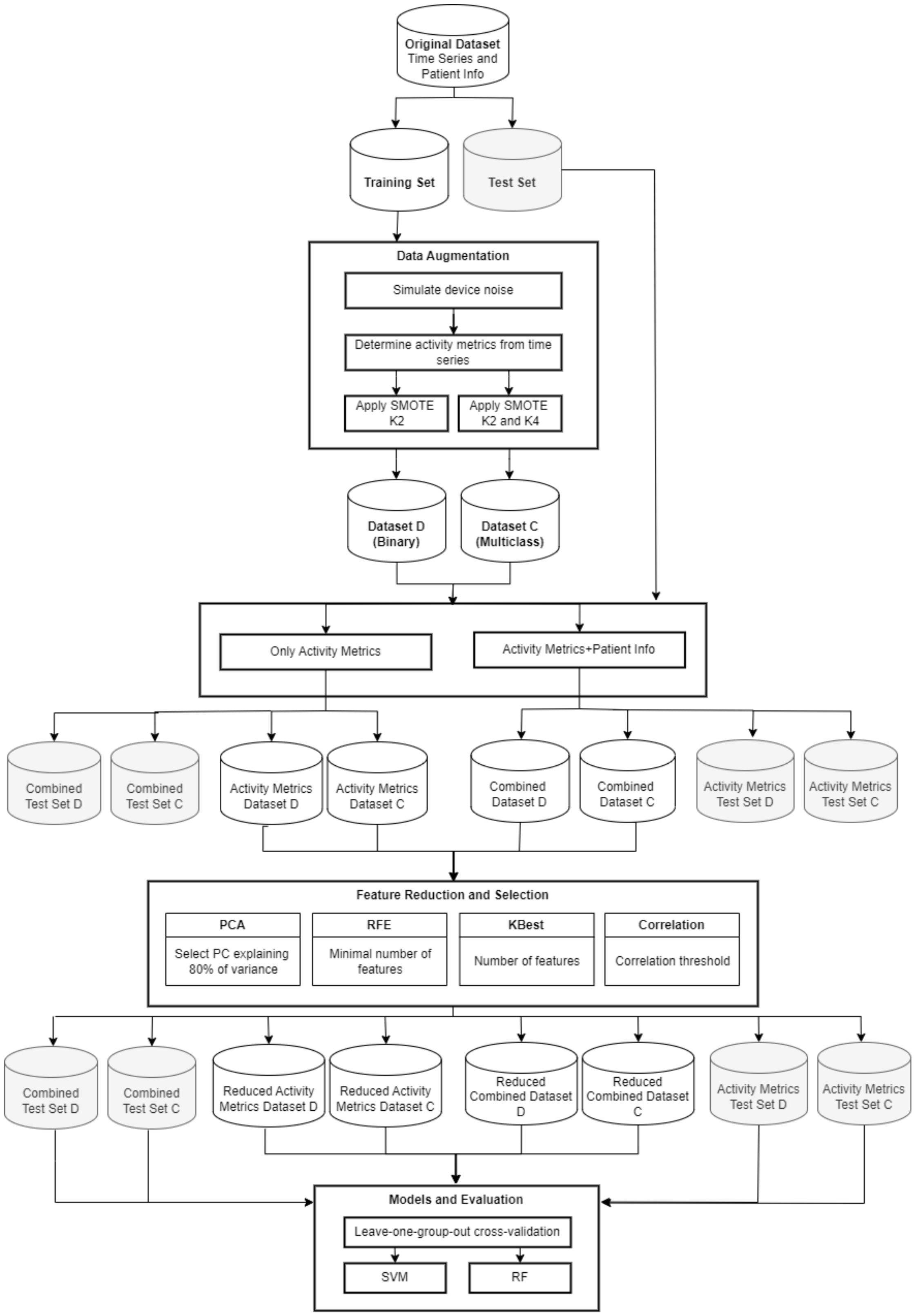


Figure 6.1: Methodology used for classifying activity levels.

Thus, in the training set, we have 29 patients and 33 samples, corresponding to 289 samples after the sliding window. Then, we apply data augmentation to the training set. With the noise addition (explained in Section 4.6) we obtain 578 samples.

Since in this dataset the clinical K-levels include K2, K3 and K4, it would be interesting to explore a multiclass classification for the prediction of these three clinical K-levels. To obtain a balanced training set for the multiclass approach, SMOTE is applied to the two minority classes: K2 and K4 (Table 6.2).

Table 6.2: Multiclass classification: number of samples in Training Set 2 before and after applying SMOTE.

	Noise addition	**SMOTE**
K2	158	336
K3	336	333
K4	84	334
Total	578	1003

As it was possible to see in Section 2.1, the main differences in reimbursement for a prosthetic device are between K2 and higher K-levels (K3 and K4). Therefore, considering the commercial point of view, it makes sense to also explore a binary classification, grouping K3 and K4 into a single class. In this case, we apply SMOTE only to K2 (Table 6.3).

To sum up, we have the following datasets:

- Dataset A: dataset obtained after sliding window, with three classes (includes Training Set and Test Set).
- Dataset B: dataset obtained after sliding window, with two classes (includes Training Set and Test Set).
- Dataset C: dataset obtained after applying data augmentation to the training set, with three classes (includes Training Set and Test Set).
- Dataset D: dataset obtained after applying data augmentation to the training set, with two classes (includes Training Set and Test Set).

Table 6.3: Binary classification: number of samples in Training Set 3 before and after applying SMOTE.

	Noise addition	**SMOTE**
K2	158	418
K3 + K4	420	420
Total	578	838

Similarly to unsupervised learning, two data approaches are taken into consideration for classification: using only activity metrics, representing 23 features, or using activity metrics combined with appointment data (including age, stride length, amputation type, and time since amputation), with a total of 27 features. Once again, the missing values in both situations must be filled in, as explained in Section 4.5. Numerical values are normalized using standard normalization to transform data so that it has a mean of zero and a standard deviation of one.

6.1.2 Feature Reduction and Selection

To reduce the number of features, we tested different methods of feature reduction and selection. Each method has parameters that need to be set, presented in Table 6.4.

Table 6.4: Feature reduction and selection methods and their parameters.

Technique	Parameters
PCA	Variance
RFE	Minimal number of features to select
KBest	Number of features to select
Correlation	Correlation threshold

For PCA, the process is the same as explained in Chapter 5. A variance of 80% should be preserved, so PCA keeps a number of principal components that satisfies that condition. Those components are used in the machine learning models.

For the application of Recursive Feature Elimination (RFE), we use the RFECV function available in the SKlearn library [1]. This method does not need a predefined number of features to select; it only needs a minimum number of features that we defined as 2. It is a dynamic process that iteratively removes features and evaluates model performance, selecting the best subset of features. This process includes cross-validation (CV) to improve the robustness. Since we have more than one sample for each patient, we have to make sure that samples from the same patient belong to the same set. To do so, we use Leave One Group Out (LOGO) CV (explained below). The optimal number of features is obtained based on the highest CV score. Since this process requires the use of a model, the model parameters are tuned before and after applying feature selection.

We also tested the application of Select KBest with F-test. It computes the F value between the features and the target (the higher the value, the stronger the relationship) and selects the K features with higher scores. To know the best number of features to select, we test the performance of the machine learning models with 10, 14, 16 and 20 features and select the number of features that correspond to the best model performance.

As we saw in Section 4.3, the activity metrics have repeated information. So, another method that is interesting to use is the Correlation Feature Selection, which removes this redundant information. This is done by removing features that are highly correlated with each other. To apply this process, it is necessary to define a correlation threshold above which features are considered redundant. We tested correlation thresholds of 0.7, 0.8 and 0.95 and selected the correlation threshold that was associated with the best model performance.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFECV.html. Accessed on 2024-05-12.

To explore the influence of these methods, we need a reference. So, we also use the models without any feature reduction or selection method (baseline).

6.1.3 Models

For the classification task, the machine learning models used are SVM and RF (explained in Chapter 3). Before training these models, it is important to tune their hyperparameters, which are settings that control the learning process of the model. Hyperparameter tuning consists in selecting the best value or process for each model's hyperparameter, leading to the best performance.

For SVM, we should take into consideration the following hyperparameters:

- C: regularization parameter which controls the trade-off between the margin and the number of training errors. A high C aims at classifying all training examples correctly, heavily penalizing the error, being better to avoid overfitting. A smaller C allows more training errors.
- Gamma: is a Kernel coefficient and it is only valid for kernel 'rbf', 'poly' and 'sigmoid'. It controls how much impact a training sample has on the decision boundary. A higher value of gamma implies that nearby support vectors have more influence.
- Kernel: pattern analysis algorithms used to capture the similarity between data points.

For RF, we should adjust the following hyperparameters:

- n_estimator: number of trees to be trained.
- max_depth: longest path between the root node and the leaf node.
- min_samples_split: minimum number of observations in any given node required to split it.
- min_samples_leaf: minimum number of samples that should be present in a leaf node after splitting a node.
- max_features: number of maximum features provided to each tree.

Grid Search is a good option to optimize hyperparameters. It consists in creating a grid with potential values for each hyperparameter and fitting the model using all possible combinations. The best combination of values is the one that is associated with the best model performance. We use GridSearchCV, which applies Grid Search combined with CV.

The most commonly used method of CV is the K-fold CV, that consists in splitting the dataset into k parts of equal size (folds). One of these folds is used for validation and the rest of the folds are used as a training set. This is repeated k times, so each fold is used as a validation set once.

To choose the type of CV to use, it is important to highlight that in the dataset we have more than one sample for each patient. Thus, when splitting the training set into training and validation set during CV, samples from the same patient should belong to the same set. This is done by using

LOGO CV. It is a variant of K-fold CV, but instead of splitting data into K-folds, it splits based on groups. So, it consists of leaving out one entire group as the validation set and training the model on the remaining data. This is repeated until all groups are used as a validation set once. For our dataset, the groups are divided based on alphanumeric identifiers that differentiate patients.

The GridSearchCV using LOGO CV is applied to SVM and RF using the potential values in Table 6.5 and 6.6, respectively.

Table 6.5: Potential values for SVM hyperparameters.

Hyperparameter	Potential values
C	0.1, 1.0, 10.0, 100.0
Gamma	'scale, 'auto'
Kernel	'rbf, 'linear', 'poly', 'sigmoid'

Table 6.6: Potential values for RF hyperparameters.

Hyperparameter	Potential values
n_estimator	100, 200, 300
max_depth	None, 10, 20
min_samples_split	2, 5, 10
min_samples_leaf	1, 2, 4
max_features	None, 'sqrt'

Having the best hyperparameters, the models are trained and the clinical K-levels are predicted.

6.1.4 Performance Evaluation

Before exploring the metrics available for performance analysis, it is important to know the following 4 concepts:

- True positive (TP): when the prediction is correct and it is a positive data point.
- True negative (TN): when the prediction is correct and it is a negative data point.
- False positive (FP): when the prediction is incorrect and it is a positive data point.
- False negative (FP): when the prediction is incorrect and it is a negative data point.

Precision focuses on the correctness of positive predictions and it is obtained using the equation:

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

On the other hand, recall emphasizes capturing all relevant instances, being calculated as:

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

The F1-score is a metric that has in consideration both precision and recall. It is obtained as:

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \cdot \text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

In this project, one of the metrics used to evaluate the performance of supervised models was the macro F1-score. It consists in taking the arithmetic mean of F1-scores of all classes. This method gives the same weight to all classes, being calculated as:

$$\text{Macro F1-score} = \frac{1}{|C|} \sum_{c \in C} \text{F1-score}_c$$

Another metric used is accuracy. It evaluates the performance of a classification model and consists in the number of correctly predicted samples of all data points:

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

When we have imbalance data, it is important to use balanced accuracy. It can be used for binary and multiclass classification and is obtained as the arithmetic mean of sensitivity and specificity:

$$\text{Balanced Accuracy} = \frac{1}{2}(\text{sensitivity} + \text{specificity}) = \frac{1}{2}(\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}})$$

A confusion matrix allows the analysis of classifier performance per class, by visualization of TP, TN, FP and FN (Figure 6.2). It is a *NxN* matrix, where N is the number of classes.

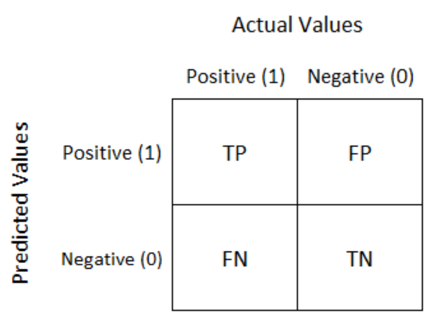


Figure 6.2: Basic Structure of a confusion matrix. Extracted from https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

6.2 Results and Discussion

6.2.1 Multiclass Classification

6.2.1.1 Approach 1: Activity Metrics Dataset C

In Table 6.7 we have results of multiclass classification using SVM with activity metrics from Dataset C, for the baseline and for each feature reduction and selection method.

Table 6.7: Performance metrics of multiclass classification using SVM with activity metrics Dataset C.

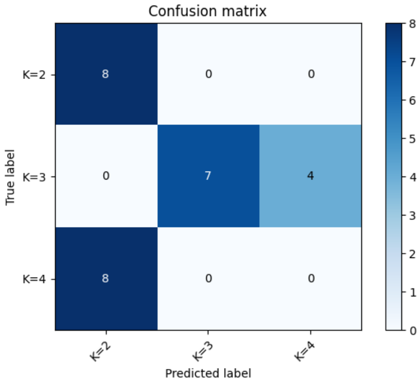
Feature Reduction and Selection	F1-Score Validation	F1-Score Test	Accuracy Test	Balanced Accuracy
Baseline	0.53	0.48	0.56	0.55
PCA	0.53	0.19	0.41	0.33
KBest	0.52	0.45	0.56	0.52
RFE	0.66	0.06	0.07	0.06
Correlation	0.52	0.48	0.56	0.54

The performance of SVM differs significantly for each feature selection methods, with similar values when no feature selection method is used and when the Correlation method is applied. PCA significantly reduced the model's performance, which suggests that in this case the principal components obtained by PCA do not capture the information needed for a correct prediction. Overall, the values of the performance metrics are quite low, around 0.50 or lower. This indicates that the model is having difficulties distinguishing between the different K-levels and feature reduction and selection methods are not improving it.

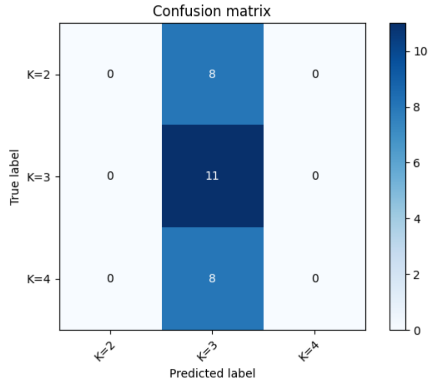
In Figure 6.3 we have the confusion matrices for SVM. For baseline and Correlation, the first class (K2) is always correctly predicted, but the prediction of third class (K4) is always incorrect (classified as K2), revealing that these models have difficulties related with this class. For PCA, all classes are predicted as K3, which is consistent with the bad performance of this model.

In Table 6.8, we have the values for the performance metrics when using RF. The baseline performance is very low, which indicates that RF model struggles with Dataset C (activity metrics). However, the performance improves with PCA and KBest, when compared to baseline. The best performance is obtained with PCA (test F1-score of 0.58, accuracy of 0.70 and balanced accuracy of 0.67), which reveals that, in this case, the principal components capture more important information for RF. When applying RFE and Correlation, the performance is also very low.

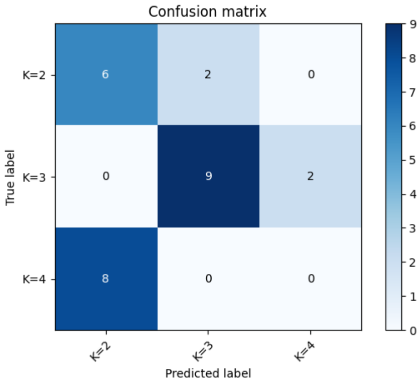
Figure 6.4 shows the confusion matrices of RF for multiclass classification with activity metrics Dataset C. As seen when analyzing the performance metrics, the best performance is obtained when using PCA and the same is observed for the confusion matrices. K2 and K3 are always predicted correctly and K4 is misclassified as K3, which is not as serious as being classified as K2 since the differences in reimbursement for a prosthetic device are mainly between K2 and higher K-levels. Baseline, RFE, and Correlation methods have a bad performance, misclassifying a significant number of samples.



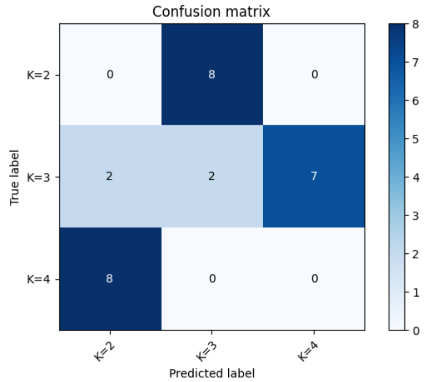
(a) Confusion matrix for the baseline model



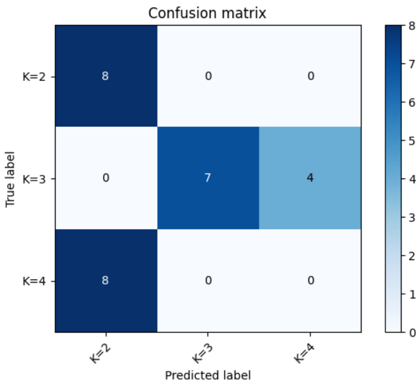
(b) Confusion matrix for PCA



(c) Confusion matrix for KBest



(d) Confusion matrix for RFE



(e) Confusion matrix for Correlation

Figure 6.3: Confusion matrices for multiclass classification using SVM with activity metrics Dataset C.

To sum up, when using activity metrics Dataset C for multiclass classification, the best performance is obtained for RF combined with PCA, showing some difficulties only in distinguish between K4 and K3.

Table 6.8: Performance metrics of multiclass classification using RF with activity metrics Dataset C.

Feature Reduction and Selection	F1-Score Validation	F1-Score Test	Accuracy Test	Balanced Accuracy
Baseline	0.60	0.17	0.26	0.21
PCA	**0.44**	**0.58**	**0.70**	**0.67**
KBest	0.60	0.51	0.56	0.54
RFE	0.60	0.14	0.26	0.21
Correlation	0.62	0.14	0.26	0.21

6.2.1.2 Approach 2: Combined Dataset C

In Table 6.9 we have the SVM performance across different feature reduction and selection methods, when using appointment data combined with activity metrics for a multiclass classification.

Table 6.9: Performance metrics of multiclass classification using SVM with combined Dataset C.

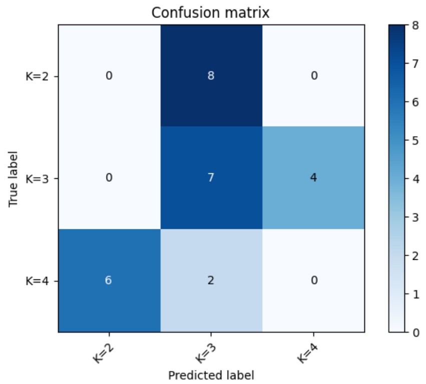
Feature Reduction and Selection	F1-Score Validation	F1-Score Test	Accuracy Test	Balanced Accuracy
Baseline	0.56	0.19	0.41	0.33
PCA	0.57	0.19	0.40	0.33
KBest	0.63	0.19	0.40	0.33
RFE	0.70	0.07	0.07	0.08
Correlation	0.59	0.19	0.40	0.33

The baseline performance was quite low, with a test F1-score of 0.19, an accuracy of 0.41, and a balanced accuracy of 0.33. Applying PCA, KBest, and Correlation did not significantly change the performance, but the RFE method drastically reduced the model's performance, with a test F1-score dropping to 0.07 and a balanced accuracy of 0.08, indicating that RFE was ineffective in this context.

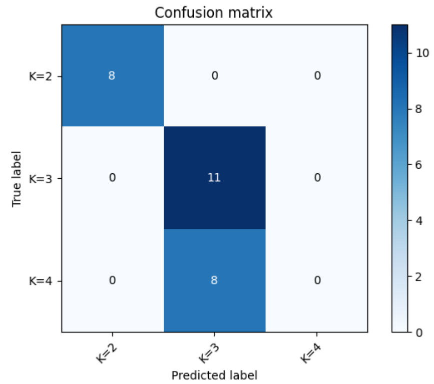
In Figure 6.5, we have the confusion matrices for SVM when using the combined Dataset C for multiclass classification. The baseline, PCA, KBest, and Correlation methods all have the same confusion matrix, where it is possible to see that all samples are predicted as K3. On the other hand, the RFE method shows a different matrix, but it reveals even more misclassification issues.

Table 6.10 presents the RF performance for different feature reduction and selection methods, when using combined Dataset C for a multiclass classification.

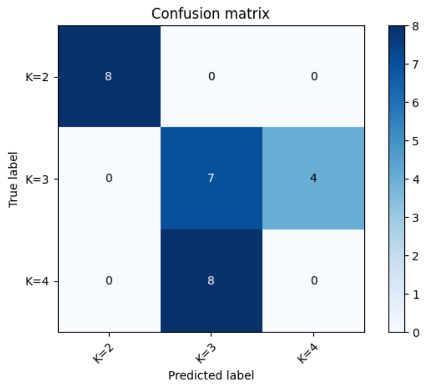
From Table 6.10 it is possible to see that the RF behaviors are slightly different than SVM. Once again, the baseline performance was low, even lower than for SVM, with an F1-score of 0.14, an accuracy of 0.26, and a balanced accuracy of 0.21. KBest, RFE, and Correlation methods maintained similar performance to the baseline. PCA improved the model's performance a little, achieving a test F1-score of 0.19 and an accuracy of 0.41, with a balanced accuracy of 0.33. However, these values are still very low and do not reflect a good performance.



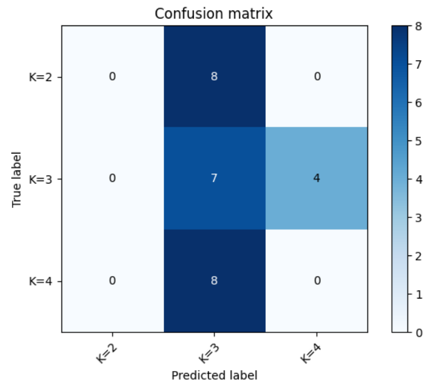
(a) Confusion matrix for the baseline model



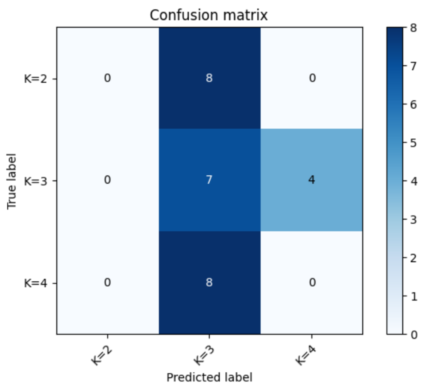
(b) Confusion matrix for PCA



(c) Confusion matrix for KBest



(d) Confusion matrix for RFE

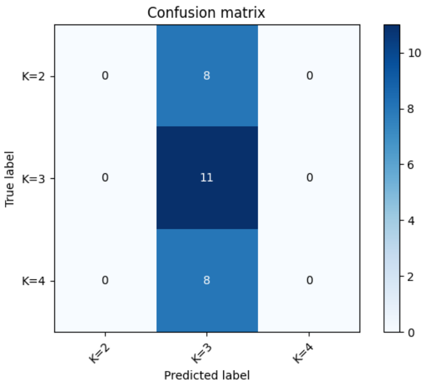


(e) Confusion matrix for Correlation

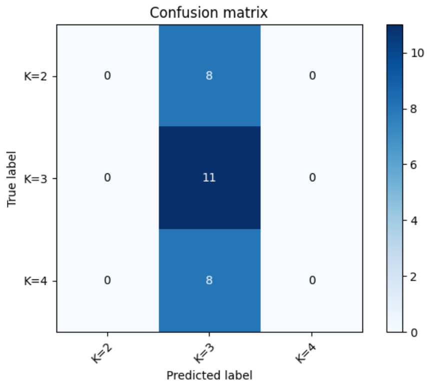
Figure 6.4: Confusion matrices for multiclass classification using RF with activity metrics Dataset C.

The confusion matrices presented in Figure 6.6 for the RF model highlight consistent misclassification, which was expected due to the metrics' values obtained. All methods classify most test samples as K3, which shows a poor performance and difficulties distinguishing the different classes.

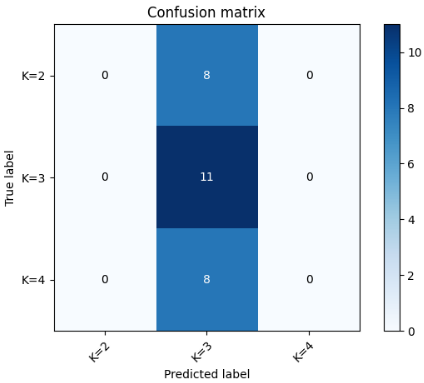
Regardless of the feature selection method, SVM and RF performed poorly, predicting nearly all samples as K3.



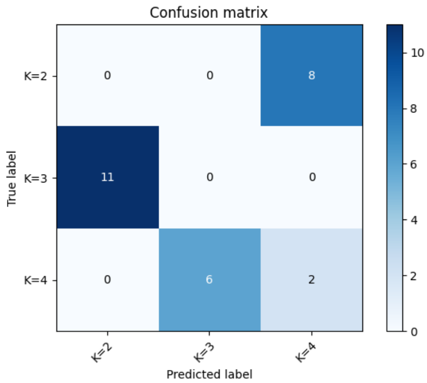
(a) Confusion matrix for the baseline model



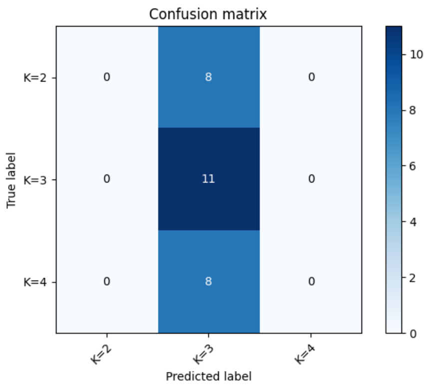
(b) Confusion matrix for PCA



(c) Confusion matrix for KBest



(d) Confusion matrix for RFE



(e) Confusion matrix for Correlation

Figure 6.5: Confusion matrices for multiclass classification using SVM with combined Dataset C.

6.2.2 Binary Classification

6.2.2.1 Approach 1: Activity Metrics Dataset D

Table 6.11 shows the results of binary classification using SVM with activity metrics Dataset D, for the baseline and for each feature reduction and selection method.

SVM presents consistently good performance for the baseline, KBest, and Correlation methods (test F1-score of 0.70, an accuracy of 0.70, and a balanced accuracy of 0.79). However, the

Table 6.10: Performance metrics of multiclass classification using RF with combined Dataset C.

Feature Reduction and Selection	F1-Score Validation	F1-Score Test	Accuracy Test	Balanced Accuracy
Baseline	0.57	0.14	0.26	0.21
PCA	**0.48**	**0.19**	**0.41**	**0.33**
KBest	0.57	0.14	0.26	0.21
RFE	0.57	0.14	0.26	0.21
Correlation	0.56	0.14	0.26	0.21

values for PCA and RFE significantly drop, showing that they reduce the model's performance. This suggests that PCA and RFE do not retain enough essential features for SVM for the binary classification task.

Table 6.11: Performance metrics of binary classification using SVM with activity metrics Dataset D.

Feature Reduction and Selection	F1-Score Validation	F1-Score Test	Accuracy Test	Balanced Accuracy
Baseline	0.75	0.70	0.70	0.79
PCA	0.61	0.39	0.63	0.45
KBest	0.76	0.70	0.70	0.79
RFE	0.65	0.29	0.41	0.29
Correlation	0.76	0.70	0.70	0.79

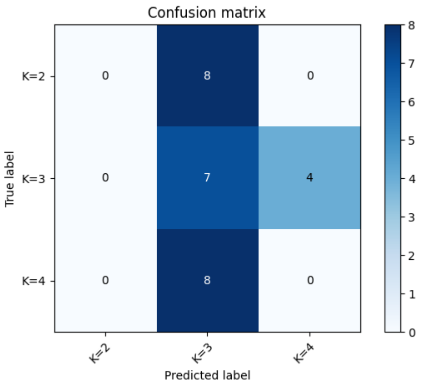
Figure 6.7 shows the confusion matrices of SVM for binary classification. For baseline, KBest, and Correlation, the model correctly classifies the positive class but misclassifies a portion of the negative class. For PCA and RFE, apart from some samples of K3+K4 being misclassified, the models do not correctly predict any sample from K2.

As seen in Table 6.12, the performance of RF is similar for baseline, KBest, RFE, and Correlation. As observed in the multiclass classification, the model's performance significantly improves when combined with PCA, with the test F1-score of 0.92, accuracy of 0.93 and balanced accuracy of 0.95. This indicates that, once again, PCA effectively captures the most relevant information for RF, enhancing its classification ability.

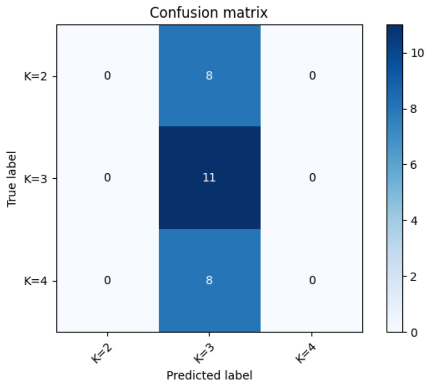
Table 6.12: Performance metrics of binary classification using RF with activity metrics Dataset D.

Feature Reduction and Selection	F1-Score Validation	F1-Score Test	Accuracy Test	Balanced Accuracy
Baseline	0.76	0.7	0.70	0.79
PCA	**0.64**	**0.92**	**0.93**	**0.95**
KBest	0.76	0.70	0.70	0.79
RFE	0.80	0.66	0.67	0.76
Correlation	0.79	0.70	0.70	0.79

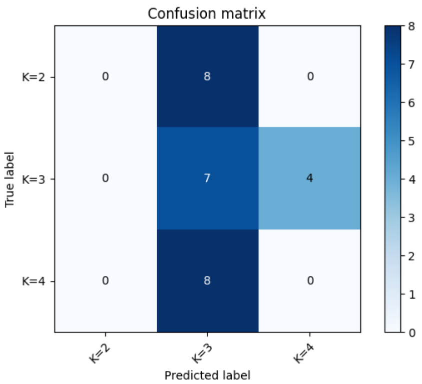
In Figure 6.8, we have the confusion matrices for RF. The confusion matrices for baseline, KBest, and Correlation show a correct classification of K2 but some samples of K3+K4 are misclassified. The RFE method has a similar confusion matrix. In contrast, the PCA method's



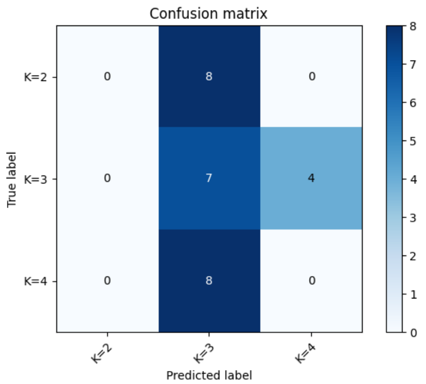
(a) Confusion matrix for the baseline model



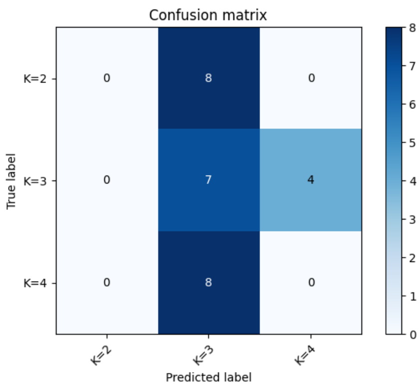
(b) Confusion matrix for PCA



(c) Confusion matrix for KBest



(d) Confusion matrix for RFE



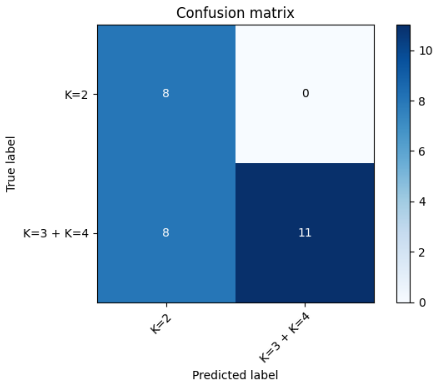
(e) Confusion matrix for Correlation

Figure 6.6: Confusion matrices for multiclass classification using RF with combined Dataset C.

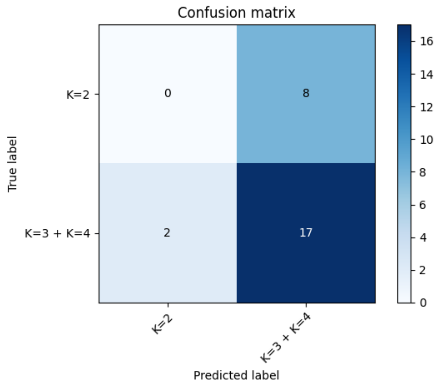
confusion matrix reveals a correct predicting for almost all samples, demonstrating a significant improvement and being the best performance among the methods.

For both models and all feature selection methods, the performances are better when using a binary classification than when doing a multiclass classification, as expected.

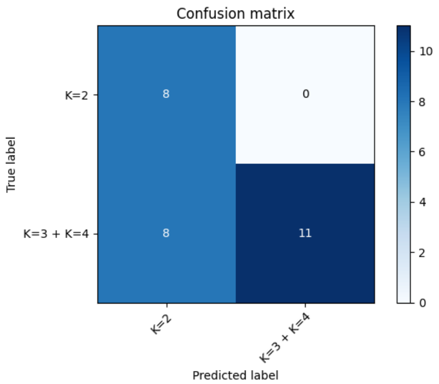
Similarly to the results obtained for multiclass classification, for this task the best performance is obtained for RF combined with PCA.



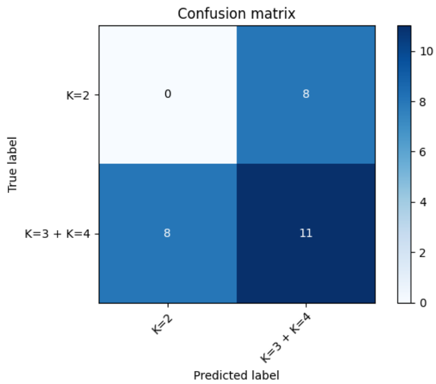
(a) Confusion matrix for the baseline model



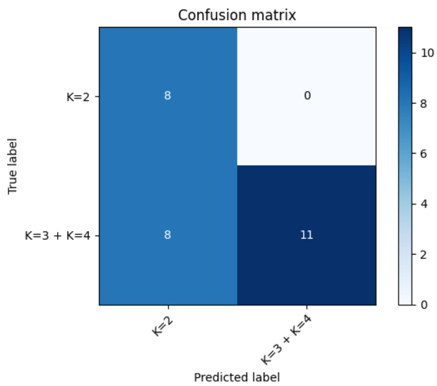
(b) Confusion matrix for PCA



(c) Confusion matrix for KBest



(d) Confusion matrix for RFE

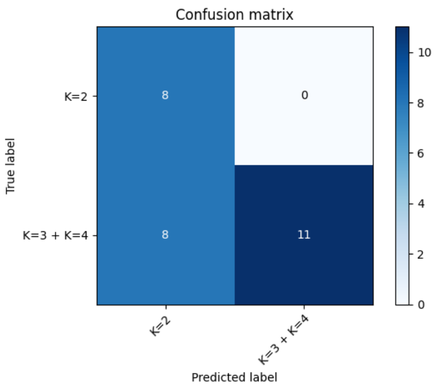


(e) Confusion matrix for Correlation

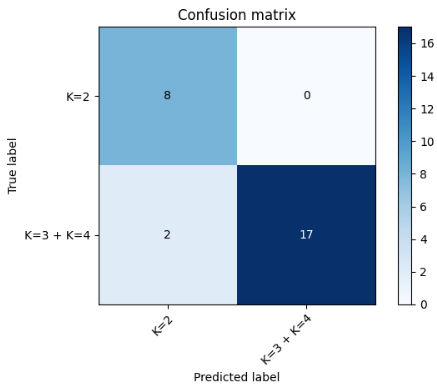
Figure 6.7: Confusion matrices for binary classification using SVM with activity metrics Dataset D.

6.2.2.2 Approach 2: Combined Dataset D

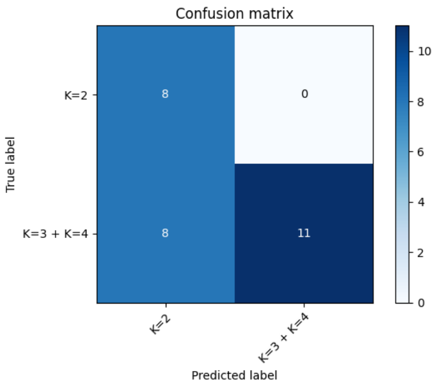
In Table 6.13, we have the results for SVM using combined Dataset D for a binary classification. The SVM model showed excellent performance across most feature reduction and selection methods. The baseline, KBest, RFE, and Correlation methods all achieved perfect performance with F1-scores, accuracy, and balanced accuracy all at 1.0. This shows that these algorithms were able



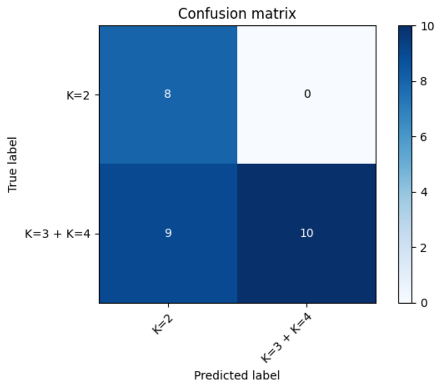
(a) Confusion matrix for the baseline model



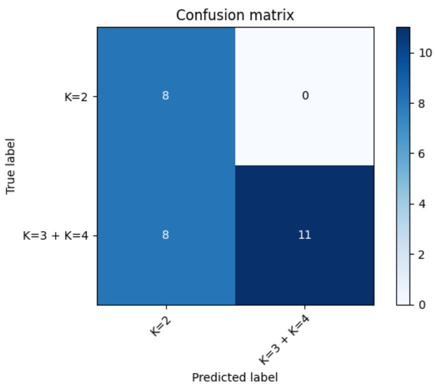
(b) Confusion matrix for PCA



(c) Confusion matrix for KBest



(d) Confusion matrix for RFE



(e) Confusion matrix for Correlation

Figure 6.8: Confusion matrices for binary classification using RF with activity metrics Dataset D.

to perfectly classify the samples in the test set. However, applying PCA slightly decreases the performance (test F1-score of 0.77, accuracy of 0.78, and balanced accuracy of 0.84). This suggests that PCA was less effective for this case compared to the other methods.

The confusion matrices of SVM using combined Dataset D for binary classification are shown in Figure 6.9. For baseline, KBest, RFE, and Correlation show correct predictions for all classes, with no misclassifications as expected from the metrics' results. For PCA, the confusion matrix indicates that there were a few misclassifications in the K3+K4 class.

Table 6.13: Performance metrics of binary classification using SVM with combined Dataset D.

Feature Reduction and Selection	F1-Score Validation	F1-Score Test	Accuracy Test	Balanced Accuracy
Baseline	**0.78**	**1.0**	**1.0**	**1.0**
PCA	0.74	0.77	0.78	0.84
KBest	**0.72**	**1.0**	**1.0**	**1.0**
RFE	**0.78**	**1.0**	**1.0**	**1.0**
Correlation	**0.80**	**1.0**	**1.0**	**1.0**

Table 6.14 presents the results for RF using combined Dataset D for a binary classification. The RF model exhibited similar performance across all feature reduction and selection methods. All algorithms achieved a test F1-score of 0.70, accuracy of 0.70, and balanced accuracy of 0.79. This consistency suggests that the RF model was robust across these methods and that removing some features does not lower the ability of the RF in performing this task.

Table 6.14: Performance metrics of binary classification using RF with combined Dataset D.

Feature Reduction and Selection	F1-Score Validation	F1-Score Test	Accuracy Test	Balanced Accuracy
Baseline	0.80	0.70	0.70	0.79
PCA	0.73	0.70	0.70	0.79
KBest	0.80	0.70	0.70	0.79
RFE	0.82	0.70	0.70	0.79
Correlation	0.80	0.70	0.70	0.79

In Figure 6.10 we have confusion matrices for RF using combined Dataset D for binary classification. All feature reduction and selection methods have equal matrices. They reveal some misclassification of K3+K4 samples that are incorrectly classified as K2, being a consistent issue.

To sum up, the best performance for the combined Dataset D was obtained using the SVM model across most feature reduction and selection methods tested (excluding PCA).

6.2.3 Comparison: Results for Datasets A and B

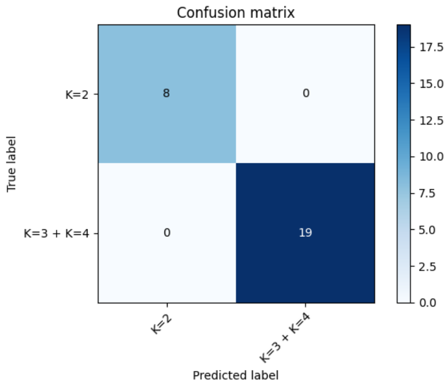
For supervised learning, the results obtained using Datasets A and B, without data augmentation, are presented in Appendix B.

In some cases, especially for multiclass task, the results are better when using datasets without data augmentation. However, for most cases, performance is better when using the datasets with data augmentation.

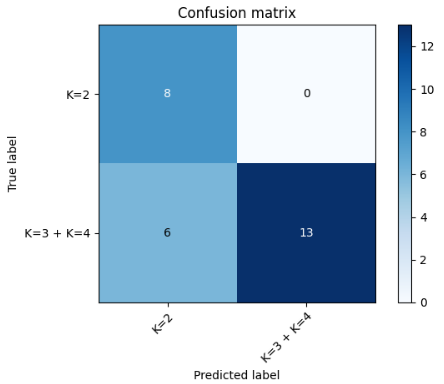
6.3 Selected Features

6.3.1 Approach 1: Activity Metrics Dataset

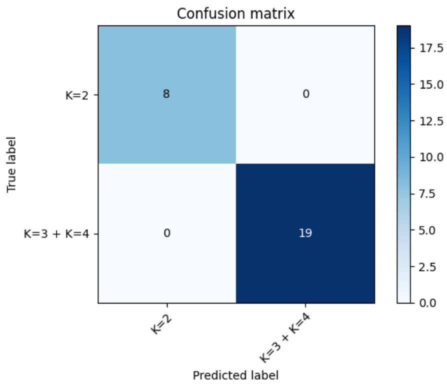
It is interesting to analyze which features are eliminated for each task and approach. Figure 6.11 shows the percentage of times each feature was eliminated by a feature selection method when



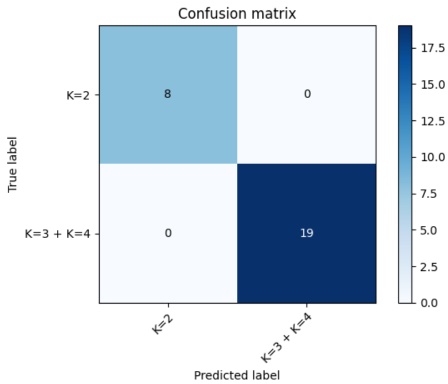
(a) Confusion matrix for the baseline model



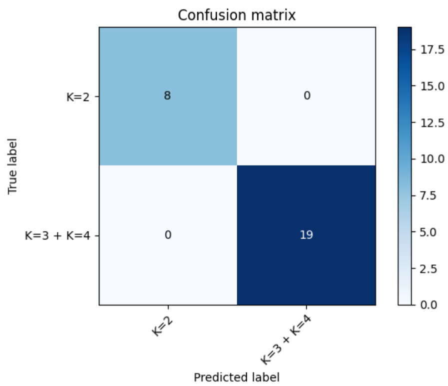
(b) Confusion matrix for PCA



(c) Confusion matrix for KBest



(d) Confusion matrix for RFE

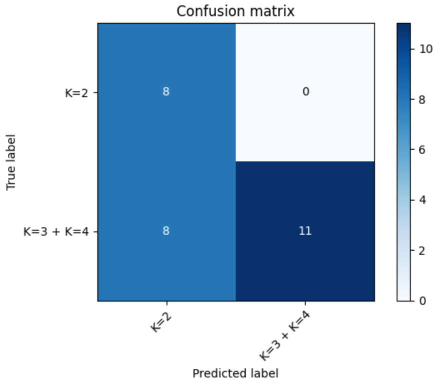


(e) Confusion matrix for Correlation

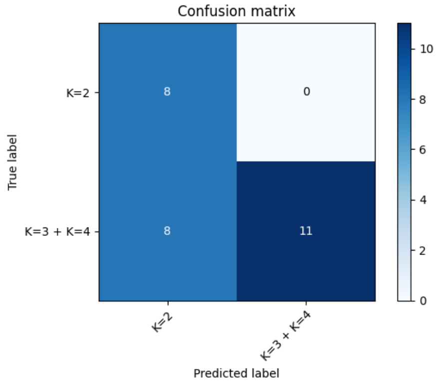
Figure 6.9: Confusion matrices for binary classification using SVM with combined Dataset D.

using the activity metrics Dataset C for multiclass classification, for both SVM and RF models.

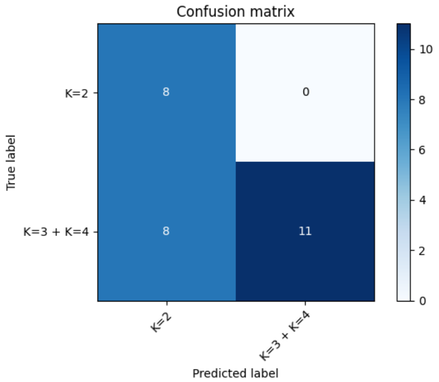
From these results, we can conclude that the most frequently eliminated feature is PeakPerformanceScore, being eliminated 75% of the time. Following, the most frequently eliminated features are EnergyExpenditureDailyAverage, EnergyExpenditureMax, AmbulationEnergyScore, CadenceVariabilityScore, ActiveTimeMediumCadenceAverage, which are eliminated 50% of the time.



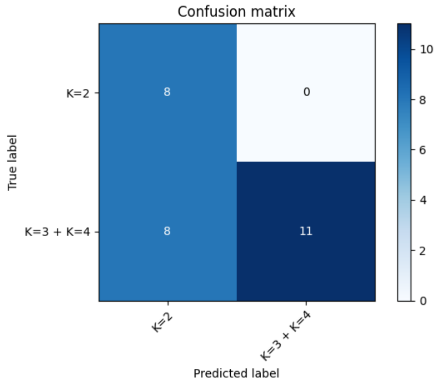
(a) Confusion matrix for the baseline model



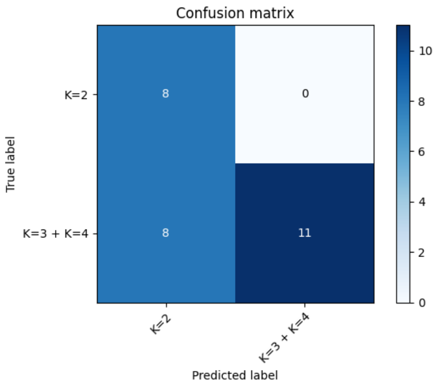
(b) Confusion matrix for PCA



(c) Confusion matrix for KBest



(d) Confusion matrix for RFE



(e) Confusion matrix for Correlation

Figure 6.10: Confusion matrices for binary classification using RF with combined Dataset D.

For the binary classification approach, when using the activity metrics Dataset D, the elimination percentage of each feature is presented in Figure 6.12.

Using activity metrics for binary classification, the most frequently eliminated feature was EnergyExpenditureDailyAverage (75%), followed by EnergyExpenditureMax, PeakPerformanceScore, ActiveTimeAverage, and ActiveTimeMediumCadenceAverage, which were eliminated by the feature selection methods 62.5% of the time.

It is important to highlight that EnergyExpenditureDailyAverage, EnergyExpenditureMax, ActiveTimeMediumCadenceAverage, and PeakPerformanceScore were eliminated in 50% or more of cases for both classification tasks. This suggests that these metrics are considered less relevant or redundant for the models.

In contrast, some features were never eliminated in both classification tasks, such as TopSpeedDailyAverage, TopSpeedMax, SixMinuteWalkTestsAverage, TwoMinuteWalkTestsAverage, TwoMinuteWalkTestsAverage_std, SixMinuteWalkTestsAverage, TenMeterWalkTestsAverage, and TenMeterWalkTestsAverage_std. This suggests that they have important information for the prediction of K-levels.

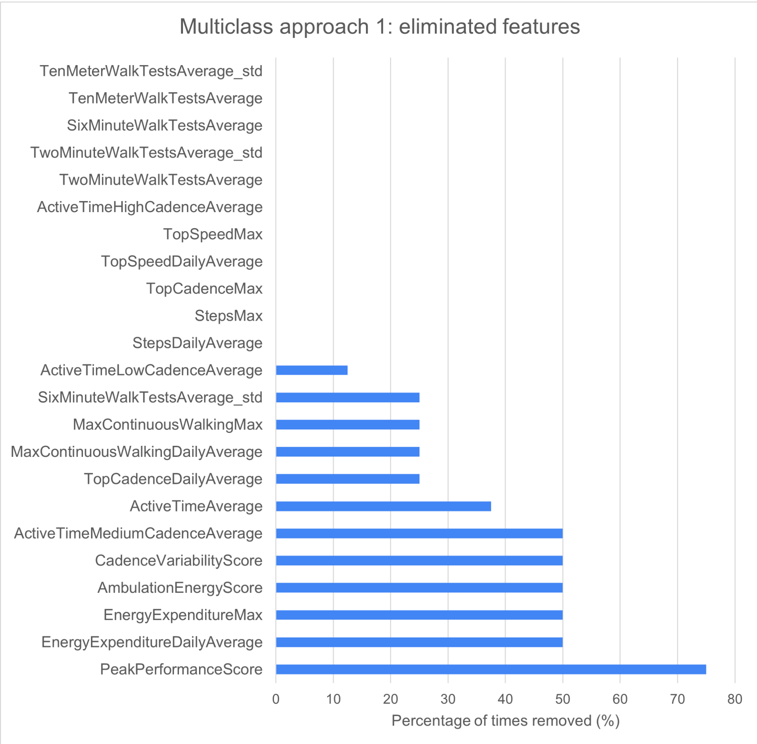


Figure 6.11: Percentage of times each feature was eliminated for multiclass classification using the activity metrics Dataset C.

6.3.2 Approach 2: Combined Dataset

For the second approach, the percentage that each feature was eliminated for multiclass classification is summarized in Figure 6.13.

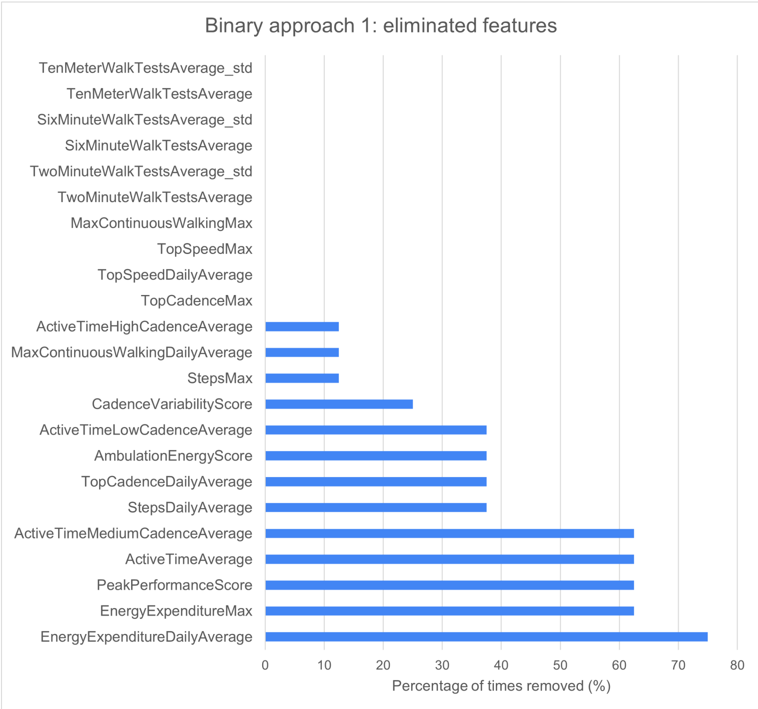


Figure 6.12: Percentage of times each feature was eliminated for binary classification using the activity metrics Dataset D.

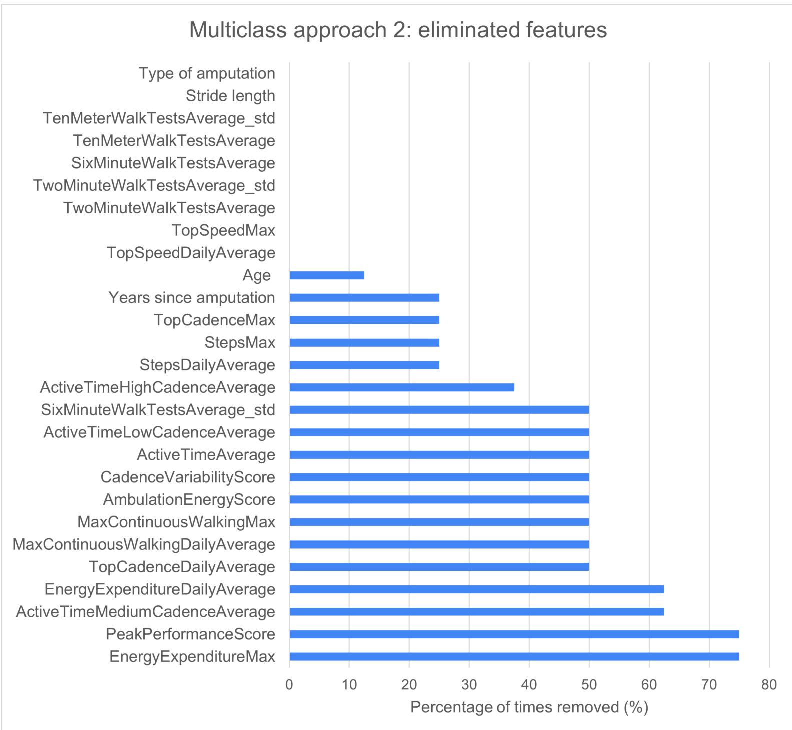


Figure 6.13: Percentage of times each feature was eliminated for multiclass classification using the combined Dataset C.

From these results, we can conclude that EnergyExpenditureMax and PeakPerformanceScore were the most frequently eliminated features (75%). Other frequently eliminated features include ActiveTimeMediumCadenceAverage and EnergyExpenditureDailyAverage (both with 62.5%) and TopCadenceDailyAverage, MaxContinuousWalkingDailyAverage, MaxContinuousWalkingMax, AmbulationEnergyScore, CadenceVariabilityScore, ActiveTimeLowCadenceAverage, ActiveTimeAverage, and SixMinuteWalkTestsAverage_std, all with 50% of elimination. From appointment data, the features eliminated were Years since amputation, eliminated twice, and Age, only eliminated once.

In Figure 6.14 is shown the percentage of times each feature was eliminated by a feature selection method for binary classification, for both SVM and RF models. The results obtained for the binary classification approach using the combined Dataset D show that EnergyExpenditureDailyAverage was the most frequently eliminated feature, with 87.5%. Other frequently eliminated features include EnergyExpenditureMax, PeakPerformanceScore, ActiveTimeAverage, and ActiveTimeMediumCadenceAverage (all with 75%). Additionally, TopCadenceDailyAverage, AmbulationEnergyScore, CadenceVariabilityScore, and ActiveTimeLowCadenceAverage were also eliminated half of the time (50%)

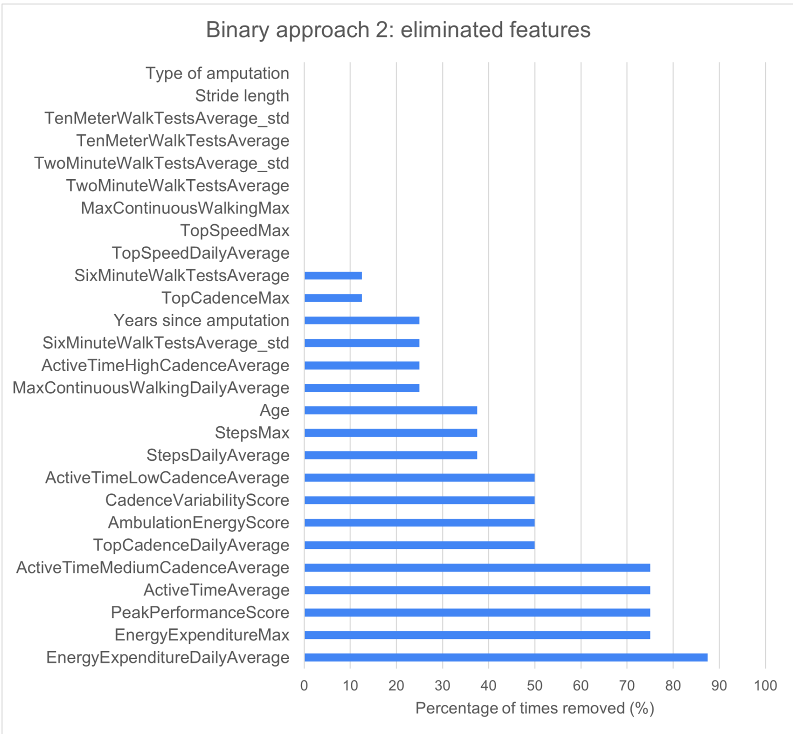


Figure 6.14: Percentage of times each feature was eliminated for binary classification using the combined Dataset D.

To sum up, some features were eliminated 50% or more of the cases for both classification tasks. These features include: EnergyExpenditureDailyAverage, EnergyExpenditureMax, PeakPerformanceScore, ActiveTimeMediumCadenceAverage, TopCadenceDailyAverage, AmbulationEnergyScore, CadenceVariabilityScore, ActiveTimeAverage, and ActiveTimeLowCadenceAverage. It is interesting to see that the features that were eliminated more often in approach 1 (EnergyExpenditureDailyAverage, EnergyExpenditureMax, ActiveTimeMediumCadenceAverage, and PeakPerformanceScore) are included in this group, reinforcing the idea that these features are less relevant or redundant.

Type of amputation, Stride length, TenMeterWalkTestsAverage_std, TenMeterWalkTestsAverage, TwoMinuteWalkTestsAverage, TwoMinuteWalkTestsAverage_std, TopSpeedDailyAverage, and TopSpeedMax are the features that were never eliminated by the feature selection methods, for both classification tasks. Apart from Type of amputation and Stride length, these activity metrics belong to the group of features that were not eliminated in approach 1, reinforcing their relevance for the prediction of K-levels. The fact that the features related to the patient (Type of amputation, Stride length, Age, and Years since amputation) were not usually eliminated indicates that they contain important information for the models in the prediction of activity levels.

6.4 Summary

This chapter evaluated the use of supervised learning to classify prosthetic users K-levels, with multiclass (K2, K3 and K4) and binary (k2 and K3+K4) tasks using SVM and RF. Different methods of feature reduction and selection were tested.

For multiclass classification using the activity metrics Dataset C, the best performance was obtained when using RF with PCA. It correctly classified all samples from K2 and K3, but misclassified K4 as K3. A K4 sample being categorized as K2 is more severe than K3 because the major differences in prosthetic device reimbursement occur between K2 and higher K-levels. However, this emphasizes the idea of approaching a binary classification, combining K3 and K4 in an unique class. For the combined Dataset C, both SVM and RF had bad performances, independently of the feature selection method, predicting almost all samples as K3.

For binary classification using the activity metrics Dataset D, the performances were consistently better than for the multiclass task and once again the best performance was obtained for RF using PCA, with a test F1-score of 0.92, accuracy of 0.93, and balance accuracy of 0.95. When using the combined Dataset D, the SVM model showed the best performance across most feature reduction and selection methods tested (excluding PCA). With F1-scores, accuracy, and balanced accuracy at 1.0, these algorithms were able to correctly classify all test samples.

It is also possible to conclude that PCA works better when combined with RF than with SVM and that binary classification generates better results than multiclass (that show poor performances).

For this supervised task, we are trying to introduce objectivity into an subjective evaluation, so there may be nuances in clinical evaluations that are not fully captured. Thus, the ideal task would involve having multiple evaluations from different clinics, though this was not feasible in this study.

In the future, an interesting approach could involve model stacking, using the class probabilities from base models as features for subsequent models. Since the binary classification problem was resolved quite effectively, using the results of this model for the multiclass problem might help address some of the ambiguity observed in the multiclass classification results.

Analyzing the percentage that each feature was eliminated for each classification task and each approach, it was possible to conclude that some features were often eliminated in both classifications tasks and approaches, such as EnergyExpenditureDailyAverage, EnergyExpenditureMax, ActiveTimeMediumCadenceAverage, and PeakPerformanceScore. This may suggest that these features are less relevant or redundant. Additionally, we also saw that some features were never eliminated by the feature selection methods, such as Type of amputation, Stride length, TenMeterWalkTestsAverage_std, TenMeterWalkTestsAverage, TwoMinuteWalkTestsAverage, TwoMinuteWalkTestsAverage_std, TopSpeedDailyAverage, and TopSpeedMax. This suggests that these features are relevant for the prediction of clinical K-levels.

Chapter 7

Conclusions and Future Work

To classify the functional ability of individuals with lower-limb amputations, the K-level system was created. People who are assessed by the clinical team as having a higher K-level are eligible for reimbursement of more advanced technology components. So, the clinical K-level is used to determine the type of prosthetic components each person is eligible for, which impacts their quality of life. A correct prescription of prosthetic components avoids issues such as discomfort, limited mobility and risk of injury, usually associated with falls. Current K-level assignment is based on patients' self-report, in-clinic performance tests and clinical experience. However, challenges associated with the K-level evaluation are associated with the subjectivity of clinical evaluations and the fact that self-reports and in-clinic performance tests do not accurately reflect daily-life activity levels.

This project aims to address these issues by using machine learning techniques to identify K-levels based on activity data collected through Adapttech's Motio StepWatch System during amputees daily-lives. It includes an analysis of the available dataset, an explanation of preprocessing and data augmentation processes used, an exploration of clustering methods to group prosthetic users with similar activity patterns and the use of supervised learning to classify prosthetic users K-levels.

Analyzing the dataset revealed that it does not have many samples and is not balanced. In the future, using a larger dataset with equal samples from K2, K3, and K4 categories would be beneficial. Additionally, it reveals that the quantity of samples related to sex is unbalanced, indicating that the majority of patients are men. Thus, it would also be interesting to obtain a similar amount of samples from male and female individuals, to make a more balanced study of the influence of sex in clinical K-level prediction using machine learning. Furthermore, incorporating evaluations from multiple clinics would also be valuable. Since clinical evaluations are subjective and different clinicians might have varying standards and biases when assessing K-levels, evaluations from various clinics can lead to a more objective assessment process, resulting in more accurate K-level classifications.

When exploring unsupervised learning, both K-means and GMM yielded similar clustering results, but given K-means' computational efficiency, it is preferable for clinical applications.

Even though none of the models are aligned with the clinical K-levels, which indicates that clinical assessments may not fully capture an individual's daily ambulatory capabilities, this analysis can be beneficial in several ways. The results showed that one cluster had higher functional ability, another low functional ability and the third had a balanced activity level. This suggests that data is clustered by activity metrics' values that are related to activity levels. So, in the future, these models can be used to create a mobile application that sends alerts to patients in the low activity cluster to encourage and guide them and can even aid clinicians in understanding and framing the patient's profile, thereby supporting decisions regarding rehabilitation and prosthetic components.

In the exploration of supervised learning for classifying prosthetic users' K-levels, the best results for multiclass classification (three classes: K2, K3 and K4) were achieved using RF with PCA with the activity metrics dataset (test F1-score of 0.58, accuracy of 0.70 and balanced accuracy of 0.67). This model successfully classified all K2 and K3 samples, but misclassified K4 as K3, which indicates difficulties in differentiating between higher K-levels. As it was expected, binary classification (combining K3 and K4 in one class) revealed to be an easier task, with more promising results. RF combined with PCA yielding high performance metrics (F1-score of 0.92, accuracy of 0.93, and balanced accuracy of 0.95) using activity metrics dataset. However, the best results were obtained for SVM using the combined dataset, achieving perfect classification scores across various feature reduction and selection methods. If the dataset was expanded and if it included a more balanced number of samples across different K-levels, it could be used to improve the robustness and generalizability of the models. These supervised learning techniques could be used to potentially enhance the accuracy of K-level assessments, especially when it comes to differentiate between K2 and higher K-levels. This would be particularly useful since reimbursement for a prosthetic device differs mainly between K2 and higher K-levels.

Since the binary classification task was resolved with high accuracy, in the future it would be interesting to explore model stacking, which would consist in using the results (class probabilities) of binary model as input for the multiclass task. This could help resolve some of the uncertainties observed in multiclass classification results.

It is possible that the extracted activity metrics, despite being logical and intuitively relevant to solving the k-level problem in which factors such as cadence variation and number of steps are relevant, may not extract the necessary characteristics of the temporal signal. To address this question, a direct analysis of activity time series and the usage of deep learning methods to extract relevant features (then used for classification problems) may be a promising approach to explore in the future.

Additionally, certain features, like EnergyExpenditureDailyAverage, EnergyExpenditureMax, ActiveTimeMediumCadenceAverage and PeakPerformanceScore, were frequently eliminated, indicating potential redundancy, while others, such as TwoMinuteWalkTestsAverage, TenMeterWalkTestsAverage_std, TopSpeedMax, Type of amputation and Stride length, proved consistently relevant.

Throughout the development of this project, several challenges were encountered. The first main difficulty was the fact that we had much less data than expected and a discrepancy in the number of samples for each K-level. This led to the necessity of applying data augmentation techniques. Initially, the Dynamic Time Warping-Synthetic Minority Over-sampling Technique was considered, applied to time series data. However, this approach revealed significant issues related to efficiency and memory consumption. Consequently, the original SMOTE applied to tabular data was chosen instead, since it allowed dataset augmentation without compromising available computational resources.

This project allowed me to have a deeper knowledge on types of amputation and prosthetic components available for each functionality level. Additionally, It also led to the understanding of the methods used to identify activity profiles among prosthesis users and how these profiles influence the assignment of prosthetic components. This project also allowed greater familiarity with data augmentation techniques and feature reduction and selection methods.

References

[1] R. Hoile, “Amputation: Surgical Practice and Patient Management,” *BMJ*, vol. 312, pp. 984–984, Apr. 1996.

[2] A. Mavroforou, S. Koutsias, F. Fafoulakis, I. Balogiannis, G. Stamatiou, and A. D. Giannoukas, “The evolution of lower limb amputation through the ages. Historical note,” *International Angiology: A Journal of the International Union of Angiology*, vol. 26, pp. 385–389, Dec. 2007.

[3] K. Ziegler-Graham, E. J. MacKenzie, P. L. Ephraim, T. G. Travison, and R. Brookmeyer, “Estimating the Prevalence of Limb Loss in the United States: 2005 to 2050,” *Archives of Physical Medicine and Rehabilitation*, vol. 89, pp. 422–429, Mar. 2008.

[4] M. S. Orendurff, S. U. Raschke, L. Winder, D. Moe, D. A. Boone, and T. Kobayashi, “Functional level assessment of individuals with transtibial limb loss: Evaluation in the clinical setting versus objective community ambulatory activity,” *Journal of Rehabilitation and Assistive Technologies Engineering*, vol. 3, p. 205566831663631, Jan. 2016.

[5] K. P. Smidt and R. Bicknell, “Prosthetics in Orthopedics,” in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2024.

[6] R. S. Gailey, K. E. Roach, E. B. Applegate, B. Cho, B. Cunniffe, S. Licht, M. Maguire, and M. S. Nash, “The Amputee Mobility Predictor: An instrument to assess determinants of the lower-limb amputee’s ability to ambulate,” *Archives of Physical Medicine and Rehabilitation*, vol. 83, pp. 613–627, May 2002.

[7] J. M. Stepien, S. Cavenett, L. Taylor, and M. Crotty, “Activity Levels Among Lower-Limb Amputees: Self-Report Versus Step Activity Monitor,” *Archives of Physical Medicine and Rehabilitation*, vol. 88, pp. 896–900, July 2007.

[8] B. Godfrey, C. Duncan, and T. Rosenbaum-Chou, “Comparison of Self-Reported vs Objective Measures of Long-Term Community Ambulation in Lower Limb Prosthesis Users,” *Archives of Rehabilitation Research and Clinical Translation*, vol. 4, p. 100220, Sept. 2022.

[9] M. C. Ashe, J. J. Eng, W. C. Miller, and J. A. Soon, “Disparity between Physical Capacity and Participation in Seniors with Chronic Disease,” *Medicine & Science in Sports & Exercise*, vol. 39, pp. 1139–1146, July 2007.

[10] A. Jayaraman, S. Deeny, Y. Eisenberg, G. Mathur, and T. Kuiken, “Global Position Sensing and Step Activity as Outcome Measures of Community Mobility and Social Interaction for an Individual With a Transfemoral Amputation Due to Dysvascular Disease,” *Physical Therapy*, vol. 94, pp. 401–410, Mar. 2014.

[11] C. Tudor-Locke, T. V. Barreira, R. M. Brouillette, H. C. Foil, and J. N. Keller, "Preliminary Comparison of Clinical and Free-Living Measures of Stepping Cadence in Older Adults," *Journal of Physical Activity and Health*, vol. 10, pp. 1175–1180, Nov. 2013.

[12] D. Trojaniello, A. Cereatti, E. Pelosin, L. Avanzino, A. Mirelman, J. M. Hausdorff, and U. Della Croce, "Estimation of step-by-step spatio-temporal parameters of normal and impaired gait using shank-mounted magneto-inertial sensors: application to elderly, hemiparetic, parkinsonian and choreic gait," *Journal of NeuroEngineering and Rehabilitation*, vol. 11, no. 1, p. 152, 2014.

[13] B. Hordacre, C. Barr, and M. Crotty, "Use of an Activity Monitor and GPS Device to Assess Community Activity and Participation in Transtibial Amputees," *Sensors*, vol. 14, pp. 5845–5859, Mar. 2014.

[14] I. Hillel, E. Gazit, A. Nieuwboer, L. Avanzino, L. Rochester, A. Cereatti, U. D. Croce, M. O. Rikkert, B. R. Bloem, E. Pelosin, S. Del Din, P. Ginis, N. Giladi, A. Mirelman, and J. M. Hausdorff, "Is every-day walking in older adults more analogous to dual-task walking or to usual walking? Elucidating the gaps between gait performance in the lab and during 24/7 monitoring," *European Review of Aging and Physical Activity*, vol. 16, p. 6, Dec. 2019.

[15] G. Hisano, S. Hashizume, Y. Kobayashi, A. Murai, T. Kobayashi, M. Nakashima, and H. Hobara, "Factors associated with a risk of prosthetic knee buckling during walking in unilateral transfemoral amputees," *Gait & Posture*, vol. 77, pp. 69–74, Mar. 2020.

[16] T. J. Reif, D. Jacobs, A. T. Fragomen, and S. R. Rozbruch, "Osseointegration Amputation Reconstruction," *Current Physical Medicine and Rehabilitation Reports*, vol. 10, pp. 61–70, June 2022.

[17] C. Jayaraman, C. K. Mummidisetty, M. V. Albert, R. Lipschutz, S. Hoppe-Ludwig, G. Mathur, and A. Jayaraman, "Using a microprocessor knee (C-Leg) with appropriate foot transitioned individuals with dysvascular transfemoral amputations to higher performance levels: a longitudinal randomized clinical trial," *Journal of NeuroEngineering and Rehabilitation*, vol. 18, p. 88, Dec. 2021.

[18] K. R. Kaufman, K. A. Bernhardt, and K. Symms, "Functional assessment and satisfaction of transfemoral amputees with low mobility (FASTK2): A clinical trial of microprocessor-controlled vs. non-microprocessor-controlled knees," *Clinical Biomechanics*, vol. 58, pp. 116–122, Oct. 2018.

[19] F. Davie-Smith and B. Carse, "Comparison of patient-reported and functional outcomes following transition from mechanical to microprocessor knee in the low-activity user with a unilateral transfemoral amputation," *Prosthetics & Orthotics International*, vol. 45, pp. 198–204, June 2021.

[20] A. Hahn and M. Lang, "Effects of Mobility Grade, Age, and Etiology on Functional Benefit and Safety of Subjects Evaluated in More Than 1200 C-Leg Trial Fittings in Germany," *JPO Journal of Prosthetics and Orthotics*, vol. 27, pp. 86–94, July 2015.

[21] M. P. Mileusnic, A. Hahn, and S. Reiter, "Effects of a Novel Microprocessor-Controlled Knee, Kenevo, on the Safety, Mobility, and Satisfaction of Lower-Activity Patients with Transfemoral Amputation," *JPO Journal of Prosthetics and Orthotics*, vol. 29, pp. 198–205, Oct. 2017.

[22] W. C. Miller, M. Speechley, and B. Deathe, "The prevalence and risk factors of falling and fear of falling among lower extremity amputees," *Archives of Physical Medicine and Rehabilitation*, vol. 82, pp. 1031–1037, Aug. 2001.

[23] M. A. Raya, SCS, ATC, R. S. Gailey, I. A. Gaunaurd, MSPT, H. Ganyard, J. Knapp-Wood, K. McDonough, DPT, T. Palmisano, and DPT, "Amputee Mobility Predictor-Bilateral: A performance-based measure of mobility for people with bilateral lower-limb loss," *Journal of Rehabilitation Research and Development*, vol. 50, no. 7, pp. 961–968, 2013.

[24] D. Brooks, J. Parsons, J. P. Hunter, M. Devlin, and J. Walker, "The 2-minute walk test as a measure of functional improvement in persons with lower limb amputation," *Archives of Physical Medicine and Rehabilitation*, vol. 82, pp. 1478–1483, Oct. 2001.

[25] L. Resnik and M. Borgia, "Reliability of Outcome Measures for People With Lower-Limb Amputations: Distinguishing True Change From Statistical Error," *Physical Therapy*, vol. 91, pp. 555–565, Apr. 2011.

[26] I. Gaunaurd, A. Kristal, A. Horn, C. Krueger, O. Muro, A. Rosenberg, K. Gruben, N. Kirk-Sanchez, P. Pasquina, and R. Gailey, "The Utility of the 2-Minute Walk Test as a Measure of Mobility in People With Lower Limb Amputation," *Archives of Physical Medicine and Rehabilitation*, vol. 101, pp. 1183–1189, July 2020.

[27] H. R. Batten, S. M. McPhail, A. M. Mandrusiak, P. N. Varghese, and S. S. Kuys, "Gait speed as an indicator of prosthetic walking potential following lower limb amputation," *Prosthetics & Orthotics International*, vol. 43, pp. 196–203, Apr. 2019.

[28] W. Dite, H. J. Connor, and H. C. Curtis, "Clinical Identification of Multiple Fall Risk Early After Unilateral Transtibial Amputation," *Archives of Physical Medicine and Rehabilitation*, vol. 88, pp. 109–114, Jan. 2007.

[29] J. Kim, N. Colabianchi, J. Wensman, and D. H. Gates, "Wearable Sensors Quantify Mobility in People With Lower Limb Amputation During Daily Life," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, pp. 1282–1291, June 2020.

[30] Z. Shelly, R. F. Burch, W. Tian, L. Strawderman, A. Piroli, and C. Bichey, "Using K-means Clustering to Create Training Groups for Elite American Football Student-athletes Based on Game Demands," *International Journal of Kinesiology and Sports Science*, vol. 8, p. 47, May 2020.

[31] G. McLachlan, *Finite Mixture Models*. Hoboken: John Wiley & Sons, 2004. OCLC: 181838458.

[32] S. Das, A. Dey, A. Pal, and N. Roy, "Applications of Artificial Intelligence in Machine Learning: Review and Prospect," *International Journal of Computer Applications*, vol. 115, pp. 31–41, Apr. 2015.

[33] M. Alauthman, A. Aldweesh, A. Al-qerem, F. Aburub, Y. Al-Smadi, A. M. Abaker, O. R. Alzubi, and B. Alzubi, "Tabular Data Generation to Improve Classification of Liver Disease Diagnosis," *Applied Sciences*, vol. 13, p. 2678, Feb. 2023.

[34] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, June 2002.

[35] R. Soundrapandiyan, A. Manickam, M. Akhloufi, Y. V. S. Murthy, R. D. M. Sundaram, and S. Thirugnanasambandam, "An Efficient COVID-19 Mortality Risk Prediction Model Using Deep Synthetic Minority Oversampling Technique and Convolution Neural Networks," *BioMedInformatics*, vol. 3, pp. 339–368, May 2023.

[36] X. Yang, Z. Zhang, X. Cui, and R. Cui, "A Time Series Data Augmentation Method Based on Dynamic Time Warping," in *2021 International Conference on Computer Communication and Artificial Intelligence (CCAI)*, (Guangzhou, China), pp. 116–120, IEEE, May 2021.

[37] G. Iglesias, E. Talavera, Ã. GonzÃąlez-Prieto, A. Mozo, and S. GÃşmez-Canaval, "Data Augmentation techniques in time series domain: a survey and taxonomy," *Neural Computing and Applications*, vol. 35, pp. 10123–10145, May 2023.

[38] T. T. Um, F. M. J. Pfister, D. Pichler, S. Endo, M. Lang, S. Hirche, U. Fietzek, and D. KuliÄĞ, "Data augmentation of wearable sensor data for parkinsonâĂŹs disease monitoring using convolutional neural networks," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, (Glasgow UK), pp. 216–220, ACM, Nov. 2017.

[39] W. Jia, M. Sun, J. Lian, and S. Hou, "Feature dimensionality reduction: a review," *Complex & Intelligent Systems*, vol. 8, pp. 2663–2693, June 2022.

[40] H. Liu and H. Motoda, eds., *Computational methods of feature selection*. Chapman & Hall/CRC data mining and knowledge discovery series, Boca Raton: Chapman & Hall/CRC, 2008. OCLC: ocn154309055.

[41] M. Doshi and S. K. Chaturvedi, "Correlation Based Feature Selection (CFS) Technique to Predict Student Perfromance," *International journal of Computer Networks & Communications*, vol. 6, pp. 197–206, May 2014.

[42] N. Omer Fadl Elssied, O. Ibrahim, and A. Hamza Osman, "A Novel Feature Selection Based on One-Way ANOVA F-Test for E-Mail Spam Classification," *Research Journal of Applied Sciences, Engineering and Technology*, vol. 7, pp. 625–638, Jan. 2014.

[43] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, pp. 273–324, Dec. 1997.

[44] M. M. Kabir, M. Shahjahan, and K. Murase, "A new local search based hybrid genetic algorithm for feature selection," *Neurocomputing*, vol. 74, pp. 2914–2928, Oct. 2011.

[45] K. Balabaeva and S. Kovalchuk, "Comparison of Efficiency, Stability and Interpretability of Feature Selection Methods for Multiclassification Task on Medical Tabular Data," in *Computational Science âĂŞ ICCS 2021* (M. Paszynski, D. KranzlmÃijller, V. V. Krzhizhanovskaya, J. J. Dongarra, and P. M. Sloot, eds.), vol. 12744, pp. 623–633, Cham: Springer International Publishing, 2021. Series Title: Lecture Notes in Computer Science.

[46] C. M. Eckhardt, S. J. Madjarova, R. J. Williams, M. Ollivier, J. Karlsson, A. Pareek, and B. U. Nwachukwu, "Unsupervised machine learning methods and emerging applications in healthcare," *Knee Surgery, Sports Traumatology, Arthroscopy*, vol. 31, pp. 376–381, Feb. 2023.

[47] Y. J. Choo and M. C. Chang, "Use of machine learning in the field of prosthetics and orthotics: A systematic narrative review," *Prosthetics & Orthotics International*, vol. 47, pp. 226–240, June 2023.

[48] X. Zhang, G. Fiedler, Z. Cao, and Z. Liu, "A support vector machine approach to detect trans-tibial prosthetic misalignment using 3-Dimensional ground reaction force features: A proof of concept," *Technology and Health Care*, vol. 26, pp. 715–721, Sept. 2018.

[49] P. Sewell, S. Noroozi, J. Vinney, R. Amali, and S. Andrews, "Static and dynamic pressure prediction for prosthetic socket fitting assessment utilising an inverse problem approach," *Artificial Intelligence in Medicine*, vol. 54, pp. 29–41, Jan. 2012.

[50] R. LeMoyne, T. Mastroianni, A. Hessel, and K. Nishikawa, "Implementation of machine learning for classifying prosthesis type through conventional gait analysis," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, (Milan), pp. 202–205, IEEE, Aug. 2015.

[51] C.-J. Huang, Y.-W. Wang, T.-H. Huang, C.-F. Lin, C.-Y. Li, H.-M. Chen, P. C. Chen, and J.-J. Liao, "Applications of machine learning techniques to a sensor-network-based prosthesis training system," *Applied Soft Computing*, vol. 11, pp. 3229–3237, Apr. 2011.

[52] N. Shawen, L. Lonini, C. K. Mummidisetty, I. Shparii, M. V. Albert, K. Kording, and A. Jayaraman, "Fall Detection in Individuals With Lower Limb Amputations Using Mobile Phones: Machine Learning Enhances Robustness for Real-World Applications," *JMIR mHealth and uHealth*, vol. 5, p. e151, Oct. 2017.

[53] N. Mathur, I. Glesk, and A. Buis, "Thermal time constant: optimising the skin temperature predictive modelling in lower limb prostheses using Gaussian processes," *Healthcare Technology Letters*, vol. 3, pp. 98–104, June 2016.

[54] L. Lonini, A. Gupta, S. Deems-Dluhy, S. Hoppe-Ludwig, K. Kording, and A. Jayaraman, "Activity Recognition in Individuals Walking With Assistive Devices: The Benefits of Device-Specific Models," *JMIR Rehabilitation and Assistive Technologies*, vol. 4, p. e8, Aug. 2017.

[55] J. D. Farah, N. Baddour, and E. D. Lemaire, "Design, development, and evaluation of a local sensor-based gait phase recognition system using a logistic model decision tree for orthosis-control," *Journal of NeuroEngineering and Rehabilitation*, vol. 16, p. 22, Dec. 2019.

[56] Y. J. Choo, J. K. Kim, J. H. Kim, M. C. Chang, and D. Park, "Machine learning analysis to predict the need for ankle foot orthosis in patients with stroke," *Scientific Reports*, vol. 11, p. 8499, Apr. 2021.

[57] C. Soto-Valero, "A Gaussian mixture clustering model for characterizing football players using the EA Sports' FIFA video game system. [Modelo basado en agrupamiento de mixturas Gaussianas para caracterizar futbolistas utilizando el sistema de videojuegos FIFA de EA Sports].," *RICYDE. Revista internacional de ciencias del deporte*, vol. 13, pp. 244–259, July 2017.

[58] E. Dolatabadi, A. Mansfield, K. K. Patterson, B. Taati, and A. Mihailidis, "Mixture-Model Clustering of Pathological Gait Patterns," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, pp. 1297–1305, Sept. 2017.

[59] D. L. Carey, K. Ong, M. E. Morris, J. Crow, and K. M. Crossley, "Predicting ratings of perceived exertion in Australian football players: methods for live estimation," *International Journal of Computer Science in Sport*, vol. 15, pp. 64–77, Dec. 2016.

[60] J. A. Albert, A. Herdick, C. M. Brahms, U. Granacher, and B. Arnrich, "Using Machine Learning to Predict Perceived Exertion During Resistance Training With Wearable Heart Rate and Movement Sensors," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, (Houston, TX, USA), pp. 801–808, IEEE, Dec. 2021.

[61] A. Alanazi, "Using machine learning for healthcare challenges and opportunities," *Informatics in Medicine Unlocked*, vol. 30, p. 100924, 2022.

[62] R. Singh, J. Hunter, A. Philip, and S. Tyson, "Gender differences in amputation outcome," *Disability and Rehabilitation*, vol. 30, pp. 122–125, Jan. 2008.

[63] M. Seth, E. H. Beisheim, R. T. Pohlig, J. R. Horne, F. B. Sarlo, and J. M. Sions, "Time Since Lower-Limb Amputation: An Important Consideration in Mobility Outcomes," *American Journal of Physical Medicine & Rehabilitation*, vol. 101, pp. 32–39, Jan. 2022.

[64] R. Sinha, W. J. Van Den Heuvel, and P. Arokiasamy, "Factors affecting quality of life in lower limb amputees," *Prosthetics & Orthotics International*, vol. 35, pp. 90–96, Mar. 2011.

[65] H. Akoglu, "User's guide to correlation coefficients," *Turkish Journal of Emergency Medicine*, vol. 18, pp. 91–93, Sept. 2018.

[66] S. N. Papageorgiou, "On correlation coefficients and their interpretation," *Journal of Orthodontics*, vol. 49, pp. 359–361, Sept. 2022.

[67] T. A. Miller, R. Paul, M. Forthofer, and S. R. Wurdeman, "Factors that influence time to prosthesis receipt after lower limb amputation: A Cox proportional hazard model regression," *PM&R*, vol. 15, pp. 474–481, Apr. 2023.

[68] E. S. Arch, J. M. Sions, J. Horne, and B. A. Bodt, "Step count accuracy of StepWatch and FitBit OneâĎć among individuals with a unilateral transtibial amputation," *Prosthetics & Orthotics International*, vol. 42, pp. 518–526, Oct. 2018.

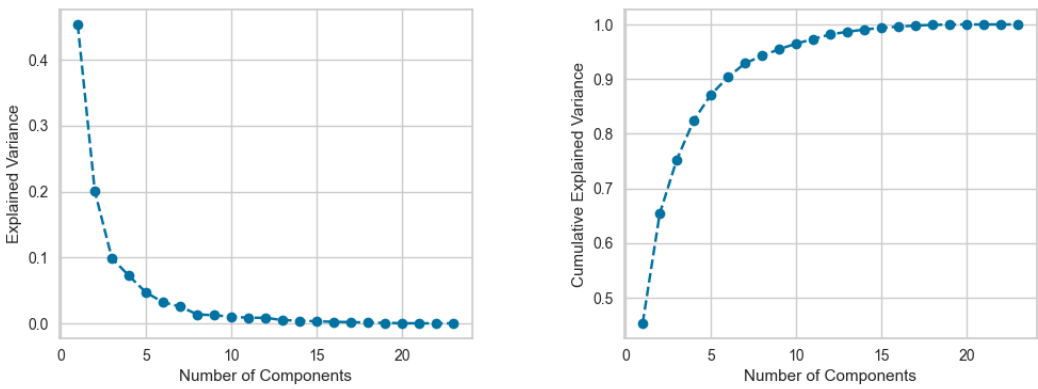
[69] J. M. Sions, E. H. Beisheim, T. J. Manal, S. C. Smith, J. R. Horne, and F. B. Sarlo, "Differences in Physical Performance Measures Among Patients With Unilateral Lower-Limb Amputations Classified as Functional Level K3 Versus K4," *Archives of Physical Medicine and Rehabilitation*, vol. 99, pp. 1333–1341, July 2018.

[70] E. Patel and D. S. Kushwaha, "Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model," *Procedia Computer Science*, vol. 171, pp. 158–167, 2020.

Appendix A

Unsupervised Learning Results: Dataset 1

A.1 Approach 1: Activity Metrics Dataset



(a) Individual explained variance

(b) Cumulative explained variance

Figure A.1: PCA using activity metrics Dataset 1.

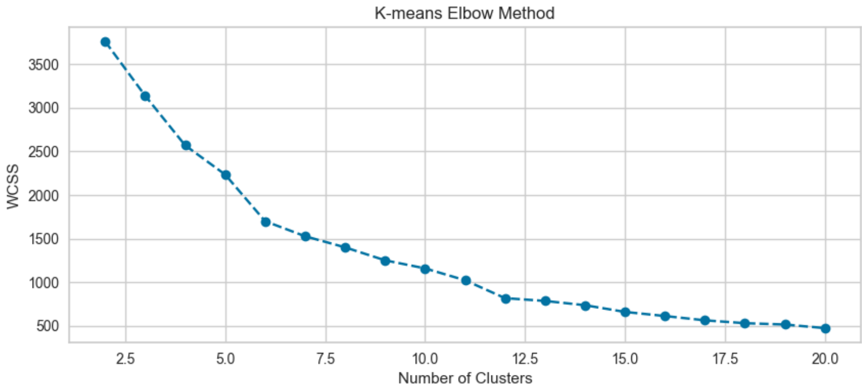
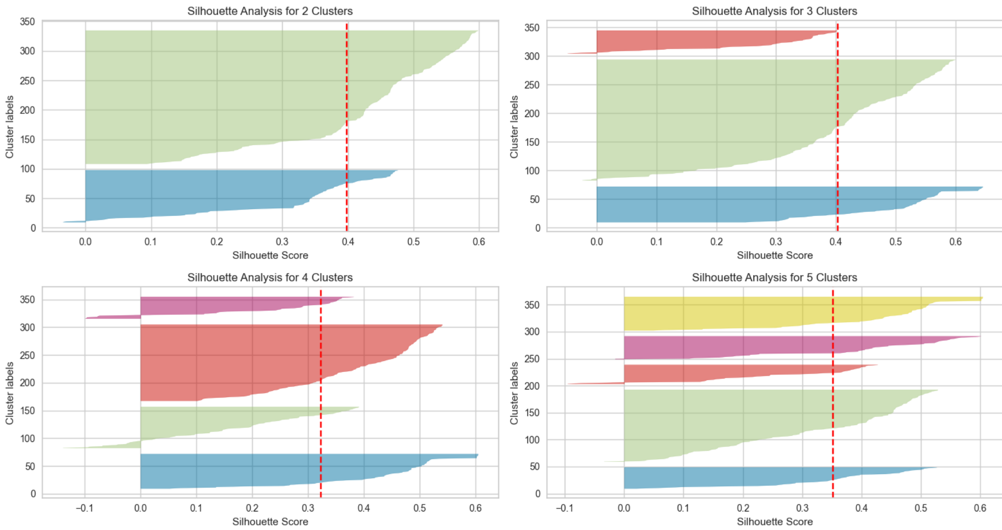
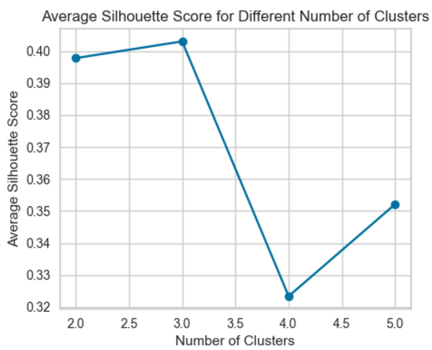


Figure A.2: K-means: Elbow method applied to K-means using activity metrics Dataset 1.

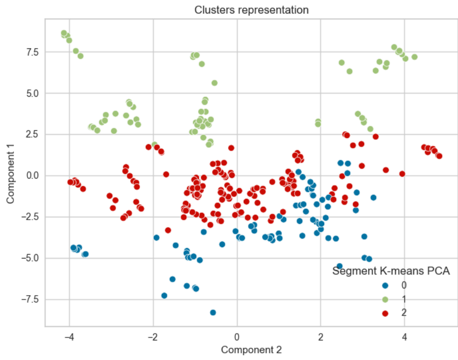


(a) Silhouette analysis plot for k=2 to k=5

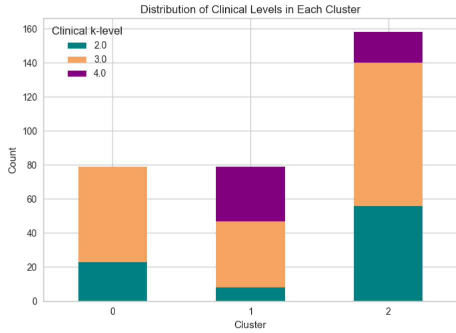


(b) Average silhouette score distribution

Figure A.3: K-means: Silhouette method applied to K-means using activity metrics Dataset 1.



(a) Cluster representation



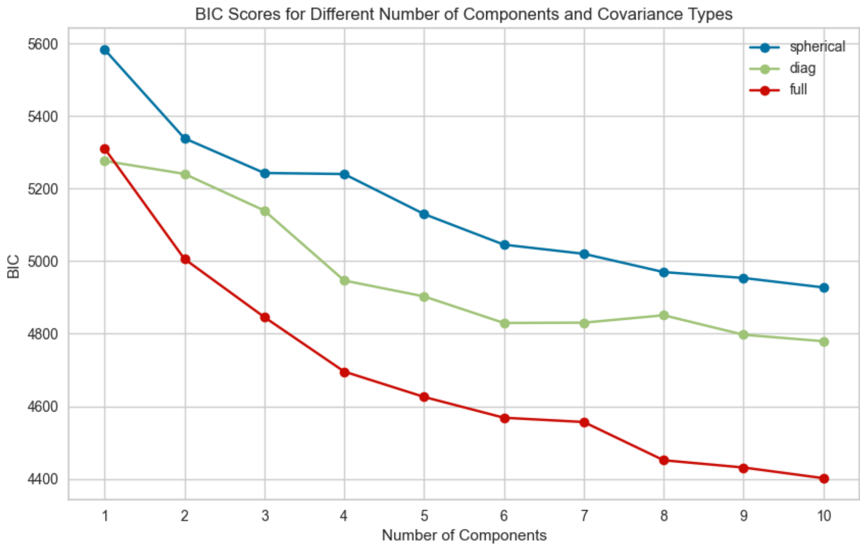
(b) K-level distribution across clusters

Figure A.4: K-means: cluster analysis (k=3) using activity metrics Dataset 1.

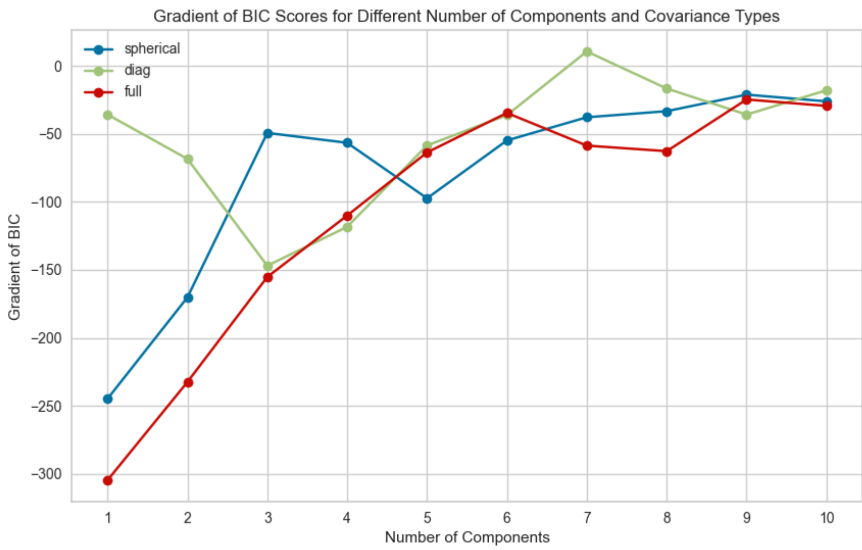
Table A.1: K-means: table with mean values for each activity metric from dataset after sliding window, by cluster.

Activity Metric	Cluster 0	Cluster 1	Cluster 2
StepsDailyAverage	1877.1	8558.9	4074.1
StepsMax	5248.4	12894.0	6135.5
TopCadenceDailyAverage	56.8	105.9	88.1
TopCadenceMax	94.4	119.9	105.6
TopSpeedDailyAverage	0.5	0.9	0.8
TopSpeedMax	0.9	1.0	0.9
EnergyExpenditureDailyAverage	32.8	134.4	66.0
EnergyExpenditureMax	84.5	200.9	97.6
MaxContinuousWalkingDailyAverage	135.6	318.3	130.7
MaxContinuousWalkingMax	402.4	739.5	264.8
AmbulationEnergyScore	2.6	4.5	4.4
PeakPerformanceScore	2.1	3.9	3.2
CadenceVariabilityScore	2.7	3.8	2.7
ActiveTimeAverage	0.6	2.4	1.4
ActiveTimeLowCadenceAverage	0.2	0.8	0.6
ActiveTimeMediumCadenceAverage	0.3	0.9	0.5
ActiveTimeHighCadenceAverage	0.1	0.7	0.3
TwoMinuteWalkTestsAverage	83.0	82.9	81.0
TwoMinuteWalkTestsAverage_std	14.2	21.2	17.3
SixMinuteWalkTestsAverage	319.4	290.9	259.5
SixMinuteWalkTestsAverage_std	18.2	38.9	13.5
TenMeterWalkTestsAverage	0.6	0.6	0.6
TenMeterWalkTestsAverage_std	0.2	0.2	0.2

Figure A.5: K-means: radar chart of mean values for each activity metric from Dataset 1.

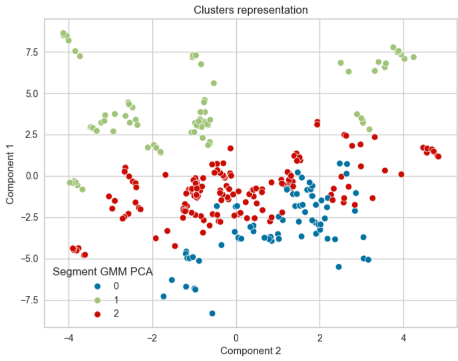


(a) BIC values

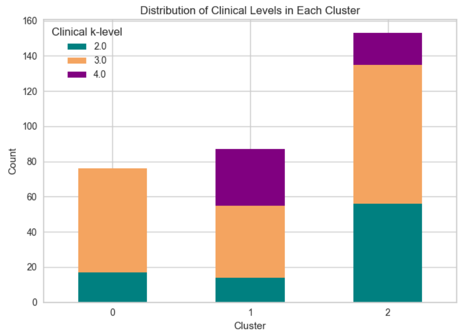


(b) Gradient of BIC

Figure A.6: GMM: BIC applied to GMM using activity metrics Dataset 1.



(a) Cluster representation



(b) K-level distribution across clusters

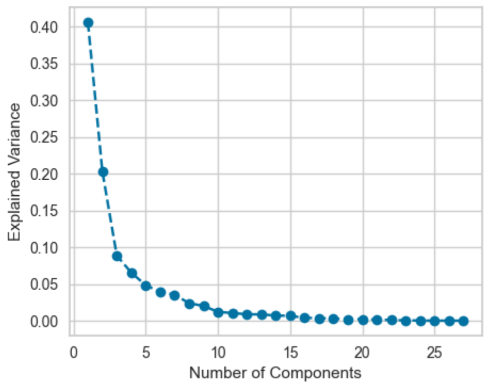
Figure A.7: GMM: cluster analysis (k=3) using activity metrics Dataset 1.

Table A.2: GMM: table with mean values for each activity metric from Dataset 1, by cluster.

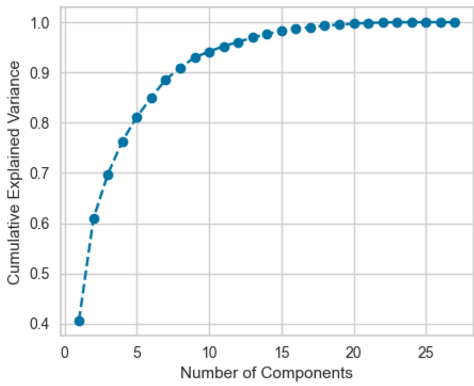
Activity Metric	Cluster 0	Cluster 1	Cluster 2
StepsDailyAverage	1897.2	8230.5	3973.3
StepsMax	5370.4	12411.3	5978.6
TopCadenceDailyAverage	57.8	104.7	86.7
TopCadenceMax	96.5	118.3	104.5
TopSpeedDailyAverage	0.6	0.9	0.8
TopSpeedMax	0.9	0.9	0.9
EnergyExpenditureDailyAverage	33.1	129.4	64.4
EnergyExpenditureMax	86.3	193.6	95.2
MaxContinuousWalkingDailyAverage	147.5	315.2	116.9
MaxContinuousWalkingMax	447.6	712.9	235.4
AmbulationEnergyScore	2.7	4.6	4.3
PeakPerformanceScore	2.1	3.9	3.1
CadenceVariabilityScore	2.8	3.8	2.6
ActiveTimeAverage	0.6	2.3	1.4
ActiveTimeLowCadenceAverage	0.2	0.7	0.6
ActiveTimeMediumCadenceAverage	0.3	0.9	0.5
ActiveTimeHighCadenceAverage	0.1	0.7	0.2
TwoMinuteWalkTestsAverage	88.4	79.9	80.1
TwoMinuteWalkTestsAverage_std	14.9	19.9	17.4
SixMinuteWalkTestsAverage	312.8	278.1	274.4
SixMinuteWalkTestsAverage_std	17.2	34.9	16.5
TenMeterWalkTestsAverage	0.7	0.6	0.6
TenMeterWalkTestsAverage_std	0.2	0.2	0.2

Figure A.8: GMM: radar chart of mean values for each activity metric from Dataset 1, by cluster.

A.2 Approach 2: Combined Dataset



(a) Individual explained variance



(b) Cumulative explained variance

Figure A.9: PCA using combined with Dataset 1.

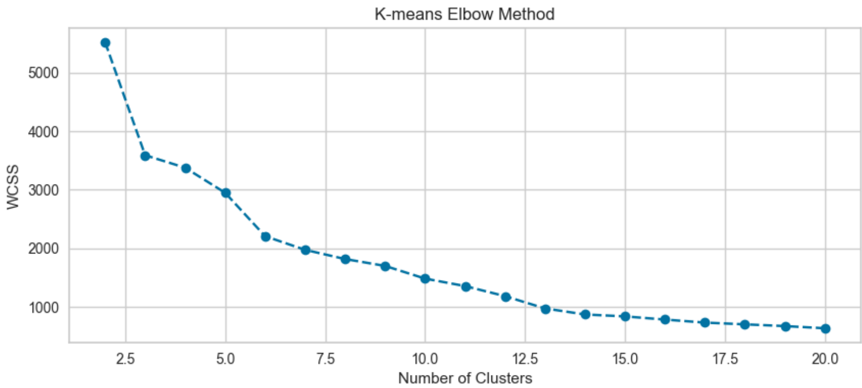
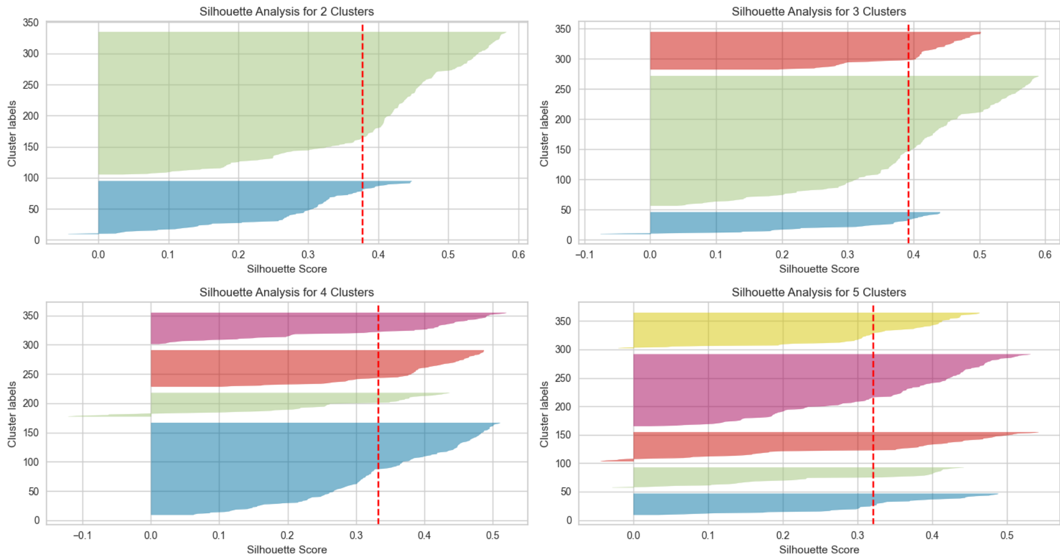
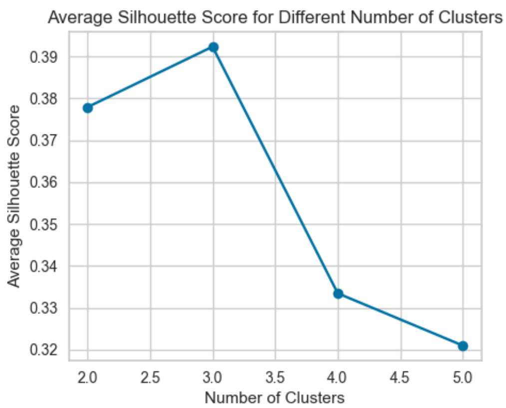


Figure A.10: K-means: Elbow method applied to K-means using combined Dataset 1.

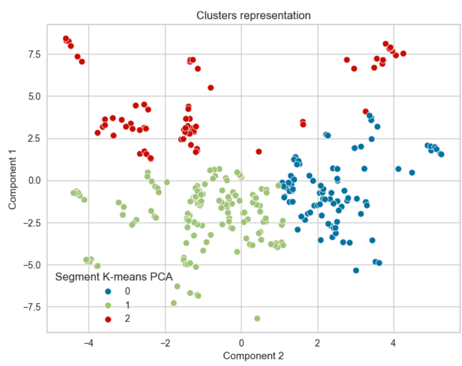


(a) Silhouette analysis plot for k=2 to k=5

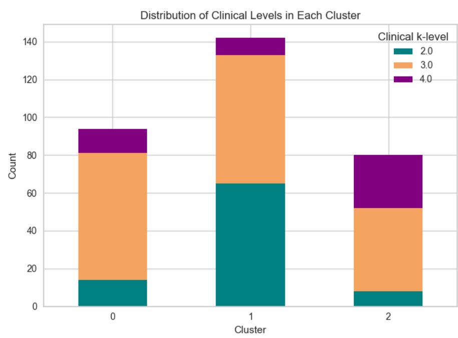


(b) Average silhouette score distribution

Figure A.11: K-means: Silhouette method applied to K-means using combined Dataset 1.



(a) Cluster representation



(b) K-level distribution across clusters

Figure A.12: K-means: cluster analysis (k=3) using combined Dataset 1.

Table A.3: K-means: table with mean values for each appointment data and activity metric from Dataset 1, by cluster.

Feature	Cluster 0	Cluster 1	Cluster 2
Age	40-49	50-59	60-69
Amputation Type	Above Knee	Above Knee	Below Knee
Years Since Amputation	17.3	8.2	19.0
Stride Length	130.4	99.3	104.3
StepsDailyAverage	3341.5	3361.0	8459.8
StepsMax	6791.6	5202.4	12818.8
TopCadenceDailyAverage	75.3	78.9	105.9
TopCadenceMax	104.0	100.26	120.1
TopSpeedDailyAverage	0.8	0.7	0.9
TopSpeedMax	1.1	0.8	1.0
EnergyExpenditureDailyAverage	54.6	55.4	132.9
EnergyExpenditureMax	107.5	83.6	199.7
MaxContinuousWalkingDailyAverage	176.0	103.5	315.9
MaxContinuousWalkingMax	473.4	202.6	734.9
AmbulationEnergyScore	3.6	3.9	4.5
PeakPerformanceScore	2.7	2.9	3.9
CadenceVariabilityScore	2.9	2.6	3.8
ActiveTimeAverage	1.0	1.1	2.4
ActiveTimeLowCadenceAverage	0.4	0.5	0.8
ActiveTimeMediumCadenceAverage	0.4	0.5	0.9
ActiveTimeHighCadenceAverage	0.2	0.2	0.7
TwoMinuteWalkTestsAverage	102.3	69.4	80.7
TwoMinuteWalkTestsAverage_std	20.3	14.0	20.4
SixMinuteWalkTestsAverage	336.4	214.4	284.7
SixMinuteWalkTestsAverage_std	17.8	13.1	36.9
TenMeterWalkTestsAverage	0.8	0.5	0.6
TenMeterWalkTestsAverage_std	0.2Âń	0.1	0.2

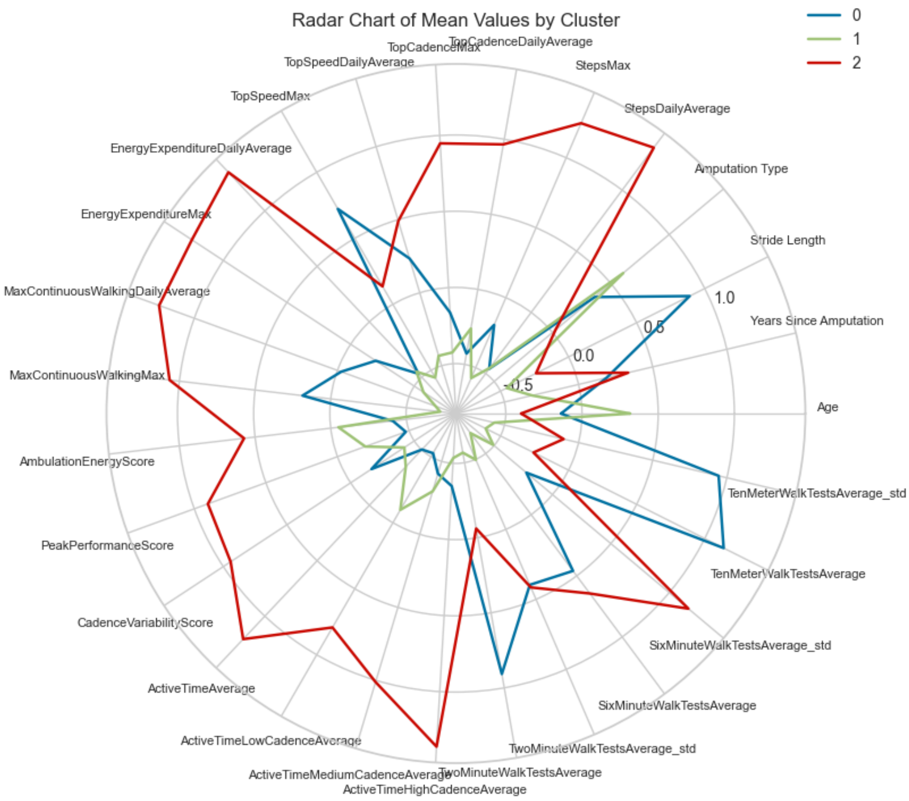
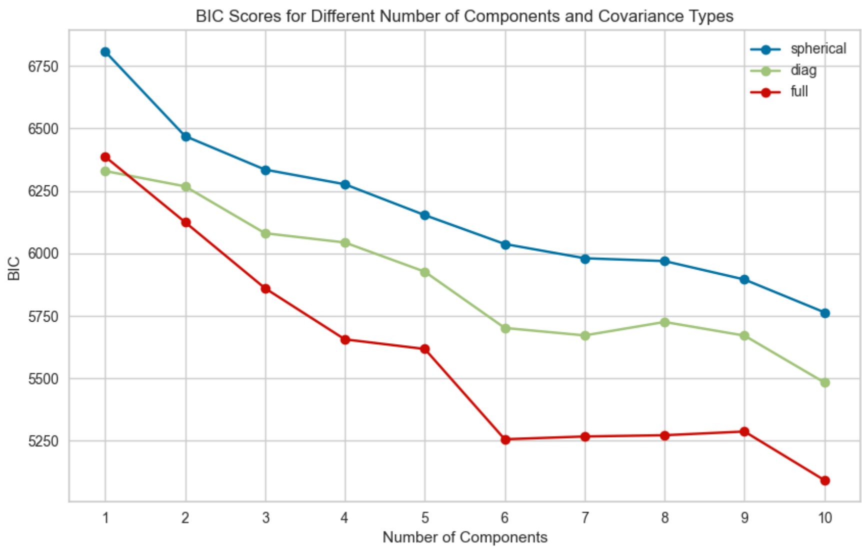
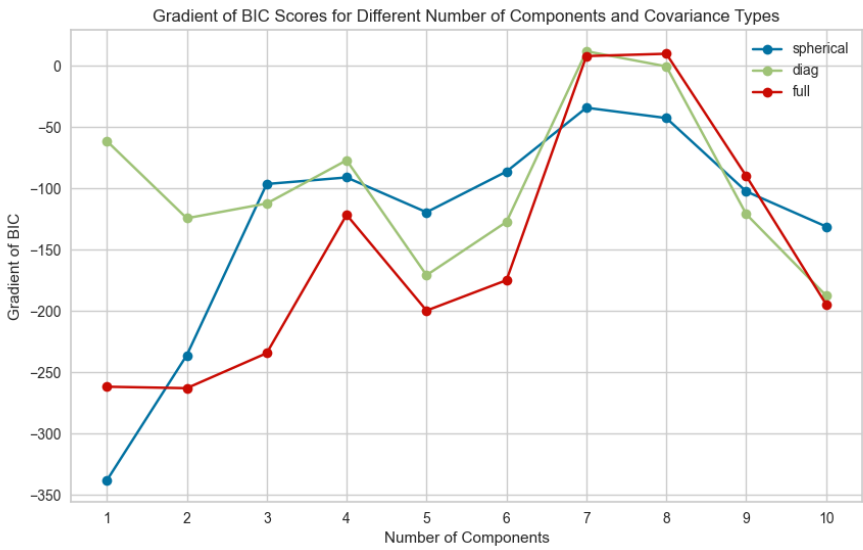


Figure A.13: K-means: radar chart of mean values for each appointment data and activity metric from Dataset 1, by cluster.

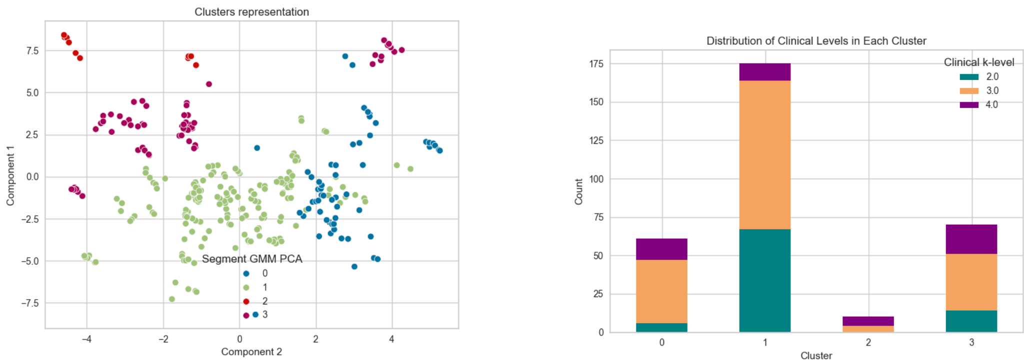


(a) BIC values



(b) Gradient of BIC

Figure A.14: GMM: BIC applied to GMM using combined Dataset 1.



(a) Cluster representation

(b) K-level distribution across clusters

Figure A.15: GMM: cluster analysis (k=3) using combined Dataset 1.

Table A.4: GMM: table with mean values for each appointment data and activity metric from Dataset 1, by cluster.

Feature	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Age	40-49	50-59	40-50	60-69
Amputation Type	Unknown	Above Knee	Below Knee	Below Knee
Years Since Amputation	26.5	9.3	3.2	19.4
Stride Length	137.0	104.9	94.4	100.5
StepsDailyAverage	3325.0	3438.0	13123.2	7606.2
StepsMax	7534.6	5378.1	21428.8	11251.2
TopCadenceDailyAverage	64.5	82.7	106.2	104.4
TopCadenceMax	98.9	103.5	112.6	119.3
TopSpeedDailyAverage	0.8	0.7	0.8	0.8
TopSpeedMax	1.2	0.9	0.9	0.9
EnergyExpenditureDailyAverage	54.2	56.5	204.4	119.9
EnergyExpenditureMax	118.8	86.3	331.7	175.8
MaxContinuousWalkingDailyAverage	210.2	103.6	395.9	308.6
MaxContinuousWalkingMax	571.6	220.1	970.1	699.6
AmbulationEnergyScore	3.2	3.9	4.6	4.6
PeakPerformanceScore	2.5	2.9	4.1	3.9
CadenceVariabilityScore	2.9	2.6	4.5	3.8
ActiveTimeAverage	1.1	1.1	3.5	2.1
ActiveTimeLowCadenceAverage	0.4	0.5	1.0	0.7
ActiveTimeMediumCadenceAverage	0.5	0.4	1.2	0.8
ActiveTimeHighCadenceAverage	0.2	0.2	1.3	0.7
TwoMinuteWalkTestsAverage	106.0	76.0	78.1	76.5
TwoMinuteWalkTestsAverage_std	19.6	16.1	16.9	19.3
SixMinuteWalkTestsAverage	344.9	249.2	264.1	270.4
SixMinuteWalkTestsAverage_std	18.5	14.6	22.4	37.1
TenMeterWalkTestsAverage	0.8	0.6	0.6	0.6
TenMeterWalkTestsAverage_std	0.2	0.2	0.2	0.2

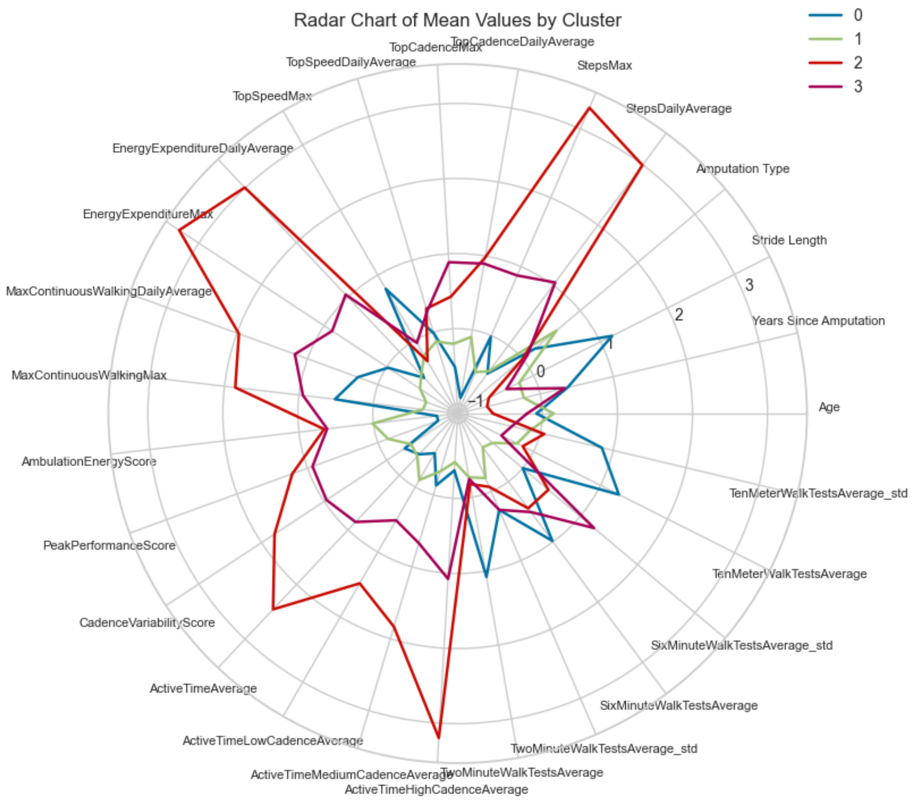


Figure A.16: GMM: radar chart of mean values for each appointment data and activity metric from Dataset 1, by cluster.

Appendix B

Supervised Learning Results: Dataset A and B

B.1 Multiclass Classification

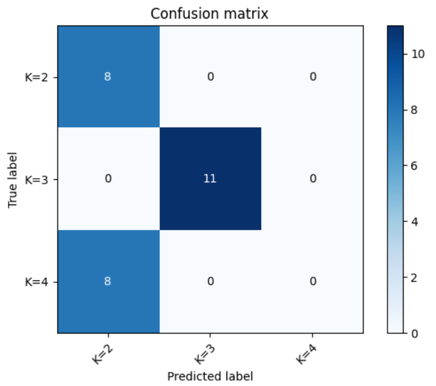
B.1.1 Approach 1: Activity Metrics Dataset A

Table B.1: Performance metrics of multiclass classification using SVM with activity metrics Dataset A.

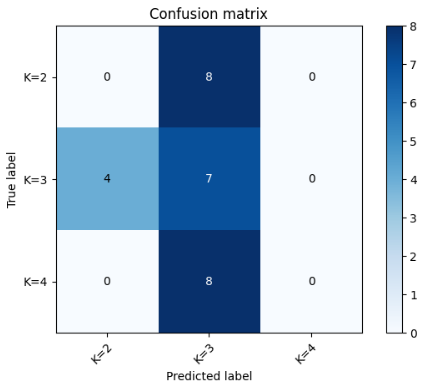
Feature Reduction and Selection	F1-Score Validation	F1-Score Test	Accuracy Test	Balanced Accuracy
Baseline	0.58	0.56	0.70	0.67
PCA	0.59	0.14	0.26	0.21
KBest	0.59	0.58	0.70	0.67
RFE	0.69	0.19	0.41	0.33
Correlation	0.61	0.56	0.70	0.67

Table B.2: Performance metrics of multiclass classification using RF with activity metrics Dataset A.

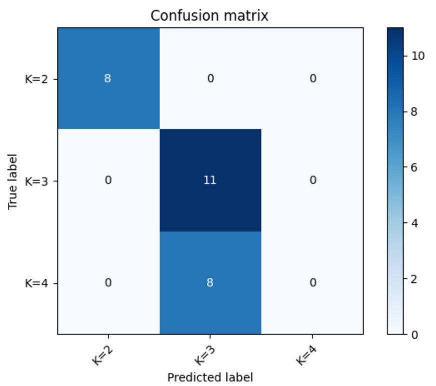
Feature Reduction and Selection	F1-Score Validation	F1-Score Test	Accuracy Test	Balanced Accuracy
Baseline	0.61	0.15	0.26	0.21
PCA	0.42	0.09	0.15	0.12
KB	0.61	0.18	0.26	0.21
RFE	0.63	0.15	0.26	0.21
Corr	0.58	0.27	0.33	0.30



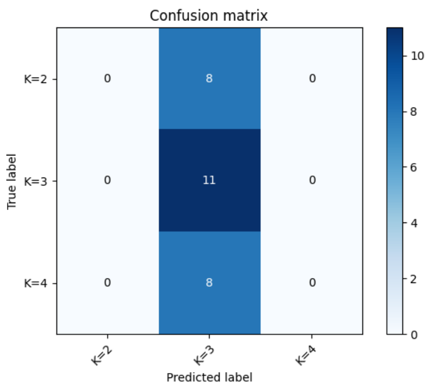
(a) Confusion matrix for the baseline model



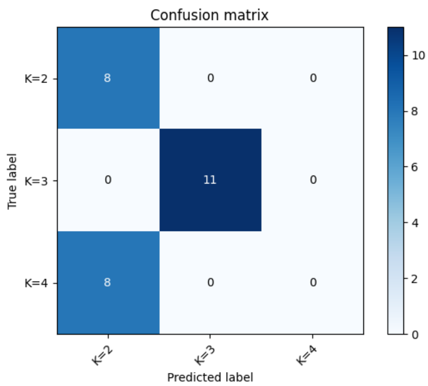
(b) Confusion matrix for PCA



(c) Confusion matrix for KBest

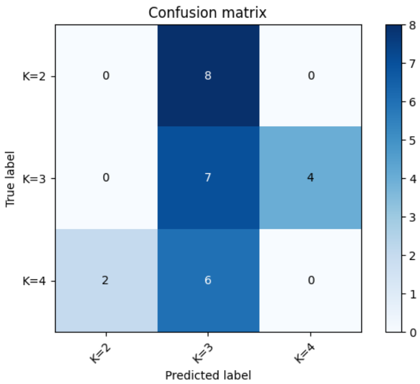


(d) Confusion matrix for RFE

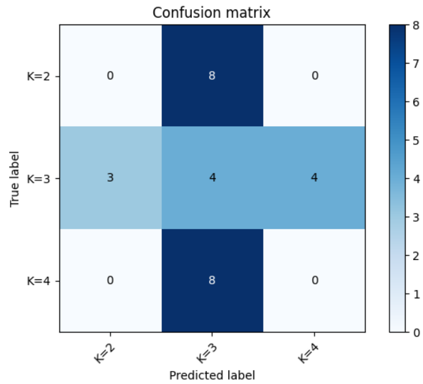


(e) Confusion matrix for Correlation

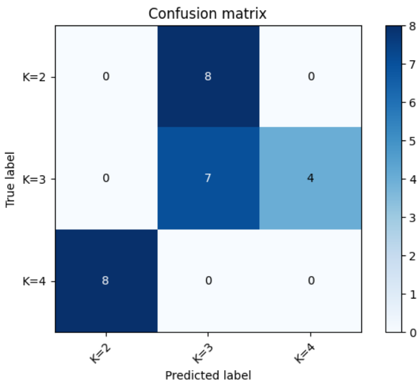
Figure B.1: Confusion matrices for multiclass classification using SVM with activity metrics Dataset A.



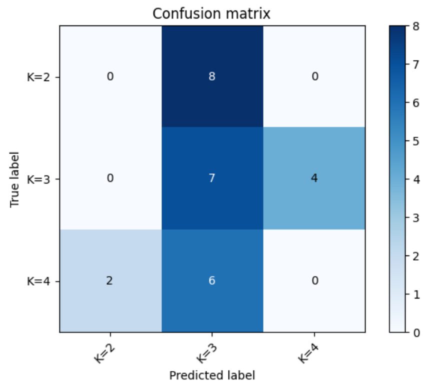
(a) Confusion matrix for the baseline model



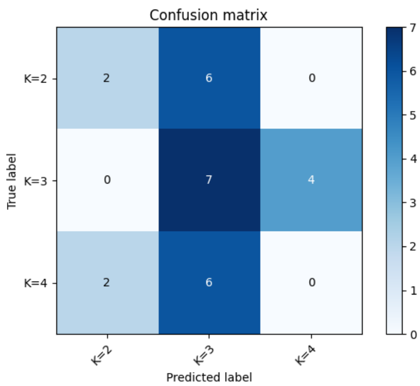
(b) Confusion matrix for PCA



(c) Confusion matrix for KBest



(d) Confusion matrix for RFE



(e) Confusion matrix for Correlation

Figure B.2: Confusion matrices for multiclass classification using RF with activity metrics Dataset A.

B.1.2 Approach 2: Combined Dataset A

Table B.3: Performance metrics of multiclass classification using SVM with combined Dataset A.

Feature Reduction and Selection	F1-Score Validation	F1-Score Test	Accuracy Test	Balanced Accuracy
Baseline	0.59	0.19	0.41	0.33
PCA	0.60	0.44	0.56	0.50
KB	0.66	0.19	0.41	0.33
RFE	0.75	0.44	0.56	0.50
Corr	0.68	0.19	0.41	0.33

Table B.4: Performance metrics of multiclass classification using RF with combined Dataset A.

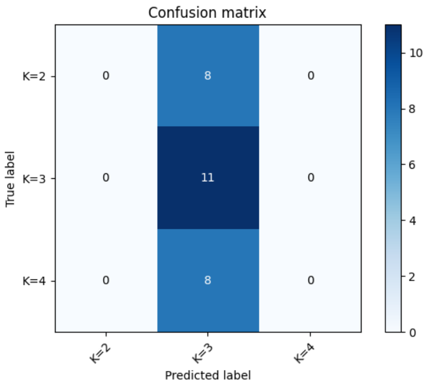
Feature Reduction and Selection	F1-Score Validation	F1-Score Test	Accuracy Test	Balanced Accuracy
Baseline	0.65	0.51	0.56	0.55
PCA	0.47	0.44	0.56	0.50
KB	0.63	0.49	0.56	0.55
RFE	0.70	0.79	0.78	0.80
Corr	0.63	0.49	0.56	0.55

B.2 Binary Classification

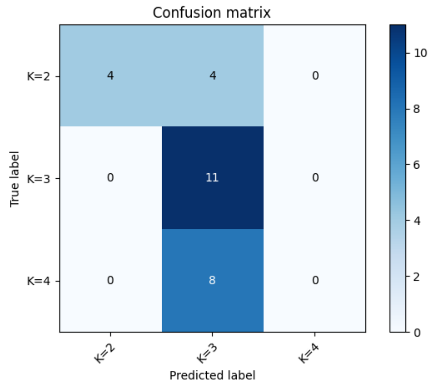
B.2.1 Approach 1: Activity Metrics Dataset B

Table B.5: Performance metrics of binary classification using SVM with activity metrics Dataset B.

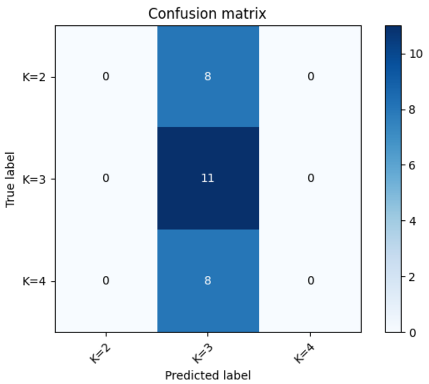
Feature Reduction and Selection	F1-Score Validation	F1-Score Test	Accuracy Test	Balanced Accuracy
Baseline	0.82	0.70	0.70	0.79
PCA	0.78	0.36	0.56	0.39
KB	0.80	0.70	0.70	0.79
RFE	0.82	0.79	0.85	0.75
Corr	0.79	0.84	0.85	0.90



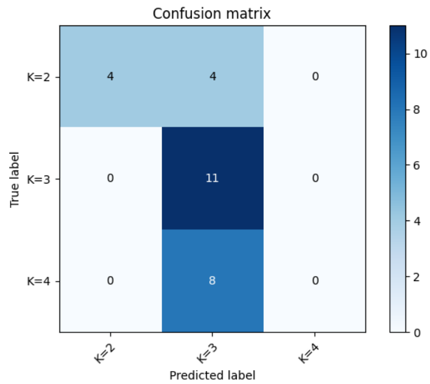
(a) Confusion matrix for the baseline model



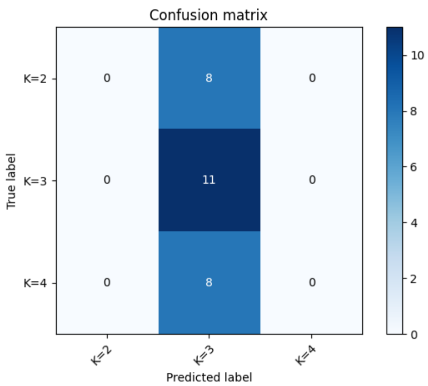
(b) Confusion matrix for PCA



(c) Confusion matrix for KBest

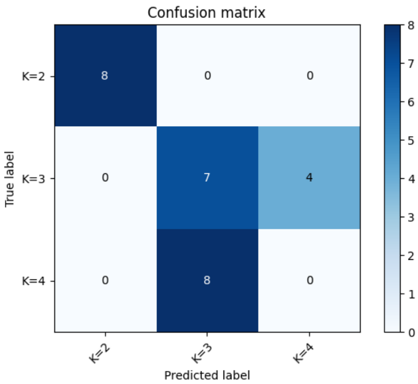


(d) Confusion matrix for RFE

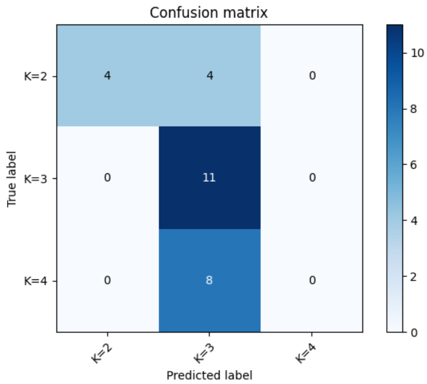


(e) Confusion matrix for Correlation

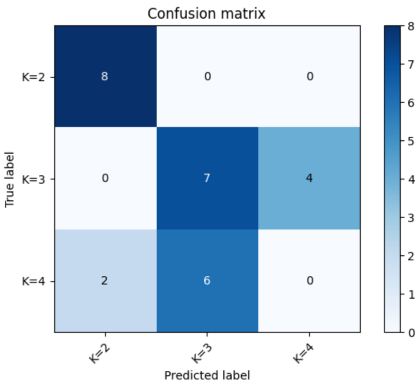
Figure B.3: Confusion matrices for multiclass classification using SVM with combined Dataset A.



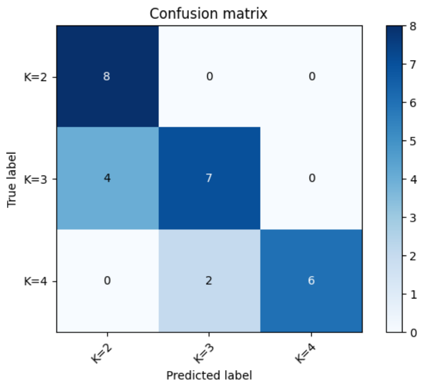
(a) Confusion matrix for the baseline model



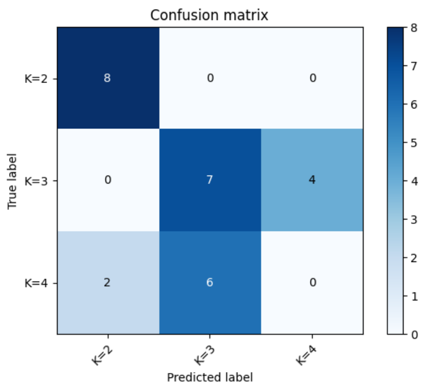
(b) Confusion matrix for PCA



(c) Confusion matrix for KBest

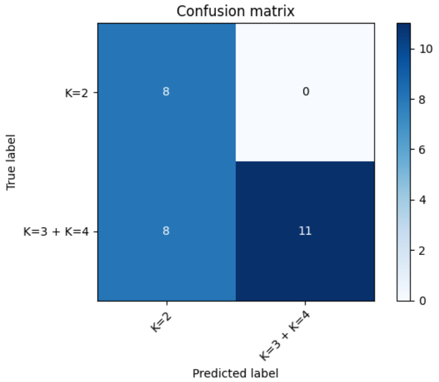


(d) Confusion matrix for RFE

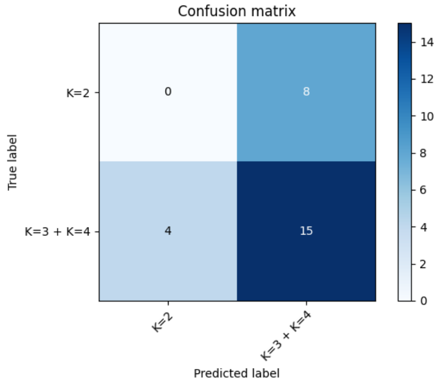


(e) Confusion matrix for Correlation

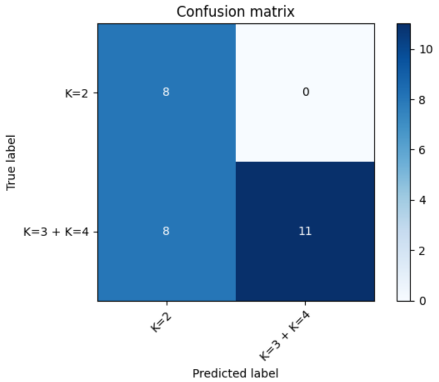
Figure B.4: Confusion matrices for multiclass classification using RF with combined Dataset A.



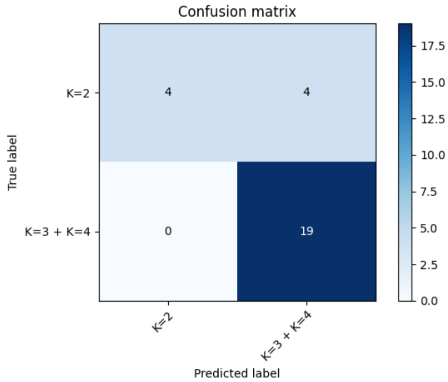
(a) Confusion matrix for the baseline model



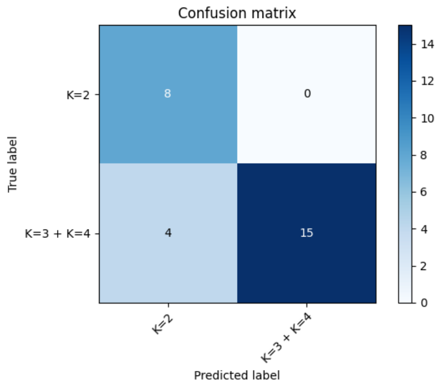
(b) Confusion matrix for PCA



(c) Confusion matrix for KBest



(d) Confusion matrix for RFE



(e) Confusion matrix for Correlation

Figure B.5: Confusion matrices for binary classification using SVM with activity metrics Dataset B.

Table B.6: Performance metrics of binary classification using RF with activity metrics Dataset B.

Feature Reduction and Selection	F1-Score Validation	F1-Score Test	Accuracy Test	Balanced Accuracy
Baseline	0.82	0.33	0.48	0.34
PCA	0.72	0.39	0.63	0.45
KB	0.80	0.39	0.63	0.45
RFE	0.82	0.39	0.63	0.45
Corr	0.78	0.41	0.70	0.50

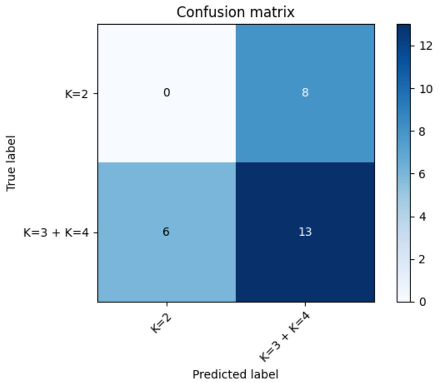
B.2.2 Approach 2: Combined Dataset B

Table B.7: Performance metrics of binary classification using SVM with combined Dataset B.

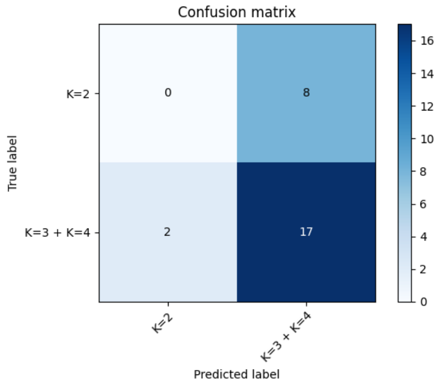
Feature Reduction and Selection	F1-Score Validation	F1-Score Test	Accuracy Test	Balanced Accuracy
Baseline	0.79	1.0	1.0	1.0
PCA	0.80	1.0	1.0	1.0
KB	0.80	0.41	0.70	0.50
RFE	0.80	0.84	0.85	0.89
Corr	0.81	0.41	0.70	0.50

Table B.8: Performance metrics of binary classification using RF with combined Dataset B.

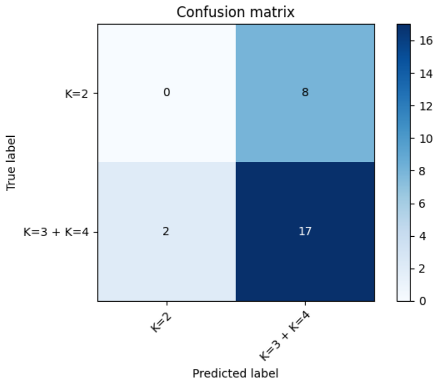
Feature Reduction and Selection	F1-Score Validation	F1-Score Test	Accuracy Test	Balanced Accuracy
Baseline	0.80	0.41	0.70	0.50
PCA	0.72	0.79	0.85	0.75
KB	0.81	0.41	0.70	0.50
RFE	0.81	0.41	0.70	0.50
Corr	0.80	0.41	0.70	0.50



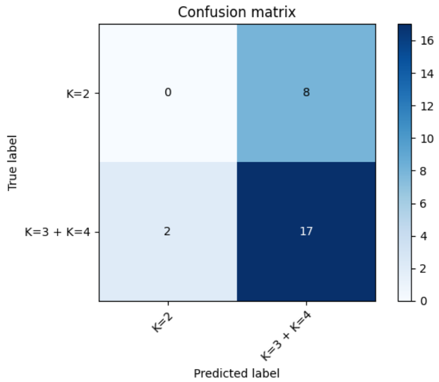
(a) Confusion matrix for the baseline model



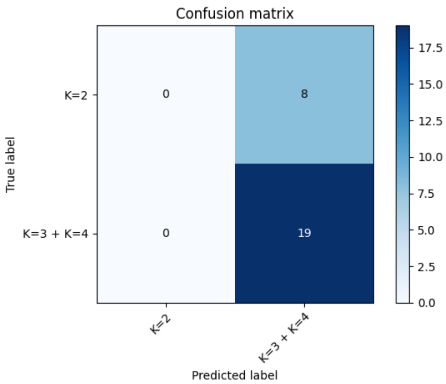
(b) Confusion matrix for PCA



(c) Confusion matrix for KBest

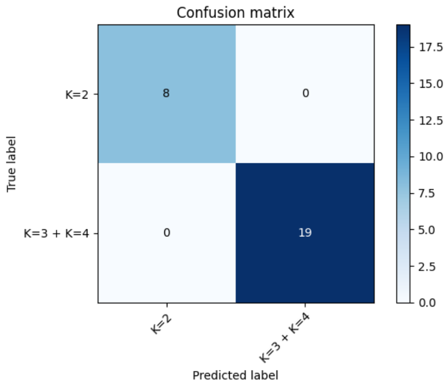


(d) Confusion matrix for RFE

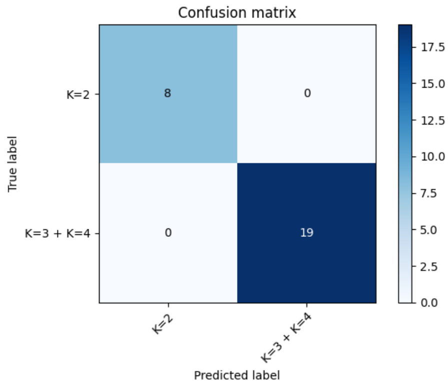


(e) Confusion matrix for Correlation

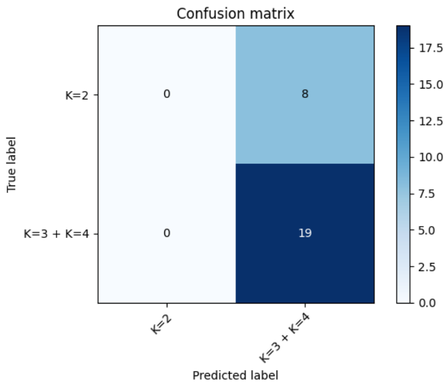
Figure B.6: Confusion matrices for binary classification using RF with activity metrics Dataset B.



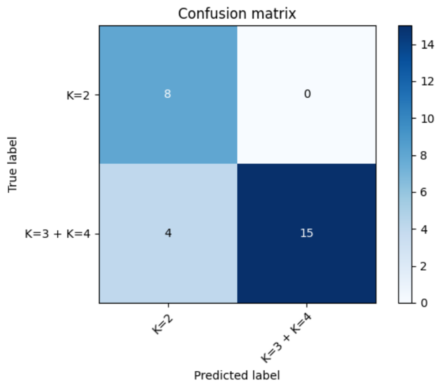
(a) Confusion matrix for the baseline model



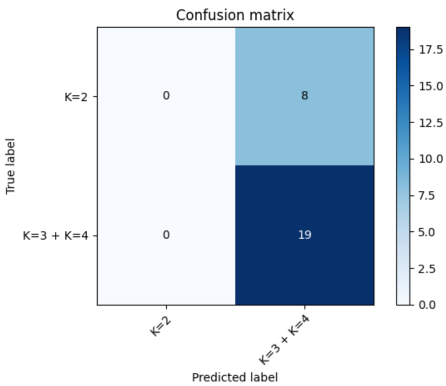
(b) Confusion matrix for PCA



(c) Confusion matrix for KBest

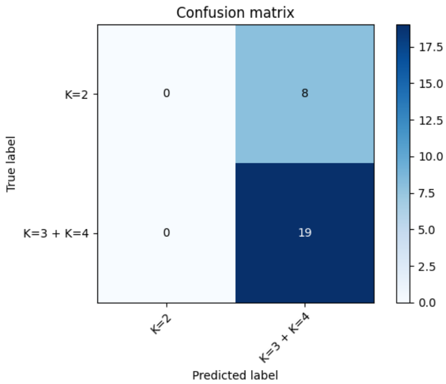


(d) Confusion matrix for RFE

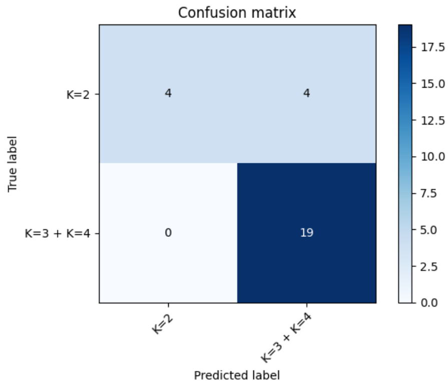


(e) Confusion matrix for Correlation

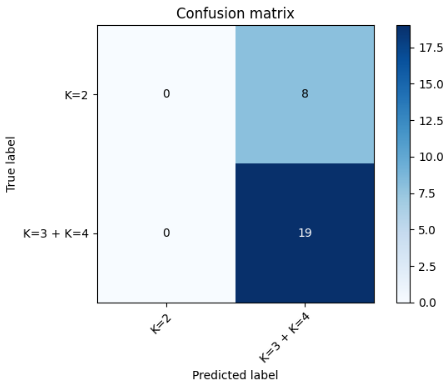
Figure B.7: Confusion matrices for binary classification using SVM with combined Dataset B.



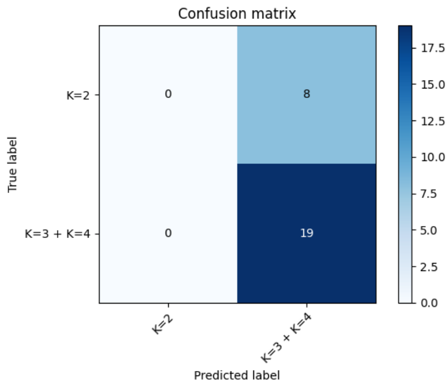
(a) Confusion matrix for the baseline model



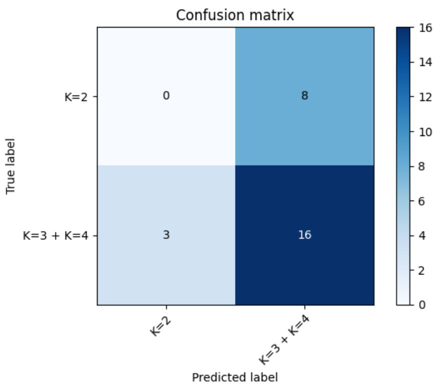
(b) Confusion matrix for PCA



(c) Confusion matrix for KBest



(d) Confusion matrix for RFE



(e) Confusion matrix for Correlation

Figure B.8: Confusion matrices for binary classification using RF with combined Dataset B.