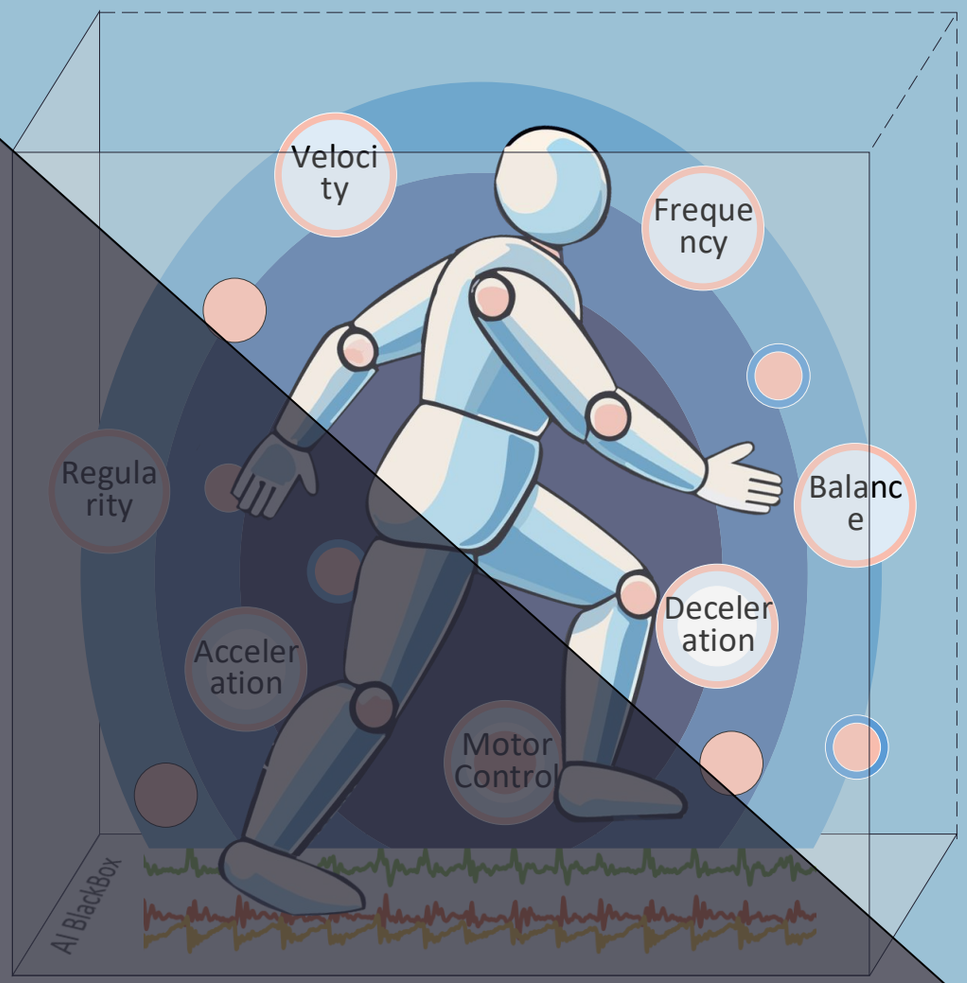
XAI

Explainable Artificial Intelligence in Movement Sciences

Focusing on gait analysis in healthy older adults and patients with back pain

Xiaoping Zheng



Research Institute SHARE

Xiaoping Zheng
Explainable Artificial Intelligence in Movement Sciences

Propositions belonging to the thesis

Explainable Artificial Intelligence in Movement Sciences

Focusing on gait analysis in healthy older adults and patients with back pain

1. Deep learning surpasses traditional machine learning in the classification of age-related gait patterns using raw acceleration data. *(this thesis)*

2. Explainable artificial intelligence can enhance the understanding of the insights gained from machine learning models in gait analysis. *(this thesis)*

3. Changes in acceleration and deceleration patterns in gait due to aging can be recognized by machine learning. *(this thesis)*

4. The presence of central sensitization in patients with chronic low back pain may be associated with alterations in gait and physical activity patterns. *(this thesis)*

5. Variations in gait and physical activity patterns could reflect changes in motor control and pain response strategies in patients with chronic low back pain and central sensitization. *(this thesis)*

6. Due to the heterogeneity within the CLBP population, a tailored rehabilitation program maybe essential for effective treatment. *(this thesis)*

7. While deep learning offers superior learning capabilities, this comes at the cost of increased model complexity and reduced transparency. *(this thesis)*

8. Choosing the appropriate machine learning models for human movement analysis requires a trade-off between accuracy and explainability. *(this thesis)*

9. “知其然, 知其所以然” (Translated and simplified into English: Know what it is, know why it is) – attributed to **Zhu Xi**, Southern Song Dynasty.

Explainable Artificial Intelligence in Movement Sciences

Focusing on gait analysis in healthy older adults and patients with back pain

Xiaoping Zheng

The experiments described in Chapters 2 and 3 were conducted at the Medical Ethical Committee of the Slotervaart Hospital (closed in 2018), the Netherlands and the University Medical Center Groningen, Groningen, the Netherlands. The experiments described in Chapters 4, 5 and 6 were conducted at the University Medical Center Groningen.

PhD training was facilitated by the research institute SHARE, part of the Graduate School of Medical Sciences Groningen.

The printing of this thesis was financially supported by the University of Groningen, University Medical Center Groningen and Research Institute SHARE.

Paranymphs: **Jorine Schoenmaker**
Jelmer Braaksma

Cover image: **The dummy was generated by ChatGPT (OpenAI).**
Cover design: **Xiaoping Zheng**
Layout and design: **Xiaoping Zheng**
Printing: **Proefschrift AIO**

Explainable Artificial Intelligence in Movement Sciences: Focusing on gait analysis in healthy older adults and patients with back pain – Xiaoping Zheng

university of groningen

Explainable Artificial Intelligence in Movement Sciences

Focusing on gait analysis in healthy older adults and patients with back pain

PhD thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus Prof. J.M.A. Scherpen
and in accordance with
the decision by the College of Deans.

This thesis will be defended in public on

Monday 29 April 2024 at 12.45 hours

by

Xiaoping Zheng

born on 8 April 1993

Supervisors

Prof. C.J.C. Lamoth
Prof. M.F. Reneman
Prof. E. Otten

Assessment Committee

Prof. N.M. Maurits
Prof. J. van Dieen
Prof. N. Strodthoff

Table of contents

Chapter 1

General Introduction

Human Movement Sciences and Gait

Human Movement Sciences represents a dynamic, multidisciplinary field dedicated to unraveling the intricate relationship between physical activity, aging, and various health conditions. Gait, the way in which individuals walk, is a fundamental component of daily physical activity. It serves as a window into the delicate control and cooperation of various systems within the human body, including the musculoskeletal, cardiorespiratory, and nervous systems. Outcomes of gait analysis have evolved into biomarkers, disclosing crucial insights into understanding age-related mobility decline and back pain related motor impairments [1, 2]. Gait analysis can be used to diagnose gait abnormalities, inform treatment strategies, guided rehabilitation regimens, and measured the effectiveness of interventions [3, 4]. In broader applications, gait analysis extends its impact by contributing to the understanding of the health among aging populations and monitoring their overall well-being [5]. In summary, Human Movement Sciences, with its focus on gait analysis, offers invaluable insights into the relationships between human movement and health. Through the lens of gait, it paves the way for advanced healthcare, rehabilitation, and health promotion.

Gait Analysis

After decades of evolution, gait analysis has emerged as a critical clinical tool for many medical and healthcare applications [6]. In the past, the focus has been on gait speed. Studies have demonstrated that, in older adults (aged over 65), gait speed is a prime predictor of mortality [7, 8]. Even an 0.1 m/s difference in speed corresponds to statistically significant changes in the expected remaining years of life [8]. Therefore, the accurate[1] estimation of gait outcomes, such as gait speed, is essential.

In current clinical settings, gait analysis is typically conducted using subjective and qualitative approaches. For instance, experts such as well-trained clinicians, can visually evaluate the gait performance of patients and assess their gait disorders. Through the observer-based timing, some qualitative gait outcomes, such as gait speed can be obtained. But the gait speed obtained via this way may lack precision[2].

To achieve more accurate estimations of gait outcomes, standard gait analysis tools such as optical motion capture systems (e.g., Vicon) are utilized in some specialized centers and clinics. Although these systems can provide highly accurate gait outcomes based on high-precision data, they are relatively costly, and involve an intrusive marker setup procedure that is time-

[1] Accuracy refers to the “closeness of agreement between a measured quantity value and the true quantity value of the measurand” [9]

[2] Precision refers to the “closeness of agreement between indications and measured quantity values obtained by replicated measurements on the same or similar objects under specified conditions” [9].

consuming and may interfere with the patient's natural movement [10]. Moreover, these systems are limited to laboratory environments for short-term walking assessment and may not accurately reflect gait in real-world settings [11]. Therefore, these systems are not pervasive enough among clinics.

Advances in wearable sensor technologies [12], particularly inertial measurement units (IMUs) equipped with accelerometers, gyroscopes, and/or magnetometers, have shown significant potential as a replacement. These devices are cost-effective, portable (non-intrusive), and capable of recording precise time-series sensor data (spanning several days or weeks) for gait analysis. Furthermore, they can be used in both clinical settings and daily living environments, and better reflect gait patterns in real-world scenarios.

Based on time-series data, accurate spatial-temporal gait outcomes such as gait speed, step length, and step width can be calculated. Additionally, based on the time information from the sensor data, dynamic gait outcomes have been explored. These gait outcomes may provide insights into how gait evolves over time [13]. Key aspects of these outcomes include gait regularity (e.g., stride regularity and gait symmetry index calculated using autocorrelation coefficients) [14], smoothness (e.g., the index of harmonicity and harmonic ratio derived from power spectral signal frequency analysis) [15], predictability (e.g., sample entropy based on information theory) [16], and stability (e.g., maximal Lyapunov exponent rooted in chaos theory) [17]. Based on these gait outcomes, different populations can be characterized.

Gait Analysis in Aging

Gait speed has been utilized as an indicator of survival rates and life expectancy in older adults and for evaluating geriatric status [7, 8]. However, gait speed alone cannot capture all the changes attributable to aging and disorders. Thus, a broader range of gait outcomes should be included for a comprehensive assessment of aging. It is observed that older adults exhibit a reduced gait speed, shorter step lengths, wider step widths, and decreased gait symmetry and regularity compared to adults [18]. Moreover, gait analysis has revealed that fallers exhibit a lower gait speed [19], lower smoothness (as indicated by harmonic ratio) [20], and lower local stability (measured by Lyapunov exponent) [21] in gait, in contrast to non-fallers.

The growing number of older adults highlights the need for monitoring gait degradation, assessing fall risks, and preventing falls for the older adults [22, 23]. To conserve limited medical resources, automatic gait analysis is becoming more and more important [24]. However, gait outcomes are related to each other in a linear way (e.g., gait speed and stride length) or interacted in non-linear ways (e.g., gait speed and gait smoothness) [25]. To handle the complex interplay within gait outcomes and accurately classify gait patterns, alternative statistical approaches may be necessary, differing from traditional approaches.

Conventional Machine Learning in Gait Analysis

Artificial intelligence (AI), including conventional machine learning (CML), is able to take the linear relationship and non-linear interaction into account [26]. Hence, it can automatically learn the gait patterns from gait outcomes and may offer valuable insights into interrelationships and interactions within gait outcomes [27]. CML has been successfully applied to classify different gait patterns related to aging and pathology [28]. For example, artificial neural networks have been used to classify age-related gait patterns with an impressive 90% accuracy [27]. Random forest has accurately distinguished fallers from non-fallers with an accuracy of 98% [29] and has supported the diagnosis of Parkinson's disease with 92.6% accuracy [30] from healthy controls.

However, CML-based gait classification depends on gait outcomes and the dependency may cause challenges of applications in clinical and daily living environments. Firstly, the used gait outcomes in CML-based gait analysis are manually designed and selected by experts, such as clinicians, rehabilitation experts, or human movement scientists. This process requires specialized knowledge and is prone to being labor-intensive [31]. Moreover, if some useful gait outcomes that can reflect the age-related changes are overlooked, it might decrease the performance of the gait classification. Secondly, to obtain stable and accurate gait outcomes such as the maximum Lyapunov exponent, certain requirements must be met, e.g., the specific signal length and sampling frequency [32]. Lastly, the computation of these handcrafted gait outcomes is often performed offline [33], meaning this approach may not be well-suited for real-time applications, such as falls detection in daily-living environments.

Deep Learning in Gait Analysis

In response to the challenges of CML-based gait analysis, deep learning (DL) has been introduced into the gait analysis. DL is an end-to-end approach [34] which means it can use time-series sensor data as input and provide predictions as output. The end-to-end character allows DL to eliminate the need for handcrafted gait outcomes, thereby avoiding the limitations of CML mentioned above. Furthermore, DL has shown superior performance over CML in various fields, such as computer vision [35] thanks to its superior learning capacity. The multi-layered structures of DL facilitate a hierarchical learning process; the initial layers extract fundamental features from sensory data, while subsequent layers progressively build upon these features to learn more abstract and high-level features [36]. This process could enable DL to learn the most suitable features for classification [31], potentially leading to more a more accurate gait classification.

Researchers have started to introduce DL into the field of gait analysis and highlight its promising performance in certain gait classification tasks [37, 38]. For instance, it has demonstrated the ability to classify fallers and non-fallers in adults with an AUC (Area under

the receiver operator curve) of 93.3% [39] and is able to classify gait of patients with Parkinson's disease with an accuracy of 89% [40]. Since aging is a continuous process and gait patterns will change gradually, classifying age-related gait patterns may be challenging. Given the growing older adult population, it is vital to explore optimal AI models for classifying age-related gait patterns. Therefore, **Chapter 2** of this thesis presented a comprehensive comparison in classification performance of CML and DL for classifying age-related gait patterns.

Black-Box Nature in AI

AI-driven gait analysis has shown promising performance, but its implementation in clinical practice is still in its infancy. Even in areas like medical imaging, where AI has already a strong tradition in research, and its performance is comparable with trained physicians [41, 42], the use of AI still suffers harsh criticisms [43]. One of the limitations for their clinical implementation arises from the black-box nature of AI models. Although the mathematical principles behind the models are well-established, the inner decision-making process of models is often opaque.

The black-box nature of AI-driven gait analysis results in a lack of interpretability and transparency, which poses challenges for patients and clinicians to trust the AI models [44]. Furthermore, this opacity may fail to meet legal requirements, such as the European General Data Protection Regulation (GDPR, EU 2016/679), which mandates transparent justifications for automated decision-making processes that have a significant impact on individuals [45]. Therefore, addressing the black-box issue in AI is crucial not only for building trust but also for ensuring ethical and legal compliance.

Explainable AI in Gait Analysis

To bridge this gap between AI and the need for interpretability and transparency, explainable AI (XAI) [46] has gained attention in the medical field. XAI focuses on revealing the underlying reasoning behind the predictions and decisions made by AI models. XAI can be broadly categorized into two main categories based on the usage stage: 1) ante-hoc explainability; and 2) post-hoc explainability [47]. Ante-hoc explainability refers to AI models that are interpretable by design, including simple CML models (such as linear regression, decision trees, and k-nearest neighbor), as well as a big part of unsupervised CML models (such as K-means, hierarchical clustering, and self-organization map) [48]. Post-hoc explainability approaches, such as SHapley Additive exPlanations (SHAP) [49], are employed to explain previously trained AI models and have thus attracted considerable interest.

Post-hoc explainability approaches have been employed in the realm of medical imaging. The explanations provided by XAI in the form of heat maps can visually highlight the salient or

important parts within an image that influence the model's decision-making process [41, 42, 50]. By comparing these salient or important parts with the assessments of medical experts, like radiologists [50], the trustworthiness of the AI model can be visually evaluated. Additionally, it is important to note that XAI explanations may not always align with domain expertise, potentially leading to discrepancies that offer fresh insights and contribute to the generation of new knowledge.

In the context of gait analysis, visualizing the important parts of time-series sensor data may not be intuitive, since it is difficult to correlate a segment of sensor signal with the domain expertise of movement scientists or clinicians. Consequently, based on the DL models discussed in chapter 2**,** the study described in **Chapter 3** of this thesis aimed to explore what AI models have learned from the accelerometer signal data for distinguishing gait patterns between adults and older adults by using post-hoc XAI, and to connect these findings with the domain expertise in movement sciences.

Gait in Back Pain

The knowledge and the methodologies of gait analysis obtained in Chapters 2 and 3 from older adults can also be applied to other populations, such as patients with back pain.

Low back pain (LBP) is the leading cause of disability [51], with the potential to progress into chronic low back pain (CLBP) when the pain persists beyond a 3-month duration. Notably, within the CLBP population, approximately 85% to 90% of patients are non-specific CLBP, as the link between known pathoanatomical factors and clinical presentations is absent [52, 53]. The CLBP population is heterogeneous [54], and the presence of central sensitization (CS) may contribute to this heterogeneity [55]. CS refers to an increased responsiveness of nociceptive neurons in the central nervous system to their normal or subthreshold afferent input [56]. Given the current limitations in directly measuring mechanisms related to CS in individual humans, the term “Human Assumed Central Sensitization” (HACS) has been introduced to describe CS. HACS is considered to play a role in the development and maintenance of CLBP [57].

Gait analysis has reported conflicting evidence in gait outcomes of patients with CLBP, including the preferred walking speed [58, 59], stride length [60, 61], and stride-to-stride variability [62, 63] when compared to healthy controls. Since movement may be changed due to pain, HACS may be a factor relating to the observed inconsistent gait patterns. To investigate the relationship between gait, CLBP, and HACS, AI-driven gait analysis could offer valuable insights. In **Chapter 4**, it was assumed that if HACS relates to changes in gait, AI-driven gait analysis could accurately classify patients with CLBP into different HACS-related groups. In this chapter, CML was selected for gait pattern classification instead of DL. Although

DL demonstrates superior performance, it comes with greater opacity [48]. Considering the trade-off between a model's performance and its transparency, CML was used. After the classification, a post-hoc XAI was employed to explain the differences in gait patterns among the HACS-related groups.

Physical Activity in Back Pain

CLBP poses substantial socioeconomic burdens and causes great individual suffering. In the Netherlands, the direct and indirect costs of back pain amount to around 0.6% to 0.9% of the gross national product [64]. Although the overall efficacy of CLBP rehabilitation programs is positive, the effect sizes are modest [65]. In the treatment of CLBP, physical exercise is often recommended [66]. However, the relationship between CLBP and physical activity is still unclear and inconsistent evidence has been shown. Some studies observed that people with CLBP exhibit lower overall physical activity intensity (PAI) during the day [67, 68], while others report no differences between patients with CLBP and healthy controls [69, 70].

Similar to Chapter 4, which assumed that gait alterations could be related to HACS, **Chapter 5** of this thesis proposed that inconsistencies in PAI evidence might also be associated with HACS. Therefore, the study described in this chapter aimed to explore PAI patterns of HACS-related subgroups in patients with CLBP using 24-hour accelerometer signal data. In this analysis, unsupervised CML with ante-hoc explainability [71] was utilized to explore and explain the differences in PAI patterns among HACS-related groups.

In Chapters 4 and 5, the HACS-related groups were determined by the central sensitization inventory (CSI) with a cut-off value of 40 [72]. However, it has been reported that the cut-off values for CSI may vary depending on different types of musculoskeletal pain [73, 74], as well as different cultural and national contexts [75]. It should be noted that the gold standard to assess HACS is currently unavailable and CSI is an indirect evaluation of the presence and severity of HACS [76]. Therefore, the study described in **Chapter 6** aimed to establish an optimal cut-off value for the Dutch-speaking population with CLBP. This chapter used unsupervised CML models with ante-hoc explainability to explore the HACS-related subgroups based on clinical outcomes, which include questionnaire data reflecting pain, physical functioning, psychological factors, and CSI values. Then, based on the found subgroups, an optimal CSI cut-off value could be established.

Aim and Outline

This thesis aims to enhance the comprehension of movement, especially gait, in healthy older adults and patients with back pain through insights from XAI. The architectural framework of the thesis is visually represented in Figure 1.

In **Chapter 2,** it was aimed to compare the performance of CML and DL in age-related gait pattern classification. Building upon the results of Chapter 2, the study described in **Chapter 3** aimed to enhance the interpretability and transparency of DL-based gait analysis in the classification of adults and older adults by using XAI. The insights gained from gait analysis in Chapters 2 and 3 could be applied to study the gait of patients with CLBP. **Chapter 4** focused on using CML to classify HACS-related subgroups based on gait outcomes collected in daily living environments, with XAI facilitating the explanation of differences in gait patterns among these subgroups. Extending the analysis beyond gait, the study described in **Chapter 5** aimed to explore and explain PAI patterns of HACS-related subgroups of patients with CLBP based on ante-hoc XAI. Additionally, the study described in **Chapter 6** aimed to establish an optimal CSI cut-off value for the Dutch-speaking population with CLBP by investigating HACS-related subgroups through clinical outcomes by using unsupervised CML with ante-hoc XAI. Lastly, **Chapter 7** provided a comprehensive discussion and conclusion, summarizing the findings and implications of this thesis.

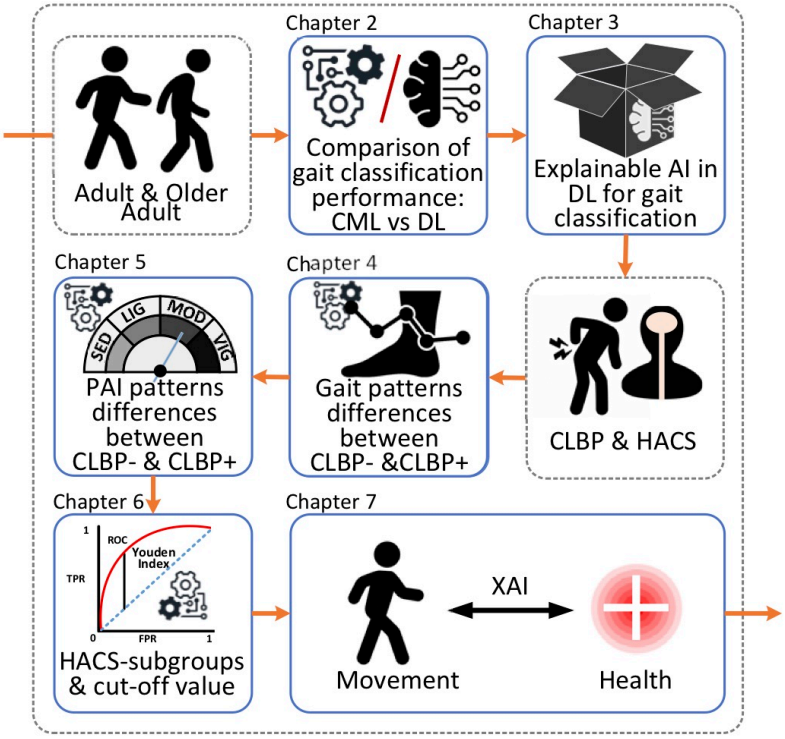


Figure 1. The architecture of this thesis. CML: conventional machine learning; DL: deep learning; AI: artificial intelligence; CLBP: chronic low back pain; HACS: human assumed central sensitization; CLBP+: chronic low back pain with high central sensitization; CLBP-: chronic low back pain with low central sensitization; PAI: physical activity intensity; XAI: explainable artificial intelligence.

Reference

[1] F. B. Horak, and M. Mancini, "Objective biomarkers of balance and gait for Parkinson's disease using body-worn sensors," *Movement Disorders,* vol. 28, no. 11, pp. 1544-1551, 2013.

[2] M. A. Brodie, N. H. Lovell, C. G. Canning, H. B. Menz, K. Delbaere, S. J. Redmond, M. Latt, D. L. Sturnieks, J. Menant, and S. T. Smith, "Gait as a biomarker? Accelerometers reveal that reduced movement quality while walking is associated with Parkinson's disease, ageing and fall risk." pp. 5968-5971.

[3] B. Debû, C. De Oliveira Godeiro, J. C. Lino, and E. Moro, “Managing gait, balance, and posture in Parkinson’s disease,” *Current neurology and neuroscience reports,* vol. 18, pp. 1-12, 2018.

[4] R. Baker, “Gait analysis methods in rehabilitation,” *Journal of neuroengineering and rehabilitation,* vol. 3, pp. 1-10, 2006.

[5] A. Godfrey, “Wearables for independent living in older adults: Gait and falls,” *Maturitas,* vol. 100, pp. 16-26, 2017.

[6] M. W. Whittle, “Clinical gait analysis: A review,” *Human movement science,* vol. 15, no. 3, pp. 369-387, 1996.

[7] N. M. Peel, S. S. Kuys, and K. Klein, “Gait speed as a measure in geriatric assessment in clinical settings: a systematic review,” *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences,* vol. 68, no. 1, pp. 39-46, 2013.

[8] S. Studenski, S. Perera, K. Patel, C. Rosano, K. Faulkner, M. Inzitari, J. Brach, J. Chandler, P. Cawthon, and E. B. Connor, “Gait speed and survival in older adults,” *Jama,* vol. 305, no. 1, pp. 50-58, 2011.

[9] I. Vim, “International vocabulary of basic and general terms in metrology (VIM),” *International Organization,* vol. 2004, pp. 09-14, 2004.

[10] S. R. Simon, “Quantification of human motion: gait analysis—benefits and limitations to its application to clinical problems,” *Journal of biomechanics,* vol. 37, no. 12, pp. 1869-1880, 2004.

[11] A. Ali, K. Sundaraj, B. Ahmad, N. Ahamed, and A. Islam, “Gait disorder rehabilitation using vision and non-vision based sensors: a systematic review,” *Bosnian journal of basic medical sciences,* vol. 12, no. 3, pp. 193, 2012.

[12] A. Muro-De-La-Herran, B. Garcia-Zapirain, and A. Mendez-Zorrilla, “Gait analysis methods: An overview of wearable and non-wearable systems, highlighting clinical applications,” *Sensors,* vol. 14, no. 2, pp. 3362-3394, 2014.

[13] J. Gervais-Hupé, J. Pollice, J. Sadi, and L. C. Carlesso, “Validity of the central sensitization inventory with measures of sensitization in people with knee osteoarthritis,” *Clinical rheumatology,* vol. 37, pp. 3125-3132, 2018.

[14] R. Moe-Nilssen, and J. L. Helbostad, “Estimation of gait cycle characteristics by trunk accelerometry,” *Journal of biomechanics,* vol. 37, no. 1, pp. 121-126, 2004.

[15] J. Bellanca, K. Lowry, J. Vanswearingen, J. Brach, and M. Redfern, “Harmonic ratios: a quantification of step to step symmetry,” *Journal of biomechanics,* vol. 46, no. 4, pp. 828-831, 2013.

[16] J. M. Yentes, and P. C. Raffalt, “Entropy analysis in gait research: methodological considerations and recommendations,” *Annals of biomedical engineering,* vol. 49, pp. 979-990, 2021.

[17] J. B. Dingwell, and H. G. Kang, “Differences between local and orbital dynamic stability during human walking,” 2007.

[18] A. Aboutorabi, M. Arazpour, M. Bahramizadeh, S. W. Hutchins, and R. Fadayevatan, “The effect of aging on gait parameters in able-bodied older subjects: a literature review,” *Aging clinical and experimental research,* vol. 28, no. 3, pp. 393-405, 2016.

[19] J. M. VanSwearingen, K. A. Paschal, P. Bonino, and T.-W. Chen, “Assessing recurrent fall risk of community-dwelling, frail older veterans using specific tests of mobility and the physical performance test of function,” *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences,* vol. 53, no. 6, pp. M457-M464, 1998.

[20] T. Doi, S. Hirata, R. Ono, K. Tsutsumimoto, S. Misu, and H. Ando, “The harmonic ratio of trunk acceleration predicts falling among older people: results of a 1-year prospective study,” *Journal of neuroengineering and rehabilitation,* vol. 10, pp. 1-6, 2013.

[21] M. J. Toebes, M. J. Hoozemans, R. Furrer, J. Dekker, and J. H. van Dieën, “Local dynamic stability and variability of gait are associated with fall history in elderly subjects,” *Gait & posture,* vol. 36, no. 3, pp. 527-531, 2012.

[22] E. Kanasi, S. Ayilavarapu, and J. Jones, “The aging population: demographics and the biology of aging,” *Periodontology 2000,* vol. 72, no. 1, pp. 13-18, 2016.

[23] F. Sun, W. Zang, R. Gravina, G. Fortino, and Y. Li, “Gait-based identification for elderly users in wearable healthcare systems,” *Information fusion,* vol. 53, pp. 134-144, 2020.

[24] S. Majumder, E. Aghayi, M. Noferesti, H. Memarzadeh-Tehran, T. Mondal, Z. Pang, and M. J. Deen, “Smart homes for elderly healthcare—Recent advances and research challenges,” *Sensors,* vol. 17, no. 11, pp. 2496, 2017.

[25] I. Hagoort, N. Vuillerme, T. Hortobágyi, and C. J. Lamoth, “Outcome-dependent effects of walking speed and age on quantitative and qualitative gait measures,” *Gait & Posture,* vol. 93, pp. 39-46, 2022.

[26] C. Prakash, R. Kumar, and N. Mittal, “Recent developments in human gait research: parameters, approaches, applications, machine learning techniques, datasets and challenges,” *Artificial Intelligence Review,* vol. 49, pp. 1-40, 2018.

[27] Y. Zhou, R. Romijnders, C. Hansen, J. van Campen, W. Maetzler, T. Hortobagyi, and C. J. C. Lamoth, “The detection of age groups by dynamic gait outcomes using machine learning approaches,” *Scientific Reports,* vol. 10, no. 1, Mar, 2020.

[28] P. Khera, and N. Kumar, “Role of machine learning in gait analysis: a review,” *Journal of Medical Engineering & Technology,* vol. 44, no. 8, pp. 441-467, 2020.

[29] R. Z. U. Rehman, Y. Zhou, S. Del Din, L. Alcock, C. Hansen, Y. Guan, T. Hortobágyi, W. Maetzler, L. Rochester, and C. J. Lamoth, “Gait analysis with wearables can accurately

classify fallers from non-fallers: a step toward better management of neurological disorders," *Sensors,* vol. 20, no. 23, pp. 6992, 2020.

[30] F. Wahid, R. K. Begg, C. J. Hass, S. Halgamuge, and D. C. Ackland, "Classification of Parkinson's disease gait using spatial-temporal gait features," *IEEE journal of biomedical and health informatics,* vol. 19, no. 6, pp. 1794-1802, 2015.

[31] F. Gu, M.-H. Chung, M. Chignell, S. Valaee, B. Zhou, and X. Liu, "A survey on deep learning for human activity recognition," *ACM Computing Surveys (CSUR),* vol. 54, no. 8, pp. 1-34, 2021.

[32] S. M. Bruijn, J. H. van Dieën, O. G. Meijer, and P. J. Beek, "Statistical precision and sensitivity of measures of dynamic gait stability," *Journal of neuroscience methods,* vol. 178, no. 2, pp. 327-333, 2009.

[33] Y. Hutabarat, D. Owaki, and M. Hayashibe, "Recent advances in quantitative gait analysis using wearable sensors: a review," *IEEE Sensors Journal*, 2021.

[34] R. Delgado-Escano, F. M. Castro, J. R. Cózar, M. J. Marín-Jiménez, and N. Guil, "An end-to-end multi-task and fusion CNN for inertial-based gait recognition," *IEEE Access,* vol. 7, pp. 1897-1908, 2018.

[35] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "Imagenet large scale visual recognition challenge," *International journal of computer vision,* vol. 115, pp. 211-252, 2015.

[36] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature,* vol. 521, no. 7553, pp. 436-444, 2015.

[37] Y. Matsushita, D. T. Tran, H. Yamazoe, and J.-H. Lee, "Recent use of deep learning techniques in clinical applications based on gait: a survey," *Journal of Computational Design and Engineering,* vol. 8, no. 6, pp. 1499-1532, 2021.

[38] C. Filipi Gonçalves dos Santos, D. d. S. Oliveira, L. A. Passos, R. Gonçalves Pires, D. Felipe Silva Santos, L. Pascotti Valem, T. P. Moreira, M. Cleison S. Santana, M. Roder, and J. Paulo Papa, "Gait recognition based on deep learning: A survey," *ACM Computing Surveys (CSUR),* vol. 55, no. 2, pp. 1-34, 2022.

[39] M. Martinez, and P. L. De Leon, "Falls risk classification of older adults using deep neural networks and transfer learning," *IEEE journal of biomedical and health informatics,* vol. 24, no. 1, pp. 144-150, 2019.

[40] J. Camps, A. Sama, M. Martin, D. Rodriguez-Martin, C. Perez-Lopez, J. M. M. Arostegui, J. Cabestany, A. Catala, S. Alcaine, and B. Mestre, "Deep learning for freezing of gait detection in Parkinson's disease patients in their homes using a waist-worn inertial measurement unit," *Knowledge-Based Systems,* vol. 139, pp. 119-131, 2018.

[41] R. Poplin, A. V. Varadarajan, K. Blumer, Y. Liu, M. V. McConnell, G. S. Corrado, L. Peng, and D. R. Webster, "Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning," *Nature biomedical engineering,* vol. 2, no. 3, pp. 158-164, 2018.

[42] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, and J. Cuadros, “Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs,” *jama,* vol. 316, no. 22, pp. 2402-2410, 2016.

[43] J.-G. Lee, S. Jun, Y.-W. Cho, H. Lee, G. B. Kim, J. B. Seo, and N. Kim, “Deep learning in medical imaging: general overview,” *Korean journal of radiology,* vol. 18, no. 4, pp. 570-584, 2017.

[44] W. J. von Eschenbach, “Transparency and the black box problem: Why we do not trust AI,” *Philosophy & Technology,* vol. 34, no. 4, pp. 1607-1622, 2021.

[45] P. Regulation, “Regulation (EU) 2016/679 of the European Parliament and of the Council,” *Regulation (eu),* vol. 679, pp. 2016, 2016.

[46] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu, "Explainable AI: A brief survey on history, research areas, approaches and challenges." pp. 563-574.

[47] T. Speith, "A review of taxonomies of explainable artificial intelligence (XAI) methods." pp. 2239-2250.

[48] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, and R. Benjamins, “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Information fusion,* vol. 58, pp. 82-115, 2020.

[49] K. Aas, M. Jullum, and A. Løland, “Explaining individual predictions when features are dependent: More accurate approximations to Shapley values,” *Artificial Intelligence,* vol. 298, pp. 103502, 2021.

[50] K. Borys, Y. A. Schmitt, M. Nauta, C. Seifert, N. Krämer, C. M. Friedrich, and F. Nensa, “Explainable AI in Medical Imaging: An overview for clinical practitioners–Saliency-based XAI approaches,” *European journal of radiology*, pp. 110787, 2023.

[51] D. Hoy, L. March, P. Brooks, F. Blyth, A. Woolf, C. Bain, G. Williams, E. Smith, T. Vos, and J. Barendregt, “The global burden of low back pain: estimates from the Global Burden of Disease 2010 study,” *Annals of the rheumatic diseases,* vol. 73, no. 6, pp. 968-974, 2014.

[52] O. Airaksinen, J. I. Brox, C. Cedraschi, J. Hildebrandt, J. Klaber-Moffett, F. Kovacs, A. F. Mannion, S. Reis, J. Staal, and H. Ursin, “European guidelines for the management of chronic nonspecific low back pain,” *European spine journal,* vol. 15, no. Suppl 2, pp. s192, 2006.

[53] J. Hartvigsen, M. J. Hancock, A. Kongsted, Q. Louw, M. L. Ferreira, S. Genevay, D. Hoy, J. Karppinen, G. Pransky, J. Sieper, R. J. Smeets, M. Underwood, and W. Lancet Low Back Pain Series, “What low back pain is and why we need to pay attention,” *Lancet,* vol. 391, no. 10137, pp. 2356-2367, Jun 9, 2018.

[54] D. R. Fourney, G. Andersson, P. M. Arnold, J. Dettori, A. Cahana, M. G. Fehlings, D. Norvell, D. Samartzis, and J. R. Chapman, “Chronic low back pain: a heterogeneous

condition with challenges for an evidence-based approach," *Spine,* vol. 36, pp. S1-S9, 2011.

[55] J. A. Echeita, H. R. S. Preuper, R. Dekker, I. Stuive, H. Timmerman, A. P. Wolff, and M. F. Reneman, "Central Sensitisation and functioning in patients with chronic low back pain: protocol for a cross-sectional and cohort study," *Bmj Open,* vol. 10, no. 3, Mar, 2020.

[56] J. D. Loeser, and R.-D. Treede, "The Kyoto protocol of IASP basic pain Terminology☆," *Pain,* vol. 137, no. 3, pp. 473-477, 2008.

[57] C. J. Woolf, "Central sensitization: implications for the diagnosis and treatment of pain," *pain,* vol. 152, no. 3, pp. S2-S15, 2011.

[58] C. J. Lamoth, O. G. Meijer, A. Daffertshofer, P. I. Wuisman, and P. J. Beek, "Effects of chronic low back pain on trunk coordination and back muscle activity during walking: changes in motor control," *European Spine Journal,* vol. 15, pp. 23-40, 2006.

[59] G. Christe, F. Kade, B. M. Jolles, and J. Favre, "Chronic low back pain patients walk with locally altered spinal kinematics," *Journal of biomechanics,* vol. 60, pp. 211-218, 2017.

[60] C. J. Lamoth, J. F. Stins, M. Pont, F. Kerckhoff, and P. J. Beek, "Effects of attention on the control of locomotion in individuals with chronic low back pain," *Journal of neuroengineering and rehabilitation,* vol. 5, no. 1, pp. 1-8, 2008.

[61] S. P. Gombatto, T. Brock, A. DeLork, G. Jones, E. Madden, and C. Rinere, "Lumbar spine kinematics during walking in people with and people without low back pain," *Gait & posture,* vol. 42, no. 4, pp. 539-544, 2015.

[62] D. Hamacher, D. Hamacher, F. Herold, and L. Schega, "Are there differences in the dual-task walking variability of minimum toe clearance in chronic low back pain patients and healthy controls?," *Gait & Posture,* vol. 49, pp. 97-101, Sep, 2016.

[63] R. Müller, T. Ertelt, and R. Blickhan, "Low back pain affects trunk as well as lower limb movements during walking and running," *Journal of biomechanics,* vol. 48, no. 6, pp. 1009-1014, 2015.

[64] L. C. Lambeek, M. W. van Tulder, I. C. Swinkels, L. L. Koppes, J. R. Anema, and W. van Mechelen, "The trend in total cost of back pain in The Netherlands in the period 2002 to 2007," *Spine,* vol. 36, no. 13, pp. 1050-1058, 2011.

[65] N. E. Foster, J. R. Anema, D. Cherkin, R. Chou, S. P. Cohen, D. P. Gross, P. H. Ferreira, J. M. Fritz, B. W. Koes, W. Peul, J. A. Turner, C. G. Maher, and W. Lancet Low Back Pain Series, "Prevention and treatment of low back pain: evidence, challenges, and promising directions," *Lancet,* vol. 391, no. 10137, pp. 2368-2383, Jun 9, 2018.

[66] M. F. Reneman, J. A. Echeita, K. van Kammen, H. R. S. Preuper, R. Dekker, and C. J. Lamoth, "Do rehabilitation patients with chronic low back pain meet World Health Organisation's recommended physical activity levels?," *Musculoskeletal Science and Practice,* vol. 62, pp. 102618, 2022.

[67] M. Soysal, B. Kara, and M. N. Arda, “Assessment of physical activity in patients with chronic low back or neck pain,” *Turk Neurosurg,* vol. 23, no. 1, pp. 75-80, 2013.

[68] C. G. Ryan, P. M. Grant, P. M. Dall, H. Gray, M. Newton, and M. H. Granat, “Individuals with chronic low back pain have a lower level, and an altered pattern, of physical activity compared with matched controls: an observational study,” *Australian Journal of Physiotherapy,* vol. 55, no. 1, pp. 53-58, 2009.

[69] J. A. Verbunt, K. R. Westerterp, G. J. van der Heijden, H. A. Seelen, J. W. Vlaeyen, and J. A. Knottnerus, “Physical activity in daily life in patients with chronic low back pain,” *Arch Phys Med Rehabil,* vol. 82, no. 6, pp. 726-30, Jun, 2001.

[70] M. G. van Weering, M. M. Vollenbroek-Hutten, T. M. Tonis, and H. J. Hermens, “Daily physical activities in chronic lower back pain patients assessed with accelerometry,” *Eur J Pain,* vol. 13, no. 6, pp. 649-54, Jul, 2009.

[71] D. van Kuppevelt, J. Heywood, M. Hamer, S. Sabia, E. Fitzsimons, and V. J. P. o. van Hees, “Segmenting accelerometer data from daily life with unsupervised machine learning,” vol. 14, no. 1, pp. e0208692, 2019.

[72] E. E. Bennett, K. M. Walsh, N. R. Thompson, and A. A. Krishnaney, “Central sensitization inventory as a predictor of worse quality of life measures and increased length of stay following spinal fusion,” *World neurosurgery,* vol. 104, pp. 594-600, 2017.

[73] R. Neblett, H. Cohen, Y. Choi, M. M. Hartzell, M. Williams, T. G. Mayer, and R. J. Gatchel, “The Central Sensitization Inventory (CSI): establishing clinically significant values for identifying central sensitivity syndromes in an outpatient chronic pain sample,” *J Pain,* vol. 14, no. 5, pp. 438-45, May, 2013.

[74] A. Mibu, T. Nishigami, K. Tanaka, M. Manfuku, and S. Yono, “Difference in the impact of central sensitization on pain-related symptoms between patients with chronic low back pain and knee osteoarthritis,” *Journal of Pain Research,* vol. 12, pp. 1757, 2019.

[75] R. Neblett, “The central sensitization inventory: A user's manual,” *Journal of Applied Biobehavioral Research,* vol. 23, no. 2, Jun, 2018.

[76] T. G. Mayer, R. Neblett, H. Cohen, K. J. Howard, Y. H. Choi, M. J. Williams, Y. Perez, and R. J. Gatchel, “The development and psychometric validation of the central sensitization inventory,” *Pain Practice,* vol. 12, no. 4, pp. 276-285, 2012.

Chapter 2

Age-related Gait Patterns Classification Using Deep Learning Based on Time-series Data from One Accelerometer

Xiaoping Zheng, Elisabeth Wilhelm, Egbert Otten, Michiel F Reneman, Claudine JC Lamoth

Submitted for Publication

Abstract

Gait pattern classification is important for healthcare. Conventional machine learning (ML) approaches based on handcrafted gait features are widely used in gait classification. However, extracting features may lead to suboptimal performance by omitting useful features. End-to-end deep learning (DL) approaches eliminate the need for feature extraction. However, some state-of-the-art DL approaches have not been explored in gait analysis. Furthermore, no consensus exists regarding the window sizes of input acceleration, which affects classification accuracy. In this study, data were collected from one accelerometer during a 3-minute indoor walking task. 267 subjects were divided into adults (18-65 years) and older adults (>65) groups. To explore age-related gait patterns classification performance, 5 DL approaches based on raw data and 4 conventional ML approaches based on handcrafted features were compared. The results show that DL outperformed conventional ML, with all AUC greater than 0.94 compared to the best conventional ML approach of 0.83. This suggests that DL may have learned important gait features related to aging that have not yet been identified by previous research. Furthermore, windows of different sizes ranging from 128 to 5120 samples were tested. The best performance of DL was achieved for a window size of 1024 (including about 20 steps). These findings indicate that the differences and relationship between gait cycles are important factors for classifying age-related gait patterns. This study could contribute to the development of more accurate gait pattern classification and assist in detecting age-related gait patterns in clinical environments.

Keywords: Accelerometers; Ageing; Deep learning; Gait classification; Machine learning.

1. Introduction

Walking is one of the most common repetitive activities of humans. Gait patterns have been acknowledged as a potential biomarker for fall risk, Parkinson's disease, and ageing [1]. Adaptive gait patterns of healthy individuals are the result of the delicate control and coordination of various systems, such as the central nervous and musculoskeletal systems. Therefore, gait analysis plays a crucial role in gait monitoring and abnormalities recognition, clinical interventions assessment, and rehabilitation programs [2].

Accelerometers have been widely used in gait assessment in clinical and daily-living environments. From accelerometer signals, comprehensive sets of gait features are calculated, such as spatial, temporal, and dynamic outcomes [3-5] to characterize the alterations in gait patterns. For instance, results from gait analysis revealed that geriatric patients exhibit slower walking speeds accompanied by less regular, less predictable, and less local stable gait patterns compared with healthy older adults [6]. In older adults (>65 years), to avoid falls and increase stability, compensatory gait patterns have been observed, including lower walking speed, shorter step length, and larger step width compared to young controls [7]. The changes in gait patterns due to aging or pathology allow for classification of different populations based on gait features [8].

Between many gait features, temporal dependencies (e.g., between walking speed and step frequency) and non-linear interactions (e.g., between walking speed and local dynamic stability) exist [9, 10]. Conventional machine learning (ML) approaches can capture the linear dependencies and non-linear interactions between gait features, and have been comprehensively explored and successfully applied to classify pathological gait patterns and age-related gait patterns [11]. Support vector machine (SVM) and random forest (RF) are often-used supervised ML approaches in gait analysis. A recent study explored the performance of 6 ML approaches (linear discriminant analysis, logistic regression (LR), naive bayes (NB), SVM, K-nearest neighbour (KNN), and RF) in classifying fallers from non-fallers based on gait features [12]. RF achieved an optimal classification accuracy of 98%. An artificial neural network (ANN) model classified geriatric patients with an accuracy of 96% by incorporating the interaction between clinical and gait variables [13]. Similar, the accurate classification of freezing of gait, a common gait characteristic seen in patients with Parkinson’s disease, has been studied by comparing 6 different ML learning approaches, namely KNN, RF, LR, NB, multilayer perceptron (MLP), and SVM [14]. The results show that SVM was the optimal classifier with 89.6% on the geometric mean of sensitivity (87.4%) and

specificity (91.7%), compared to MLP, the second best classifier, which achieved 85.16%.

Although the above studies demonstrate that gait classification benefits from conventional ML approaches, there are several drawbacks of these ML approaches for clinical application. First, they rely on handcrafted gait features for the input. The design and selection of handcrafted gait features requires expert knowledge and may omit some important features leading to suboptimal performance. Second, the calculation of gait features is performed most often offline and is laborious [3]. Third, to obtain stable and accurate gait features, such as maximum Lyapunov index as an index of stability, some strict signal requirements need to be met, e.g., regarding signal length and sample frequency, which may hamper clinical applications [15].

Unlike conventional ML approaches, deep learning (DL) approaches include feature extraction as a part of the model and are able to learn suitable features from the raw data automatically. Apart from this, the multiple processing layers of DL approaches allow the progressive extraction of higher-level features from the accelerometer signal [16]. Moreover, DL approaches have been shown to outperform conventional ML approaches in several fields (e.g., visual recognition and text analytics), especially in time series classification tasks [17]. In order to eliminate the need of handcrafted features and improve classification performance, DL approaches have been explored for clinical gait classification [18]. For instance, recurrent neural network (RNN) models (long short-term memory (LSTM) and bidirectional LSTM (BiLSTM)) have been applied on IMU (inertial measurement units; acceleration and gyro) data obtained during 1-minute walking to classify fallers and non-fallers in patients with multiple sclerosis [19]. With a window size of 1 minute, the results of this study show that DL approach (BiLSTM) outperformed conventional ML approaches, SVM and LR, with 0.88 area under the receiver operator curve (AUC), compared to the best performance of conventional ML which was 0.79. For classification into fallers and non-fallers based on acceleration data during daily life, the performance of LSTM and gated recurrent unit (GRU) was compared using a 1.28-second window size. Both LSTM and GRU provided classification results around 0.96 accuracy [20].

The studies mentioned above show the potential of DL, especially RNN, for gait analysis. However, other state-of-the-art DL approaches (e.g., convolutional neural network (CNN) and hybrid neural network (HNN)) were not included [21]. RNN is capable of learning temporal relationships from accelerometer data, while CNN is widely known for its feature extraction

capability and has been extensively used in time series classification tasks [21]. The combination of the deep structures of CNN and RNN called HNN, such as convolutional LSTM (ConvLSTM), may benefit from both advantages. These models have been successfully employed in some tasks similar to gait patterns classification, such as human activity recognition (HAR) based on inertial sensors [22]. Moreover, window size is discussed in other fields, such as in HAR [23] but not in gait analysis yet.

Considering the cyclic nature of walking, with repetitive gait cycles (spatial-temporal characteristics of swing and stand-phases), the choice of the window size and time span is crucial, since it will affect classification results. Additionally, in a given time window, different step frequencies from subjects may lead to disparities in the number of steps, which may, in turn, impact other gait features, and result in an unfair comparison between subjects. To mitigate these potential biases, acceleration data can be normalized based on subject's step frequency. This normalization process serves to standardize the number of steps within each time window, thereby facilitating more accurate and equitable comparisons between subjects.

This study aims to compare different ML and DL approaches to classify age groups (adult vs. older adult) based on acceleration time-series obtained during 3-minute walking. More specifically, we will: 1) compare the classification performance of DL approaches (CNN, LSTM, GRU, and ConvLSTM) based on acceleration time-series data with conventional ML approaches (RF, SVM, NB, and KNN) that use handcrafted gait features as input; 2) explore how window size affects the classification results in DL approaches; 3) study the effect of step frequency in DL classification by normalizing the data in the time window by step frequency.

2. Methods

2.1. Subjects, equipment, and data collection

Accelerometer data obtained during walking from different studies [6, 24-27] were pooled to create the present dataset. Herein 267 out of 394 participants were included in this study. Participants were excluded because of cognitive impairment (n=104), no walking data (n=19; power spectrum values smaller than 0.5 HZ), and insufficient length of walking data (n=4; less than 100 seconds walking data within the 3-minute recording). Participants were divided into two sub-groups: the adult group (18–65) (74:56, Female: Male; mean age 38.3, SD: 15.5), the older adult group (>65) (60:77, Female: Male; mean age 77.1, SD: 6.1). Subjects were asked to walk for 3 minutes at a comfortable walking speed. During walking, one accelerometer

(iPod, Dynaport, or ActiGraph; 100 Hz sampling frequency) was fixed with a belt near the level of lumbar segment L3.

Data were obtained between 2008 and 2022. The Medical Ethical Committee of the University Medical Centre Groningen and the Medical Ethical Committee of the Slotervaart Hospital (closed in 2018) approved the studies and all participants signed written informed consent. This study was conducted according to the principles expressed in the Declaration of Helsinki.

2.2. Data preprocessing

A median filter (with window size of 5 samples) was used to remove spike noise and an additional low pass filter (5rd order Butterworth cut-off frequency 20 Hz) was applied to remove the high frequency noise. To remove the data collected during the sensor installation or uninstallation, the first and last 5 seconds of data were discarded. To ensure that only walking data were included in the analysis, data recorded during turning were removed.

2.3. Segmentation and datasets splitting

In order to extract accurate gait features for conventional ML(e.g., maximum Lyapunov exponent and sample entropy), longer accelerometer signal series are needed. Therefore, 1024-sample window size was used. To enable a fair comparison between ML and DL approaches, the filtered walking accelerometer data of each subject were split into windows of 1024-samples.

The adult and older adult subjects were proportionally randomly divided into training, testing, and validation sets (186: 54: 27). To ensure fair consideration of each subject and make full use of the data, 10 corresponding segments were randomly sampled to form the dataset for each subject.

To study the impact of window size on the classification performance of DL approaches, various window sizes, including 128, 256, 2048, and 5120 samples, were compared.

To investigate the impact of step frequency on DL in gait classification, the following processes were conducted. For subject i, the averaged step frequency f_i was calculated using FFT based on the corresponding walking accelerometer data. So, the length of one step in data points was defined as s_i,

$$s_i = \frac{1}{f_i} \times 100,$$

where 100 was the sampling frequency. For a 128-sample window size, the data points within the range of $[0, 1.2s_i]$ were selected. Because the step frequency tends to vary slightly during walking and to ensure that the data points of one complete step were included, the coefficient 1.2 was used for this window size. Then these data points were interpolated by a 1-D smoothing spline (with a smoothing factor 0) into a size of 128 to replace the corresponding segments. For other window sizes, the coefficients were set as

$$coefficient = \frac{window\ size}{128} \times 1.2.$$

2.4. Deep learning

In this study, the whole dataset was normalized to get rid of the heterogeneity of accelerometers.

1) Convolutional Neural Network (CNN) is one of the earliest successfully used DL approaches. Because of its excellent capacity for feature extraction, it has been widely used for human movement recognition [28]. CNN includes two parts: 1) convolution layer and pooling layer for feature extraction; 2) fully connected layer and detector layer for classification. These layers can be stacked to form a deep CNN.

2) Recurrent Neural Network (RNN) is a family of neural networks that have recurrent connections. The recurrent connections enable RNN to keep the “memory” from the input of the previous moment and use it to influence the output of the current input. Because it can learn sequential dependencies, RNN outperforms in dealing with sequential problems compared with general neural networks. Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are the two mostly used variants of the RNN architecture in movement recognition [22]. LSTM has memory cells comprising inputs, forget, and outputs gates to control store or forget information. GRU has a similar gated structure to adaptively capture sequential dependencies, and it is more computationally efficient than LSTM. Bi-Directional LSTM (BiLSTM) is a variant of sequentially stacked LSTM layers. It has two hidden states that allow the model to use information from the past and the future, of each input, which has been shown to improve performance for some classification tasks [19].

3) Hybrid Models such as ConvLSTM combine the deep structures of CNN and LSTM. CNN is known for its feature extraction capability and LSTM is capable of learning temporal dynamics. In order to obtain the dual advantages of CNN and LSTM, hybrid models are

proposed.

To further improve generalization performance, dropout and early stopping regularization techniques were used. Dropout removes nonoutput units randomly from the network. Thus, the dropout technique can reduce the scale and complexity of the neural network and eventually avoid overfitting. Early stopping is a simple regularization technique. It will stop the training when the validation error which is used as an estimate of the generalization error has no improvement in k epochs.

2.5. Conventional machine learning

In order to compare the performance of DL approaches based on acceleration signal with conventional ML approaches based on gait feature input, 4 approaches [4] were employed, including nonlinear models (support vector machine (SVM), naive bayes (NB), K-nearest neighbour (KNN)) and tree ensemble models (Random Forest (RF)).

Herein 36 gait features were extracted from the accelerometer data and were used to train the conventional ML approaches. These gait features represent the pace (walking speed, stride length, stride time, stride frequency, and acceleration root mean square), regularity (stride regularity and gait symmetry index), smoothness (index of harmonicity and harmonic ratio), predictability (sample entropy), and stability (maximal Lyapunov exponent, maximal Lyapunov exponent normalized per stride by time) of gait and have been described previously [4, 24]. The gait features were normalized and a kernel principal component analysis (KPCA) was employed for dimensionality reduction. To increase generalization, the parameters from the training dataset were utilized to normalize the testing and validation datasets.

2.6. Fine-tuning

The classification performance of ML approaches, especially DL approaches, is highly sensitive to the hyperparameter setting. The conventionally used 5-fold cross validation and randomized search has been widely used to find the best hyperparameters for conventional ML approaches. However, for DL approaches and big datasets, cross validation which repeats training iterations would lead to exploding computation costs. Hence, it is important to find the global optimum in a minimum number of steps. Bayesian Optimization (BO) has become a state-of-the-art solution [29]. It incorporates prior performance of the hyperparameters and updates the new hyperparameters to achieve better performance. BO uses an acquisition function that directs sampling to areas where an improvement over the current best

observation is likely. Thus, in this study, BO was used to find the best hyperparameters for the DL and conventional ML approaches. For each approach, 18 parameter combination trails were searched.

2.7. Hyperparameter space

The details of the hyperparameter space of conventional ML approaches for BO are as follows:

SVM) The “rbf” kernel was used. The box constraint parameter (C) was varied from 1 to 250. The degree was set from 1 to 50. The gamma was set from values 0.01 to 10 with increments of 0.05. The “True” and “False” shrinking were also considered.

NB) The portion of the largest variance of all features was set from -11 to -7 power of 10 with -10 power of 10 step.

KNN) The number of neighbors was varied from 1 to 15. Different algorithms which compute the nearest neighbors were explored, “ball_tree”, “kd_tree”, and “brute”. The weight functions “uniform” and “distance”, were explored.

RF) Different numbers of trees in the range of 10–1000 with increments of 10 were explored. For the trees, the hyperparameters were set as: the number of maximum depths varied from 3 to 25; the maximum number of leaf nodes was varied from 5 to 50. The minimum number of samples required to be at a leaf node was varied between 5 and 50.

The structure of DP approaches is shown in Fig. 1. After the input layer, the optimal number of the deep learning layer stack was tuned by BO. Each stack consists of a Deep layer (Conv1D, GRU, LSTM, Bidirectional-LSTM, or ConvLSTM2D layer) to extract features, a Batch Normalization layer to make the training faster and more stable, a Pooling layer to reduce the number of parameters, and a Dropout layer to avoid overfitting. Then, a Dense layer was used to fully connect the output of the previous layers and a flatten layer was used to flatten the dimensions of the output. Another Dropout layer was added after the flatten layer. The final layer was an output layer which was a combination of a Dense layer and a softmax activation. This Dense layer had 2 units to classify the output into 2 classes. The hyperparameter space of DL approaches for BO is shown in Table 1.

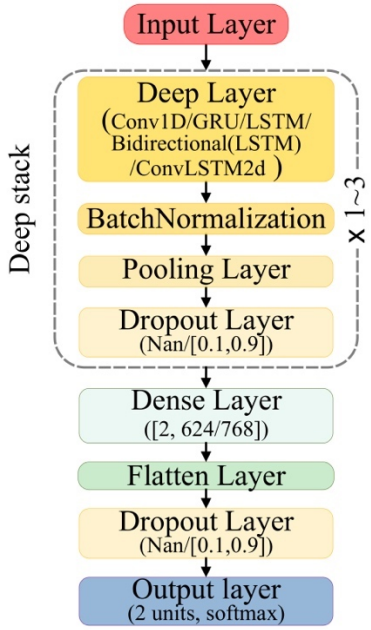


Figure 1. General architecture of the deep learning algorithms used in this study. A detailed description of the individual layers can be found in Table 1. Conv1D: 1-dimensional convolution neural network; GRU: gated recurrent unit; LSTM: long short-term memory; ConvLSTM2d: 2-dimensional convolution long short-term memory

Table 1. Hyperparameters space for deep learning approaches

Layer	Parameters		CNN	GRU	LSTM	BiLSTM	ConvLSTM
Input							
Stack	Stack number		1-3				
	Deep layer	Layer name	Conv1D	GRU	LSTM	Bidirectional (LSTM)	ConvLSTM2 D
		Unit/ Filter	[2, 768]				[2, 624]
		Kernel size	[1, 15]	-			[1, 8]
		Activation	"ReLU"		"tanh"		
	Batch Normalization		-				
	Pooling	Kernel size	-				[1, 8]
	Dropout	Rate	Nan or [0.1, 0.9]				
Dense	Unit		[2, 768]				[2, 624]
Flatten			-				
Dropout	Rate		Nan or [0.1, 0.9]				
Output	Dense		2 units and "softmax" activation				
	Learning rate		[1e-5, 1e-2]				

CNN: convolutional neural network; GRU: gate recurrent unit; LSTM: long short-term memory; BiLSTM: bi-directional long short-term memory; ConvLSTM: convolutional long short-term memory; ReLU: rectified linear unit; tanh: hyperbolic tangent function.

2.8. Evaluation

Classification performance was evaluated with commonly used evaluation metrics, including

accuracy, recall (sensitivity), precision, and F1 score (harmonic mean of sensitivity and precision). The receiver operating characteristic curves were compiled and the AUC was reported.

To allow for the verification and reproduction of results, the project repository has been made publicly available (https://github.com/xzheng93/Age-related_gait_classification). It includes the source code, dataset, log files of all the experiments, and optimal models for each approach. The overall data processing pipeline is illustrated in Fig. 2.

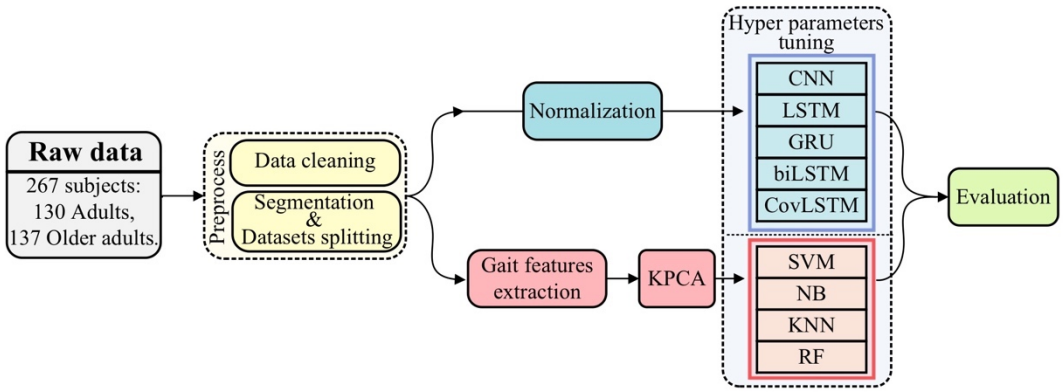


Figure 2. Data processing pipeline for age-related gait patterns classification. KPCA: kernel principal component analysis; CNN: convolutional neural network; GRU: gate recurrent unit; LSTM: long short-term memory; BiLSTM: bi-directional long short-term memory; ConvLSTM: convolutional long short-term memory; SVM: support vector machine; NB: naive bayes; KNN: k-nearest neighborhood; RF: random forest.

3. Results

After hyperparameter tuning, the best architectures and hyperparameters of each approach and the tuning logs were stored in https://github.com/xzheng93/Age-related_gait_classification/tree/main/result/logs.

3.1. Classification performance comparison for 1024 window size

The complete list of performance metrics calculated for each algorithm based on a window size is depicted in Table 2. With F1-scores ranging from 0.86 to 0.9, the DL approaches outperformed the ML approaches (F1 scores ranging from 0.69 to 0.74).

To enable an in-depth look at the individual performance of the classifiers, their confusion matrices are presented in Fig. 3 and Fig. 4. GRU achieved the highest overall accuracy (89.3%) and was able to correctly classify 82.3% and 95.7% of the adult and older adult samples, respectively. With 73.9% and 73.5%, SVM and RF achieved the highest accuracies among the conventional ML

approaches. The accuracy difference between the DL approaches with the highest accuracy and the ML approaches with the highest accuracy was 15.4%.

Table 2. The performance metrics of dl and conventional ml approaches

	Models	Acc	Pre	Recall	F1	AUC
Deep learning	CNN	0.88	0.83	0.98	0.90	0.96
	LSTM	0.85	0.81	0.94	0.87	0.94
	GRU	0.89	0.85	0.96	0.90	0.96
	BiLSTM	0.87	0.84	0.93	0.88	0.94
	ConvLSTM	0.85	0.85	0.86	0.86	0.95
Conventional machine learning	SVM	0.74	0.74	0.74	0.74	0.83
	NB	0.72	0.74	0.72	0.72	0.75
	KNN	0.68	0.79	0.68	0.69	0.76
	RF	0.74	0.74	0.74	0.74	0.82

CNN: convolutional neural network; GRU: gate recurrent unit; LSTM: long short-term memory; BiLSTM: bi-directional long short-term memory; ConvLSTM: convolutional long short-term memory; SVM: support vector machine; NB: naive bayes; KNN: k-nearest neighborhood; RF: random forest.

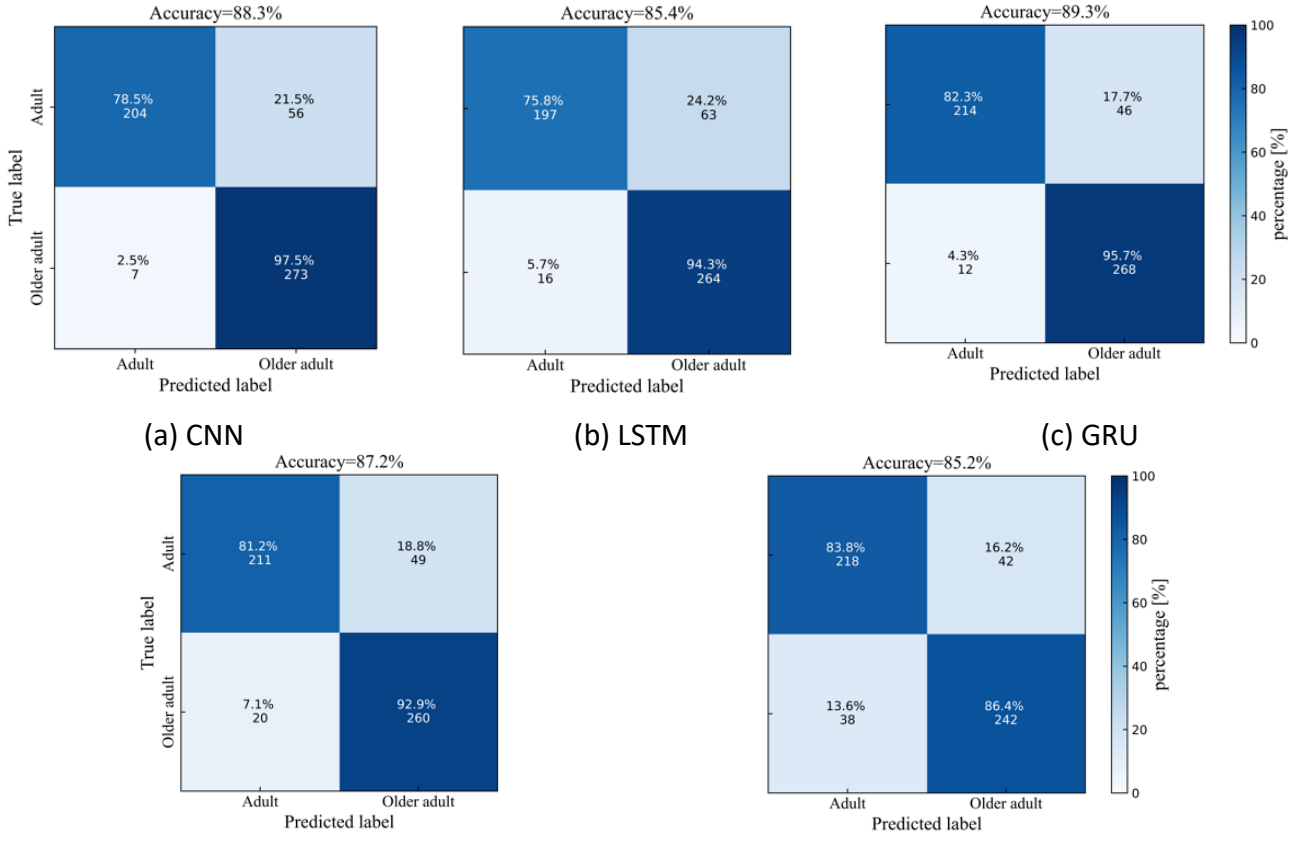
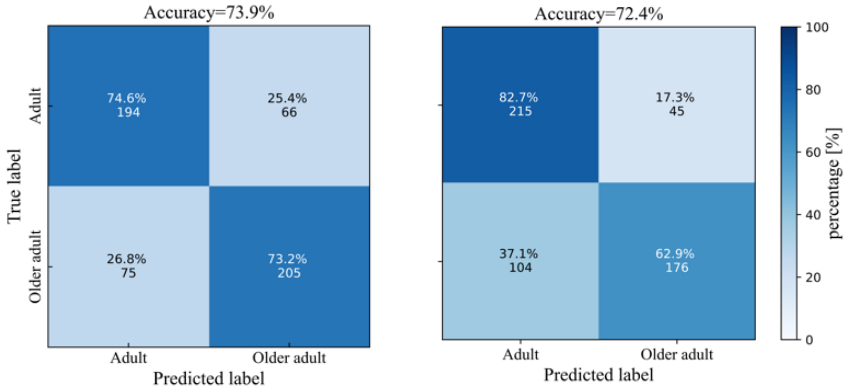
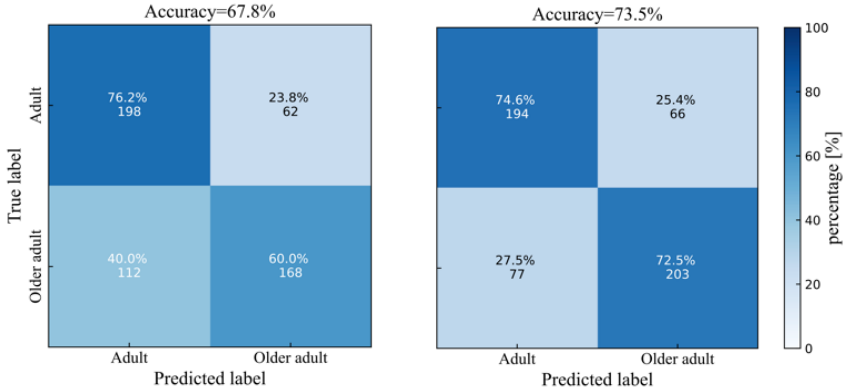


Figure 3. Confusion matrices of DL approaches: (a) CNN, (b) LSTM, (c) GRU, (d) BiLSTM, (e) ConvLSTM. For testing, 54 subjects and their corresponding 10 segments (n=540) were used. A: adult group (n=26); OA: older adult group (n=28). CNN: convolutional neural network; GRU: gate recurrent unit; LSTM: long short-term memory; BiLSTM: bi-directional long short-term memory; ConvLSTM: convolutional long short-term memory.



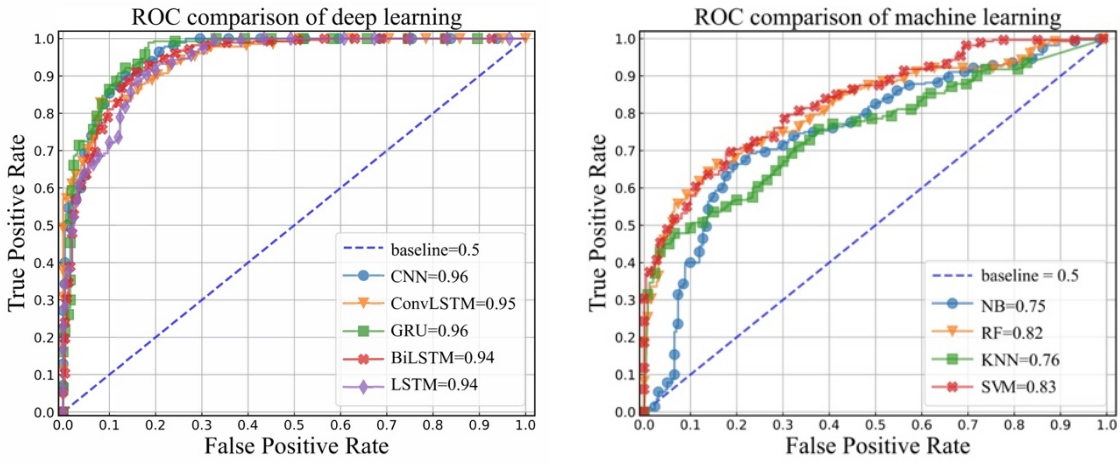
(a) SVM (b) NB



(c) KNN (d) RF

Figure 4. Confusion matrices of conventional ML approaches: (a) SVM, (b) NB, (c) KNN, (d) RF. For testing, 54 subjects and their corresponding 10 segments (n=540) were used. A: adult group (n=26); OA: older adult group (n=28). SVM: support vector machine; NB: naive bayes; KNN: k-nearest neighbour; RF: random forest.

To further demonstrate the characteristics of the classifiers, the ROC curves are presented in Fig. 5. As depicted in Fig. 5, DL approaches had an AUC higher than 0.94, while the ML classifiers only achieved AUC values of 0.75 to 0.83.



(a) DL approaches (b) conventional ML approaches

Figure 5. Receiver operating characteristic curves (ROC) and area under curve (AUC) of (a)

DL approaches and (b) conventional ML approaches. CNN: convolutional neural network; GRU: gate recurrent unit; LSTM: long short-term memory; BiLSTM: bi-directional long short-term memory; ConvLSTM: convolutional long short-term memory; SVM: support vector machine; NB: naive bayes; KNN: k-nearest neighbour; RF: random forest.

3.2. The effect of different window sizes and step frequency on DL

The mean step frequencies for the adult and older adult groups were 1.72 (SD: 0.08) and 1.97 (SD: 0.20) step/second, respectively. Given window sizes ranging from 128 to 5120 samples, each segment may include approximately 1, 4, 8, 20, 35, or 88 steps.

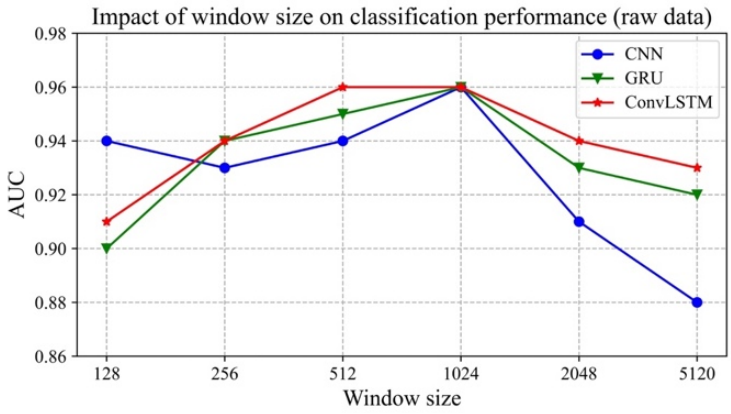
The different window sizes for the DL approaches were investigated and the AUC results are shown in Fig. 6 (a). Additional accuracy metrics can be found in Appendix A. To represent the performance of convolutional, recurrent, and hybrid models, CNN, GRU, and ConvLSTM were selected, respectively. When the window size increased from 128 to 1024, the performance of GRU and ConvLSTM improved, while the performance of CNN fluctuated with the window sizes. However, when the window sizes increased to 2048 and 5120, the AUC values decreased in all approaches, especially in CNN. This may be attributed to the decreased number of segments resulting from larger window sizes. Specifically, as the window size increased from 128 to 5120, the number of segments sharply decreased from 26700 to 534.

The classification results based on step frequency normalization data are shown in Fig. 6 (b) (details in the appendix B). All approaches exhibited a similar trend as with raw data. CNN and GRU achieved comparable results compared to raw data, while ConvLSTM showed slightly worse performance.

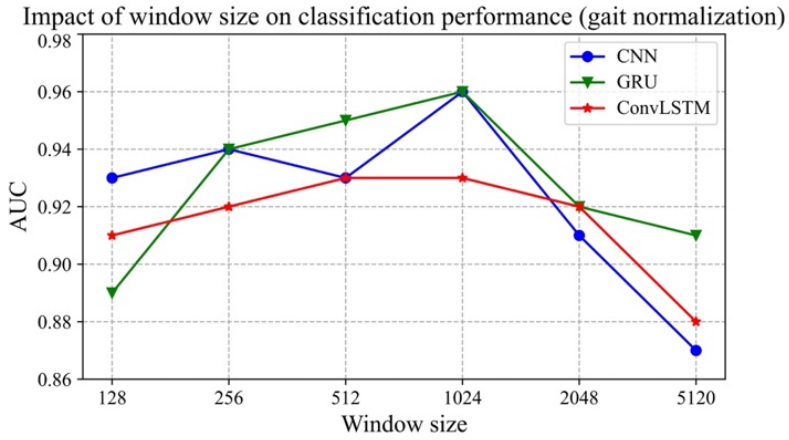
4. Discussion

This study compared the classification capacity of 5 DL and 4 conventional ML approaches in classifying age-related gait patterns in healthy adults based on acceleration time series data collected by one 3D-accelerometer for a 3-minute walk. The results show that the DL approaches achieved better classification performance by a remarkable margin (all AUC greater than 0.94 for DL vs. 0.83 for SVM (best in conventional ML)). The study also explored the effective window size for DL and studied the effect of step frequency on DL classification. It was found that when the window size increased from 128 to 1024, the classification performance of GRU and ConvLSTM increased, while the performance of CNN fluctuated. However, when the window sizes were larger than 1024, the performance of all DL approaches decreased. Based on the normalized gait data, CNN and GRU exhibited a similar

trend and performance as with raw data, while ConvLSTM showed slightly worse performance.



(a) based on raw data



(b) based on step frequency normalization data

Figure 6. Impact of window size on classification performance. CNN: convolutional neural network; GRU: gate recurrent unit; ConvLSTM: convolutional long short-term memory.

In this study, DL approaches exhibited superior performance compared to conventional ML approaches for the 1024-sample window size. This can be attributed to two main reasons. First, DL has superior learning capabilities. By using multi stacks of simple layers, DL is able to model a high degree of nonlinearity in the input data. Second, DL has better features extraction capacity and is able to select the most suitable features for classification because of its end-to-end characteristic which integrates feature extraction, selection, and classification within a neural network model [22]. DL can learn large amounts of features from raw data, including high-level features constructed from features learned in lower layers [16]. In contrast, the conventional ML approaches rely on handcrafted gait features that require domain expertise and may miss some useful features [30]. Although this study considered a comprehensive set of gait features, including pace, regularity, smoothness, and stability, the

substantial performance gap between DL and conventional ML suggests that important features reflecting changes in gait due to aging may not have been included in the handcrafted features. Furthermore, the better performance of DL suggests that handcrafted gait features are not necessary for age-related gait pattern classification.

Signal segmentation plays a crucial role for gait classification as it not only affects the classification performance [31], but also helps to understand how the data size and time span influence the classification results [23]. GRU is a kind of RNN approach and is designed to learn both short- and long-term dependent features [22]. ConvLSTM combines CNN and RNN structures and is expected to benefit from larger window sizes, since bigger window sizes provide richer features, especially time-dependent features, which are valuable for gait analysis. The AUC scores of GRU and ConvLSTM show remarkable increases when the window size increases from 128 (including about 1 step) to 1024 (including about 20 steps). Although walking is a repeating activity, the relationship and slight differences between gait cycles may disclose the postural control ability of subjects. Changes in postural control due to aging are associated with alterations in sensory, musculoskeletal, and neuromuscular systems [32]. Gait analysis shows that because of the decrease of postural control capacity, older adults have less regularity and higher variability [33], lower local stability (higher largest Lyapunov exponent) [34], worse gait symmetry (lower symmetry index) [35], and less complexity (lower sample entropy) in gait [36]. These gait features are designed to measure the time dependent changes of gait cycles. Therefore, a larger window size contains this kind of information and may allow GRU and ConvLSTM to discriminate age groups. The sharp improvements of GRU and CovLSTM from 128- to 1024-sample window size may indicate that the time dependent changes in gait cycles are important for age-related gait pattern discrimination.

CNN is well known for its local temporal and spatial features extraction capacity, allowing it to capture the repeating gait patterns in the data and achieve good performance. However, it cannot learn long-term dependencies from the data, which means that larger window sizes may not provide additional features for CNN. Additionally, the accelerometer data in this study were collected from participants walking in a straight and empty hallway without external perturbation. Thus, the gait cycle is highly repetitive and presumably predictable, as evident by the mean sample entropy values of 0.275. These may explain the small fluctuations in classification performance among window sizes ranging from 128 to 1024. For the 128-sample window size, CNN achieved accurate performance (AUC=0.94). This result may indicate that the shape of one or two steps contains rich information that can be used to

discriminate the age-related gait patterns. Indeed, gait features such as root mean square (gait intensity), rhythm (the proportion of stance and swing phase), and harmonic index (the smoothness of the acceleration curve) are designed to assess the shape of acceleration in gait, and have been utilized to characterize different age populations [36]. Therefore, CNN specifically capture this type of information to discriminate the age-related gait patterns.

When the window size increased beyond 1024, the number of segments decreased, leading to a decline in classification performance for all approaches. Moreover, using a 5120-sample window size resulted in an insufficient number of segments (n=523), which decreased the generalization of the model and led to a sharp decline in classification performance. Despite this, GRU and ConvLSTM demonstrated only slightly decreased performance, due to their ability to benefit from longer data segments compared to CNN.

CNN and GRU exhibited similar classification performance with both raw data and the normalization of segment data by step frequency. ConvLSTM exhibited slightly inferior performance when using normalized data compared to using raw data. This may be due to the normalization process which excluded step frequency information and only used a partial data from the original segment. These results may indicate that data normalization based on step frequency may not be necessary for gait classification using DL.

In this study, all the DL approaches got best classification performance for the 1024-sample window size. This window size may be also applicable for gait pattern classification in daily living environments, where walking bouts are typically short (60% of bouts less than 30s) [37].

5. Limitation

Although DL approaches yield promising results in age-related gait pattern classification performance, the black-box nature of these approaches is often seen as a limitation for clinical applications [33]. While the underlying mathematical principles of these approaches are understood, it is unclear why a particular prediction has been made. Additionally, features extracted by DL cannot be linked to physical explanations and models. Identifying age-related gait features that were not been covered by the current gait analysis may help to gain knowledge about age-related changes and assist clinicians in identifying people with altered gait patterns. Facing these challenges, the field of explainable artificial intelligence has gained increasing attention in recent years [38]. Further research should focus on determining the justification of classification predictions and making the prediction processes comprehensible

to clinical experts. For the next step, popular methods, such as Local Interpretable Model-Agnostic Explanation [39], could be used to interpret and explain DL in performing gait pattern classification. Apart from this, the proposed approaches should be validated on data recorded outside lab environments before they are deployed in a daily life environment. This step is necessary, because the environment will influence walking (e.g., stride length variability [40]). Since the dataset size used in this study is limited, it is not possible to further deduce whether using a larger window size (e.g., 5120) would result in improved classification performance.

6. Conclusion

This study compared 4 conventional ML and 5 DL approaches for classifying the age-related gait patterns of healthy adults, based on one accelerometer sensor. The results show that DL approaches based on raw accelerometer data outperformed the conventional ML approaches based on handcrafted features. This suggests that DL may have captured certain gait features related to aging that were not previously reported. The investigation into the impact of different window sizes on classification performance indicates that the shape of acceleration and the relationship as well as differences between gait cycles are important factors for age-related gait pattern classification. The results show that a 1024-sample window size was suitable for gait pattern classification as it yielded the best performance across all tested DL approaches. Notably, the study discovered that normalizing data by step frequency did not affect classification performance, suggesting that this step frequency normalization may not be necessary in a healthy population.

The present study highlights the potential of DL approaches for accurately classifying age-related gait patterns using accelerometer data. The findings suggest that subtle variations and interconnections between gait cycles, as well as the shape of one step acceleration are important factors for age-related pattern classification, contributing to a better understanding of the postural control changes associated with aging. This study could serve as one of the first stepping stones towards future studies in the development of more accurate gait pattern classification, facilitating abnormal gait patterns diagnosis in the clinical environment and gait monitoring in the daily living environment.

Reference

[1] M. A. Brodie, N. H. Lovell, C. G. Canning, H. B. Menz, K. Delbaere, S. J. Redmond, M. Latt, D. L. Sturnieks, J. Menant, and S. T. Smith, "Gait as a biomarker? Accelerometers

reveal that reduced movement quality while walking is associated with Parkinson's disease, ageing and fall risk." pp. 5968-5971.

[2] D. Jarchi, J. Pope, T. K. Lee, L. Tamjidi, A. Mirzaei, and S. Sanei, “A review on accelerometry-based gait analysis and emerging clinical applications,” *IEEE reviews in biomedical engineering,* vol. 11, pp. 177-194, 2018.

[3] Y. Hutabarat, D. Owaki, and M. Hayashibe, “Recent advances in quantitative gait analysis using wearable sensors: a review,” *IEEE Sensors Journal*, 2021.

[4] X. Zheng, M. F. Reneman, J. A. Echeita, R. H. S. Preuper, H. Kruitbosch, E. Otten, and C. J. Lamoth, “Association between central sensitization and gait in chronic low back pain: Insights from a machine learning approach,” *Computers in biology and medicine,* vol. 144, pp. 105329, 2022.

[5] Y. H. Zhou, R. Z. U. Rehman, C. Hansen, W. Maetzler, S. Del Din, L. Rochester, T. Hortobagyi, and C. J. C. Lamoth, “Classification of Neurological Patients to Identify Fallers Based on Spatial-Temporal Gait Characteristics Measured by a Wearable Device,” *Sensors,* vol. 20, no. 15, Aug, 2020.

[6] L. Kikkert, N. Vuillerme, J. P. van Campen, B. A. Appels, T. Hortobagyi, and C. J. Lamoth, “Gait characteristics and their discriminative power in geriatric patients with and without cognitive impairment,” *Journal of Neuroengineering and Rehabilitation,* vol. 14, Aug, 2017.

[7] A. Aboutorabi, M. Arazpour, M. Bahramizadeh, S. W. Hutchins, and R. Fadayevatan, “The effect of aging on gait parameters in able-bodied older subjects: a literature review,” *Aging clinical and experimental research,* vol. 28, no. 3, pp. 393-405, 2016.

[8] A. Mannini, D. Trojaniello, A. Cereatti, and A. M. Sabatini, “A machine learning framework for gait classification using inertial sensors: Application to elderly, post-stroke and huntington’s disease patients,” *Sensors,* vol. 16, no. 1, pp. 134, 2016.

[9] Y. Zhou, R. Romijnders, C. Hansen, J. van Campen, W. Maetzler, T. Hortobagyi, and C. J. C. Lamoth, “The detection of age groups by dynamic gait outcomes using machine learning approaches,” *Scientific Reports,* vol. 10, no. 1, Mar, 2020.

[10] I. Hagoort, N. Vuillerme, T. Hortobágyi, and C. J. Lamoth, “Outcome-dependent effects of walking speed and age on quantitative and qualitative gait measures,” *Gait & Posture,* vol. 93, pp. 39-46, 2022.

[11] C. Prakash, R. Kumar, and N. Mittal, “Recent developments in human gait research: parameters, approaches, applications, machine learning techniques, datasets and challenges,” *Artificial Intelligence Review,* vol. 49, pp. 1-40, 2018.

[12] R. Z. U. Rehman, Y. Zhou, S. Del Din, L. Alcock, C. Hansen, Y. Guan, T. Hortobágyi, W. Maetzler, L. Rochester, and C. J. Lamoth, "Gait analysis with wearables can accurately classify fallers from non-fallers: a step toward better management of neurological disorders," *Sensors,* vol. 20, no. 23, pp. 6992, 2020.

[13] Y. Zhou, J. van Campen, T. Hortobágyi, and C. J. Lamoth, "Artificial neural network to classify cognitive impairment using gait and clinical variables," *Intelligence-Based Medicine,* vol. 6, pp. 100076, 2022.

[14] A. Samà, D. Rodríguez-Martín, C. Pérez-López, A. Català, S. Alcaine, B. Mestre, A. Prats, M. C. Crespo, and À. Bayés, "Determining the optimal features in freezing of gait detection through a single waist accelerometer in home environments," *Pattern Recognition Letters,* vol. 105, pp. 135-143, 2018.

[15] S. M. Bruijn, J. H. van Dieën, O. G. Meijer, and P. J. Beek, "Statistical precision and sensitivity of measures of dynamic gait stability," *Journal of neuroscience methods,* vol. 178, no. 2, pp. 327-333, 2009.

[16] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature,* vol. 521, no. 7553, pp. 436-444, 2015.

[17] I. H. Sarker, "Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions," *SN Computer Science,* vol. 2, no. 6, pp. 1-20, 2021.

[18] Y. Matsushita, D. T. Tran, H. Yamazoe, and J.-H. Lee, "Recent use of deep learning techniques in clinical applications based on gait: a survey," *Journal of Computational Design and Engineering,* vol. 8, no. 6, pp. 1499-1532, 2021.

[19] B. M. Meyer, L. J. Tulipani, R. D. Gurchiek, D. A. Allen, L. Adamowicz, D. Larie, A. J. Solomon, N. Cheney, and R. S. McGinnis, "Wearables and deep learning classify fall risk from gait in multiple sclerosis," *IEEE journal of biomedical and health informatics,* vol. 25, no. 5, pp. 1824-1831, 2020.

[20] F. Luna-Perejón, M. J. Domínguez-Morales, and A. Civit-Balcells, "Wearable fall detector using recurrent neural networks," *Sensors,* vol. 19, no. 22, pp. 4885, 2019.

[21] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Deep learning for time series classification: a review," *Data mining and knowledge discovery,* vol. 33, no. 4, pp. 917-963, 2019.

[22] F. Gu, M.-H. Chung, M. Chignell, S. Valaee, B. Zhou, and X. Liu, "A survey on deep learning for human activity recognition," *ACM Computing Surveys (CSUR),* vol. 54, no. 8, pp. 1-34, 2021.

[23] M. Jaén-Vargas, K. M. R. Leiva, F. Fernandes, S. B. Gonçalves, M. T. Silva, D. S. Lopes, and J. J. S. Olmedo, "Effects of sliding window variation in the performance of acceleration-based human activity recognition using deep learning models," *PeerJ Computer Science,* vol. 8, pp. e1052, 2022.

[24] N. M. Kosse, S. Caljouw, D. Vervoort, N. Vuillerme, and C. J. Lamoth, "Validity and reliability of gait and postural control analysis using the tri-axial accelerometer of the iPod touch," *Annals of biomedical engineering,* vol. 43, no. 8, pp. 1935-1946, 2015.

[25] C. J. Lamoth, F. J. van Deudekom, J. P. van Campen, B. A. Appels, O. J. de Vries, and M. Pijnappels, "Gait stability and variability measures show effects of impaired cognition and dual tasking in frail people," *Journal of neuroengineering and rehabilitation,* vol. 8, no. 1, pp. 1-9, 2011.

[26] M. H. de Groot, H. C. van der Jagt-Willems, J. P. van Campen, W. F. Lems, J. H. Beijnen, and C. J. Lamoth, "A flexed posture in elderly patients is associated with impairments in postural control during walking," *Gait & posture,* vol. 39, no. 2, pp. 767-772, 2014.

[27] T. IJmker, and C. J. Lamoth, "Gait and cognition: the relationship between gait stability and variability with executive function in persons with and without dementia," *Gait & posture,* vol. 35, no. 1, pp. 126-130, 2012.

[28] A. Murad, and J. Y. Pyun, "Deep Recurrent Neural Networks for Human Activity Recognition," *Sensors (Basel),* vol. 17, no. 11, Nov 6, 2017.

[29] J. Wu, X.-Y. Chen, H. Zhang, L.-D. Xiong, H. Lei, and S.-H. Deng, "Hyperparameter optimization for machine learning models based on Bayesian optimization," *Journal of Electronic Science and Technology,* vol. 17, no. 1, pp. 26-40, 2019.

[30] F. Gu, K. Khoshelham, S. Valaee, J. Shang, and R. Zhang, "Locomotion activity recognition using stacked denoising autoencoders," *IEEE Internet of Things Journal,* vol. 5, no. 3, pp. 2085-2093, 2018.

[31] O. Banos, J.-M. Galvez, M. Damas, H. Pomares, and I. Rojas, "Window size impact in human activity recognition," *Sensors,* vol. 14, no. 4, pp. 6474-6499, 2014.

[32] D. Borah, S. Wadhwa, U. Singh, S. L. Yadav, M. Bhattacharjee, and V. Sindhu, "Age related changes in postural stability," *Indian J Physiol Pharmacol,* vol. 51, no. 4, pp. 395-404, 2007.

[33] M. L. Callisaya, L. Blizzard, M. D. Schmidt, J. L. McGinley, and V. K. Srikanth, "Ageing and gait variability—a population-based study of older people," *Age and ageing,* vol. 39, no. 2, pp. 191-197, 2010.

[34] S. Mehdizadeh, "The largest Lyapunov exponent of gait in young and elderly individuals: A systematic review," *Gait & posture,* vol. 60, pp. 241-250, 2018.

[35] H. Kobayashi, W. Kakihana, and T. Kimura, “Combined effects of age and gender on gait symmetry and regularity assessed by autocorrelation of trunk acceleration,” *Journal of neuroengineering and rehabilitation,* vol. 11, no. 1, pp. 1-6, 2014.

[36] N. M. Kosse, N. Vuillerme, T. Hortobagyi, and C. J. C. Lamoth, “Multiple gait parameters derived from iPod accelerometry predict age-related gait changes,” *Gait & Posture,* vol. 46, pp. 112-117, May, 2016.

[37] M. S. Orendurff, J. A. Schoen, G. C. Bernatz, A. D. Segal, and G. K. Klute, “How humans walk: bout duration, steps per bout, and rest duration,” *Journal of Rehabilitation Research & Development,* vol. 45, no. 7, 2008.

[38] A. Adadi, and M. Berrada, “Peeking inside the black-box: a survey on explainable artificial intelligence (XAI),” *IEEE access,* vol. 6, pp. 52138-52160, 2018.

[39] N. B. Kumarakulasinghe, T. Blomberg, J. Liu, A. S. Leao, and P. Papapetrou, "Evaluating local interpretable model-agnostic explanations on clinical machine learning classification models." pp. 7-12.

[40] E. Warmerdam, J. M. Hausdorff, A. Atrsaei, Y. Zhou, A. Mirelman, K. Aminian, A. J. Espay, C. Hansen, L. J. Evers, and A. Keller, “Long-term unsupervised mobility assessment in movement disorders,” *The Lancet Neurology,* vol. 19, no. 5, pp. 462-470, 2020.

Appendix A.

Classification performance of DL approaches for different window sizes based on raw data.

Table 1. The performance metrics of DL for the 128 sample-window size

Models	Accuracy	Precision	Recall	F1-score	AUC
CNN	0.87	0.83	0.94	0.88	0.94
GRU	0.81	0.78	0.89	0.83	0.90
ConvLSTM	0.84	0.79	0.95	0.86	0.90

Table 2. The performance metrics of DL for the 256 sample-window size

Models	Accuracy	Precision	Recall	F1-score	AUC
CNN	0.88	0.83	0.96	0.89	0.93
GRU	0.87	0.82	0.95	0.88	0.94
ConvLSTM	0.88	0.84	0.95	0.89	0.93

Table 3. The performance metrics of DL for the 512 sample-window size

Models	Accuracy	Precision	Recall	F1-score	AUC
CNN	0.86	0.84	0.90	0.87	0.94
GRU	0.86	0.88	0.86	0.87	0.95
ConvLSTM	0.89	0.85	0.94	0.90	0.96

Table 4. The performance metrics of DL for the 2048 sample-window size

Models	Accuracy	Precision	Recall	F1-score	AUC
CNN	0.89	0.82	1	0.90	0.91
GRU	0.87	0.84	0.93	0.88	0.93
ConvLSTM	0.90	0.84	1	0.92	0.94

Table 5. The performance metrics of DL for the 5120 sample-window size

Models	Accuracy	Precision	Recall	F1-score	AUC
CNN	0.88	0.81	1	0.90	0.88
GRU	0.81	0.8	0.86	0.83	0.92
ConvLSTM	0.84	0.83	0.88	0.85	0.93

CNN: convolutional neural network; GRU: gate recurrent unit; ConvLSTM: convolutional long short-term memory.

Appendix B.

Classification performance of DL approaches for different window sizes based on gait normalization data.

Table 1. The performance metrics of DL for the 128 sample-window size

Models	Accuracy	Precision	Recall	F1-score	AUC
CNN	0.86	0.83	0.93	0.87	0.93
GRU	0.82	0.81	0.85	0.83	0.89
ConvLSTM	0.83	0.82	0.87	0.84	0.91

Table 2. The performance metrics of DL for the 256 sample-window size

Models	Accuracy	Precision	Recall	F1-score	AUC
CNN	0.88	0.83	0.95	0.89	0.94
GRU	0.87	0.82	0.96	0.89	0.94
ConvLSTM	0.84	0.82	0.89	0.86	0.92

Table 3. The performance metrics of DL for the 512 sample-window size

Models	Accuracy	Precision	Recall	F1-score	AUC
CNN	0.85	0.82	0.91	0.86	0.93
GRU	0.88	0.84	0.94	0.89	0.95
ConvLSTM	0.86	0.86	0.88	0.87	0.93

Table 4. The performance metrics of DL for the 1024 sample-window size

Models	Accuracy	Precision	Recall	F1-score	AUC
CNN	0.89	0.88	0.91	0.89	0.96
GRU	0.89	0.85	0.95	0.9	0.96
ConvLSTM	0.84	0.78	0.98	0.87	0.93

Table 5. The performance metrics of DL for the 2048 sample-window size

Models	Accuracy	Precision	Recall	F1-score	AUC
CNN	0.83	0.79	0.9	0.84	0.91
GRU	0.84	0.8	0.91	0.85	0.92
ConvLSTM	0.84	0.81	0.91	0.86	0.92

Table 6. The performance metrics of DL for the 5120 sample-window size

Models	Accuracy	Precision	Recall	F1-score	AUC
CNN	0.81	0.77	0.89	0.83	0.87
GRU	0.84	0.79	0.95	0.86	0.91
ConvLSTM	0.82	0.84	0.82	0.83	0.88

CNN: convolutional neural network; GRU: gate recurrent unit; ConvLSTM: convolutional long short-term memory.

Chapter 3

Explaining Deep Learning Models for Age-related Gait Classification Based on Acceleration Time Series

Xiaoping Zheng, Egbert Otten, Michiel F Reneman, Claudine JC Lamoth

Submitted for Publication

Abstract

Background:

Gait analysis holds significant importance in monitoring daily health, particularly among older adults. Advancements in sensor technology enable the capture of movement in real-life environments and generate big data. Machine learning, notably deep learning (DL), shows promise to use these big data in gait analysis. However, the inherent black-box nature of these models poses challenges for their clinical application. This study aims to enhance transparency in DL-based gait classification for aged-related gait patterns using Explainable Artificial Intelligence, such as SHapley Additive exPlanations (SHAP).

Methods:

A total of 244 subjects, comprising 129 adults and 115 older adults (age>65), were included. They performed a 3-minute walking task while one accelerometer was affixed to the lumbar segment L3. DL models, convolutional neural network (CNN) and gated recurrent unit (GRU), were trained using accelerations contained 1-stride and 8-stride, respectively, to classify adult and older adult groups. SHAP was employed to explain the models' predictions.

Results:

CNN achieved a satisfactory performance with an accuracy of 81.4% and an AUC of 0.89, and GRU demonstrated promising results with an accuracy of 84.5% and an AUC of 0.94. SHAP analysis revealed that both CNN and GRU assigned higher SHAP values to the data from vertical and walking directions, particularly emphasizing data around heel contact, spanning from the terminal swing to loading response phases. Furthermore, SHAP values indicated that GRU did not treat every stride equally.

Conclusion:

CNN accurately distinguished between adults and older adults based on the characteristics of a single stride's data. GRU achieved accurate classification by considering the relationships and subtle differences between strides. In both models, data around heel contact emerged as most critical, suggesting differences in acceleration and deceleration patterns during walking between different age groups.

Keywords: Accelerometers; Ageing; Deep learning; Gait classification; Machine learning; Explainable Artificially intelligence.

1. Introduction

Gait analysis plays a significant role in monitoring the quality of life, particularly among older individuals, since maintaining mobility and independence in later years is essential [1]. Gait performance can provide insights into the control and coordination of various systems, such as the neuromusculoskeletal system and the nervous system. Aging, as a continuous process, is often associated with the loss of muscle mass, decreased bone density, and declining nerve function, which can result in an altered gait pattern [2].

With the development of miniaturization of sensors (e.g., accelerometers), modern movement tracking systems can provide vast amounts of reliable data about human movements [3], allowing for the diagnosis, monitoring, and rehabilitation of gait patterns in daily living environments. Given the high variability, dimensionality, non-linear interactions, and temporal dependencies of the data collected during walking, traditional statistical approaches have limited capabilities [4]. Therefore, machine learning (ML) approaches have gained importance in clinical gait analysis due to their ability to handle complex data.

Machine learning has demonstrated promising results in clinical gait classification tasks. For example, Artificial Neural Network (ANN) achieved high accuracy (90%) in classifying different age groups gait based on handcrafted gait outcomes [5]. The design and selection of handcrafted gait outcomes require expert knowledge and are laborious. Deep learning (DL) can perform gait classification based on raw sensor signals and has demonstrated superior performance [6]. A recent study compared the classification performance of recurrent neural networks (bidirectional long short-term memory) and conventional machine learning (support vector machine (SVM) and linear regression) in classifying fallers and non-fallers in patients with multiple sclerosis [7]. The results of this study indicated that the deep learning approach outperformed conventional machine learning (area under the curve: 0.88 vs. 0.79 for SVM).

However, many machine learning models suffer from a lack of transparency and interpretability due to their black-box nature [8]. It is often unclear why a specific prediction has been made, even though the mathematical principles underlying these methods are well-established and well-understood. This opacity makes it challenging for patients and clinicians to trust the models, and strongly limits their practical applications in clinical contexts. Furthermore, this lack of transparency does not comply with the requirements of the European General Data Protection Regulation (GDPR, EU 2016/679) [9], which mandates the explanation of the logic behind any automated decision-making process that significantly affects individuals. Apart from this, the black-box nature makes it impossible to know what the model has truly learned, consequently obstructing the potential for generating new knowledge and a better understanding of human gait movement.

To overcome these limitations, Explainable Artificial Intelligence (XAI) has gained attention in the field of medicine. XAI is an approach aimed at revealing the reasoning behind a system's predictions and decisions, which becomes even more critical when handling sensitive and personal health data. XAI can be broadly categorized into two main categories based on the stage of use: 1) ante-hoc explainability; and 2) post-hoc explainability [10]. Ante-hoc explainability refers to simple models that are interpretable by design, such as linear regression models, decision trees, k-nearest neighbour models, and Bayesian models [11]. However, it is often assumed that ante-hoc explainable models do not achieve satisfactory performance; therefore, opaque models (such as deep learning) are frequently employed [11]. This leads to post-hoc explainability approaches, which can be used to explain a previously trained model or its prediction.

Layer-wise Relevance Propagation (LRP) [12, 13], Local Interpretable Model-Agnostic Explanations (LIME) [14], and SHapley Additive exPlanations (SHAP) [15] are the popular post-hoc explainability approaches. LRP propagates relevance scores from the output layer back to the input layer to determine the relevance of each input variable to the output decision. LIME perturbs the original data to observe how it affects predictions and aims to provide interpretable and faithful explanations, but it suffers from instability. It has been reported that two very close input samples may get greatly varied explanations in a simulated setting [16]. SHAP utilizes SHapley values to represent the contribution of each input variable to a certain prediction and this approach has strong theoretical backing and ensures that the contribution of each input variable is fairly and efficiently distributed among all the input variables of the instance [17]. This efficiency property distinguishes SHAP from other approaches and suggests that it might be the only approach that provides a full and fair explanation for the prediction of a machine learning model. This property may highlight the potential advantages of SHAP values in terms of fairness and legal compliance in certain situations.

The application of XAI approaches in deep learning-based clinical gait analysis is still in early stages. One study used LRP to explain a convolutional neural network (CNN) model in classifying individual gait patterns based on one stride data collected by ground reaction forces and full-body joint angles [12]. In another study, CNN was used to classify the walking of healthy subjects while performing four different dual tasks. The classification was based on data from no more than two strides, derived from ground reaction force and plastic optical fiber distributed sensors. [13]. In this study, LRP was used to indicate which parts of the signal had the heaviest influence on the gait classification. Both studies applied a CNN model-based classification, which is famous for extraction of local temporal and spatial features. However, since only 2 strides of data were used as input, the long-term dependent changes in gait, which provide information about postural control ability [18], were not considered.

The present study aims to improve the transparency of DL-based gait classification, with acceleration time series obtained during a 3-minute walking task as input. The goal is to differentiate between the gait patterns of adults and older adults. More specifically, we will: 1) employ a CNN and a gated recurrent unit (GRU) designed to learn long-term dependent features; 2) utilize the SHAP approach to indicate the importance of the input signal in classification for both models. This study can contribute to improving the transparency and interpretability of deep learning-based gait analysis and potentially lead to better clinical decision-making.

2. Methods

The overview of data acquisition and analysis in this study is presented in Fig. 1, with a CNN model being used as the example. The data were collected during a 3-minute walking task (Fig. 1(a)). Stride data from all subjects underwent preprocessing before being employed to train the CNN model which aimed to classify subjects into adult and older adult groups (Fig. 1(b)). To interpret the CNN model, the SHAP approach was applied. The SHAP values were visualized using a colour spectrum to illustrate the contributions of input data to the classification process (Fig. 1(c)). In this representation, deeper red indicates a higher contribution.

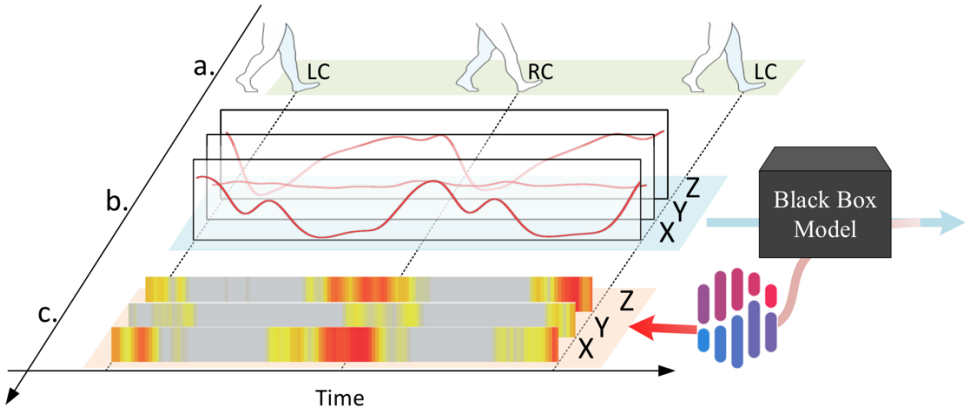


Figure 1. Overview of data acquisition and analysis of CNN. (a): walking data collection; (b): preprocessing stride data and training CNN based on one stride data; (c) interpreting the CNN model by SHAP, deeper red colour represents a higher contribution to the classification process. CNN: convolutional neural network; LC: left contact; RC: right contact

2.1. Subjects, equipment, and data collection

In this study, the dataset consisted of 386 subjects, which were derived by merging data from existing datasets [19-23]. Subjects with cognitive impairment (n=104) and those with insufficient walking data (n=38, having fewer than 10 segments of 8 consecutive strides) were excluded. The remaining subjects (n = 244) were divided into two groups: adults (ages 18-65, n=129) and older adults (ages >65, n=115). The mean ages were 38.3 (SD: 15.4) and 76.7 (SD:

5.9), respectively. During the study, subjects were instructed to walk at a comfortable speed for 3 minutes while wearing an accelerometer (iPod, Dynaport, or ActiGraph) fixed to a belt near the lumbar segment L3. When assuming a standing and upright position, the orientation of the axes was as follows: the X-axis pointed toward the ground (representing the vertical direction, V), the Y-axis faced the walking direction (indicating the anteroposterior direction, AP), and the Z-axis was perpendicular to the walking direction, extending from the patient's left to right (representing the mediolateral direction, ML). The sampling frequency was 100 Hz.

The studies were conducted between 2008 and 2022 and were approved by the Medical Ethical Committee of the University Medical Centre Groningen and the Medical Ethical Committee of the Slotervaart Hospital. All subjects provided written informed consent in accordance with the Declaration of Helsinki.

2.2. Data preprocessing and data splitting

To remove high-frequency noise, a second-order Butterworth low-pass filter with a cut-off frequency of 10 Hz was employed. The resulting filtered signal was normalized to a range of −1 to 1. A stride was defined as the period from the first left heel contact to the right contact and back to the second left heel contact. Heel contact was detected based on the peaks of both AP- and V-axis acceleration data, with the left or right foot determined by the values in the acceleration of ML direction. During the walking, the sensor sways with the movement of the body in the ML directions, and left heel contacts show higher readings in the ML direction than right contacts. To align the starting and ending timing of different strides, the data from each stride were interpolated to a uniform length of 128 samples. The segments, each with a length of 128, were employed in the CNN model, and for each subject, the initial 80 segments were chosen. In the case of the GRU model, 8 consecutive segments (stride) were merged into a singular segment comprising 1024 samples, yielding a total of 10 segments for each subject. The adult and older adult subjects were randomly and proportionally divided into training, testing, and validation sets at a ratio of 146:49:49. Their corresponding data were used as training, testing, and validation data set.

2.3. Classifiers

CNN and GRU were utilized in this study because of excellent capacity of CNN for local special and temporal feature extraction and the outperformance of GRU in learning long-term dependent features [24].

The interpolated one-stride segments for CNN and interpolated eight-stride segments for GRU were organized in x-, y-, and z-axis order, to generate a single signal data with 3 channels (128*3 and 1024*3 respectively). The optimal hyperparameters for both the CNN and GRU

models were tuned by Bayesian Optimization (BO) [25]. Unlike conventional techniques such as randomized search cross-validation, BO considers the prior performance of the hyperparameters and updates them to achieve better performance. This allows BO to find the global optimum with a minimum number of steps. For each model, 15 parameter combinations were tested. Detailed information about the hyperparameter space settings can be referenced in Table 1, while the learning rate, which was also optimized using BO, was configured within the range of [1e-5, 1e-2].

Table 1. Hyperparameters space for CNN and GRU

	Layer		CNN	GRU
Input			128X3	1024X3
Stack (Stack number: [1,3])	Deep layer	Layer name	Conv1D	GRU
		Unit/ Filter	[2, 768]	
		Kernel size	[1, 15]	-
		Activation	"ReLU"	"tanh"
	Batch Normalization		-	
	Pooling		-	
	Dropout	Rate	Nan or [0.1, 0.9]	
	Dense	Unit	[2, 768]	
	Flatten		-	
	Dropout	Rate	Nan or [0.1, 0.9]	
Output	Dense		2 units and "softmax" activation	

CNN: convolutional neural network; GRU: gate recurrent unit; ReLU: rectified linear unit; tanh: hyperbolic tangent function.

2.4. Evaluation

The assessment of classification performance was conducted using widely recognized evaluation measures such as accuracy, recall (sensitivity), precision, and F1 score (the harmonic mean of sensitivity and precision). Receiver operating characteristic (ROC) curves were generated and the area under the curve (AUC) was calculated as well.

To ensure transparency and reproducibility of the findings, we have made the project repository publicly accessible at https://github.com/xzheng93/Explainable_DL. The repository contains the source code, dataset, log files of experiments.

2.5. SHAP

The SHAP approach [17] was used to explain the prediction of a signal segment x in the given model f (CNN or GRU) based on the SHapley values from coalitional game theory. The original input segment x was mapped through the function $h_x(z')$ to get the input for the SHAP explanation $g(\bullet)$. $z' \in \{0,1\}^N$, where N is the number of features (data points or sets of data points in the signal segment) of x and, 0 and 1 mean the absence or presence of features in x. Applying to the model f: $f(h_x(z'))$, the SHAP explanation [17] can be defined as:

$$g(z') = f(h_x(z')) = \emptyset_0 + \sum_{i=1}^{N} \emptyset_i z_i'$$

where $\emptyset_i$ is the SHapley value of a feature i in the segment x.
The definition of SHapley values $\emptyset_i$ is as follows:

$$\emptyset_i = \frac{1}{|N|!} \sum_{\{i\} \in s \; and \; s \subseteq N} (|s| - 1)! \, (|N| - |s|)! \, [f(s) - f(s - \{i\})]$$

where s is the segment which data features i is present. $|\bullet|$ represents the length of a segment except absent features. The definition of SHapley value ensures the efficiency, symmetry, dummy, and additivity properties.

The efficiency property can represented as:

$$\sum_{i \in N} \emptyset_i = f(x)$$

The sum of the SHapley values of all separated features equals the value of the coalition of all the features (the whole signal segment). Therefore, all the gain is distributed among the segment.

The symmetry property means that if the contributions of two features i and j are equal, they will contribute equally to all possible coalitions. This can be represented as, if $\emptyset_i = \emptyset_j$, then

$$f(s \cup \{i\}) = f(s \cup \{j\})$$

where $s \subseteq N$ and $\{i, j\} \notin s$.

The dummy property entails that a feature i does not change the predicted value:

$$f(s \cup \{i\}) = f(s)$$

Then its SHapley value $\emptyset_i$ equals 0.

Regarding the additivity property, if a coalition game employs two gain functions f' and $f^{"}$, the SHapley values are additive:

$$\emptyset_i(f' + f^{"}) = \emptyset_i(f') + \emptyset_i(f^{"})$$

The SHapley value is built based on a solid theory. The properties of SHapley value give the explanation a reasonable foundation and distinguish the SHAP from other methods such as LIME [26]. The SHAP explanation might be the only legally compliant method to meet the law requirement of GDPR [26].

3. Results

The optimal hyperparameters and architecture for both CNN and GRU were determined through BO, and the results are presented in Fig. 2 and Fig. 3. In the case of CNN, the training

dataset comprised 11680 (146*80) one-stride data (segments), the testing dataset included 3920 (49*80) segments, and the validation dataset contained 3920 (49*80) one-stride segments. For GRU, the training dataset consisted of 1460 eight-stride segments (146*10), the testing dataset encompassed 490 eight-stride segments (49*10), and the validation dataset comprised 490 eight-stride segments (49*10).

The CNN architecture included three 1D convolutional layers followed by batch normalization, max-pooling, and dropout layers. The first convolutional layer consisted of 88 filters with a kernel size of 13, while the second convolutional layer comprised 336 filters with a kernel size of 5. The third convolutional layer contained only 2 filters with a kernel size of 1. The dropout rates were 0.3, 0.6, and 0 for the first, second, and third dropout layers, respectively. A dense layer with 74 units, a fully connected layer, and a dropout layer with a rate of 0.5 were also incorporated into the architecture. Finally, a softmax activation function was employed for classification. The Adam optimization was used, and the optimal learning rate of 0.0015 was discovered via BO. The model was trained for 150 epochs with early stopping based on the validation accuracy with a patience of 20 epochs.

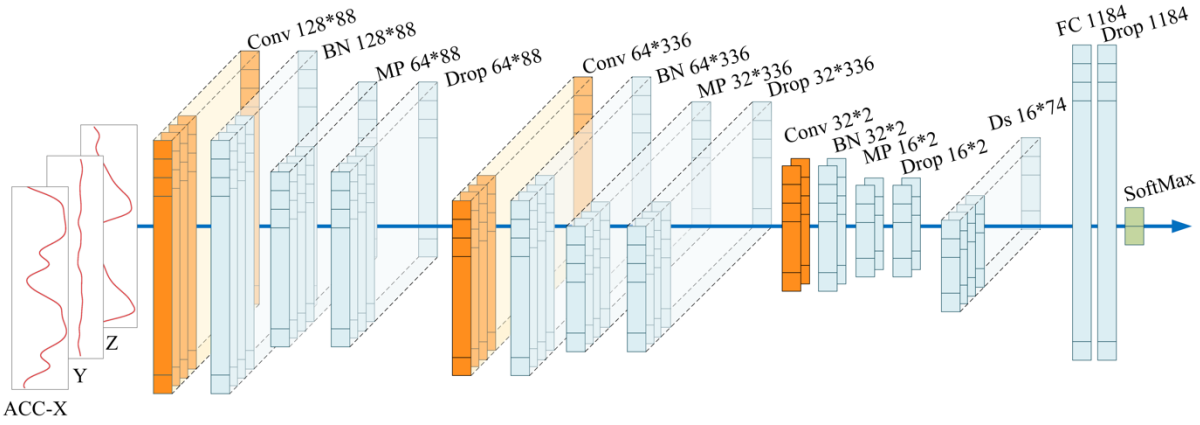


Figure 2. The architecture and optimal hyperparameter of CNN. ACC: acceleration; CNN: convolutional neural network; Conv: 1-dimension convolutional layer (in orange); BN: batch normalization layer; MP: max-pooling layer; Drop: dropout layer; Ds: dense layer; SoftMax: softmax activation (in green).

Fig. 3 graphically illustrates the architecture and optimal hyperparameters of the proposed GRU model. The GRU architecture included three GRU layers, each followed by batch normalization, max-pooling, and dropout layers for regularization. The first GRU layer had 666 filters, the second had 438 filters, and the third had 2 filters. The three dropout layers had a rate of 0.5, 0.7, and 0, respectively. A dense layer with 676 units and a dropout layer with a rate of 0.1 were also incorporated into the architecture. The final layer of the GRU consisted of a fully connected layer followed by a softmax activation function for classification. A graphical representation of the GRU architecture is shown in Fig. 3. Adam optimization was

employed with a learning rate of 0.0003. The same training epoch setting and early stopping as for CNN were used.

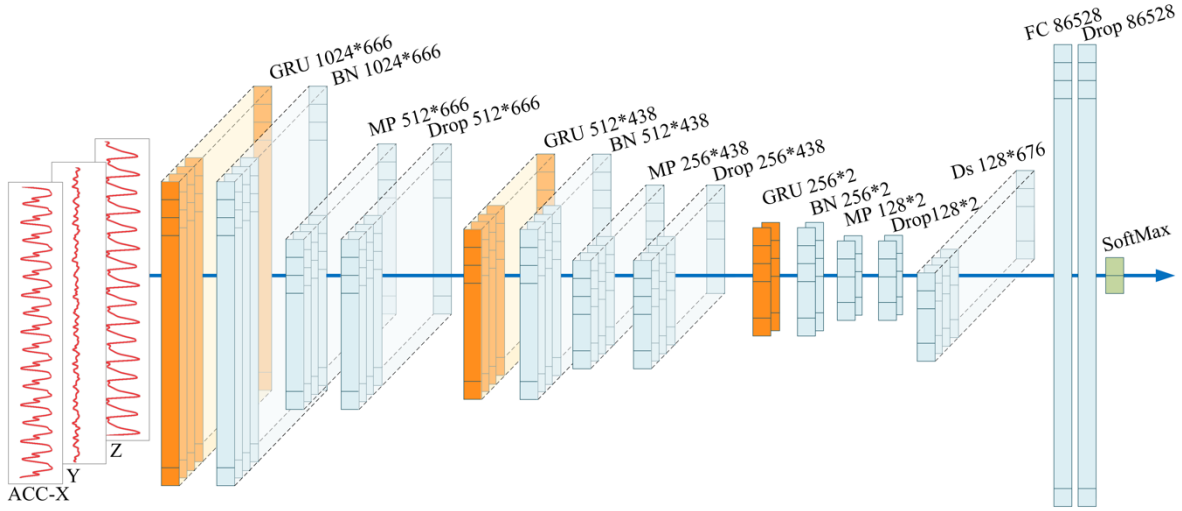


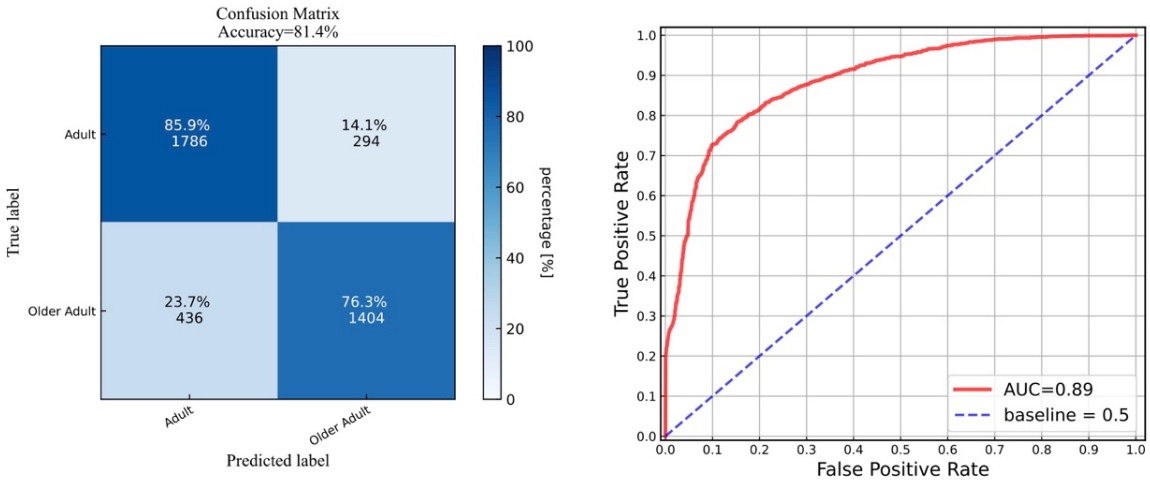
Figure 3. The architecture and optimal hyperparameter of GRU. ACC: acceleration; GRU: gate recurrent unit layer (in orange); BN: batch normalization layer; MP: max-pooling layer; Drop: dropout layer; Ds: dense layer; SoftMax: softmax activation (in green).

Based on the testing data, the classification performance of the CNN is summarized in Table 2, achieving an accuracy of 81.4%, precision of 82.7%, recall of 76.3%, F1-score of 79.3%, and an AUC of 0.89. Detailed classification results for the CNN model, including the confusion matrix and ROC curve, are presented in Fig. 4. In the adult group, 85.9% of data samples were correctly classified, while 14.1% were incorrectly classified as older adults. In the older adult group, 76.3% of the samples were correctly classified, while 23.7% were incorrectly classified as adults.

Table 3 displays the classification performance of the GRU, with further details provided in Fig. 5. The GRU model achieved an accuracy of 84.5%, precision of 79.4%, recall of 90.4%, F1-score of 84.6%, and an AUC of 0.94. The confusion matrix in Fig. 5(a) illustrates that 79.2% of adults and 90.4% of older adults were correctly classified.

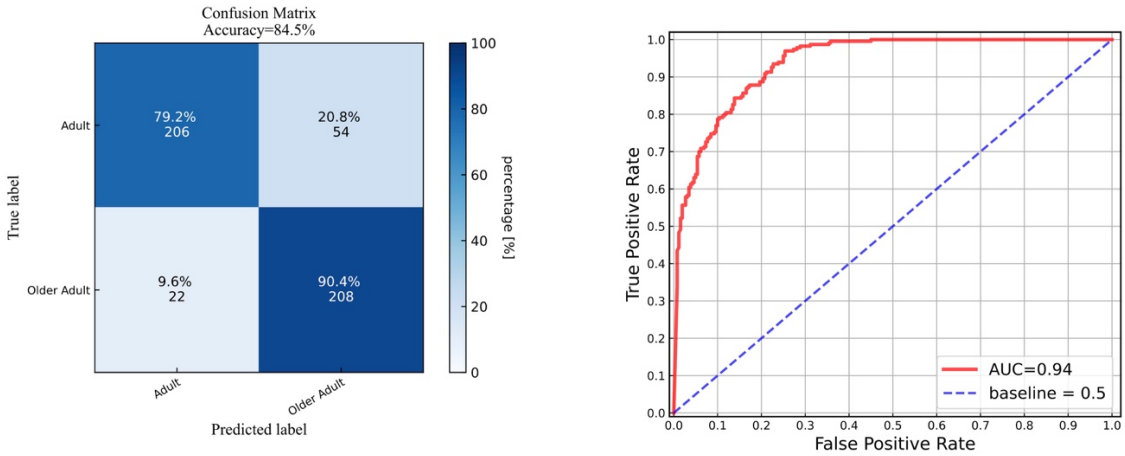
After the evaluation, the mean absolute SHAP values for the testing data of both CNN and GRU models were computed and visualized in Fig. 6. In this representation, a deeper red colour signifies a greater contribution to the classification process.

In Fig. 6(a), an abundance of red colour is observed in the V and AP directions, particularly around heel contact. A similar pattern is evident in Fig. 6(b), with a prevalence of red colour in the V and AP directions, centered around heel contact event. Notably, Fig. 6(b) reveals that not all gait cycles are equally significant, as some exhibit a higher degree of red colour, indicating greater importance in the classification process.



(a) Confusion matrix (b) Receiver operating characteristic curve

Figure 4. (a) Confusion matrix and (b) Receiver operating characteristic curve for CNN. CNN: convolutional neural network; AUC: area under the curve.



(a) Confusion matrix (b) Receiver operating characteristic curve

Figure 5. (a) Confusion matrix and (b) Receiver operating characteristic curve for GRU. GRU: gate recurrent unit layer; AUC: area under the curve.

Table 2. The performance metrics of CNN

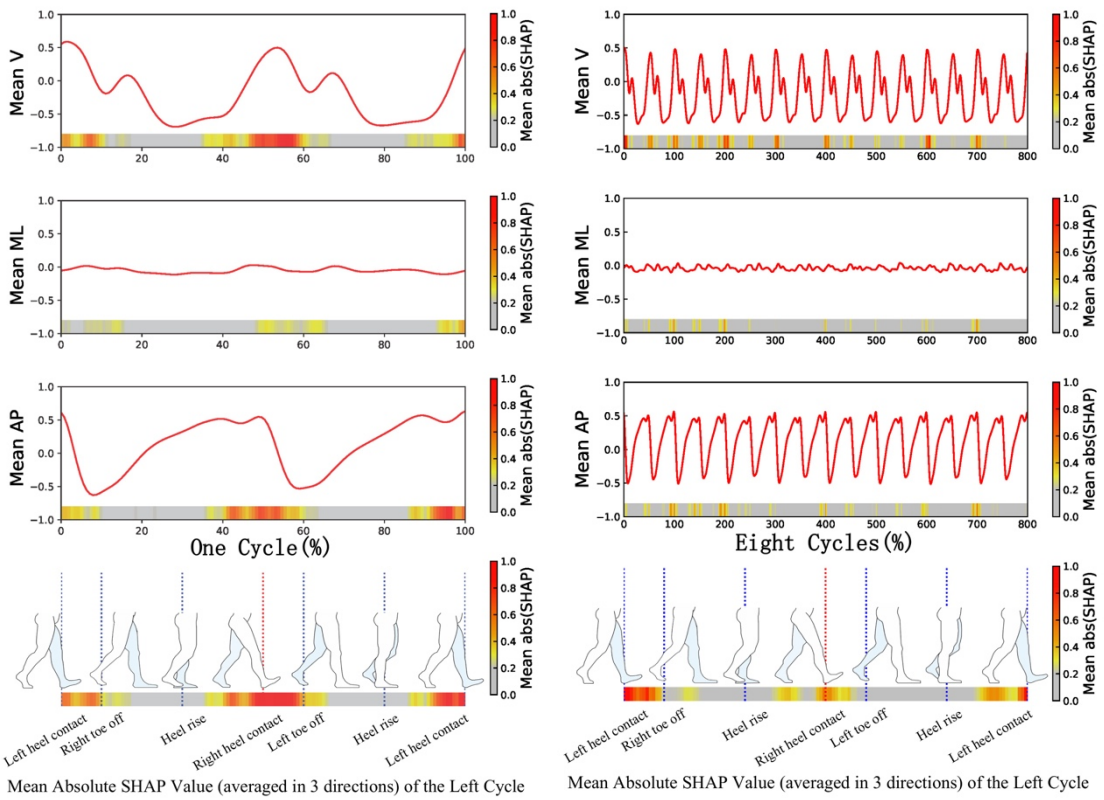
	Accuracy	Precision	Recall	F1-score	AUC
CNN	81.4%	82.7%	76.3%	79.3%	0.89

CNN: convolutional neural network; AUC: area under the curve.

Table 3. The performance metrics of GRU

	Accuracy	Precision	Recall	F1-score	AUC
GRU	84.5%	79.4%	90.4%	84.6%	0.94

GRU: gate recurrent unit layer; AUC: area under the curve.



(a) CNN (b) GRU

Figure 6. SHAP values results of (a) CNN and (b) GRU. CNN: convolutional neural network; GRU: gate recurrent unit; V: vertical direction; AP: anteroposterior direction; ML: mediolateral direction; SHAP: SHapley Additive exPlanations. The acceleration data in each panel were normalized for both amplitude (ranging from -1 to 1) and time (using left heel contact as the reference point).

4. Discussion

The primary aim of this study is to increase the transparency of non-linear DL models in gait analysis. For this purpose, state-of-the-art DL models, specifically CNN and GRU, were explained by using a cutting-edge XAI approach (SHAP). These models are applied to the task of classifying individuals into two distinct groups: adults and older adults, based on acceleration time series data collected during 3 minutes walking. The results indicate that the CNN model achieved satisfying classification performance, with an accuracy of 81.4% and an AUC of 0.89, despite being trained on data from one stride. The GRU model exhibited promising classification capabilities, achieving an accuracy of 84.5% and an AUC of 0.94, utilizing eight-stride data. To attain an understanding and interpretation of the proposed DL models in the context of gait classification, the SHAP approach was employed. The SHAP values shed light on the models' decision-making processes, revealing a predominant reliance on acceleration data from the AP and V directions, rather than the ML direction, for the classification task. Specifically, data surrounding gait events such as heel contact in the AP and V directions emerged as the most influential inputs contributing to the differentiation between adults and older adults.

In this study, CNN and GRU were employed. CNN is renowned for its exceptional capacity to extract local spatial-temporal features. It has the potential to capture time-independent gait features, such as root mean square (indicative of gait intensity), rhythm (reflecting the proportion of stance and swing phases), and harmonic index (measuring the smoothness of the acceleration curve). These features have previously been successfully utilized in characterizing age-related gait differences in other studies [5, 27]. The promising accuracy achieved by CNN suggests that even single stride data contains rich information that can effectively distinguish age-related gait patterns. The SHAP values underscore that data corresponding to the heel contact event play an important role in this discrimination. On the other hand, GRU is designed to capture both short- and long-term dependent features. It may learn time-dependent gait features that can reflect the intricate relationships and subtle differences between gait cycles. Time-dependent gait features, including regularity, variability, local stability (as measured by the largest Lyapunov exponent), gait symmetry (using the symmetry index), and complexity (evaluated through sample entropy), are crucial in age-related gait classification. These features offer insights into changes in postural control that arise due to aging [28, 29]. The SHAP values presented in Fig. 6 (b) highlight that not all gait cycles were treated equally by GRU, as some gait cycles exhibit higher SHAP values. This observation suggests that GRU takes the relationships and slight variations between gait cycles into account when classifying individuals into the adult and older adult groups.

Aging is an ongoing process often accompanied by a gradual decline in balance [30]. Changes within the neuromusculoskeletal system, such as the loss of muscle fibers and reduced muscle force production, can result in diminished muscle strength and flexibility [31]. These alterations may impact functionality and contribute to reduced mobility. Furthermore, the sensory systems that are critical for effective postural control, including the visual, vestibular, and proprioceptive systems, tend to deteriorate with age, further affecting one's balance [32]. However, the SHAP results from this study indicate that DL models predominantly rely on data from the AP and V directions, instead of the ML direction, which is more closely associated with balance capacity. This observation aligns with a recent study, emphasizing the significance of dynamic gait parameters in the AP and V directions for classifying age-related gait patterns [5]. These parameters include Root Mean Square in AP and V directions, Lyapunov Exponent in the V direction, step regularity in the V direction, Cross Entropy in both V and ML directions, and gait speed when utilizing artificial neural networks for classification [5]. Given that the study [5] only utilized one accelerometer, it is possible that the similar results may be attributed to the limited sensitivity of a single accelerometer in detecting balance-related postural control information. An additional explanation for the limited contribution of ML direction data to the classification process could be attributed to the simplicity of the task undertaken in this study. Subjects engaged in a 3-minute walking task within a clean and well-lit hallway, walking at their preferred pace without any perturbations.

Consequently, this task may not effectively capture the variations in balance capacity between adults and older adults in more challenging real-life environments, which may explain why ML-direction data did not play as a significant role in the classification process as AP and V direction data.

The colour spectrum of SHAP values in Fig. 6 not only reveals which axes contribute more to the classification process but also identifies specific gait events that play a crucial role. It shows that data spanning from the terminal swing to the loading response phase consistently yield higher SHAP values, particularly around the event of heel contact. The terminal swing is a phase before the heel contacts the ground. It involves the final preparations of the leg and foot for ground contact. For example, muscles around the ankle and knee are activated to ensure that the limb is prepared to provide the necessary stability during heel contact and loading response [33]. During this phase, the leg starts to slow down its forward swing to prevent excessive force upon heel strike. The acceleration, during this phase, in the walking direction (AP) has no obvious increase. The loading response phase starts with the initial contact of the heel with the ground and ends with toe-off of the opposite limb. After the heel contact, the body undergoes shock absorption and weight acceptance as the body's weight is transferred onto the stance limb [34]. During this phase, the acceleration readings in the AP direction reach their maximum around heel contact, and sharply decrease to reach their minimum around toe-off. The acceleration readings around the terminal swing phase effectively represent how an individual prepares for deceleration, the moment of heel contact illustrates the process of deceleration, and the toe-off event signifies the initiation of acceleration [34].

The acceleration readings disparities observed by SHAP in these gait events/phases may indicate that adults and older adults exhibited different acceleration and deceleration patterns during walking. It can be attributed to the changes in kinematic and kinetic factors associated with aging which were observed by previous studies. Research has indicated that older adults tend to exhibit a reduced knee extension angle [35] and moment [36] at the point of heel contact, which can be closely linked to weaker muscles, such as the quadriceps [37, 38]. These changes may result in a reduced absorption force in the knee joint [39]. These alterations can lead to compensatory gait adjustments aimed at alleviating joint discomfort or stiffness, particularly in the hip and knee joints. Compared to adults, older adults exhibit limited capabilities in limb advancement during the push-off period. Research shows that, during the loading response phase, older adults often demonstrate reduced hip extension and moment [36, 40-42], particularly during the toe-off phase. This reduction may be indicative of decreased power in the hip extensors [41, 43, 44], which could imply weakness in swinging and kicking the lower limbs to generate forward propulsive force while walking. Furthermore, older adults tend to exhibit decreased independent movement of the

subtendons [45] and reduced plantar flexor moments [42], contributing to lower propulsive power at the ankle [46]. The current study has insufficient data to examine whether these kinematic and kinetic factors are responsible for the distinct acceleration and deceleration patterns. Further studies are necessary in this regard.

XAI, such as the SHAP approach, provides explainability results based on input and model output data. Alterations in input signals can yield divergent outputs, and these changes may be influenced not only by aging but also by independent parameters, such as sensor brands. This study utilized three different types of accelerometers which have different dynamic ranges and accuracy. To minimize potential biases introduced by these independent parameters in prediction explanations, signal amplitude standardization and gait cycle normalization were used. It is important to note that while these techniques mitigate the bias, they may inadvertently remove valuable information, such as information related to gait intensity and walking speed. Although a prior study has demonstrated that gait cycle normalization has only a marginal impact on age-related gait patterns classification performance [6]. Notably, XAI provides explanations based on correlations and associations within the data rather than revealing causal relationships. Consequently, explanations offered by XAI may not always align with human intuition or domain expertise. These disparities can offer novel insights, or lead to misunderstandings or mistrust, particularly in critical domains like healthcare. Unfortunately, a ground truth for evaluating the quality of XAI explanations remains absent. Hence, the explainability results should be interpreted cautiously. Additionally, it is worth mentioning that the SHAP approach, while effective, can be computationally demanding and may not be optimally scalable for large datasets or real-time applications. These computational constraints limit their practical use.

5. Conclusion

The present study enhances the transparency and interpretability of the proposed DL in gait analysis by incorporating the SHAP approach. The results demonstrate that CNN can accurately distinguish between adults and older adults based on data from one single stride. The key factors contributing to this classification were the accelerations around heel contact in the AP and V directions. GRU also exhibited promising classification performance, leveraging data from eight consecutive strides. The SHAP results from GRU suggest that it may capture the relationships and subtle variations between gait cycles, particularly the accelerations around heel contact in the AP and V directions. These findings imply that adults and older adults exhibit distinct acceleration and deceleration patterns during 3 minutes of walking.

This study underscores the potential of methods that enable understanding and interoperation of machine learning predictions, such as SHAP, in advancing the application of

machine learning in gait analysis. Consequently, XAI holds the promise of facilitating the implementation of machine learning-based decision-support systems in clinical practice.

References

[1] A. Hanley, C. Silke, and J. Murphy, “Community-based health efforts for the prevention of falls in the elderly,” *Clinical interventions in aging*, pp. 19-25, 2011.

[2] M. Intriago, G. Maldonado, R. Guerrero, O. Messina, and C. Rios, “Bone mass loss and Sarcopenia in Ecuadorian patients,” *Journal of Aging Research,* vol. 2020, 2020.

[3] D. Kobsar, J. M. Charlton, C. T. Tse, J.-F. Esculier, A. Graffos, N. M. Krowchuk, D. Thatcher, and M. A. Hunt, “Validity and reliability of wearable inertial sensors in healthy adult walking: A systematic review and meta-analysis,” *Journal of neuroengineering and rehabilitation,* vol. 17, pp. 1-21, 2020.

[4] T. Chau, “A review of analytical techniques for gait data. Part 1: fuzzy, statistical and fractal methods,” *Gait & posture,* vol. 13, no. 1, pp. 49-66, 2001.

[5] Y. Zhou, R. Romijnders, C. Hansen, J. van Campen, W. Maetzler, T. Hortobagyi, and C. J. C. Lamoth, “The detection of age groups by dynamic gait outcomes using machine learning approaches,” *Scientific Reports,* vol. 10, no. 1, Mar, 2020.

[6] X. Zheng, E. Wilhelm, M. F. Reneman, E. Otten, and C. J. Lamoth, “Age-related Gait Patterns Classification Using Deep Learning Based on Time-series Data from One Accelerometer,” 2023.

[7] B. M. Meyer, L. J. Tulipani, R. D. Gurchiek, D. A. Allen, L. Adamowicz, D. Larie, A. J. Solomon, N. Cheney, and R. S. McGinnis, “Wearables and deep learning classify fall risk from gait in multiple sclerosis,” *IEEE journal of biomedical and health informatics,* vol. 25, no. 5, pp. 1824-1831, 2020.

[8] A. Adadi, and M. Berrada, “Peeking inside the black-box: a survey on explainable artificial intelligence (XAI),” *IEEE access,* vol. 6, pp. 52138-52160, 2018.

[9] P. Regulation, “Regulation (EU) 2016/679 of the European Parliament and of the Council,” *Regulation (eu),* vol. 679, pp. 2016, 2016.

[10] T. Speith, "A review of taxonomies of explainable artificial intelligence (XAI) methods." pp. 2239-2250.

[11] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, and R. Benjamins, “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Information fusion,* vol. 58, pp. 82-115, 2020.

[12] F. Horst, S. Lapuschkin, W. Samek, K.-R. Müller, and W. I. Schöllhorn, “Explaining the unique nature of individual gait patterns with deep learning,” *Scientific reports,* vol. 9, no. 1, pp. 1-13, 2019.

[13] A. S. Alharthi, A. J. Casson, and K. B. Ozanyan, “Spatiotemporal analysis by deep learning of gait signatures from floor sensors,” *IEEE Sensors Journal,* vol. 21, no. 15, pp. 16904-16914, 2021.

[14] C. Dindorf, W. Teufl, B. Taetz, G. Bleser, and M. Fröhlich, “Interpretability of input representations for gait classification in patients after total hip arthroplasty,” *Sensors,* vol. 20, no. 16, pp. 4385, 2020.

[15] X. Zheng, M. F. Reneman, J. A. Echeita, R. H. S. Preuper, H. Kruitbosch, E. Otten, and C. J. Lamoth, “Association between central sensitization and gait in chronic low back pain: Insights from a machine learning approach,” *Computers in biology and medicine,* vol. 144, pp. 105329, 2022.

[16] D. Alvarez-Melis, and T. S. Jaakkola, “On the robustness of interpretability methods,” *arXiv preprint arXiv:1806.08049*, 2018.

[17] S. M. Lundberg, and S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in neural information processing systems,* vol. 30, 2017.

[18] D. Borah, S. Wadhwa, U. Singh, S. L. Yadav, M. Bhattacharjee, and V. Sindhu, “Age related changes in postural stability,” *Indian J Physiol Pharmacol,* vol. 51, no. 4, pp. 395-404, 2007.

[19] L. Kikkert, N. Vuillerme, J. P. van Campen, B. A. Appels, T. Hortobagyi, and C. J. Lamoth, “Gait characteristics and their discriminative power in geriatric patients with and without cognitive impairment,” *Journal of Neuroengineering and Rehabilitation,* vol. 14, Aug, 2017.

[20] C. J. Lamoth, F. J. van Deudekom, J. P. van Campen, B. A. Appels, O. J. de Vries, and M. Pijnappels, “Gait stability and variability measures show effects of impaired cognition and dual tasking in frail people,” *Journal of neuroengineering and rehabilitation,* vol. 8, no. 1, pp. 1-9, 2011.

[21] N. M. Kosse, S. Caljouw, D. Vervoort, N. Vuillerme, and C. J. Lamoth, “Validity and reliability of gait and postural control analysis using the tri-axial accelerometer of the iPod touch,” *Annals of biomedical engineering,* vol. 43, no. 8, pp. 1935-1946, 2015.

[22] M. H. de Groot, H. C. van der Jagt-Willems, J. P. van Campen, W. F. Lems, J. H. Beijnen, and C. J. Lamoth, “A flexed posture in elderly patients is associated with impairments in postural control during walking,” *Gait & posture,* vol. 39, no. 2, pp. 767-772, 2014.

[23] T. IJmker, and C. J. Lamoth, “Gait and cognition: the relationship between gait stability and variability with executive function in persons with and without dementia,” *Gait & posture,* vol. 35, no. 1, pp. 126-130, 2012.

[24] F. Gu, M.-H. Chung, M. Chignell, S. Valaee, B. Zhou, and X. Liu, “A survey on deep learning for human activity recognition,” *ACM Computing Surveys (CSUR),* vol. 54, no. 8, pp. 1-34, 2021.

[25] J. Wu, X.-Y. Chen, H. Zhang, L.-D. Xiong, H. Lei, and S.-H. Deng, "Hyperparameter optimization for machine learning models based on Bayesian optimization," *Journal of Electronic Science and Technology,* vol. 17, no. 1, pp. 26-40, 2019.

[26] K. Aas, M. Jullum, and A. Løland, "Explaining individual predictions when features are dependent: More accurate approximations to Shapley values," *Artificial Intelligence,* vol. 298, pp. 103502, 2021.

[27] N. M. Kosse, N. Vuillerme, T. Hortobagyi, and C. J. C. Lamoth, "Multiple gait parameters derived from iPod accelerometry predict age-related gait changes," *Gait & Posture,* vol. 46, pp. 112-117, May, 2016.

[28] K. K. Patterson, N. K. Nadkarni, S. E. Black, and W. E. McIlroy, "Gait symmetry and velocity differ in their relationship to age," *Gait & posture,* vol. 35, no. 4, pp. 590-594, 2012.

[29] D. Kobsar, C. Olson, R. Paranjape, T. Hadjistavropoulos, and J. M. Barden, "Evaluation of age-related differences in the stride-to-stride fluctuations, regularity and symmetry of gait using a waist-mounted tri-axial accelerometer," *Gait & posture,* vol. 39, no. 1, pp. 553-557, 2014.

[30] R. W. Bohannon, P. A. Larkin, A. C. Cook, J. Gear, and J. Singer, "Decrease in timed balance test scores with aging," *Physical therapy,* vol. 64, no. 7, pp. 1067-1070, 1984.

[31] C. A. Laughton, M. Slavin, K. Katdare, L. Nolan, J. F. Bean, D. C. Kerrigan, E. Phillips, L. A. Lipsitz, and J. J. Collins, "Aging, muscle activity, and balance control: physiologic changes associated with balance impairment," *Gait & posture,* vol. 18, no. 2, pp. 101-108, 2003.

[32] T. Coelho, Â. Fernandes, R. Santos, C. Paúl, and L. Fernandes, "Quality of standing balance in community-dwelling elderly: Age-related differences in single and dual task conditions," *Archives of gerontology and geriatrics,* vol. 67, pp. 34-39, 2016.

[33] A. Schmitz, A. Silder, B. Heiderscheit, J. Mahoney, and D. G. Thelen, "Differences in lower-extremity muscular activation during walking between healthy older and young adults," *Journal of electromyography and kinesiology,* vol. 19, no. 6, pp. 1085-1091, 2009.

[34] A. Kharb, V. Saini, Y. Jain, and S. Dhiman, "A review of gait cycle and its parameters," *IJCEM International Journal of Computational Engineering & Management,* vol. 13, pp. 78-83, 2011.

[35] E. Chung, S.-H. Lee, H.-J. Lee, and Y.-H. Kim, "Comparative study of young-old and old-old people using functional evaluation, gait characteristics, and cardiopulmonary metabolic energy consumption," *BMC geriatrics,* vol. 23, no. 1, pp. 1-11, 2023.

[36] D. C. Kerrigan, M. K. Todd, U. Della Croce, L. A. Lipsitz, and J. J. Collins, "Biomechanical gait alterations independent of speed in the healthy elderly: evidence for specific limiting impairments," *Archives of physical medicine and rehabilitation,* vol. 79, no. 3, pp. 317-322, 1998.

[37] B. Holm, M. T. Kristensen, J. Bencke, H. Husted, H. Kehlet, and T. Bandholm, “Loss of knee-extension strength is related to knee swelling after total knee arthroplasty,” *Archives of physical medicine and rehabilitation,* vol. 91, no. 11, pp. 1770-1776, 2010.

[38] J. W. Kwon, S. M. Son, and N. K. Lee, “Changes of kinematic parameters of lower extremities with gait speed: a 3D motion analysis study,” *Journal of Physical Therapy Science,* vol. 27, no. 2, pp. 477-479, 2015.

[39] M. Bendall, E. Bassey, and M. Pearson, “Factors affecting walking speed of elderly people,” *Age and ageing,* vol. 18, no. 5, pp. 327-332, 1989.

[40] C. F. Oliveira, E. R. Vieira, F. M. Machado Sousa, and J. P. Vilas-Boas, “Kinematic changes during prolonged fast-walking in old and young adults,” *Frontiers in medicine,* vol. 4, pp. 207, 2017.

[41] W. S. Kim, and E. Y. Kim, “Comparing self-selected speed walking of the elderly with self-selected slow, moderate, and fast speed walking of young adults,” *Annals of rehabilitation medicine,* vol. 38, no. 1, pp. 101-108, 2014.

[42] S.-u. Ko, J. M. Hausdorff, and L. Ferrucci, “Age-associated differences in the gait pattern changes of older adults during fast-speed and fatigue conditions: results from the Baltimore longitudinal study of ageing,” *Age and ageing,* vol. 39, no. 6, pp. 688-694, 2010.

[43] D. A. Winter, *Biomechanics and motor control of human gait: normal, elderly and pathological*, 1991.

[44] P. Morfis, and M. Gkaraveli, “Effects of aging on biomechanical gait parameters in the healthy elderly and the risk of falling,” *Journal of Research & Practice on the Musculoskeletal System (JRPMS),* vol. 5, no. 2, 2021.

[45] J. R. Franz, and D. G. Thelen, “Imaging and simulation of Achilles tendon dynamics: implications for walking performance in the elderly,” *Journal of Biomechanics,* vol. 49, no. 9, pp. 1403-1410, 2016.

[46] K. Rasske, and J. R. Franz, “Aging effects on the Achilles tendon moment arm during walking,” *Journal of biomechanics,* vol. 77, pp. 34-39, 2018.

Chapter 4

Association between Central Sensitization and Gait in Chronic Low Back Pain: Insights from a Machine Learning Approach

Xiaoping Zheng, Michiel F Reneman, Jone Ansuategui Echeita, Rita HR Schiphorst Preuper, Herbert Kruitbosch, Egbert Otten, Claudine JC Lamoth

Computers in Biology and Medicine (2022) 144, 105329

Abstract

Background: Central sensitization (CS) is often present in patients with chronic low back pain (CLBP). Gait impairments due to CLBP have been extensively reported; however, the association between CS and gait is unknown. The present study examined the association between CS and CLBP on gait during activities of daily living.

Method: Forty-two patients with CLBP were included. CS was assessed through the Central Sensitization Inventory (CSI), and patients were divided in a low and high CS group (23 CLBP- and 19 CLBP+, respectively). Patients wore a tri-axial accelerometer device for one week. From the acceleration signals, gait cycles were extracted and 36 gait outcomes representing quantitative and qualitative characteristics of gait were calculated. A Random Forest was trained to classify CLBP- and CLBP+ based on the gait outcomes. The maximum Youden index was computed to measure the diagnostic test's ability and SHapley Additive exPlanations (SHAP) indexed the gait outcomes' importance to the classification model.

Results： The Random Forest accurately (84.4%) classified the CLBP- and CLBP+. Youden index was 0.65, and SHAP revealed that the gait outcomes important to the classification model were related to gait smoothness, stride frequency variability, stride length variability, stride regularity, predictability, and stability.

Conclusions: CLBP- and CLBP+ patients had different motor control strategies. Patients in the CLBP- group presented with a more “loose control”, with higher gait smoothness and stability, while CLBP+ patients presented with a “tight control”, with a more regular, less variable, and more predictable gait pattern.

Keywords： Low back pain, Central sensitization, Supervised machine learning, Gait, Daily life.

1. Introduction

Chronic low back pain (CLBP) is one of the most prevalent chronic musculoskeletal pains [1]. It is responsible for high treatment costs, sick leave and individual suffering and it represents a significant socioeconomic burden [2]. For 85%–90% of patients with CLBP, the relation between pathoanatomical and clinical presentations is absent [3] and, therefore, it is classified as nonspecific CLBP [4]. In CLBP, and other chronic musculoskeletal disorders, central sensitization (CS) might be present (reviewed in Ref. [5]). CS is defined as "increased responsiveness of nociceptive neurons in the central nervous system to their normal or subthreshold afferent input" [6] and manifests as mechanical hypersensitivity, allodynia and hyperalgesia [7]. A considerable number of people need treatment for CLBP. Although the overall efficacy of CLBP rehabilitation programs is positive, the effect sizes are modest [8].

Correctly recognizing the physical and psychosocial factors perpetuating pain and physical disability of patients with CLBP remains a challenge [9]. Altered motor control of patients with CLBP could possibly contribute to the persistence of CLBP [10]. Altered motor control could affect daily-living activities, as patients with CLBP often exhibit altered movement patterns and motor control strategies; probably to avoid painful movement, such as walking [11]. Many clinicians may intuitively identify "abnormal" gait patterns in patients with CLBP, but identification and objectifying of specific "abnormal" gait outcomes is challenging. During walking, it is suggested that patients often adopt a "protective guarding" or "splinting" strategy [12] to avoid painful movements of the spine. These adaptations may lead to a slower and less flexible gait pattern [13]. Evidence for this, however, is ambiguous. Studies between patients with CLBP and healthy controls, observed inconsistent evidence regarding preferred walking velocity [13, 14], stride length [15, 16], and stride-to-stride variability [17, 18].

A possible explanation for these inconsistencies might be an unknown heterogeneity within the samples, such as the presence of CS. CS could plausibly be related to the inconsistent results, because the presence of high CS levels is associated with long-lasting chronic pain [19] and movement may be changed due to pain. Also, general gait outcomes such as walking speed and stride length, might not be sensitive enough to detect small differences between patients with low or high levels of CS. In addition to stride related parameters, gait outcomes that reflect gait quality in terms of regularity, synchronization, smoothness, local stability, and predictability, are sensitive to detect differences in gait performance. These gait outcomes were successfully used to detect the differences between age groups [20], older adults with and without fall risk [21], and patients with and without Parkinson's disease [22]. Even though the effects of CLBP on gait have been frequently investigated in controlled laboratory studies, there are no studies about the relationship between CS levels and gait performance under daily-living environments.

Advances in wearable technology and machine learning approaches offer new opportunities in gait data collection and analysis. Wearable technology allows researchers to record patients' physical activities in unobserved, daily-living environments over extended periods of time. These data can reflect the real gait performance of the patients, since being observed may change the performance of patients under the controlled laboratory environment [23]. The successful employment of machine learning approaches in gait analysis makes it possible to extract the most informative gait outcomes from the accelerometer sensor data [20]. If patients with low and high levels of CS walk differently, machine learning approaches will be able to successfully recognize these differences and can classify patients with low and high CS levels based on their gait outcomes. Many gait outcomes are not independent and interact with each other, such as gait speed and step regularity. Machine learning approaches such as random forest (RF), are able to process high dimensional and non-linear data structures and take the interrelation and interaction of the gait outcomes into consideration [20].

Therefore, the aim of this study was to analyze whether and how the presence of CS is related to differences in gait performance of patients with CLBP during daily life by using a machine learning approach. It was hypothesized that patients with CLBP and higher CS levels show differences in daily life gait performance, compared with those with lower CS levels.

2. Methods

2.1. Patients

This study included patients with primary CLBP who were recruited from the outpatient Pain Rehabilitation Department of the Center for Rehabilitation of the University Medical Center Groningen (CvR-UMCG). Primary CLBP is defined as low back pain persistent for more than three months, with pain not being the result of any other diagnosis. The patients were selected according to the following inclusion criteria: (a) age between 18 and 65 years old at the time of recruitment; (b) admitted to the interdisciplinary pain rehabilitation program; (c) could follow instructions; (d) signed informed consent. Additionally, patients were excluded if they: (a) had a specific diagnosis that would better account for the symptoms (e.g. cancer, inflammatory diseases and/or spinal fractures); (b) had neuralgia and/or radicular pain in the legs; (c) were pregnant; (d) were in an acute phase of pain.

The study was approved by the Medical Research Ethics Committee of the University Medical Center Groningen (METc 2016/702) and conducted according to the principles expressed in the Declaration of Helsinki. The data used in this paper were derived from a larger study, of which protocol details are described elsewhere [19].

2.2. Data collection

Demographics were collected and standard clinical test were applied as part of the usual care of CLBP patients that are referred to the outpatient Pain Rehabilitation Department of the Center for Rehabilitation. Assessments included: Visual Analogue Scale for pain intensity (VAS Pain; 0–10), the Dictionary of Occupational Titles (DOT), the Pain Disability Index (PDI; 0–70), the physical functioning subscale of the Rand36 questionnaire (Rand36-PF; 0–100), the Pain Catastrophizing Scale (PCS, 0–52), the Injustice Experience Questionnaire (IEQ, 0–48), and the Brief Symptom Inventory (BSI global severity index t-score) (see Table 3).

Central sensitization (CS). The presence of CS-related manifestations was assessed with section A of the Central Sensitization Inventory (CSI) [24]. Section A has 25-items to assess the presence of common CS-related symptoms. Scores can range from 0 to 100 where a higher scoring represents a higher level of CS. A score lower than 40 indicates lower CS levels (CLBP- group) and a score of 40–100 is interpreted as higher CS levels (CLBP+ group) [25].

Accelerometer data. The accelerometer data were collected between 2017 and 2019. Patients were instructed to wear a tri-axial accelerometer (ActiGraph GT3X, Actigraph Corporation, Pensacola, FL) at all times for about one week, excluding sleeping or bathing times. The accelerometer was worn at the front right hip of the patient (at the anterior superior iliac spine). Assuming a standing and upright position, the Y-axis pointed to the ground (vertical direction, V), Z-axis faced the walking direction (anteroposterior direction, AP), and the X-axis was perpendicular to the walking direction, pointing from a patient's right to left (mediolateral direction, ML). These directions are approximate only. The sampling frequency of the accelerometer was set to 100 Hz and the dynamic range was ± 6 gravity.

2.3. Data processing and analysis

2.3.1. Raw data segmentation

Accelerometer data of each patient were segmented into 24h span data segments (from 12:00 p.m. to next day 11:59 a.m.) to represent the activities during the days. Because the measurement started at 12:00 p.m., to make full use of the data, the 24h span was between 12:00 p.m. until next day 12:00 p.m. Data that did not completely cover this 24h span was discarded from the analysis. Because of technical errors or personal reasons, a full week of data could not be collected from all patients. To compare the data between different patients fairly, 4 segments (representing 4 days) of each patient were included in the analysis. Therefore, 7 patients who had less than 4 segments, were excluded. From patients with more than 4 segments, 4 segments were randomly sampled. Fig. 1(a) graphically shows the process of the raw data segmentation.

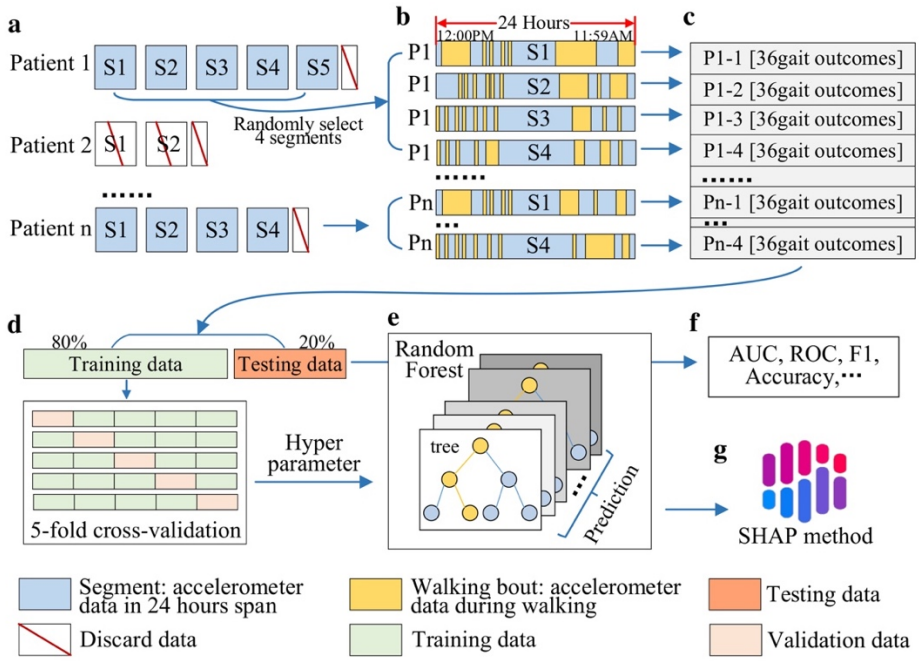


Figure 1. The data processing and analysis: (a) raw data segmentation, (b) walking bouts extraction, (c) gait outcome vectors, (d) training and testing data preparation, (e) Random Forest classifier, (f) accuracy evaluation, (g) feature importance.

2.3.2. Walking bouts extraction

The accelerometer data of the 4 segments were first smoothed by a low-pass 2nd order Butterworth filter with a 20 Hz cut-off frequency. Subsequently, potential walking events were detected by the Fast Fourier Transform (FFT) method [26], which identified periods with 0.5–3.0 Hz power spectrum values. To remove false walking events from the potential walking periods, the zero-cross method [27] was employed. If the time interval between any two adjacent walking events was shorter than 2 seconds, these two walking events were merged into one walking bout. Finally, the walking bouts in each segment were extracted and their gait outcomes were calculated. Fig. 1(b) presents the walking bouts as the yellow vertical bars in the rectangle.

2.3.3. Gait outcomes

All walking bouts in one 24 hours segment were used to determine the total duration of walking, the total number of steps, the maximum duration of a walking bout and the maximum number of steps of a walking bout. Subsequently, all walking bouts exceeding 10 seconds were selected and cut into non-overlapping 10-second windows [28]. From the segment, each 10-second window was used to calculate different gait outcomes, and these values were averaged over all 10-second windows in the segment representing the patient's gait performance on that day.

Gait outcomes were divided into two categories, quantitative and qualitative gait outcomes. From one segment, we obtained one gait outcome vector, including 36 gait outcomes, based on the walking bouts (see Fig. 1(c)). The detailed descriptions of the quantitative and qualitative gait outcomes are presented in Table 1 and Table 2 —for extended explanation of variables, see reference [29].

Pearson correlation coefficient was calculated to examine relationship of gait outcomes between weekdays and weekend. The Pearson correlation coefficient ranges from -1 to 1, where 1 represents a perfect correlation.

The Mann-Whitney U test was used to statistically test the differences between CLBP- and CLBP+ groups for demographics and CSI scores. To separate CLBP- and CLBP+ groups by gait outcomes, RF was used.

Table 1. Quantitative Gait Outcomes.

Catalog	Gait characteristic	Description and method
Pace	Total duration of walking in the day	The accumulated time (in seconds) of the walking bouts in one segment.
	Total number of steps in the day	The accumulated steps of walking bouts in one segment.
	Maximum duration of a walking bout	Duration (in seconds) of longest walking bout in one segment.
	Maximum number of steps of a walking bout	Maximum number of steps of one walking bout in one segment.
	Walking speed (WS; mean, variability)	$WS = D/T$, where D is the distance (in meters) and equals the accumulated of step length; T is the corresponding time (in seconds).
	Stride length (SL; mean, variability)	SL $= 2\sqrt{2lh - h^2}$, where h is the change in height (in meters), l equals leg length (in meters). h was calculated by a double integration of the accelerometer signal in vertical direction. SL is the sum of the adjacent two step lengths.
	Stride time (ST; mean, variability)	$ST = n/f$, where f is the sample frequency (in Hertz) and n is the number of samples per dominant period derived from autocorrelation.
	Stride frequency (SF; mean, variability-V/ML/AP)	$SF = f/n$.
	Root mean square of the variability of the amplitude of accelerations (RMS),	$RMS = \sqrt[2]{\frac{1}{n}(x^2 + y^2 + z^2)}$, where x, y, z represent the accelerometer signal (in meters per second squared) in x, y, z axis and n is the number of samples.

2.3.4. Random forest classifier

RF is considered as the optimal machine learning classification approach for the present data, because it performs well with (a) nonlinear and linear data; (b) high dimensional data; and (c) unbalanced and small datasets [30]. Apart from this, a comparison of different machine learning classifiers was performed to help to select RF as the best classifier for this study (details in Appendix A).

Table 2. Qualitative Gait Outcomes.

Catalog	Gait characteristic	Description and method
Regularity	Stride regularity (SR; V, ML, AP, All)	SR is computed by using the unbiased autocorrelation coefficient: $Ad(m) = \frac{1}{N-\|m\|}\sum_{i=1}^{N-\|m\|} Acc(i) \cdot Acc(i+m)$, where $Acc(i)$ is the sample acceleration signal, N the number of samples, and m the number of time lag. The first peak of $Ad(m)$ is Ad_1 and it represents the stride regularity. Higher values (maximum 1.0) reflect repeatable patterns between strides.
	Gait symmetry index (GSI)	GSI quantifies the ratio of the first and second peak of the $Ad(m)$, as Ad_1/Ad_2. It is a measure of the degree of symmetry of the left and right lower limbs during walking.
Smoothness	Index of harmonicity (IH; V, ML, AP, All)	$IH = \frac{P_0}{\sum_{i=0}^{6} P_i}$. It is the ratio of the power spectral density of the fundamental frequency P_0 and the sum of the power spectral density of the first six frequency P_i. IH quantifies gait smoothness, with higher values representing a smoother (max 1.0) gait pattern.
	Harmonic ratio (HR; V, ML, AP)	$HR = \frac{\sum P_a}{\sum P_b}$. In VT and AP directions, $\sum P_a$ = the sum of even power spectral and $\sum P_b$ = the sum of odd power spectral. In ML direction, P_a is odd and P_b is even. It reflects the rhythmicity of the walking patterns. Higher values mean more rhythmic
Predictability	Sample entropy (Sen; V, ML, AP)	$\mathrm{Sen} = -\ln\left(\frac{A}{B}\right)$, with $A = d[Acc_{m+1}(i), Acc_{m+1}(j)] < r$, $\mathrm{B} = d[Acc_m(i), Acc_m(j)] < r$. $Acc_m(i)$ means the accelerometer signal vector from time i to $m+i-1$. $d[\,]$ is the Chebyshev distance, and r was set to 0.3. Sen quantifies the predictability of a time series. Smaller values (minimum 0) indicate better synchronization between acceleration signals.

	Maximal Lyapunov exponent (max LyE; V, ML, AP)	Max LyE, as calculated by the Rosenstein algorithm, quantifies the local stability of trunk acceleration patterns. The fitting window length was $60/100 * f$, where f is the sample frequency, and the embedding dimension was set to 7. The overall max LyE were calculated and normalized per stride. Higher values represent greater sensitivity to local perturbations.
Stability	Maximal Lyapunov exponent normalized per stride by time (max LyE per stride; V, ML, AP)	

The input data of this approach was $< S, L >$. S represents the gait outcome vectors of all patients and L was its corresponding label. The definition of S is: $S = \{s_1, s_2, \dots, s_i, \dots, s_m\}$ and $s_i = [d_1, \dots., d_k]$, where s_i represents a gait outcome vector i and m is the number of all gait outcome vectors, d represents a gait outcome and $k = 36$. $L = l_1, \dots., l_m$, where $l \in \{CLBP-, CLBP+\}$.

RF is constructed in four steps. Step one: Randomly sample n gait outcome vectors from S and n corresponding labels from L, with replacement. These new sets of gait outcome vectors and labels are called S_b and L_b. In S_b, s_i may appear more than one time or not appear. Step two: In S_b, randomly sample j ($j \leq k$) gait outcomes from s. Therefore, $s_i' = d_1', \dots., d_j'$ and $S_b' = s_1', \dots, s_n'$. Step three: Training a decision tree f_b on S_b', L_b. Step four: Repeat steps one to three 1000 times and combine the decision trees into an ensemble, called RF, that predicts by voting (see Fig. 2).

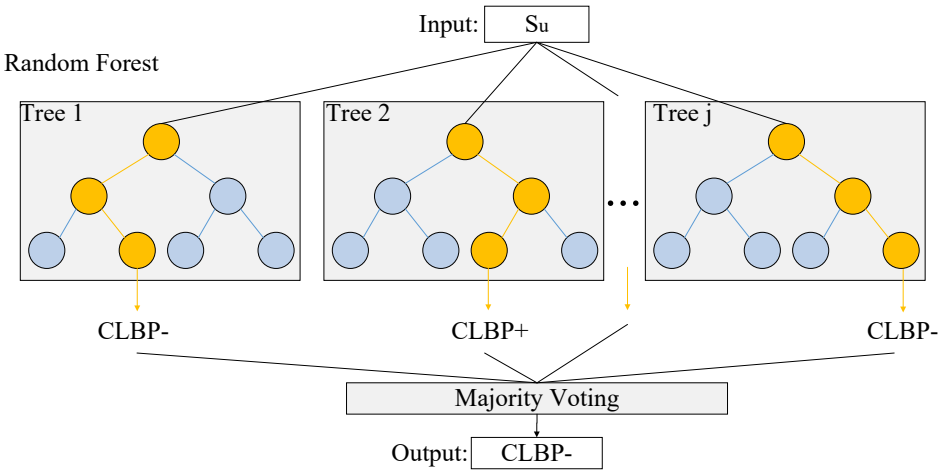


Figure 2. Architecture of the Random Forest classifier.

Before training RF, 80% of patients were randomly selected and their 4 corresponding gait outcome vectors were used as the training data. The gait outcome vectors of the remaining 20% of patients were used as the testing data. To avoid overfitting of the hyperparameters, a 5-fold cross-validation approach was used to estimate them, as shown in Fig. 1(d). Four folds

were used to train the model and the remaining fold was used to estimate the performance of the current hyperparameters in RF. The performance reported by the 5-fold cross-validation was the average of the values computed in the 5 splits. After the best hyperparameters were determined, the testing dataset was used to evaluate the generalizability of the model.

2.3.5. Accuracy evaluation

Accuracy, sensitivity, specificity, precision, F1-score, and maximum Youden index were calculated to evaluate the performance of the classification (Fig. 1(f)). In this study, CLBP+ was considered as the positive case and CLBP- was the negative case. Correct predictions of CLBP+ and CLBP- patients are called true positives (TP) and true negatives (TN), respectively. Incorrect classifications of CLBP- patients as CLBP+ or of CLBP+ patients as CLBP-, are called false positives (FP) and false negatives (FN) respectively.

Accuracy was the proportion of all the correct classification results.

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Sensitivity represents the proportion of positive cases that are correctly assigned (true positive rate).

$$sensitivity = \frac{TP}{TP + FN}$$

Specificity refers to the rate of correctly predicted negative cases in all negative cases (true negative rate).

$$specificity = \frac{TN}{TN + FP}$$

Precision is the ratio of the correctly predicted positive cases in all predicted positive cases.

$$precision = \frac{TP}{TP + FP}$$

F1-score is the harmonic mean (average) of the precision and sensitivity.

$$F1 = \frac{2 \times precision \times sensitivity}{precision + sensitivity}$$

The receiver operating characteristic (ROC) curve was calculated to evaluate the performance of RF. The Y-axis of this curve represents the true positive rate (sensitivity) and the X-axis means false positive rate (1-specificity). The overall classification performance of RF was evaluated by the area under the ROC curve (AUC). A classification model with a larger AUC value has a higher correct rate, and AUC = 1 represents perfect performance. The maximum Youden index was computed to measure the diagnostic test's ability.

$$J(c) = Max\{sensitivity(c) - (1 - specificity(c))\}$$

where c is the cut-point. When the value J is maximum, the corresponding c is the optimal cut-point.

2.3.6. Feature importance

SHapley Additive exPlanations (SHAP) [31] was used to assess the gait outcomes' importance to the classification model. SHAP connects optimal credit allocation with local explanations using the classic Shapley values from game theory. Shapley values, $\emptyset_i$, explains the importance of gait outcome i for RF and is defined as:

$$\emptyset_i = \frac{1}{|N|!} \sum_{\{i\} \in s \; and \; s \subseteq N} (|s| - 1)! \, (|N| - |s|)! \, [R(s) - R(s - \{i\})]$$

where N is the size of the full set of gait outcomes, s is the subset that includes i in N, and $R(\,)$ is the accuracy of RF of the input gait outcomes. Since computing the exact Shapley values is computationally expensive, SHAP uses a tree explainer to exploit the information stored in the tree structure to calculate the SHAP values which are highly approximate Shapley values. Therefore, higher SHAP values represent higher impact to classify the CLBP- and CLBP+ groups.

3. Results

Demographic characteristics are provided in Table 3. Out of a total of 60 patients, 11 were excluded because essential parts of their dataset were incomplete (CSI scores or/and accelerometer data), 7 were excluded because they had less than 4 segments data (3 had 1 segment, 2 had 2 segments, and 2 had 3 segments). Therefore, 42 patients were included in the data analysis. Differences between CLBP+ and CLBP- group characteristics (Table 3) were not statistically significant ($p>0.05$), with the exception of CSI ($p<0.001$) and BSI ($p=0.01$).

Because 42 patients (23 CLBP- and 19 CLBP+) were included, and for every patient, 4 segments were randomly selected, the total accelerometer data segments were 168. Therefore, the scales of training and testing dataset were 136 and 32. The mean Pearson correlation coefficient between workdays and weekend was 0.983, indicating almost perfect correlation.

Testing data were used to evaluate the generalizability of RF and the confusion matrix is shown in Fig. 3. From the confusion matrix, accuracy, sensitivity, specificity, precision, and the F1-score were calculated to evaluate performance of the model. RF achieved an accurate classification-result (84.4% accuracy), and the sensitivity and specificity were 75.0% and 93%, respectively. The precision was 92% and the F1-score was 82.6%. The ROC curve is presented in Fig. 4, showing that RF achieved an AUC of 0.83 and the maximum Youden index was 0.69.

The importance of the gait outcomes for RF is shown in Fig. 5. Based on the SHAP values, the top 10 gait outcomes (above the red line in Fig. 5) were considered as important to the classification model. For the gait outcomes below the red line, the SHAP values were too low. Important gait outcomes are IH-V, SF variability-ML/AP, SR-ML, Max LyE-V/ML, Sen-AP, Max LyE per stride-V, HR-ML and SL variability.

Table 3. Patient characteristics (n=42).

	CLBP- (n=23)	CLBP+ (n=19)	All (n=42)	P-Value
Gender	15W / 8M	12W / 7M	27W / 15M	
Age, years	40.8 ± 12.8	38.1 ± 12.7	39.6 ± 12.6	
Height, cm	173.5 ± 10.6	175.7 ± 8.8	174.5 ± 9.8	
Weight, kg	87 ± 17.7	85.4 ± 15.1	86.3 ± 16.4	
Body mass index, kg/m^2	28.9 ± 5.3	27.7 ± 4.4	28.3 ± 4.9	
Central Sensitization Inventory (0-100)	31± 4.8	48.7 ± 8.7	39.0 ± 11.2	< 0.0001
Time since pain onset (years)	4.5 ± 6.1	3.5 ± 3.1	4.1 ± 4.9	
Educational Level	17S / 6H	10S / 9H	26S / 15H	
Physical demands at work (DOT; Se/Li/Me/He)	3/11/8/1	4/7/7/1	7/18/15/2	
Patient-reported Pain Intensity (VAS, 0-10)	5.5 ± 2	5.2 ± 1.8	5.4 ± 1.9	
Disability (PDI, 0-70)	33.6 ± 11.2	26.8 ± 11.9	31.0 ± 11.7	
Work Ability (WAS, 0-10)	4.5 ± 2.3	4.9 ± 2.8	4.6 ± 2.5	
Physical Functioning (Rand36-PF, 0-100)	49.8 ± 22.3	63.3 ± 16.1	54.7 ± 21.1	
Catastrophizing (PCS, 0-52)	16.3 ± 8.9	20.3 ± 11.1	18.1 ± 10	
Injustice (IEQ, 0-48)	15.2 ± 8.9	18.5 ± 8.5	16.7 ± 8.8	
Psychological traits Screening (BSI, t-score)	34.4 ± 4.9	41.5 ±5.8	37.6 ± 6.4	= 0.01

Except gender, all results represent mean ± standard deviation. CLBP-, CLBP+: Patients with chronic low back pain with low (-) and high (+) central sensitization levels. W: Women; M: Men. H: Higher education; S: Secondary education. Se: Sedentary; Li: Light; Me: Medium; He: Heavy. DOT: Dictionary of Occupational Titles. VAS: Visual Analogue Scale. PDI: Pain Disability Index. WAS: Work Ability Score. Rand36-PF: Rand 36-Physical Functioning subscale. PCS: Pain Catastrophizing Scale. IEQ: Injustice Experience Questionnaire. BSI: Brief Symptom Inventory.

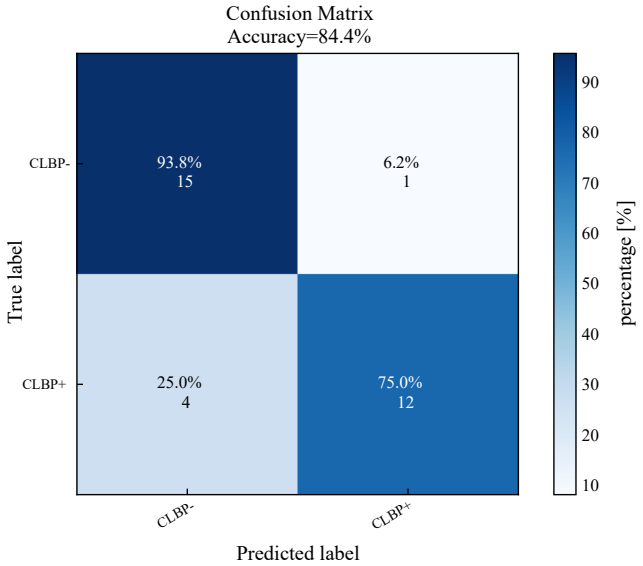


Figure 3. Classification results for Random Forest, and the mean accuracy was 84.4%. CLBP-, CLBP+: Patients with chronic low back pain with lower (-) and higher (+) central sensitization levels.

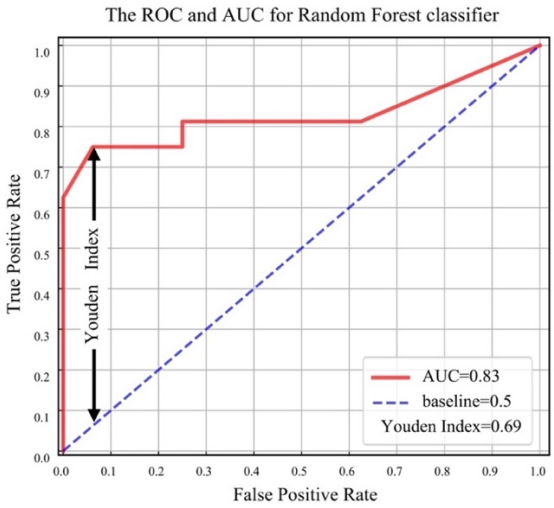


Figure 4. The receiver operating characteristic (ROC) curve (in red) for Random Forest classifier. AUC: area under the curve.

Fig. 6 shows the violin-box plot of the 10 most important gait outcomes. Violin-box plots are a hybrid of kernel density and box plots, and the dots show the individual data. A box plot contains a set of whiskers, a box and a horizontal line in the middle of the box, representing the minimum, maximum, first quartile, third quartile and median in the data, respectively. From this figure, it is easy to distinguish the differences of the median between groups. It shows that the CLBP- group has higher IH-V, HR-ML (better smoothness); higher SF-variance-ML, SF-variance-AP, SL-variance (higher variability); lower SR-ML (lesser regularity), lower Max LyE-V, Max LyE-per-stride-V, slightly lower Max LyE-ML (better stability); and slightly higher Sen-AP (lesser predictability). Although the differences of medians between 2 groups in Sen-AP and Max LyE-ML are small, their distributions are different. In Sen-AP, data of CLBP- had a wider distribution and CLBP+ shows more data at the bottom. In the Max LyE-ML, data of CLBP- is concentrated around the median, while CLBP+ has a wide distribution and a lower peak. For other gait outcomes, the distributions are also different. In IH-V, distributions of CLBP- and CLBP+ all showed a bimodal distribution, but the peaks of distribution are different. In SF Variability-ML and SF Variability-AP, CLBP+ has a larger peak at the bottom while CLBP- has a wide range distribution. Similarly, in SR-ML, CLBP+ has a concentrating distribution while the peak of CLBP- is lower. In Max LyE-V and Max LyE per stride -V, CLBP- shows a log-normal distribution while CLBP+ shows a wider distribution. In HR-ML and SL Variability, the distributions are similar but CLBP+ has more outliers.

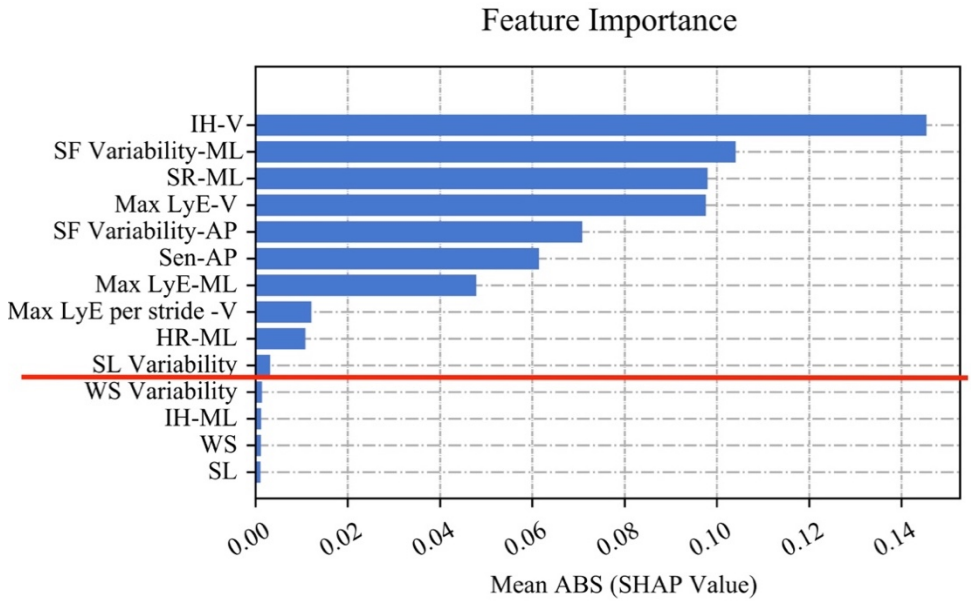


Figure 5. Features importance of the Random Forest classifier. The 10 gait outcomes above the red line are: index of harmonicity in vertical direction (IH-V), variability of stride frequency in mediolateral/anteroposterior direction (SF variability-ML/AP), stride regularity in mediolateral direction (SR-ML), Maximal Lyapunov exponent in vertical/mediolateral direction (Max LyE-V/ML), sample entropy in anteroposterior direction (Sen-AP), Max LyE-V: Maximal Lyapunov exponent per stride in vertical direction, harmonic ratio in mediolateral direction (HR-ML) and variability of stride length (SL variability). The remaining gait outcomes below the red line are: WS variability: variability of walking speed, IH-ML: index of harmonicity in mediolateral direction, WS: mean walking speed and SL: mean stride length. ABS: absolute value. SHAP: SHapley Additive exPlanations.

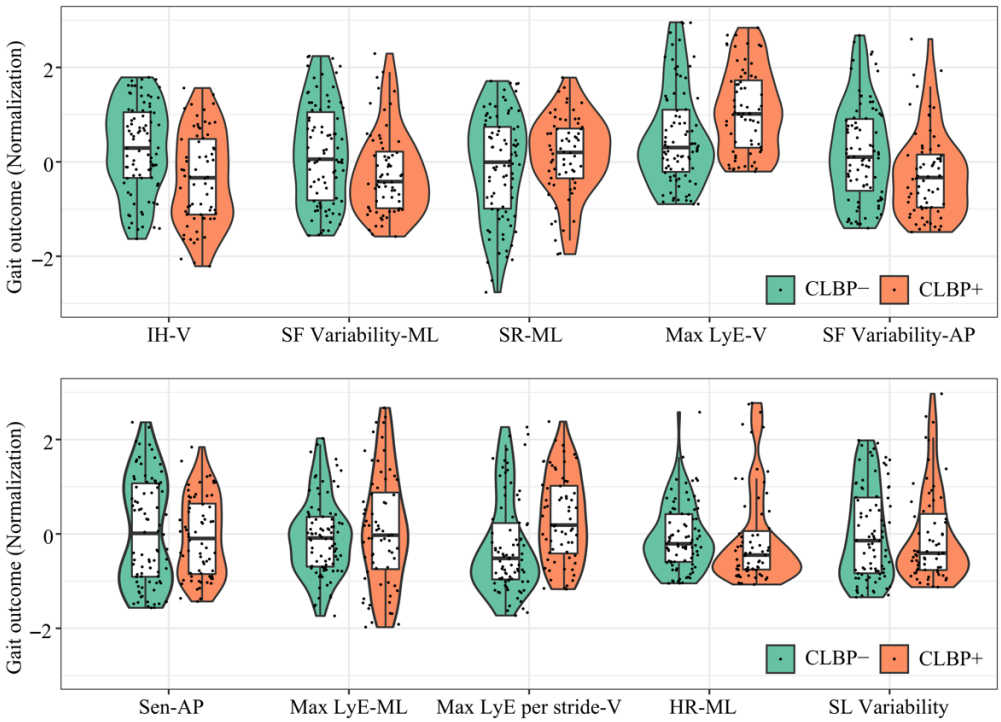


Figure 6. Violin-box plot for the 10 gait outcomes. Dots show the individuals data. CLBP-, CLBP+: Patients with chronic low back pain with low (-) and high (+) central sensitization levels. IH-V: index of harmonicity in vertical direction, SF variability-ML/AP: variability of stride frequency in mediolateral/ anteroposterior direction, SR-ML: stride regularity in mediolateral direction, Max LyE-V/ML: Maximal Lyapunov exponent in vertical/mediolateral direction, Sen-AP: sample entropy in anteroposterior direction and HR-ML: harmonic ratio in mediolateral direction.

4. Discussion

The aim of this study was to analyze whether and how the presence of CS was related to differences in gait performance of patients with CLBP during daily life by using a machine learning approach. Based on quantitative and qualitative gait outcomes, using a RF approach, the two groups (CLBP- and CLBP+) could be classified with a high accuracy. The classification results indicated that patients with CLBP- walk differently from patients with CLBP+. Furthermore, the SHAP values showed that the differences between the CLBP- and CLBP+ groups were present in gait outcomes that represented smoothness, stability, predictability, regularity, and variability.

In the present study, we addressed the walking measurement of patients with CLBP in a daily-living environment. Walking in a controlled laboratory or during a clinical assessment is different from self-initiated gait, during activities of daily living. Walking in daily life, might be subject to environmental perturbations, quick changes while performing a task, and often involves the performance of several actions at the same time [32], e.g., walking when carrying a cup of coffee. These influences on gait are not present in controlled studies and are not captured by conventional gait outcomes that average outcomes over stride cycles, such as mean step length, mean step time, and number of steps. Therefore, the present study included gait outcomes that take into account the interdependency of gait cycles and how gait cycles evolve over time, e.g., using sample entropy as a measure of predictability of the gait pattern, the maximal Lyapunov exponent as quantification of local stability and correlation-based measures [33].

The accuracy value of 84.4% shows that RF has a high classification accuracy. The specificity of RF reveals that 93% of the samples (15 samples, true negative) are correctly classified as the member of CLBP- without a high CS level, but it misses 7% (1 sample, false positive). The sensitivity shows that 75% of samples (12 samples, true positive) of the CLBP+ group were assigned to this group, however 25% were wrongly classified as belonging to the CLBP- group (4 samples, false negative). Decreasing the possibility of false positive will increase the possibility of false negative, and vice versa. The F1-score was calculated to take false positive and false negative into consideration at the same time by computing their harmonic mean. The high F1-score (82.6%) of RF implies that the model has a good and balanced performance. The Youden Index (0.69) was higher than 0.5 which means that RF has a diagnostic test's

ability to balance sensitivity and specificity. The AUC indicates that RF has a 83% chance to distinguish CLBP+ and CLBP- correctly. Based on these performance measures of RF, this study leads us to conclude that the CLBP- and CLBP+ groups had different gait patterns, and that the gait outcomes important to the classification model identified by SHAP are trustworthy.

In the present study RF was applied for classification, among many available machine learning approaches, such as K-Nearest Neighbour (KNN), Naive Bayes (NB), Artificial Neural Network (ANN), Support Vector Machine (SVM). In general, machine learning approaches can take the interaction of gait outcomes into consideration. KNN and NB are instance-based learning approaches which imply they do not learn from training data [34]. Our choice for RF was based on the results of a previous study that compared RF, ANN, and SVM to classify different age groups on similar gait outcomes. It showed that all approaches had a good overall classification accuracy [20]. Moreover, for the current dataset, our preliminary empirical work in which we compared the performance of different machine learning classifiers, showed that RF and ANN had the best performance compared to SVM, NB and KNN (details were in Appendix A). A drawback of ANN is that it requires a large data set to find the optimal activation function and avoid overfitting [35]. With a limited size of dataset, both SVM and RF are good choices. Considering the clinical aim of the study, namely to investigate the relationship between CLBP, CS, and gait patterns, it is important that the results of the machine learning can be translated into meaningful outcomes that can support clinical decision making. SVM can deal with non-linear data by using kernel functions; however, choosing an appropriate kernel function could be difficult for clinicians. Additionally, it implicitly maps gait outcomes to a high-dimensional features space. This mapping changes the structure of gait outcomes and makes it hard to explain which gait outcomes contribute most to the classification model. Similarly, ANN uses various activation functions (e.g., Tanh, Sigmoid), and makes the interactions of the gait outcomes invisible. On the contrary, RF is an ensemble of decision trees. Decision trees can incorporate gait outcomes interactions naturally in the classification process. For example, a decision tree with depth 2 from a RF, with the father node IH-V and the son node Sen-AP, can describe an interactive gait pattern: if IH-V >* and Sen-AP >*, the data belong to CLBP-. Because RF includes multiple decision trees it can capture the complex interaction of gait outcomes with good accuracy. Each tree is built based on a random subset of gait outcomes and the samples in the dataset can be repeatedly selected when training. Consequently, it can help to reduce the chance of overfitting and provide a generalized model. RF can incorporate gait outcomes interactions naturally in the classification process. SHAP can use this information that is stored in the tree structure to disclose which gait outcomes are different between the CLBP- and CLBP+ groups. These differences in terms of gait regularity, smoothness, and stability are meaningful to the clinicians.

In this study, SHAP was used to evaluate the importance of each gait outcome, instead of the conventionally used Gini impurity and information entropy. The value of Gini impurity is based on the tree structure in RF and information entropy reflects the level of "information" of a gait outcome. Gait outcomes are interrelated and interact in a complex nonlinear manner [33]. SHAP is based on game theory and evaluates the contribution of each gait outcome to the classification accuracy by computing all possible combinations between gait outcomes. Therefore, SHAP provides a good method to explain the importance of gait outcomes to RF. The SHAP values suggest that the differences between CLBP- and CLBP+ groups are reflected in smoothness, stability, predictability, regularity, and variability of gait. Compared with the CLBP- group, the CLBP+ group exhibited lower smoothness and local stability of gait, while the CLBP+ group exhibited a more regular, less variable, and more predictable gait pattern.

Gait patterns of patients with CLBP, are usually compared with the gait pattern of healthy persons. To the best of our knowledge, this is the first study in patients with CLBP that addresses the differences in gait pattern between two CLBP groups based on low and high CS levels, which makes a direct comparison with other studies intricate. The results of different gait patterns between low and high CS levels support the notion that within the heterogenous CLBP group, different motor control strategies are adopted. Two motor control strategies on a continuum have been suggested with “tight control” and “loose control” at each end, and normal trunk control in the middle [36].

The gait patterns of the CLBP+ group might suggest that patients with CLBP+ adopt a more “tight control”. The “tight control” involves increased trunk muscle activation and enhanced muscle co-contraction, might enhance control over trunk posture and movement [36]. Increased muscle activation and enhanced co-contraction would help individuals to maintain the stability of the lumbar spine [37] by restricting the movement amplitude of the lumbar spine. However, in a complex daily-living environment, this strategy might impair patients' ability to maintain balance during walking because of the unstable surfaces and environmental perturbations [38], and therefore has a lower gait stability (compared with patients with CLBP-). Increased co-contraction would reduce the demand for the intricate control of the sequences of muscle activation. It might avoid the potential error raised by inaccurate sensory feedback of CLBP [38]. This might allow patients to control their trunks’ movement precisely [39] and, therefore, result in a lower variability and a higher regularity of gait. Our results might infer that the CLBP+ group exhibited a more “tight control”. Therefore, the lower stability and variability, higher regularity and predictability in gait of the CLBP+ group could be the result of the adoption of “tight control”.

The gait patterns of the CLBP- group, on the other hand, might be explained by a “loose control” strategy. The “loose control” that involves reduced muscle excitability, might reduce

the control over trunk movements [36]. The spine of which each spinal unit has 6 degrees of freedom, is controlled by its surrounding musculature. Reduced muscular excitability, leads to a reduced control over the spinal muscle, to larger amplitude movements, and to more movement variability during repeated tasks [36]. The increased variability in gait of the CLBP- group might support this idea. Additionally, increased variability would lead to a lower regularity in gait which was also found in the CLBP- group. Apart from this, increased motor variability might probably prevent muscle fatigue [40] because it allows sharing the load between different structures or tissues. Motor variability makes it possible to explore new pain-free motor control solutions [41]. This is a possible explanation for the higher smoothness in gait of patients with CLBP-, because it allowed them to flexibly adapt to the complex daily-living environment by using different movement solutions. So, the higher variability and smoothness, and lower regularity in gait patterns might hint that the CLBP- group adopted a more “loose control” compared with the CLBP+ group.

Although the “tight control” adapted strategy might have short-term benefits, it may also contribute to a higher level of CS. The “tight control” present in patients with CLBP+ presumably increases muscle activation and co-contraction, and leads to larger forces acting on the spine and higher spinal loading. Moreover, it has been shown that even when patients are at rest, muscle co-contraction can be continuous [42]. These long-lasting peripheral noxious stimuli might explain the development and/or persistence of CS [5]. Additionally, it has been reported that a “tight control” strategy relates to negative pain cognitions [43], a psychological process that also might contribute to higher CS scores of the CLBP+ group.

Clinically, the gait outcomes identified as important to the classifier, may assist clinicians in providing them with a more accurate understanding of the gait performance of patients with CLBP, with low or high CS levels, and with an explicit operationalization of the observed “abnormal” gait pattern of patients with chronic pain. Whether “abnormal” should be interpreted as a functional or a dysfunctional motor control strategy in the short or long term, remains to be studied. RF and SHAP used in this study have presented a novel way to identify interacting features, and therefore, can be used for further studies. The presented accurate classification could become meaningful if this would lead to effective treatment approaches. The differences in gait patterns between the CLBP- and CLBP+ groups may reflect variations in motor control adaptation strategies. These strategies could be the causes and consequences of differences in central sensitization levels among patients with CLBP, or both. While this cross-sectional study has objectified a relation between CS and gait outcomes, the causality of this relation is unknown. Follow-up studies would benefit from a longitudinal design with multiple measurements to help further unravelling of this relation, as well as the relation to disability.

In line with most studies on walking and CLBP, we used cross-sectional data, thus we are not allowed to infer causality between motor control changes, CS and CLBP. Some patients had analgesic or anti-inflammatory treatment at the beginning of the study, and how these medicines interact with CS and gait outcomes is unknown. Moreover, we labelled the groups based on CSI scores with a cut-off value from a previous study [25]. It should also be noted that a gold standard measure to diagnose CS is unavailable. CSI is regarded as an indirect measure of CS, because higher scores are associated with the presence of CS syndromes [25]. In addition to gait assessment, it would be interesting to explore differences in physical activities between the CLBP- and CLBP+ groups, because several studies reported that the relationship between CLBP and physical activity levels is heterogeneous [44].

5. Conclusion

The present study analysed gait data during daily living of patients with CLBP and low or high CS levels. RF and SHAP were applied for classification and for assessing the contribution of gait outcomes to the model. This analytic approach demonstrated that RF has the ability to accurately classify subgroups of patients with CLBP and low or high CS levels based on differences in gait outcomes. The SHAP results showed that the differences in gait outcomes between low and high CS levels were in gait regularity, variability, predictability, smoothness, and stability. This may imply that patients with low and high CS levels adopted different motor control strategies. Patients with CLBP and low CS level (CLBP-) use a “loose control” and, therefore, exhibited more smoothness and stability in gait patterns. Patients with CLBP and high CS level (CLBP+) adopted a “tight control” and showed a more regular, less variable and more predictable gait pattern.

The results of this study may contribute to a better understanding of gait characteristics in patients with CLBP, its association with CS, and may in the future assist in better-personalized rehabilitation interventions [45].

Reference

[1] M. S. Thiese, K. T. Hegmann, E. M. Wood, A. Garg, J. S. Moore, J. Kapellusch, J. Foster, and U. Ott, “Prevalence of low back pain by anatomic location and intensity in an occupational population,” *Bmc Musculoskeletal Disorders,* vol. 15, Aug 21, 2014.

[2] S. Dagenais, J. Caro, and S. Haldeman, “A systematic review of low back pain cost of illness studies in the United States and internationally,” *Spine Journal,* vol. 8, no. 1, pp. 8-20, Jan-Feb, 2008.

[3] J. Hartvigsen, M. J. Hancock, A. Kongsted, Q. Louw, M. L. Ferreira, S. Genevay, D. Hoy, J. Karppinen, G. Pransky, J. Sieper, R. J. Smeets, M. Underwood, and W. Lancet Low Back Pain Series, “What low back pain is and why we need to pay attention,” *Lancet,* vol. 391, no. 10137, pp. 2356-2367, Jun 9, 2018.

[4] O. Airaksinen, J. I. Brox, C. Cedraschi, J. Hildebrandt, J. Klaber-Moffett, F. Kovacs, A. F. Mannion, S. Reis, J. Staal, and H. Ursin, "European guidelines for the management of chronic nonspecific low back pain," *European spine journal,* vol. 15, no. Suppl 2, pp. s192, 2006.

[5] J. Nijs, S. Z. George, D. J. Clauw, C. Fernández-de-las-Peñas, E. Kosek, K. Ickmans, J. Fernández-Carnero, A. Polli, E. Kapreli, and E. Huysmans, "Central sensitisation in chronic pain conditions: latest discoveries and their potential for precision medicine," *The Lancet Rheumatology,* vol. 3, no. 5, pp. e383-e392, 2021.

[6] M. W. van Tulder, B. W. Koes, and L. M. Bouter, "A cost-of-illness study of back pain in The Netherlands," *Pain,* vol. 62, no. 2, pp. 233-240, 1995.

[7] R. H. Jenkinson, "Central sensitization," *Pain: A Review Guide*, pp. 45-48, 2019.

[8] N. E. Foster, J. R. Anema, D. Cherkin, R. Chou, S. P. Cohen, D. P. Gross, P. H. Ferreira, J. M. Fritz, B. W. Koes, W. Peul, J. A. Turner, C. G. Maher, and W. Lancet Low Back Pain Series, "Prevention and treatment of low back pain: evidence, challenges, and promising directions," *Lancet,* vol. 391, no. 10137, pp. 2368-2383, Jun 9, 2018.

[9] C. Maher, M. Underwood, and R. Buchbinder, "Non-specific low back pain," *Lancet,* vol. 389, no. 10070, pp. 736-747, Feb 18, 2017.

[10] M. Götze, M. Ernst, M. Koch, and R. Blickhan, "Influence of chronic back pain on kinematic reactions to unpredictable arm pulls," *Clinical Biomechanics,* vol. 30, no. 3, pp. 290-295, 2015.

[11] T. Giesecke, R. H. Gracely, M. A. Grant, A. Nachemson, F. Petzke, D. A. Williams, and D. J. Clauw, "Evidence of augmented central pain processing in idiopathic chronic low back pain," *Arthritis Rheum,* vol. 50, no. 2, pp. 613-23, Feb, 2004.

[12] D. K. Ahern, M. J. Follick, J. R. Council, N. Laser-Wolston, and H. Litchman, "Comparison of lumbar paravertebral EMG patterns in chronic low back pain patients and non-patient controls," *Pain,* vol. 34, no. 2, pp. 153-160, 1988.

[13] C. J. Lamoth, O. G. Meijer, A. Daffertshofer, P. I. Wuisman, and P. J. Beek, "Effects of chronic low back pain on trunk coordination and back muscle activity during walking: changes in motor control," *European Spine Journal,* vol. 15, pp. 23-40, 2006.

[14] G. Christe, F. Kade, B. M. Jolles, and J. Favre, "Chronic low back pain patients walk with locally altered spinal kinematics," *Journal of biomechanics,* vol. 60, pp. 211-218, 2017.

[15] C. J. Lamoth, J. F. Stins, M. Pont, F. Kerckhoff, and P. J. Beek, "Effects of attention on the control of locomotion in individuals with chronic low back pain," *Journal of neuroengineering and rehabilitation,* vol. 5, no. 1, pp. 1-8, 2008.

[16] S. P. Gombatto, T. Brock, A. DeLork, G. Jones, E. Madden, and C. Rinere, "Lumbar spine kinematics during walking in people with and people without low back pain," *Gait & posture,* vol. 42, no. 4, pp. 539-544, 2015.

[17] D. Hamacher, D. Hamacher, F. Herold, and L. Schega, “Are there differences in the dual-task walking variability of minimum toe clearance in chronic low back pain patients and healthy controls?,” *Gait & Posture,* vol. 49, pp. 97-101, Sep, 2016.

[18] R. Müller, T. Ertelt, and R. Blickhan, “Low back pain affects trunk as well as lower limb movements during walking and running,” *Journal of biomechanics,* vol. 48, no. 6, pp. 1009-1014, 2015.

[19] J. A. Echeita, H. R. S. Preuper, R. Dekker, I. Stuive, H. Timmerman, A. P. Wolff, and M. F. Reneman, “Central Sensitisation and functioning in patients with chronic low back pain: protocol for a cross-sectional and cohort study,” *Bmj Open,* vol. 10, no. 3, Mar, 2020.

[20] Y. Zhou, R. Romijnders, C. Hansen, J. van Campen, W. Maetzler, T. Hortobagyi, and C. J. C. Lamoth, “The detection of age groups by dynamic gait outcomes using machine learning approaches,” *Scientific Reports,* vol. 10, no. 1, Mar, 2020.

[21] K. S. van Schooten, M. Pijnappels, S. M. Rispens, P. J. M. Elders, P. Lips, and J. H. van Dieen, “Ambulatory Fall-Risk Assessment: Amount and Quality of Daily-Life Gait Predict Falls in Older Adults,” *Journals of Gerontology Series a-Biological Sciences and Medical Sciences,* vol. 70, no. 5, pp. 608-615, May, 2015.

[22] S. Del Din, A. Godfrey, B. Galna, S. Lord, and L. Rochester, “Free-living gait characteristics in ageing and Parkinson's disease: impact of environment and ambulatory bout length,” *Journal of Neuroengineering and Rehabilitation,* vol. 13, May, 2016.

[23] J. Vickers, A. Reed, R. Decker, B. P. Conrad, M. Olegario-Nebel, and H. K. Vincent, “Effect of investigator observation on gait parameters in individuals with and without chronic low back pain,” *Gait & Posture,* vol. 53, pp. 35-40, Mar, 2017.

[24] van Wilgen CP, Meeus M, Descheemaeker F, and C. B. "Central Sensitization Inventory – Nederlandse consensusvertaling," 11, 2021.

[25] R. Neblett, H. Cohen, Y. Choi, M. M. Hartzell, M. Williams, T. G. Mayer, and R. J. Gatchel, “The Central Sensitization Inventory (CSI): establishing clinically significant values for identifying central sensitivity syndromes in an outpatient chronic pain sample,” *J Pain,* vol. 14, no. 5, pp. 438-45, May, 2013.

[26] N. Ichinoseki-Sekine, Y. Kuwae, Y. Higashi, T. Fujimoto, M. Sekine, and T. Tamura, “Improving the accuracy of pedometer used by the elderly with the FFT algorithm,” *Med Sci Sports Exerc,* vol. 38, no. 9, pp. 1674-81, Sep, 2006.

[27] J. Qian, L. Pei, J. Ma, R. Ying, and P. Liu, “Vector graph assisted pedestrian dead reckoning using an unconstrained smartphone,” *Sensors,* vol. 15, no. 3, pp. 5032-5057, 2015.

[28] K. S. van Schooten, M. Pijnappels, S. M. Rispens, P. J. M. Elders, P. Lips, A. Daffertshofer, P. J. Beek, and J. H. van Dieen, “Daily-Life Gait Quality as Predictor of Falls in Older People: A 1-Year Prospective Cohort Study,” *Plos One,* vol. 11, no. 7, Jul, 2016.

[29] L. Kikkert, N. Vuillerme, J. P. van Campen, B. A. Appels, T. Hortobagyi, and C. J. Lamoth, "Gait characteristics and their discriminative power in geriatric patients with and without cognitive impairment," *Journal of Neuroengineering and Rehabilitation,* vol. 14, Aug, 2017.

[30] G. Biau, and E. Scornet, "A random forest guided tour," *Test,* vol. 25, pp. 197-227, 2016.

[31] S. M. Lundberg, and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems,* vol. 30, 2017.

[32] C. J. Lamoth, E. Ainsworth, W. Polomski, and H. Houdijk, "Variability and stability analysis of walking of transfemoral amputees," *Medical engineering & physics,* vol. 32, no. 9, pp. 1009-1014, 2010.

[33] L. Quach, A. M. Galica, R. N. Jones, E. Procter-Gray, B. Manor, M. T. Hannan, and L. A. Lipsitz, "The nonlinear relationship between gait speed and falls: the maintenance of balance, independent living, intellect, and zest in the elderly of Boston study," *Journal of the American Geriatrics Society,* vol. 59, no. 6, pp. 1069-1073, 2011.

[34] B. Pogorelc, Z. Bosnić, and M. Gams, "Automatic recognition of gait-related health problems in the elderly using machine learning," *Multimedia tools and applications,* vol. 58, pp. 333-354, 2012.

[35] E. Halilaj, A. Rajagopal, M. Fiterau, J. L. Hicks, T. J. Hastie, and S. L. Delp, "Machine learning in human movement biomechanics: Best practices, common pitfalls, and new opportunities," *Journal of biomechanics,* vol. 81, pp. 1-11, 2018.

[36] J. H. van Dieen, N. P. Reeves, G. Kawchuk, L. R. van Dillen, and P. W. Hodges, "Motor Control Changes in Low Back Pain: Divergence in Presentations and Mechanisms," *Journal of Orthopaedic & Sports Physical Therapy,* vol. 49, no. 6, pp. 370-379, Jun, 2019.

[37] J. Cholewicki, A. P. Simons, and A. Radebold, "Effects of external trunk loads on lumbar spine stability," *Journal of biomechanics,* vol. 33, no. 11, pp. 1377-1385, 2000.

[38] N. W. Mok, and P. W. Hodges, "Movement of the lumbar spine is critical for maintenance of postural recovery following support surface perturbation," *Experimental brain research,* vol. 231, pp. 305-313, 2013.

[39] W. Van den Hoorn, S. Bruijn, O. Meijer, P. Hodges, and J. Van Dieën, "Mechanical coupling between transverse plane pelvis and thorax rotations during gait is higher in people with low back pain," *Journal of biomechanics,* vol. 45, no. 2, pp. 342-347, 2012.

[40] P. Madeleine, "On functional motor adaptations: from the quantification of motor strategies to the prevention of musculoskeletal disorders in the neck–shoulder region," *Acta Physiologica,* vol. 199, pp. 1-46, 2010.

[41] M. L. Meier, A. Vrana, and P. Schweinhardt, "Low Back Pain: The Potential Contribution of Supraspinal Motor Control and Proprioception," *Neuroscientist*, pp. 1073858418809074, Nov 2, 2018.

[42] A. Schinkel-Ivy, B. C. Nairn, and J. D. Drake, "Investigation of trunk muscle co-contraction and its association with low back pain development during prolonged sitting," *Journal of Electromyography and Kinesiology,* vol. 23, no. 4, pp. 778-786, 2013.

[43] G. B. Ross, P. J. Sheahan, B. Mahoney, B. J. Gurd, P. W. Hodges, and R. B. Graham, "Pain catastrophizing moderates changes in spinal control in response to noxiously induced low back pain," *Journal of biomechanics,* vol. 58, pp. 64-70, 2017.

[44] R. Parker, E. Bergman, A. Mntambo, S. Stubbs, and M. Wills, "Levels of physical activity in people with chronic pain," *South African Journal of Physiotherapy,* vol. 73, no. 1, pp. 1-7, 2017.

[45] J. H. van Dieen, N. P. Reeves, G. Kawchuk, L. R. van Dillen, and P. W. Hodges, "Analysis of Motor Control in Patients With Low Back Pain: A Key to Personalized Care?," *Journal of Orthopaedic & Sports Physical Therapy,* vol. 49, no. 6, pp. 380-388, Jun, 2019.

Appendix A.

We did empirical work of comparing often applied approaches for classification: random forest (RF), artificial neural network (ANN), support vector machine (SVM), naive bayes (NB), and K-nearest neighbours (KNN). As can be seen in the Table below, RF and ANN had the best performance compared to SV, NB, and KNN. RF performed better in precision and specificity, while ANN performed better in sensitivity.

Table 1. Classification performance comparison

	RF	ANN	SVM	NB	KNN (n=3)
Accuracy	84.4%	81.2%	68.8%	62.5%	62.5%
Sensitivity	75%	93%	56%	50%	50%
Specificity	93%	68.8%	81.3%	75%	75%
Precision	92%	75%	75%	66.7%	66.7%
F1-score	82.6%	83%	64.1%	57%	57%
AUC	0.83	0.85	0.67	0.62	0.62

RF: random frost; ANN: artificial neural network; SVM: support vector machine; NB: naive bayes; and KNN: K-nearest neighbors.

Chapter 5

Relationship Between Physical Activity and Central Sensitization in Chronic Low Back Pain: Insights from Machine Learning

Xiaoping Zheng, Michiel F Reneman, Rita HR Schiphorst Preuper,

Egbert Otten, Claudine JC Lamoth

Computer Methods and Programs in Biomedicine (2023) 232, 107432

Abstract

Background:

Chronic low back pain (CLBP) is a leading cause of disability. The guidelines for the management of CLBP often recommend optimizing physical activity (PA). Among a subsample of patients with CLBP, central sensitization (CS) is present. However, knowledge about the association between PA intensity patterns, CLBP, and CS is limited. The objective PA computed by conventional approaches (e.g., cut-points) may not be sensitive enough to explore this association. This study aimed to investigate PA intensity patterns in patients with CLBP and low or high CS (CLBP-, CLBP+ respectively) by using an advanced unsupervised machine learning approach, Hidden semi-Markov Model (HSMM).

Methods:

Forty-two patients were included (23 CLBP-, 19 CLBP+). CS-related symptoms (e.g., fatigue, sensitivity to light, psychological features) were assessed by a CS Inventory. Patients wore a standard 3D-accelerometer for one week and PA was recorded. The conventional cut-points approach was used to compute the time accumulation and distribution of PA intensity levels in a day. For the two groups, two HSMMs were developed to measure the temporal organization of and transition between hidden states (PA intensity levels), based on the accelerometer vector magnitude.

Results:

Based on the conventional cut-points approach, no significant differences were found between CLBP- and CLBP+ groups ($p=0.87$). In contrast, HSMMs revealed significant differences between the two groups. For the 5 identified hidden states (rest, sedentary, light PA, light locomotion, and moderate-vigorous PA), the CLBP- group had a higher transition probability from rest, light PA, and moderate-vigorous PA states to the sedentary state ($p<0.001$). In addition, the CBLP- group had a significantly shorter bout duration of the sedentary state ($p<0.001$). CLBP+ group exhibited longer durations of active ($p<0.001$) and inactive states ($p=0.037$) and had higher transition probabilities between active states ($p<0.001$).

Conclusions:

HSMM discloses the temporal organization and transitions of PA intensity levels based on accelerometer data, yielding valuable and detailed clinical information. The results imply that patients with CLBP- and CLBP+ have different PA intensity patterns. Patients with CLBP+ may adopt a distress-endurance response pattern with a prolonged bout duration of activity engagement.

Key words: Low back pain; Physical activity; Central Sensitization; Accelerometer; Daily life; Hidden semi-Markov Model, Chronic pain; Avoidance-endurance model.

1. Introduction

Chronic low back pain (CLBP) is a globally prevalent chronic musculoskeletal disorder [1] and is costly due to medical treatments and work productivity loss [2]. Additionally, it is a leading cause of high levels of disability [3]. In the management of CLBP, optimizing physical activity (PA) is often recommended [4]. However, the results of studies that examine PA intensity levels between patients with CLBP and healthy controls are inconsistent. Some studies suggest that people with CLBP exhibit lower overall PA intensity levels (averaged over one day or several days), compared to matched controls [5, 6], while others have reported that overall PA intensity levels do not differ significantly between patients with CLBP and healthy controls [7, 8]. With respect to the PA distribution during a day, one study reports that patients with CLBP have significantly lower overall PA intensity levels in the morning compared to healthy controls [8] while another study observes that during the evening, CLBP patients are significantly less active than healthy controls [9].

One explanation for the inconsistencies might be the heterogeneity within the population of CLBP. A subsample of patients with CLBP might have central sensitization (CS) [10]. CS is the increased responsiveness to noxious and non-noxious stimuli [11]. A recent paper proposes two major sub-types of CS: a "bottom-up" type and a “top-down” type [12]. The “bottom-up” type may be driven by persistent peripheral noxious input, causing an imbalance in excitatory and inhibitory central neurotransmitters, and altering gene regulations in the central nervous system, leading to central hyperexcitability. The “top-down” type is suggested to be driven without ongoing nociceptive input and the primary contribution may originate in supraspinal structures, including the psychosocial symptoms, such as fear-avoidance beliefs and pain catastrophizing thoughts. These symptoms may enhance forebrain activity, i.e., the neuronal activation of the prefrontal cortex and the limbic system, leading to central hyperexcitability. Central hypersensitivity eventually can spread and expand to multiple brain regions [12]. Consequently, patients with CS may be present with widespread hyperalgesia, fear-avoidance beliefs, pain catastrophizing thoughts, anxiety, and depression [12] The relation between CS and PA is still unclear but the plausibility of this relation can be derived from relations between fear-avoidance beliefs, pain catastrophizing thoughts, and pain levels on the one hand and CS on the other [13]. Both the fear-avoidance model [14] and the avoidance-endurance model (AEM) [15] postulated that fear beliefs and catastrophizing thoughts could eventually lead to physical inactivity and lower PA intensity levels [16]. However, their relationship with the measured PA is still inconclusive. The evidence on the relation between fear beliefs and PA is inconsistent [17-19]. To evaluate the relationship between AEM and PA in CLBP, one study compares the PA (measured by accelerometer) between avoiders and persisters (labelled by Patterns of Activity Measure-Pain questionnaire), but the results do not show differences between these two groups [17]. This finding is supported by another study which reports that, among a CLBP sample, none of the objective measures of PA were

associated with fear beliefs and this finding does not support one aspect of the fear-avoidance model that fear beliefs are associated with physical inactivity [18]. This might be due to the fact that PA intensity is studied in terms of average intensity levels over, for instance, days, which lacks sensitivity to detect small differences, since a recent study reports that fear beliefs and catastrophizing are associated with the distribution of PA during a day instead of the average PA levels [19].

Therefore, to obtain more in-depth knowledge about the impact of CLBP and CS levels on PA, the temporal organization (distribution) as well as the transitions between PA intensity levels provide important information. For instance, with the same accumulated time of sedentary activity, it makes a difference if someone takes regular small bouts of activity in-between sedentary times or not. Small bouts of activity can counteract the harmful effects of a prolonged sedentary period [20]. Similar, frequent short bursts of moderate-vigorous PA may have smaller health benefits than less frequent but longer periods of moderate-vigorous PA [21]. Knowing the differences in the PA intensity patterns of patients with CLBP and low or high CS will provide us with knowledge regarding the possible underlying mechanism of CLBP and assist in the development of tailored rehabilitation programs.

The well-established approach to study PA intensity levels is the cut-points approach which is based on predefined acceleration thresholds (e.g., activity counts/per minutes) to divide the acceleration data into sedentary (<100), light (100-1951), and moderate to vigorous (>1951) PA intensity levels based on healthy populations [22]. By using the predefined thresholds, this approach provides a standardized method for classifying PA intensity levels, allowing for comparability across studies. However, predefined thresholds for PA intensity levels are most often based on healthy populations which may not accurately reflect small changes in PA intensity levels of patient populations. For instance, PA levels of patients with lumbar spinal stenosis disease are all identified as sedentary and light, despite small changes in PA intensity having significant clinical implications [23]. Furthermore, the thresholds may vary with age, and may not provide an accurate and precise PA intensity level assessment for every individual. To get more accurate PA intensity levels, supervised machine learning approaches [24], such as Naive Bayes, K-nearest neighbours [25], random forest [26], support vector machines [27], and artificial neural network [28], are used to classify the PA intensity levels or predict the continued metabolic equivalent of task (MET) values based on the labelled accelerometer data. Although supervised machine learning approaches show potential for accurate PA intensity assessment, they require data that are labelled with different PA intensity levels or METs prior to the model training. To get rid of the need of a-priori information (predefined thresholds or labelled data), unsupervised machine learning approaches (such as K-means [29], density-based spatial clustering of applications with noise [30], and hierarchical clustering [29]) are used to cluster the PA intensity levels from

accelerometer data. However, the outcomes of the above approaches are usually used to compute average or summary statistics from one- or several-time windows instead of providing more in-depth information about the temporal organization and transitions between PA intensity levels. Hidden semi-Markov model (HSMM) [31] is considered as an unsupervised machine learning approach as well as a probabilistic graphical model [32]. The characteristics of the probabilistic graphical model enable HSMM to disclose the temporal organization and transition information of PA intensity levels. Therefore, HSMM has several advantages. First, it is a data-driven approach. It can get rid of expensive calibration or data labelling. Second, HSMM has higher resolution by clustering PA intensity levels (hidden states) from acceleration than the cut-points approach, instead of 4 levels of intensity (sedentary, light, moderate, and vigorous PA). Third, more in-depth information, such as temporal organization and transitions of PA intensity levels may provide a more parsimonious and biologically informative description of the data. Fourth, differences in temporal organization of and transitions between PA intensity levels from HSMM between groups can be statistically tested, and therefore, differences between groups may be recognized.

This study aimed to investigate whether low or high CS levels are associated with different PA intensity patterns (temporal organization of and the transitions between PA intensity levels) of patients with CLBP, using HSMM (based on daily-living accelerometer data). The highlights and contributions of this paper can be summarized as follows:

- The paper shows that the application of HSMM provides detailed and valuable clinical information about temporal organization and transitions of PA intensity levels based on accelerometer data.
- The analytic strategy used in this study provides a deeper understanding of the relationship between PA, CS, and CLBP.
- A comparison of the HSMM results with the conventional cut-points approach showed that the cut-points approach did not find statistical differences in patients with high or low levels of CS while HSMM did.
- Different PA intensity and transition patterns between patients with high levels and low levels of CS were found, implying the need for different intervention strategies

2. Material and Methods

2.1. Patients

Patients aged between 18 and 65 years old who had primary CLBP were recruited from the outpatient Pain Rehabilitation Department of the Center for Rehabilitation of the University Medical Center Groningen (CvR-UMCG). Primary CLBP is defined as low back pain persistent for more than three months, with pain not being the result of any other diagnosis [33]. Criteria for excluding patients from these studies were: (a) having a specific diagnosis that would

better account for the symptoms (e.g., cancer, inflammatory diseases and/or spinal fractures); (b) having neuralgia and/or radicular pain in the legs; (c) being pregnant.

The study was approved by the Medical Research Ethics Committee of the University Medical Center Groningen (METc 2016/702). Informed consent was obtained from all subjects and this study was conducted according to the principles expressed in the Declaration of Helsinki. The data used in this paper were derived from a larger study, for which protocol details were described elsewhere [34]. From the same sample, a study about gait analysis was published [35].

2.2. Data collection

Central sensitization (CS). The presence of CS-related manifestations was assessed with section A of the Central Sensitization Inventory (CSI) questionnaire [36]. Section A has 25-items to assess the presence of common CS-related symptoms. Scores can range from 0-100 where a higher score represents a higher level of CS. A score lower than 40 indicates low CS (CLBP- group) and a score of 40-100 indicates moderate-high CS (CLBP+ group) [37].

Secondary measures. Clinical data were measured via self-reported questionnaires. Visual Analogue Scale assesses the pain intensity (VAS Pain; 0-10). VAS scores lower than 3.4 represent mild pain, between 3.5 to 7.4 represent moderate pain, and higher than 7.5 represent severe pain [38]. Pain Disability Index (PDI; 0-70) with higher scores reflect higher interference of pain with daily activities. The physical functioning subscale of the Rand36 questionnaire (Rand36-PF; 0-100) with higher scores represent lesser disability. Pain Catastrophizing Scale (PCS, 0-52), Injustice Experience Questionnaire (IEQ, 0-48), and Brief Symptom Inventory (BSI global severity index t-score) were used to assess psychological-related symptoms. PCS scores [39] and IEQ scores [40] over 30 are associated with clinical relevance, and higher scores of BSI relate to more severe psychological symptoms.

Accelerometer data collection and formatting. Patients were instructed to wear a tri-axial accelerometer (ActiGraph GT3X, Actigraph Corporation, Pensacola, FL) at all times for about one week, excluding sleeping or bathing time. The accelerometer was worn at the front right hip of the patient (at the anterior superior iliac spine). The sampling frequency of the accelerometer was set to 100 Hz and the dynamic range was $\pm$6 gravity.

The raw data of each patient were firstly segmented into 24-h span data segments (from 12:00 P.M. to the next day 11:59 A.M). This time span was used because the measurement started at 12:00 P.M. Secondly, the data that did not completely cover this 24-h span were discarded. For technical or patient-related reasons, most of the patients did not have a full week of data. In order to compare the data between different patients fairly, 4 segments (representing 4 days) of each patient were randomly selected.

2.3. Conventional cut-points approach

The cut-points approach were set as sedentary (<100 counts/min), light physical activity (100 to 1951 counts/min), and moderate-to-vigorous physical activity (> 1952 counts/min). These thresholds have been reported by a former study [22]. The activity count [41] is an objective index to assess the energy expenditure of activities. It is generated from the accelerometer data following frequency filtering, thresholding, rectification, and integration processing. The accelerometer data were resampled to 30 Hz. Then, the other bandpass Butterworth filter with 4 orders was applied. Another filter with coefficient matrices which were found by [41] was used. The filtered data was truncated to 0 if the value was smaller than the threshold of 0.068. Finally, consecutive samples were accumulated into the 5-s epoch counts data. In this study, the activity count was computed based on accelerometer data by using the package ActigraphCounts [42].

2.4 Hidden semi-Markov model

For each 24-h span data segment, the gravity effects were removed from the 3-axis raw accelerometer data (details of this process are provided in Appendix A) and then the vector magnitude was computed as: $\text{acceleration} = \sqrt{accX^{\,2} + accY^{\,2} + accZ^{\,2}}$. This vector magnitude was simply called acceleration in this study. Next, the acceleration during unworn time in the night was removed from each segment. The unworn time was defined as the period in which acceleration values were all 0 and lasting longer than 3h. Finally, the acceleration was averaged across every 5s. These acceleration segments were used to train HSMMs later.

The acceleration segments of the CLBP- and CLBP+ group were separately used to train two HSMMs, respectively. HSMM was used to explore the different PA intensity levels from acceleration. It is an unsupervised machine learning approach, and it can automatically cluster hidden states from the observation data (acceleration). Hidden states are not directly observed from the acceleration but are inferred from clusters of the acceleration. The HSMM cannot provide the physical meaning of hidden states, but hidden states can be interpreted as PA intensity levels because they were clustered from the acceleration and contained characteristics of acceleration values and duration.

Before training the HSMM, the number of hidden states M should be determined. On the one hand, a high number of states may fit the observation data better but will increase the model complexity and the risk of overfitting. On the other hand, a small number of states is easier to interpret and more computationally efficient but may lead to underfitting. So, estimating M is a matter of balancing between model complexity and model performance. The Bayesian Information Criterion (BIC) has been widely used to help model selection and it takes model

complexity and model performance into consideration [43]. In this study, BIC was used to help to determine the number of states (More details of BIC are shown in Appendix B).

Let $S = \{s_1, s_2, \dots, s_M\}$ be a set of hidden states, and $O = \{o_1, o_2, \dots, o_T\}$ be the set of observations (acceleration). The HSMM is defined as $\lambda = (\pi, A, B, C)$:

- π is the initial probability distribution of hidden states. $\pi = \{\pi_1, \pi_2, \dots, \pi_M\}$, where $\sum_i \pi_i = 1$.
- A is the hidden states transition matrix and its elements are denoted by a_{ij},where $i \neq j$ and $\sum_i a_{ij} = 1$. a_{ij} represents the probability of hidden state s_i transiting to hidden state s_j.
- B is the emission probability matrix. The emission probability of observation data from time a to b, given the current state in s_i is $b_i(o_a^b)$. In this study, emission probability was modeled as Gaussian distributions.
- C is the duration probability matrix and it is represented by $c_i(d)$, where d is the duration of current hidden state s_i, $d \in \{1,2, \dots, D\}$. The durations are modelled as a discrete Poisson distribution. In order to reduce the training time, the maximum duration D is set to 2 hours.

The parameters of HSMM were learned by Bayesian estimation with a Gibbs sampler. The model was trained until the hamming distance between the assigned states of two consecutive iterations was smaller than 0.05. Then, the model was treated as convergent and had a stable parameter set. The implementation of HSMM was based on the package Pyhsmm [44].

After the model training, every acceleration segment has a corresponding hidden state sequence which shows the temporal distribution of PA intensity levels. The hidden state sequences are called fingerprints in this study because they represent the PA intensity distribution of individual patients. In order to show the full 24h fingerprint of each segment, the unworn time was added to the fingerprint based on the original timestamp. The input and output of the HSMM training process are shown in Fig. 1.

2.5 Statistical analysis

An independent T-test was applied to examine the differences in the overall or day-averaged PA intensity levels between the CLBP- and CLBP+ groups. These PA intensity levels were calculated using the cut-point approach.

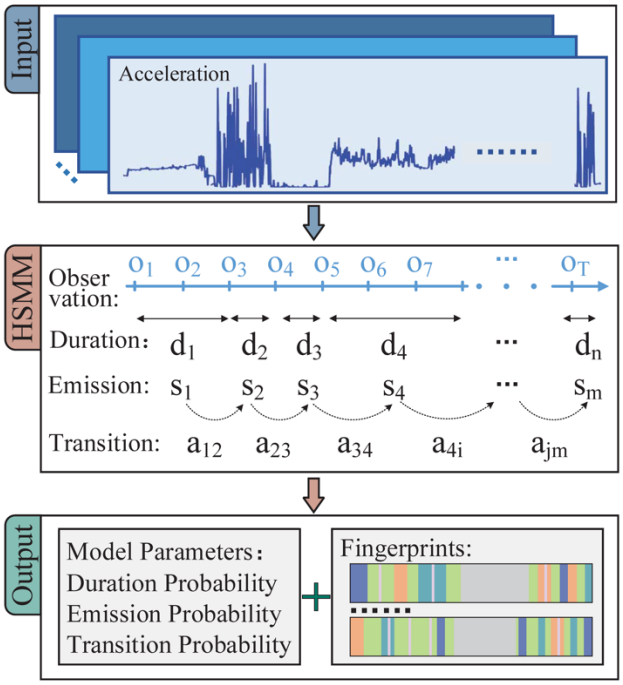


Figure. 1. HSMM training process. Input: the input data of the model; HSMM: hidden semi-Markov model; Output: the output after the model training; O: observation (acceleration) data; d: duration of the observation; s: hidden sate; a: transition probability.

Significant differences in outcomes between the two HSMMs trained by acceleration of the CLBP- and CLBP+ groups were assessed using independent T-tests (duration and accumulation of time of hidden states) or binomial proportion tests [45] (hidden state transitions) (details of binomial proportion tests are shown in Appendix C).

For all analyses, the significant level was set to 0.05. The Jensen–Shannon divergence (JSD) was used to measure the similarity between the emission probabilities. Smaller values (minimum 0) indicate more similarity. In this study, all analyses were conducted offline and were performed using Python.

To interpret the physical meaning of hidden states, the walking events were identified from the accelerometer data. Walking is a common activity during the day, and it is reported that the metabolic equivalent of task (MET) value of moderate pace walking (1.25–1.43 m/s) is 3.5 METs [46]. The details of walking event detection, step frequency, and walking speed calculations are shown in Appendix D.

3. Results

3.1. Demographic results

Of the 60 patients, 7 patients were excluded due to lack of measurements for sufficient numbers of days (<4 days) by accelerometer; 11 patients were excluded due to the missing

CSI scores. Finally, 42 patients were included in the data analysis. Patients in CLBP- and CLBP+ groups had similar demographic and clinical characteristics, with the exception of CSI score ($p < 0.0001$) and BSI ($p = 0.01$) (see Table 1 for demographics, for a more detailed demographics see [35]). CLBP- and CLBP+ had moderate pain and poor work ability. Their PCS scores [39] and IEQ scores [40] were lower than the clinically relevant level.

Table 1. Patient characteristics (n=42).

	CLBP- (n=23)	CLBP+ (n=19)	All (n=42)	P-Value
Sex	15W / 8M	12W / 7M	27W / 15M	
Age, years	40.8 ± 12.8	38.1 ± 12.7	39.6 ± 12.6	
Height, cm	173.5 ± 10.6	175.7 ± 8.8	174.5 ± 9.8	
Weight, kg	87 ± 17.7	85.4 ± 15.1	86.3 ± 16.4	
Body mass index, kg/m^2	28.9 ± 5.3	27.7 ± 4.4	28.3 ± 4.9	
Central Sensitization Inventory (0-100)	31± 4.8	48.7 ± 8.7	39.0 ± 11.2	< 0.01
Patient-reported Pain Intensity (VAS, 0-10)	5.5 ± 2	5.2 ± 1.8	5.4 ± 1.9	
Disability (PDI, 0-70)	33.6 ± 11.2	26.8 ± 11.9	31.0 ± 11.7	
Physical Functioning (Rand36-PF, 0-100)	49.8 ± 22.3	63.3 ± 16.1	54.7 ± 21.1	
Catastrophizing (PCS, 0-52)	16.3 ± 8.9	20.3 ± 11.1	18.1 ± 10	
Injustice (IEQ, 0-48)	15.2 ± 8.9	18.5 ± 8.5	16.7 ± 8.8	
Psychological Traits Screening (BSI, t-score)	34.4 ± 4.9	41.5 ±5.8	37.6 ± 6.4	= 0.01

Except sex, all results represent mean ± standard deviation. CLBP-, CLBP+: Patients with chronic low back pain with low (-) and moderate-high (+) central sensitization levels. W: Women; M: Men. VAS: Visual Analogue Scale. PDI: Pain Disability Index. Rand36-PF: Rand 36-Physical Functioning subscale. PCS: Pain Catastrophizing Scale. IEQ: Injustice Experience Questionnaire. BSI: Brief Symptom Inventory.

Because 23 patients with CLBP- and 19 patients with CLBP+ were included and 4 segments of every patient were randomly selected, 92 (29 on weekend) and 76 (24 on weekend) acceleration segments were used for the cut-points approach and for training of HSMMs for CLBP- and CLBP+ respectively.

3.2. Results from cut-points approach

Based on the cut-points approach, no statistically significant differences were found in the averaged PA intensity levels calculated over every hour of the day or over the whole day ($p = 0.87$) (see Fig. 2).

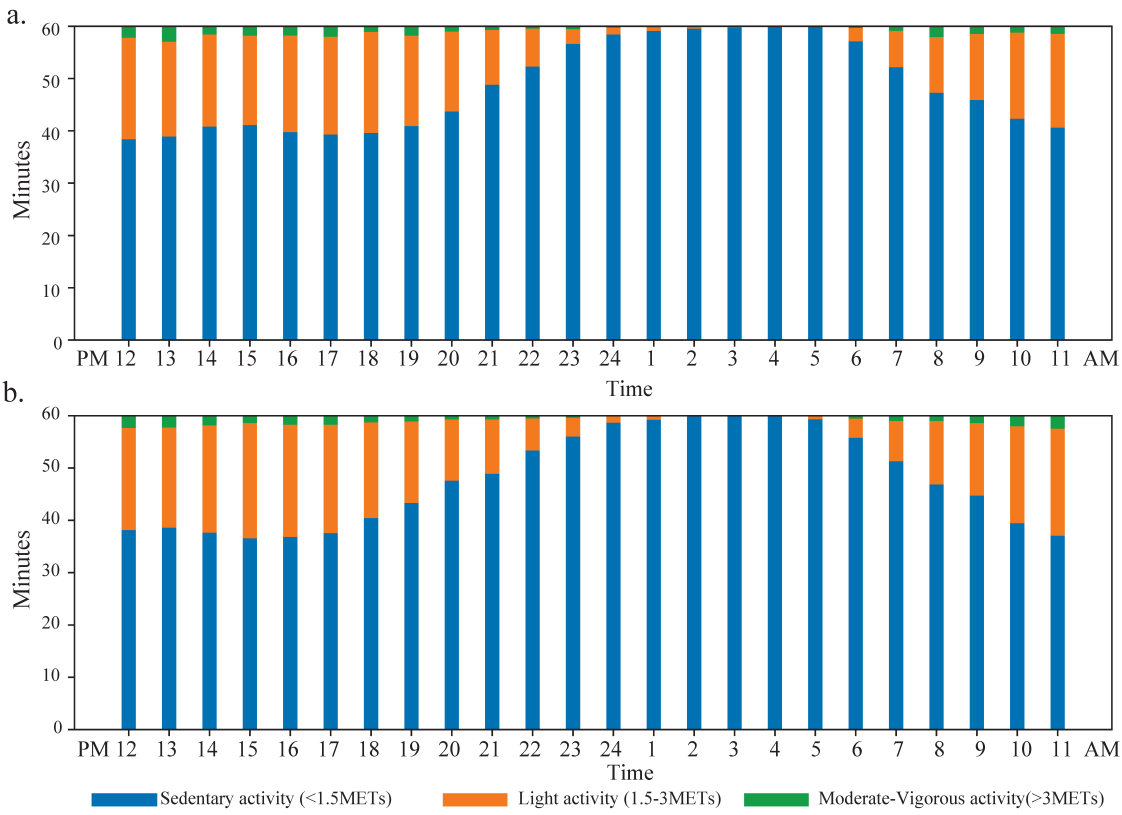


Figure. 2. Overall physical activity intensity levels of each hour during a day, based on the cut-points approach, of (a) the CLBP- group and (b) the CLBP+ group. CLBP-, CLBP+: Patients with chronic low back pain with low (-) or moderate-high (+) central sensitization levels.

3.3 Results from HSMM

The 4, 5, 6 hidden states HSMM were considered and BIC scores were computed. Taking the model fitting, model complexity, and model uncertainty into consideration, the number of hidden states was set to 5. More details are shown in Appendix E.

After the HSMMs training, the model parameters (emission probability, duration probability, and transition probability) were learned from the data.

3.3.1. Emission probabilities of hidden states

The emission probability distributions (modelled as Gaussian distributions) of hidden states 1 to 5 showed clear separation, with mean values dispersing approximately from 0 to 80 milli-gravity (mg). The differences of corresponding hidden states between CLBP- and CLBP+ were assessed by JSD values. All the JSD values are close to 0, suggesting similar emission probability distribution (see Table 2). Therefore, a direct comparison of hidden state sequences between the groups is allowed.

Table 2. Emission probability distribution per hidden state and Jensen–Shannon divergence of hidden state distribution between CLBP- and CLBP+.

Group	CLBP-		CLBP+		
	Mean (mg)	SD	Mean (mg)	SD	JSD
State 1	0.19	0.76	0.8	4.85	0.31
State 2	6.37	4.91	8.46	6.69	0.04
State 3	20.84	7.31	25.96	12.95	0.08
State 4	40.17	12.57	44.88	9.21	0.05
State 5	63.47	55.38	83.01	52.89	0.02

CLBP-, CLBP+: Patients with chronic low back pain with low (-) and moderate-high (+) central sensitization levels. SD: standard deviation. JSD: Jensen–Shannon divergence. mg: milli-gravity.

For both groups, the mean acceleration values of hidden state 1 were close to 0 mg which implied that patients were inactive, such as lying down (≈1.0 METs). The mean acceleration values of hidden state 2 were lower than 10 mg and may correspond to sedentary activities, such as desk work (≈1.3 METs). For hidden state 3, the mean values of acceleration were relatively low (smaller than 30 mg) and therefore, patients were assumed to do light PA during this state, for example, standing with manual applications (≈2.0 METs).

Based on the detection of walking events, it was shown that hidden states 4 and 5 contained 82.3% of the total walking events (35.9% and 46.4%, respectively), with walking speeds of 1.18 and 1.33 m/s, respectively. The speed of walking events in hidden state 4 was lower than 1.25 m/s, so hidden state 4 was interpreted as light locomotion state. The mean walking speed in hidden state 5 was within 1.25–1.43 m/s (1.33 m/s). It corresponded to 3.5 METs and was interpreted as moderate-vigorous PA state (> 3.0 METs).

Therefore, the 5 hidden states were specified as rest state (≈1.0 METs), sedentary state (≈1.5 METs), light PA state (≈2.0 METs), light locomotion state (≈2.0-3.0 METs), and moderate-vigorous PA state (>3.0 METs).

3.3.2. Duration and accumulation time of hidden states

Comparing the bout duration time of hidden states between the two groups, the CLBP- group exhibited a significantly shorter bout duration in the rest state ($P<0.001$) and the sedentary state ($P<0.001$). This group stayed for a longer bout duration in the light PA state ($P<0.001$) and the moderate-vigorous PA state ($P<0.001$) (see Fig. 3(a)).

In a day, the CLBP- group spent less accumulated time in the rest state ($P=0.002$) and the light PA state ($P<0.001$) and spent more time in the light locomotion state ($P=0.002$) and the moderate-vigorous PA state ($P<0.001$) (see Fig. 3(b)). Because the CLBP- and CLBP+ groups spend a similar accumulated time in the sedentary activity state ($P=0.4$) in a day, but the CLBP- group had a shorter bout duration in the sedentary activity state, the CLBP- group had more

frequent and shorter sedentary states. This was also shown in the transition probability matrixes.

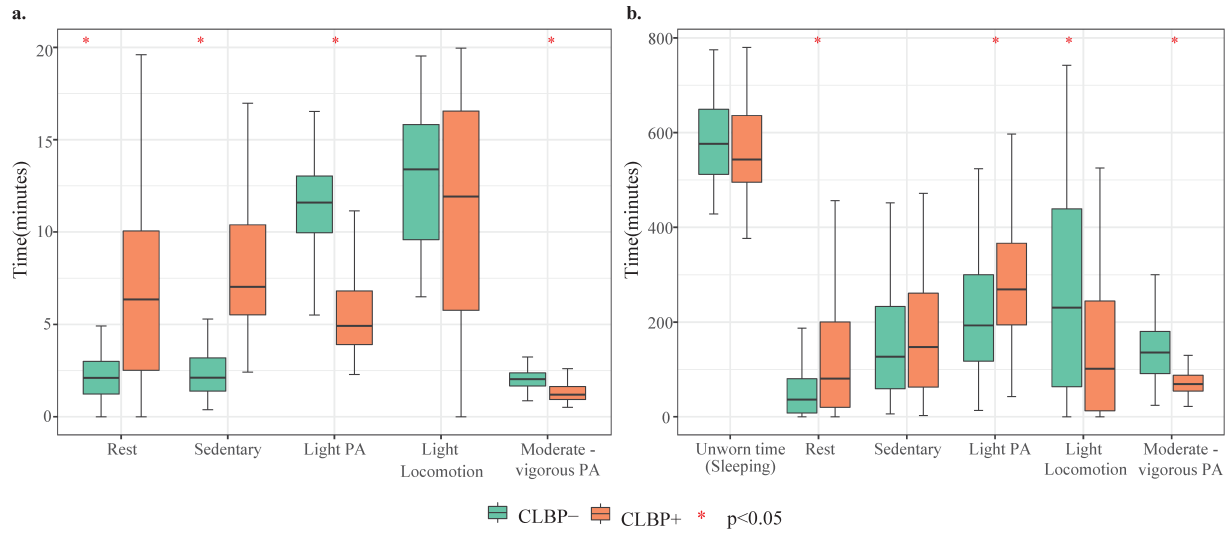


Figure. 3. Boxplots for (a) Bout duration of each hidden state, (b) Accumulated time of each hidden state and unworn time in a day. P value obtained by T-test. CLBP-, CLBP+: Patients with chronic low back pain with low (-) and moderate-high (+) central sensitization levels; PA: physical activity.

3.3.3. Transition probabilities of hidden states

The statistical differences in transition probabilities between the CLBP- and CLBP+ groups show that the CLBP- group had more frequent transitions from the rest, light PA, and moderate-vigorous PA states to the sedentary state (with P<0.001). The CLBP+ group, on the other hand, had more frequent transitions from the rest, sedentary, light PA, and moderate-vigorous PA states to the light PA state and/or the moderate-vigorous PA state (with P<0.001) (see Table 3, in red). Higher transitions from the light PA and moderate-vigorous PA states to the sedentary state may suggest that patients with CLBP- performed more frequent rest after being active. On the contrary, patients with CLBP+ had higher transitions from the active states (light PA and moderate-vigorous PA states) to other active states (moderate-vigorous PA states and light PA), which may suggest that they more often persist for a long period of activity.

To further investigate whether the patients in the CLBP+ group had longer bouts of activity than those in the CLBP- group, the light PA, light locomotion, and moderate-vigorous PA states were merged into a new state, named active state. The rest and sedentary states were merged as the inactive state. Fig. 4 shows that the CLBP+ group persisted significantly longer in both the inactive (P<0.001) and active states (P=0.037) than the CLBP- group.

Table 3. Transition probability matrix of CLBP- and CLBP+ groups.

From State	Group	To Hidden State				
		Rest	Sedentary	Light PA	Light locomotion	Moderate-vigorous PA
Rest	CLBP-	-	0.88	0.01	0.01	0.11
	CLBP+		0.40	0.21	0.00	0.38
Sedentary	CLBP-	0.39	-	0.10	0.03	0.47
	CLBP+	0.27		0.37	0.01	0.35
Light PA	CLBP-	0.01	0.42	-	0.14	0.42
	CLBP+	0.01	0.15		0.07	0.71
Light locomotion	CLBP-	0.01	0.14	0.17	-	0.67
	CLBP+	0.01	0.02	0.44		0.52
Moderate-vigorous PA	CLBP-	0.05	0.60	0.13	0.20	-
	CLBP+	0.10	0.13	0.68	0.09	

Statistically significant differences are printed in red (P<0.05). CLBP-, CLBP+: Patients with chronic low back pain with low (-) and moderate-high (+) central sensitization levels; PA: physical activity.

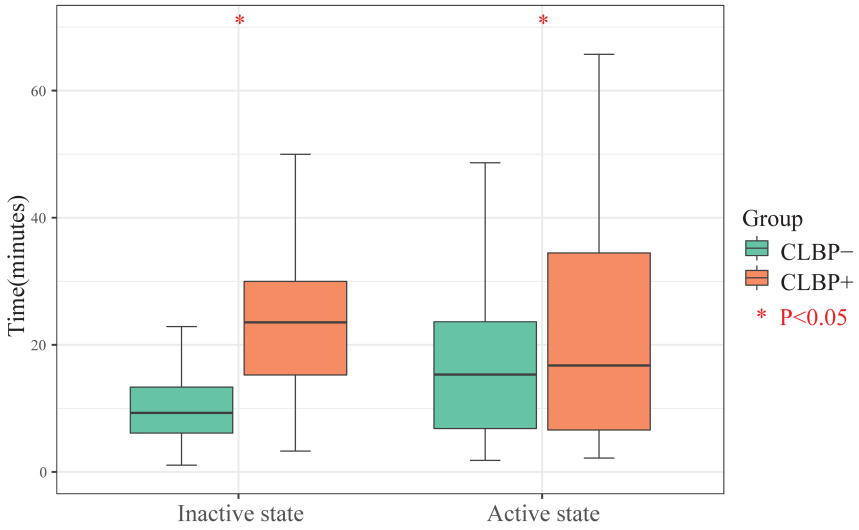


Figure. 4. Bout duration of inactive state and active state of CLBP- and CLBP+ groups. P value obtained by T-test. CLBP-, CLBP+: Patients with chronic low back pain with low (-) and moderate-high (+) central sensitization levels.

3.3.4. Group and individual levels HSMM information

The group level accumulated time of each hidden state and transition probabilities between hidden states can be graphically organized as a HSMM graph, or signature of the group; as shown in Fig. 5. Every circle represents a hidden state, and the area of the circle represents the accumulated time in a day. The figure shows that in the CLBP- group, the sedentary state was the most frequently present state (with 5 red arrows) while in the CLBP+ group, there were more red arrows between active states.

Apart from visualizing the outcomes of the HSMM at the group level, it can also provide details about individual level's hidden states organization information; the PA fingerprints (two examples are shown in Appendix F).

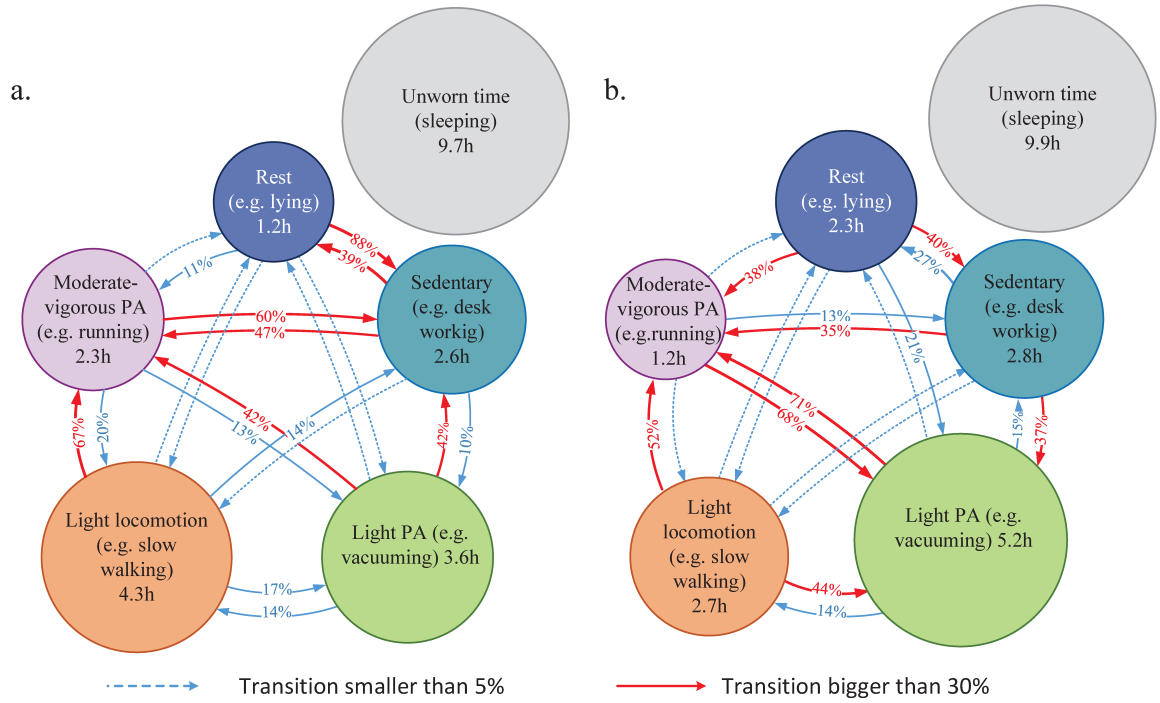


Figure. 5. Five states graphs (and unworn time) of (a) CLBP- and (b) CLBP+. The arrows between 2 circles represent the transition probability (in percentage) between these 2 corresponding circles. When the transition probability is higher than 30%, the arrow is colored red. CLBP-, CLBP+: Patients with chronic low back pain with low (-) and moderate-high (+) central sensitization levels; HSMM: Hidden semi-Markov models; PA: physical activity.

4. Discussion

The aim of this study was to explore in detail the relationship between PA, CS, and CLBP by analysing the temporal organization of and transition between PA intensity levels (PA intensity patterns) using HSMM. While the conventional cut-points approach did not show significant differences in averaged or overall PA intensity levels between the CLBP- and CLBP+ groups in a day, the data-driven approach, HSMM, revealed different PA intensity patterns between the two groups. The CLBP- group had a significantly shorter bout duration of the sedentary state and higher transition probabilities from the rest, light PA, and moderate-vigorous PA states to the sedentary state. On the contrary, the CLBP+ group exhibited a longer bout duration of the active state and higher transition probabilities between active states (light PA state, light locomotion state, and moderate-vigorous PA state) and longer bout duration of inactive states (rest and sedentary states).

Significant differences in temporal organization of and transition between PA intensity levels may suggest that patients with CLBP- and CLBP+ had different PA intensity patterns. The CLBP- group exhibited higher transition probabilities from the light PA and moderate-vigorous PA states to the sedentary state, and had shorter bout durations of the sedentary state compared with CLBP+ group. These findings may imply that patients in the CLBP- group broke tasks into smaller bouts and took frequent and short rests. Alternatively, patients in the CLBP+ group exhibited higher transition probabilities between active states and had longer bout durations of the active and inactive states. This may imply that patients in CLBP+ group exhibited a prolonged period of activity engagement and had a long period of rest. The PA intensity pattern of CLBP+ group may be in line with the endurance [47] type from the Avoidance-endurance model (AEM) [15]. AEM postulates that a subgroup of patients shows a pattern of fear-avoidance responses caused by high catastrophizing, fear of pain, and avoidance behaviour; another subgroup of patients shows a pattern of endurance responses with overuse and overload of physical structures, despite having pain. The PA intensity pattern of patients in the CLBP- group may not be assigned to fear-avoidance responses because of the lower PCS score; although frequent short rests are regarded as an avoidance strategy [48] to some degree. AEM distinguishes 2 types of endures responses, namely: distress-endurers (DE), who tend to have more pessimistic thoughts and feel more negative, and eustress-endurers, characterized by a positive mood. Endurance responses were assumed to exhibit a prolonged period of activity engagement which is in line with the results of CLBP+ group. The higher BSI score may suggest that patients with CLBP+ adopted the DE pattern. However, because the data in this study were derived from a larger study which lacks measurements of fear-avoidance beliefs and distress-endurance (e.g., avoidance-endurance questionnaire), the interpretation of the association between PA intensity patterns and the AEM model should be treated with caution. Consistent with others, the conventional cut-point approach with summary statistics analysis was not sufficiently sensitive to observe differences in objective PA levels [18]. The HSMM and analytic strategy used in this study show the potential to gain a better understanding of the relationship between daily PA intensity patterns, CS, and psychosocial factors, such as fear-avoidance or catastrophizing beliefs.

The findings of the current study do not support or refute the hypothesis that because of higher catastrophizing thought, fear of pain, or pain, patients with higher CS exhibit a fear-avoidance pattern. In the present study, catastrophizing thoughts and self-reported pain intensity between the CLBP+ and CLBP- groups were not significantly different. This finding is in line with the results of a previous study that observed that the majority of maladaptive responses in CLBP are characterized by endurance instead of fear-avoidance (i.e., 35.6% vs 9.6%) [49].

The current findings may imply the existence of a "bottom-up" type of CS in patients of the CLBP+ group [12]. Patients with the DE response pattern might expose the muscles to continuous stress and repetitive strain causing microdamage, laxity, and inflammation [50]. Ongoing nociceptive input may drive CS, which could explain why patients in CLBP+ group with DE pattern exhibited a higher level of CS. The CLBP- group showed more resting in-between activity periods, which may contribute to partially remitting features of CS because of the removal of ongoing nociceptive input [12], and consequently, CLBP- patients exhibited a lower level of CS. However, this study is cross-sectional, longitudinal studies are needed to gain insight in the causality or temporal relations between PA and CS in patients with CLBP.

No statistical differences were found in the accumulated time of the sedentary state between CLBP- and CLBP+ groups. This finding supports evidence from previous observations that sitting time did not correlate with CS [51]. The overall intensity levels of PA over a day between CLBP- and CLBP+ groups were not significantly different. Patients in the CLBP- group spent more time in the light-locomotion and moderate-vigorous PA states while patients in CLBP+ group spent more time in the light PA state. These findings support the suggestion that the quality and intensity of activities instead of the overall amount of PA is associated with levels of CS in patients with CLBP [51].

The population with CLBP is heterogeneous. For effective clinical interventions, it is important to gain more knowledge about subgroups of patients in this group. HSMM is applied to identify subgroups based on their PA intensity patterns and these subgroups may be linked to pain-related features, such as CS. This will contribute to a better understanding of underlying functional mechanisms of the development and maintenance of chronic pain and pain-related disability. In the present study, HSMM can not only provide information at the group level about PA temporal association, but also generate PA fingerprints for individual patients. These fingerprints may help patients and clinicians visualize the patients' everyday PA organization. Based on the specific PA intensity patterns from the subject's fingerprint, clinicians may adjust therapy accordingly and personalization of interventions may increase the effectiveness of treatment.

5. Limitations

Although accelerometer devices can nowadays be commercially purchased at a low cost, and algorithms and source code are available via open-source platforms, the data processing still requires an advanced level of knowledge and skills that are typically not held by clinicians. Further developments should be geared toward making these analyses more user-friendly to enable routine clinical use. The present study derived hidden states from one accelerometer sensor for feasibility reasons. Using multiple sensors would provide more information, but may cause more effort on data collection (e.g., subjects may be reluctant to wear for an

extended period of time because of lack of comfort). It should also be noted that a gold standard measure to diagnose CS is unavailable [52]. CSI is a broad assessment tool to indirectly measure CS, because higher scores are associated with the presence of CS syndromes [37]. Fifty-three acceleration segments were collected during weekends and PA may vary across workdays and weekends. However, both groups had almost the same proportion of weekend data (30%). The weekend data may not affect the comparison. Lifestyle factors (e.g., type of physical activity, sleep quality, and work load) and psychosocial (e.g., stress; anxiety, depression, and pain catastrophizing) factors are contributing factors that exacerbate CS [51]. Identifying the relationship between these factors with CS and CLBP may help to identify the most important modifiable factors that influence CS. This remains an important topic for future studies.

6. Conclusion

In this study, the results showed that the CLBP- and CLBP+ groups had different PA intensity and transition patterns. Patients in the CLBP- group had a higher transition probability from the rest, light PA, and moderate-vigorous PA states to the sedentary state and had a significantly shorter bout duration of the sedentary state. Conversely, patients in CLBP+ group exhibited longer durations of active and inactive states, and had higher transition probabilities between active states, which may support the suggestion that patients in CLBP+ group adopted a DE pattern. The results of this study may contribute to a better understanding of the relationship between PA, CS, and CLBP and will contribute to improve personalized rehabilitation prevention and interventions, and the development of CLBP-specific physical activity guidelines.

HSMM is able to automatically cluster the PA intensity levels from accelerometer data and provide detailed information about temporal organization and transitions of PA intensity levels. Hence, it can recognize the differences between the CLBP- and CLBP+ groups while the conventional cut-points approach is not sufficiently sensitive. This study highlights the potential use of HSMM in future research to explore the relationship between PA, CS, and CLBP, shedding a new light on the dynamics of PA intensity and transition patterns.

Reference

[1] M. S. Thiese, K. T. Hegmann, E. M. Wood, A. Garg, J. S. Moore, J. Kapellusch, J. Foster, and U. Ott, “Prevalence of low back pain by anatomic location and intensity in an occupational population,” *Bmc Musculoskeletal Disorders,* vol. 15, Aug 21, 2014.

[2] S. Dagenais, J. Caro, and S. Haldeman, “A systematic review of low back pain cost of illness studies in the United States and internationally,” *Spine Journal,* vol. 8, no. 1, pp. 8-20, Jan-Feb, 2008.

[3] T. S. Carey, J. K. Freburger, G. M. Holmes, L. Castel, J. Darter, R. Agans, W. Kalsbeek, and A. Jackman, “A Long Way to Go Practice Patterns and Evidence in Chronic Low Back Pain Care,” *Spine,* vol. 34, no. 7, pp. 718-724, Apr 1, 2009.

[4] M. F. Reneman, J. A. Echeita, K. van Kammen, H. R. S. Preuper, R. Dekker, and C. J. Lamoth, “Do rehabilitation patients with chronic low back pain meet World Health Organisation's recommended physical activity levels?,” *Musculoskeletal Science and Practice,* vol. 62, pp. 102618, 2022.

[5] M. Soysal, B. Kara, and M. N. Arda, “Assessment of physical activity in patients with chronic low back or neck pain,” *Turk Neurosurg,* vol. 23, no. 1, pp. 75-80, 2013.

[6] C. G. Ryan, P. M. Grant, P. M. Dall, H. Gray, M. Newton, and M. H. Granat, “Individuals with chronic low back pain have a lower level, and an altered pattern, of physical activity compared with matched controls: an observational study,” *Australian Journal of Physiotherapy,* vol. 55, no. 1, pp. 53-58, 2009.

[7] J. A. Verbunt, K. R. Westerterp, G. J. van der Heijden, H. A. Seelen, J. W. Vlaeyen, and J. A. Knottnerus, “Physical activity in daily life in patients with chronic low back pain,” *Arch Phys Med Rehabil,* vol. 82, no. 6, pp. 726-30, Jun, 2001.

[8] M. G. van Weering, M. M. Vollenbroek-Hutten, T. M. Tonis, and H. J. Hermens, “Daily physical activities in chronic lower back pain patients assessed with accelerometry,” *Eur J Pain,* vol. 13, no. 6, pp. 649-54, Jul, 2009.

[9] C. D. Spenkelink, M. M. Hutten, H. J. Hermens, and B. O. Greitemann, “Assessment of activities of daily living with an ambulatory monitoring system: a comparative study in patients with chronic low back pain and nonsymptomatic controls,” *Clin Rehabil,* vol. 16, no. 1, pp. 16-26, Feb, 2002.

[10] D. J. Clauw, “Diagnosing and treating chronic musculoskeletal pain based on the underlying mechanism(s),” *Best Pract Res Clin Rheumatol,* vol. 29, no. 1, pp. 6-19, Feb, 2015.

[11] T. Giesecke, R. H. Gracely, M. A. Grant, A. Nachemson, F. Petzke, D. A. Williams, and D. J. Clauw, “Evidence of augmented central pain processing in idiopathic chronic low back pain,” *Arthritis Rheum,* vol. 50, no. 2, pp. 613-23, Feb, 2004.

[12] S. E. Harte, R. E. Harris, and D. J. Clauw, “The neurobiology of central sensitization,” *Journal of Applied Biobehavioral Research,* vol. 23, no. 2, pp. e12137, 2018.

[13] E. Huysmans, K. Ickmans, D. Van Dyck, J. Nijs, Y. Gidron, N. Roussel, A. Polli, M. Moens, L. Goudman, and M. De Kooning, “Association Between Symptoms of Central Sensitization and Cognitive Behavioral Factors in People With Chronic Nonspecific Low Back Pain: A Cross-sectional Study,” *Journal of Manipulative and Physiological Therapeutics,* vol. 41, no. 2, pp. 92-101, Feb, 2018.

[14] J. W. S. Vlaeyen, and S. J. Linton, “Fear-avoidance and its consequences in chronic musculoskeletal pain: a state of the art,” *Pain,* vol. 85, no. 3, pp. 317-332, Apr, 2000.

[15] M. I. Hasenbring, and J. A. Verbunt, "Fear-avoidance and Endurance-related Responses to Pain: New Models of Behavior and Their Consequences for Clinical Practice," *Clinical Journal of Pain,* vol. 26, no. 9, pp. 747-753, Nov-Dec, 2010.

[16] B. Elfving, T. Andersson, and W. J. Grooten, "Low levels of physical activity in back pain patients are associated with high levels of fear-avoidance beliefs and pain catastrophizing," *Physiother Res Int,* vol. 12, no. 1, pp. 14-24, Mar, 2007.

[17] I. P. Huijnen, J. A. Verbunt, M. L. Peters, R. J. Smeets, H. P. Kindermans, J. Roelofs, M. Goossens, and H. A. Seelen, "Differences in activity-related behaviour among patients with chronic low back pain," *European Journal of Pain,* vol. 15, no. 7, pp. 748-755, 2011.

[18] F. A. Carvalho, C. G. Maher, M. R. Franco, P. K. Morelhao, C. B. Oliveira, F. G. Silva, and R. Z. Pinto, "Fear of movement is not associated with objective and subjective physical activity levels in chronic nonspecific low back pain," *Archives of physical medicine and rehabilitation,* vol. 98, no. 1, pp. 96-104, 2017.

[19] M. B. Miller, M. J. Roumanis, L. Kakinami, and G. C. Dover, "Chronic pain patients' kinesiophobia and catastrophizing are associated with activity intensity at different times of the day," *Journal of pain research,* vol. 13, pp. 273, 2020.

[20] Y. Kim, G. J. Welk, S. I. Braun, and M. Kang, "Extracting objective estimates of sedentary behavior from accelerometer data: measurement considerations for surveillance and research applications," *PLoS One,* vol. 10, no. 2, pp. e0118078, 2015.

[21] S. J. Strath, R. G. Holleman, C. R. Richardson, D. L. Ronis, and A. M. Swartz, "Peer reviewed: objective physical activity accumulation in bouts and nonbouts and relation to markers of obesity in US adults," *Preventing chronic disease,* vol. 5, no. 4, 2008.

[22] P. S. Freedson, E. Melanson, and J. Sirard, "Calibration of the computer science and applications, inc. accelerometer," *Medicine and science in sports and exercise,* vol. 30, no. 5, pp. 777-781, 1998.

[23] J. Norden, M. Smuck, A. Sinha, R. Hu, and C. Tomkins-Lane, "Objective measurement of free-living physical activity (performance) in lumbar spinal stenosis: are physical activity guidelines being met?," *The Spine Journal,* vol. 17, no. 1, pp. 26-33, 2017.

[24] V. Farrahi, M. Niemelä, M. Kangas, R. Korpelainen, and T. Jämsä, "Calibration and validation of accelerometer-based activity monitors: A systematic review of machine-learning approaches," *Gait & posture,* vol. 68, pp. 285-299, 2019.

[25] L. Atallah, B. Lo, R. King, and G.-Z. Yang, "Sensor positioning for activity recognition using wearable accelerometers," *IEEE transactions on biomedical circuits and systems,* vol. 5, no. 4, pp. 320-329, 2011.

[26] D. John, J. Sasaki, J. Staudenmayer, M. Mavilia, and P. S. Freedson, "Comparison of raw acceleration from the GENEA and ActiGraph™ GT3X+ activity monitors," *Sensors,* vol. 13, no. 11, pp. 14754-14763, 2013.

[27] A. Mannini, S. S. Intille, M. Rosenberger, A. M. Sabatini, and W. Haskell, “Activity recognition using a single accelerometer placed at the wrist or ankle,” *Medicine and science in sports and exercise,* vol. 45, no. 11, pp. 2193, 2013.

[28] A. H. Montoye, J. M. Pivarnik, L. M. Mudd, S. Biswas, and K. A. Pfeiffer, “Validation and comparison of accelerometers worn on the hip, thigh, and wrists for measuring physical activity and sedentary behavior,” *AIMS Public Health,* vol. 3, no. 2, pp. 298, 2016.

[29] P. Jones, E. M. Mirkes, T. Yates, C. L. Edwardson, M. Catt, M. J. Davies, K. Khunti, and A. V. Rowlands, “Towards a portable model to discriminate activity clusters from accelerometer data,” *Sensors,* vol. 19, no. 20, pp. 4504, 2019.

[30] C. Dobbins, and R. Rawassizadeh, "Towards clustering of mobile and smartwatch accelerometer data for physical activity recognition." p. 29.

[31] D. van Kuppevelt, J. Heywood, M. Hamer, S. Sabia, E. Fitzsimons, and V. J. P. o. van Hees, “Segmenting accelerometer data from daily life with unsupervised machine learning,” vol. 14, no. 1, pp. e0208692, 2019.

[32] P. Larrañaga, and S. Moral, “Probabilistic graphical models in artificial intelligence,” *Applied soft computing,* vol. 11, no. 2, pp. 1511-1528, 2011.

[33] S. Rozenberg, “Chronic low back pain: definition and treatment,” *La Revue du praticien,* vol. 58, no. 3, pp. 265-272, 2008.

[34] J. Ansuategui Echeita, H. R. Schiphorst Preuper, R. Dekker, I. Stuive, H. Timmerman, A. P. Wolff, and M. F. Reneman, “Central Sensitisation and functioning in patients with chronic low back pain: protocol for a cross-sectional and cohort study,” *BMJ Open,* vol. 10, no. 3, pp. e031592, Mar 8, 2020.

[35] X. Zheng, M. F. Reneman, J. A. Echeita, R. H. S. Preuper, H. Kruitbosch, E. Otten, and C. J. Lamoth, “Association between central sensitization and gait in chronic low back pain: Insights from a machine learning approach,” *Computers in biology and medicine,* vol. 144, pp. 105329, 2022.

[36] T. G. Mayer, R. Neblett, H. Cohen, K. J. Howard, Y. H. Choi, M. J. Williams, Y. Perez, and R. J. Gatchel, “The development and psychometric validation of the central sensitization inventory,” *Pain Practice,* vol. 12, no. 4, pp. 276-285, 2012.

[37] R. Neblett, H. Cohen, Y. Choi, M. M. Hartzell, M. Williams, T. G. Mayer, and R. J. Gatchel, “The Central Sensitization Inventory (CSI): establishing clinically significant values for identifying central sensitivity syndromes in an outpatient chronic pain sample,” *J Pain,* vol. 14, no. 5, pp. 438-45, May, 2013.

[38] A. M. Boonstra, H. R. S. Preuper, G. A. Balk, and R. E. Stewart, “Cut-off points for mild, moderate, and severe pain on the visual analogue scale for pain in patients with chronic musculoskeletal pain,” *Pain®,* vol. 155, no. 12, pp. 2545-2550, 2014.

[39] M. J. Sullivan, S. R. Bishop, and J. Pivik, “The pain catastrophizing scale: development and validation,” *Psychological assessment,* vol. 7, no. 4, pp. 524, 1995.

[40] M. J. Sullivan, "User Manual for the Injustice Experience Questionnaire IEQ," McGill University Montreal, 2008.

[41] S. L. Kozey, K. Lyden, C. A. Howe, J. W. Staudenmayer, and P. S. Freedson, "Accelerometer output and MET values of common physical activities," *Med Sci Sports Exerc,* vol. 42, no. 9, pp. 1776-84, Sep, 2010.

[42] J. C. Brønd. "ActigraphCounts," 11, 2021.

[43] A. A. Neath, and J. E. Cavanaugh, "The Bayesian information criterion: background, derivation, and applications," *Wiley Interdisciplinary Reviews: Computational Statistics,* vol. 4, no. 2, pp. 199-203, 2012.

[44] Matt J, Alex W, Yarden K, Chia-ying L, Scott L, Kevin S, and N. F. "Bayesian inference in HSMMs and HMMs," 11, 2021.

[45] V. Carola, O. Mirabeau, and C. T. Gross, "Hidden Markov Model Analysis of Maternal Behavior Patterns in Inbred and Reciprocal Hybrid Mice," *Plos One,* vol. 6, no. 3, Mar 8, 2011.

[46] B. E. Ainsworth, W. L. Haskell, S. D. Herrmann, N. Meckes, D. R. Bassett, C. Tudor-Locke, J. L. Greer, J. Vezina, M. C. Whitt-Glover, and A. S. Leon, "2011 Compendium of Physical Activities: A Second Update of Codes and MET Values," *Medicine and Science in Sports and Exercise,* vol. 43, no. 8, pp. 1575-1581, Aug, 2011.

[47] C. Eccleston, and G. Crombez, "Pain demands attention: A cognitive-affective model of the interruptive function of pain," *Psychological Bulletin,* vol. 125, no. 3, pp. 356-366, May, 1999.

[48] P. A. Karsdorp, and J. W. Vlaeyen, "Active avoidance but not activity pacing is associated with disability in fibromyalgia," *Pain®,* vol. 147, no. 1-3, pp. 29-35, 2009.

[49] M. I. Hasenbring, D. Hallner, B. Klasen, I. Streitlein-Bohme, R. Willburger, and H. Rusche, "Pain-related avoidance versus endurance in primary care patients with subacute back pain: Psychological characteristics and outcome at a 6-month follow-up," *Pain,* vol. 153, no. 1, pp. 211-217, Jan, 2012.

[50] M. I. Hasenbring, N. E. Andrews, and G. Ebenbichler, "Overactivity in Chronic Pain, the Role of Pain-related Endurance and Neuromuscular Activity An Interdisciplinary, Narrative Review," *Clinical Journal of Pain,* vol. 36, no. 3, pp. 162-171, Mar, 2020.

[51] K. Moriki, E. Tushima, H. Ogihara, R. Endo, T. Sato, and Y. Ikemoto, "Combined effects of lifestyle and psychosocial factors on central sensitization in patients with chronic low back pain: A cross-sectional study," *J Orthop Sci*, Aug 14, 2021.

[52] I. Schuttert, H. Timmerman, K. K. Petersen, M. E. McPhee, L. Arendt-Nielsen, M. F. Reneman, and A. P. Wolff, "The Definition, Assessment, and Prevalence of (Human Assumed) Central Sensitisation in Patients with Chronic Low Back Pain: A Systematic Review," *Journal of Clinical Medicine,* vol. 10, no. 24, pp. 5931, 2021.

Appendix A.

The accelerometer data was smoothed by the moving mean function (movmean() function in MATLAB 2017a), where AccX was determined as:

$$AccX'_i = \frac{1}{k} \sum_{n=i-\frac{2}{k}+1}^{i+\frac{2}{k}} AccX_n, n \in [0, length\ of\ AccX]$$

Here the moving window k was set to 100, the same as the sampling frequency. The window size is automatically truncated at the endpoints when there are not enough elements to fill the window.

The smoothed acceleration vector was defined as $v_i = [AccX'_i, AccY'_i, AccZ'_i]$, and u_i is the corresponding unit vector. To rotate the smooth acceleration vector into a vertical vector, the cross product of u_i and $[0,0,-1]$ was computed as attitude vector c_i and its unit vector is cu_i. Then, the angle between unit attitude vector cu_i and down vertical vector $[0,0,-1]$ was derived from their dot product. With this angle, the rotation matrix m was obtained using the atv2mat function (MATLAB 2017a). The rotation matrix was used to rotate the data as follows:

$$p_i = [AccX_i, AccY_i, AccZ_i]$$
$$pr_i = (m * p_i^T)^T = [x, y, z]$$
$$psr_i = (m * p_i^T)^T = [x', y', z']$$

The accelerometer data without gravity components was $[x, y, z - z']$.

Appendix B.

The Bayesian Information Criterion (BIC) is defined as:

$$BIC = -2logL + Plog(T)$$

where $logL$ is the logarithmized likelihood of the model, T indicates the length of the observation time-series, P denotes the number of independent parameters of the model. P is defined as:

$$P = m^2 + k * m - 1$$

where m is the umber of hidden states and k is the number of parameters of the underlying distribution of the observation process (e.g., k=2 for the normal distribution (mean and standard deviation)).

A bigger $logL$ and smaller P will lead to a smaller BIC which hints a better model. Here, a bigger $logL$ means the model is fitting well and a smaller P means the model has fewer parameters which may avoid overfitting. However, a bigger $logL$ is generated by a bigger P and a smaller P will lead to a smaller $logL$. Therefore, the BIC score is a trade-off and the model selection is not a straightforward procedure in the context of HMM. The choices of the number of hidden states remain subjective.

Appendix C.

A Binomial proportion test is conducted as follows:

$$Z = \frac{f_2 - f_1}{\sqrt{P(1-P)(\frac{1}{N_1} + \frac{1}{N_2})}}$$

$$P = \frac{f_2 N_2 + f_1 N_1}{N_1 + N_2}$$

where N_1, N_2 are the numbers of a hidden state i in the CLBP- and CLBP+ group, and $f_1 N_1$, $f_2 N_2$ are the numbers of hidden state transitions ij in 2 groups. Z scores can be converted to p values easily.

Appendix D.

a. Walking event detection

The accelerometer data was smoothed by a 20 Hz low-pass filter with a 2nd order Butterworth filter. The Fast Fourier Transform (FFT) [1] based approach was used to find the dominant frequency of the data in sliding windows and the length of the sliding window was set to 6 seconds. The data which 0.5–3.0 Hz was treated as potential walking events and the zero-cross approach [2] was employed to verify it.

b. Step Frequency

The step frequency is $SF = f/n$, where f is the sample frequency (in Hertz) and n is the number of samples per dominant period derived from autocorrelation.

c. Walking Speed

The walking speed is $WS = SF * l$, where l is the leg length. l was estimated as 48% of the body height [3].

[1] A. Chiarotto, R. W. Ostelo, M. Boers, and C. B. Terwee, "A systematic review highlights the need to investigate the content validity of patient-reported outcome measures for physical functioning in patients with low back pain," *Journal of Clinical Epidemiology,* vol. 95, pp. 73-93, Mar, 2018.

[2] J. Qian, L. Pei, J. Ma, R. Ying, and P. Liu, "Vector graph assisted pedestrian dead reckoning using an unconstrained smartphone," *Sensors,* vol. 15, no. 3, pp. 5032-5057, 2015.

[3] A. R. Frisancho, "Relative leg length as a biological marker to trace the developmental history of individuals and populations: growth delay and increased body fat," *American Journal of Human Biology,* vol. 19, no. 5, pp. 703-710, 2007.

Appendix E.

The relevant number of hidden states was calculated starting with a minimum of 4 states, since the cut-points approach uses 4 PA intensity levels (sedentary, light, moderate, and vigorous activity). To determine the actual number of hidden states, BIC scores were

computed. The averaged BIC scores of CLBP- and CLBP+ for the 4, 5, and 6 state HSMMs were 500578, 328256, and 267840, respectively. From 4 to 5 hidden state HSMMs, the BIC scores decreased by 34% while from 5 to 6 hidden state HSMMs, there was only an 18% decrease. Taking the model fitting, model complexity, and model uncertainty into consideration, the number of hidden states was set to 5.

Appendix F.

Fig. 1 shows two examples PA fingerprints from CLBP- and CLBP+. The fingerprint of a patient with CLBP- showed more frequent and shorter bouts of sedentary states between active states, while the fingerprint of a patient with CLBP+ showed prolonged bouts of active and inactive states.

Figure. 1. Example fingerprints from (a) CLBP- group and (b) CLBP+ group. CLBP-, CLBP+: Patients with chronic low back pain with low (-) and moderate-high (+) central sensitization levels; PA: physical activity.

Chapter 6

Establishing Central Sensitization Inventory Cut-off Values in Patients with Chronic Low Back Pain by Unsupervised Machine Learning

Xiaoping Zheng, Claudine JC Lamoth, Hans Timmerman, Ebert Otten, and Michiel F Reneman

Submitted for Publication

Abstract

Background:

Central sensitization (CS) cannot be directly demonstrated in humans and it is proposed to be referred to as Human Assumed Central Sensitization (HACS). HACS is involved in the development and maintenance of chronic low back pain (CLBP). Identifying HACS in individuals is crucial for tailoring appropriate treatment strategies, but there is no gold standard for assessing HACS. The Central Sensitization Inventory (CSI) was developed to evaluate the presence of HACS, with a cut-off value of 40/100 based on patients with chronic pain. However, various factors including pain conditions (e.g., CLBP, migraine, etc.), and gender may influence this cut-off value. For chronic pain conditions such as CLBP, unsupervised clustering approaches can take these factors into consideration and automatically learn HACS-related patterns. Therefore, this study aimed to determine the cut-off values for a Dutch-speaking population with CLBP, considering the total group and stratified by gender based on unsupervised machine learning.

Methods:

In this cross-sectional study, questionnaire data covering pain, physical, and psychological aspects were collected from patients with CLBP and aged-matched pain-free adults (referred to as healthy controls, HC). Four clustering approaches were applied to identify HACS-related clusters based on the questionnaire data and gender. The clustering performance was assessed using internal and external indicators. Subsequently, receiver operating characteristic (ROC) analysis was conducted on the best clustering results to determine the optimal cut-off values.

Results:

The study included 151 subjects, consisting of 63 HCs and 88 patients with CLBP. Hierarchical clustering yielded the best results, identifying three clusters: healthy group, CLBP with low HACS level, and CLBP with high HACS level groups. Based on the low HACS level group (including HC and CLBP with low HACS level) and high HACS level group, the cut-off values for the overall groups were 35 (sensitivity 0.76, specificity 0.76), 34 for females (sensitivity 0.72, specificity 0.69), and 35 for males (sensitivity 0.92, specificity 0.81).

Conclusion:

The findings suggest that the optimal cut-off value for CLBP is 35. The gender-related cut-off values should be interpreted with caution due to the unbalanced gender distribution in the sample. The methodology employed in this study may provide new insights into identifying HACS-related patterns and contributes to establishing accurate cut-off values.

Keywords: Unsupervised machine learning; central sensitization; human assumed central sensitization; low back pain; central sensitization inventory; cutoff value.

1. Introduction

Central Sensitization (CS) refers to an increased responsiveness of nociceptive neurons in the central nervous system to their normal or subthreshold afferent input [1]. However, due to the inability to measure the mechanisms related to CS in individual humans [2], a term known as Human Assumed Central Sensitization (HACS) [2] has been proposed to refer to CS. HACS has been implicated in the development and maintenance of various chronic pain conditions, such as Chronic Low Back Pain (CLBP), fibromyalgia, and osteoarthritis [3]. CLBP is a leading contributor to global disability [4]. While the overall efficacy of rehabilitation for patients with CLBP is generally positive, the average effect sizes are modest [5]. The possible presence of HACS is one of the key factors contributing to the complexity of CLBP [6] which could be among the factors responsible for the modest treatment effects [7]. Recognizing HACS in individuals with CLBP is crucial for tailoring appropriate treatment strategies, as interventions targeting CS may differ from those addressing peripheral mechanisms [8, 9].

Despite its importance in recognizing HACS in CLBP, there is currently no universally accepted gold standard for diagnosing HACS [2]. The Central Sensitization Inventory (CSI) questionnaire was developed as a self-report questionnaire to screen for the presence and severity of HACS in individuals experiencing musculoskeletal pain [10]. The CSI has demonstrated good psychometric properties in various pain conditions [11]. A cut-off value of 40 out of 100 was established based on a study involving patients with chronic pain with CS syndromes ((CSS), e.g., fibromyalgia, chronic fatigue syndrome, etc.), and it has demonstrated good sensitivity (81%) and specificity (75%) [12]. However, it has been observed that the cut-off values for CSI vary across different types of musculoskeletal pain, ranging from 11 to 40 [12-15], as well as across different cultural and national contexts [16]. This variability highlights the need for establishing context-specific cut-off values to improve the utility of the CSI in diverse populations. Moreover, gender plays a significant role in pain, potentially affecting the presentation and severity of HACS and ultimately influencing the determination of the cut-off value [17, 18]. To the best of our knowledge, there is no cut-off value established for Dutch-speaking patients with CLBP. Because of the lack of gold standard, the relationship between the CSI and HACS remains ambiguous. It is uncertain whether the CSI indicates enhanced nociceptive responses or also a psychological hypervigilance [19, 20].

To address these challenges, unsupervised machine learning [21, 22] may be a possible approach. Unsupervised clustering approaches are data-driven, and can automatically learn the relationships between variables and explore the possible HACS-related subgroups based on the questionnaire data that reflect pain, physical functioning, psychological factors, and HACS. These clustering approaches do not rely on prior knowledge or assumptions about the underlying structure of the data and can identify distinct groups within the data based on patterns of HACS. Based on the clustering results, researchers can uncover the optimal cut-

off value that best differentiates individuals with low and high levels of HACS in a data-driven and context-specific manner. Apart from this, these approaches are flexible and can be applied to various types of data [23], such as demographic, cultural, and psychosocial factors, making them suitable for analyzing the complex and multidimensional nature of HACS. Additionally, the good scalability [23] of these approaches makes them easily scalable to accommodate large datasets, such as electronic health record systems. In the future, by collecting more diverse and representative samples of Dutch-speaking patients with CLBP, this scalability ensures that the established cut-off value is robust and generalizable to the broader population with CLBP.

In this study, by using questionnaires which provide information about pain, physical, and psychological aspects, we aim to 1) explore the HACS-related subgroups based on unsupervised clustering approaches; 2) establish the optimal cut-off values of CSI within the Dutch-speaking population with CLBP based on clustering results; and examine gender differences in optimal cut-off values.

2. Methods

2.1. Participants

The data of Dutch-speaking patients with CLBP utilized in the present study was extracted from an existing dataset of a broader study [24]. Data collection took place from September 2017 to September 2019, and comprehensive protocol details have been previously documented [24]. The aged-matched Dutch-speaking healthy controls (HC) were recruited by advertisements on social media and flyers.

Patients with CLBP were recruited from the outpatient Pain Rehabilitation Department at the Center for Rehabilitation of the University Medical Center Groningen (CvR-UMCG). CLBP is characterized by recurring pain in the lower back lasting for over 3 months. This pain is associated with emotional distress and/or functional disability and is not caused by any other diagnosis [25]. Inclusion criteria were as follows: 1) age ≥ 18 years; 2) admission to the interdisciplinary pain rehabilitation program; 3) ability to follow instructions; 4) signed informed consent. Patients were excluded if they: 1) had a specific diagnosis that better accounted for their CLBP symptoms (e.g., cancer, inflammatory diseases, or spinal fractures); 2) experienced neuralgia and/or radicular pain in the legs (examination by physiatrist); 3) were pregnant. The presence of comorbidities related to HACS (e.g., fibromyalgia, osteoarthritis or chronic fatigue syndrome) is no reason for exclusion from the study. The HCs were included if they: 1) were aged ≥ 18 years; 2) could follow instructions; 3) provided signed informed consent. Exclusion criteria for healthy controls: 1) report more than mild pain (evaluated by Visual Analogue Scale, see below); 2) use of antidepressant or antiepileptic drugs at the time of completing the questionnaire.

The Dutch-speaking patients with CLBP were collected with the approval of the Medical Research Ethics Committee of the University Medical Center Groningen (METc 2016/702). All procedures were conducted in accordance with the ethical principles outlined in the Declaration of Helsinki.

2.2. Measures

In this study, eight questionnaires assessed central HACS-related factors, including pain, physical functioning, psychological aspects, and HACS.

Pain was measured by Visual Analogue Scale (VAS), with values ranging from 0 to 100 mm. Values below 44 mm represent mild pain, 45 to 74 mm indicate moderate pain, and above 75 mm signify severe pain [26].

Functioning was evaluated using the Pain Disability Index (PDI) [27], the physical functioning subscale of the Rand36 questionnaire (Rand36-PF) [28], and the Work Ability Score (WAS) [29]. Higher PDI values (0-70) reflect greater pain interference with daily activities, while higher Rand36-PF values (0-100) indicate lower disability. WAS assessed self-reported work ability, with higher values representing better work ability.

Psychological Aspects were measured using the Pain Catastrophizing Scale (PCS, 0-52) [30], the Injustice Experience Questionnaire (IEQ, 0-48) [31], and the Brief Symptom Inventory (BSI global severity index t-score). PCS and IEQ values over 30 are clinically relevant, and higher BSI values denote more severe psychological symptoms.

HACS was evaluated by the CSI part A. CSI values can range from 0-100, with higher values assuming a higher level of CS [32]. Only section A was utilized in this study.

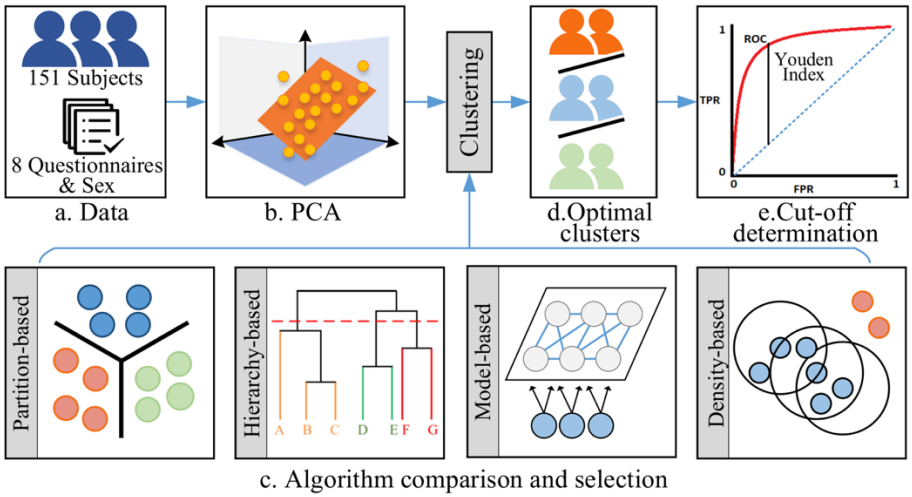


Figure 1. The data processing and analysis pipeline: (a) data collection; (b) PCA; (c) Clustering algorithms comparison and selection; (d) Optimal clusters; (e) Cut-off determination.

The data processing pipeline is depicted in Figure 1. Initially, questionnaire data and gender information were obtained from 151 subjects (Fig. 1(a)). Subsequently, the data were standardized using the Z-score approach, and Principal Component Analysis (PCA) was employed to reduce dimensionality (Fig. 1(b)). The first four components, accounting for 80% of the variance, were utilized. At the end, the features in each data sample which represents each subject were reduced from 9 to 4. Four kinds of clustering approaches were applied (Fig. 1(c)) to identify potential HACS-related groups. The optimal clusters were determined based on the most effective clustering results (Fig. 1(d)). Lastly, receiver operating characteristic (ROC) analysis was conducted on these clusters to ascertain the best cut-off values for CSI (Fig. 1(e)).

2.3. Clustering approach

After pre-processing, to find the most suitable clustering approach for this study, 4 kinds of clustering approaches were included: K-means (partition based), Hierarchical clustering (hierarchy based), Self-organizing map (model-based), and Density-based spatial clustering of applications with noise (DBSCN) (Density-based), see Fig. 1(c).

K-Means is one of the most commonly used clustering approaches based on partition [33]. The basic idea of this kind of clustering algorithm is to regard the center of data points as the center of the corresponding cluster. The algorithm can be summarized by the following procedure.

1) Randomly select K data samples as the centroid of K clusters and form a cluster prototype matrix as $M = [m_1, m_2, \dots, m_k]$. In this study, the cluster number K was set to 3.
2) Assign each data sample in the dataset to the nearest cluster.
3) Recalculate the centroid of the K clusters based on the current partition and update the cluster prototype matrix $M' = [m'_1, m'_2, \dots, m'_K]$..

$$m'_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$$

where C_i is the clusters i with the centroid m_1, x is the data sample in the dataset.

4) Repeat steps 2)-3) until there is no change for each cluster.

At the end, all data samples will be assigned to one of the three clusters to represent different HACS-related patterns. The K-means algorithm works well for compact and hyper spherical clusters.

Hierarchical clustering is a hierarchy-based clustering approach, which constructs the hierarchical relationship among data in order to cluster [34]. In this study, clusters are formed by iteratively dividing the patterns using a bottom-up approach. The agglomerative clustering can be summarized by the following procedure.

1) Start with N atomic clusters which each of them includes exactly one data sample. N represents the number of data samples of the whole data set, where $N = 151$. Calculate the Euclidean distance between any 2 clusters and form a proximity matrix.
2) Search the minimal distance d_{ij}in the proximity matrix and combine C_i and C_j to form a new cluster $C_{i\cup j}$.
3) Update the proximity matrix by computing the distances $d_{k(ij)}$ between the new cluster and the other clusters. $d_{k(ij)}$ can be computed by the Lance–Williams algorithm as follows.

$$d_{k(ij)} = \alpha_i d_{ik} + \alpha_j d_{jk} + \beta d_{ij} + \gamma \left| d_{ik} - d_{jk} \right|$$

$$\alpha_i = \frac{n_i + n_k}{n_i + n_j + n_k}, \alpha_j = \frac{n_j + n_k}{n_i + n_j + n_k}, \beta = \frac{-n_k}{n_i + n_j + n_k}, \gamma = 0$$

where $n_i, n_j$and n_k are the sizes of the disjoint clusters C_i, C_j and C_k.

4) Repeat steps 2)-3) until all objects are in the same clusters.

The results of hierarchical clustering are depicted by a dendrogram. The dendrogram describes the similarity between different data samples. The ultimate clustering results can be obtained by cutting the dendrogram at different levels.

Self-organizing map (SOM) is a model-based approach which is based on neural network learning approaches [35]. The core idea of SOM is to build a map of dimension reduction from the input space of high dimension to output space of low dimension on the assumption that there exists topology in the input data. The algorithm can be summarized by the following procedure.

1) Define the topology of the SOM with size $X * Y$.

$$X = Y = \sqrt{5\sqrt{N}}$$

where N is the number of data samples (subjects) of the whole dataset (N=151).

2) Randomly initialize the prototype weight matrix $W(0)_{X*Y*D}$ for the SOM network, where D is the number of features in the input data. In this study, D is equal 4 since the PCA reduced the dimension of each data sample from 9 to 4.
3) Calculate the distance between an input data sample x and the nodes of network. The node $r_c$which is closest to x is chosen as the winning node.
4) Update the weight matrix $W(t)$ as

$$w_{(i,j)}(t+1) = w_{(i,j)}(t) + \eta g_{(i,j)} \cdot (x - w_{(i,j)}(t))$$

where $w_{(i,j)}(t)$ is the weight of node at location (i,j) at time t, ηg represents the neighborhood function and η is the learning rate. $g_{(i,j)}$ is defined by the Gaussian method as

$$g_{(i,j)} = \exp\left(\frac{-||r_c - r_{(i,j)}||^2}{2\sigma^2}\right)$$

where $r_{(i,j)}$is the node at the location (i, j) in network.

5) Repeat steps 3)-4) until no change of neuron position that is more than a small positive number is observed.

Density-based spatial clustering of applications with noise (DBSCAN) is a clustering approach based on density, proposed by Ester Martin. It can find a cluster with any shape upon one density condition [36]. DBSCAN has the following basic concepts:

1) Set the radius of DBSCAN Algorithm Analysis neighborhood as ε, and set the minimum number of data sample sets as $MinPts$. In this study, ε and $MinPts$ were empirically determined and set to 15.
2) Randomly select one unvisited data sample P and mark it as visited. If $N_\varepsilon(P) > MinPts$, then mark this data sample P as core data sample. The $N_\varepsilon(P)$ is defined as

$$N_\varepsilon(P) = \{q \in D | dist(p, q) < \varepsilon\}$$

where q is a data sample from the dataset D, $dist(p, q)$ means the distance between P and q. If $dist(p, q) < \varepsilon$, q and P are directly density-reachable, such that $N_\varepsilon(P)$ is the number of data sample directly density-reachable from P.

3) Find out all the density-reachable data sample from P, mark them as visited and merge to the same cluster as P. The definition of density-reachable is: if there is a chain of objects $P_1, \ldots, P_n$, where $P_1 = P$, $P_n = Q$, and P_{i+1} is directly density-reachable from P_i, such that P is density-reachable from Q.
4) Repeat steps 2)-3), until all the objects are visited. The data samples which are visited but not in the clusters are noise data samples.

2.4. Clustering performance evaluation

To evaluate the clustering results of the unsupervised clustering approaches, internal and external validation measures were used. Internal validation measures, including the silhouette coefficient, Davies-Bouldin index, and Calinski-Harabasz index, evaluate cluster quality based on clustering results. For external validation, clustering results were compared with known labels (HC). External validation assists in determining clustering accuracy and ensuring the meaningfulness of clustering outcomes.

The *Silhouette Coefficient* indicates the cohesion of an object within its own cluster, and the separation of this object and other clusters [37]. A value close to 1 means clusters are well apart from each other and clearly distinguished. The definition of the silhouette coefficient SC is:

$$SC = \max_K \tilde{s}(K)$$

where K represents the number of clusters and $\tilde{s}(K)$ computes the mean value of $s(i)$ in cluster K. The definition of $s(i)$ is:

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

where,

$$a(i) = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(i,j)$$

$$b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i,j)$$

where $|C_I|$ represent the size of cluster I, $d(i,j)$ represents the distance between data sample i, and j.

The *Davies-Bouldin index* is a function of the ratio of the within-cluster scatter, to the between-cluster separation [38]. A lower value will mean that the clustering is better. The scatter of cluster I is defined as:

$$S_I = \frac{1}{|C_I|} \sum_{i=1}^{|C_I|} dis(i, A_I)$$

where A_I is the centroid of C_I. The definition of separation between clusters I and J is:

$$M_{I,J} = dis(A_I, A_j)$$

Davies-Bouldin index of K clusters results can be computed as:

$$DB = \frac{1}{K} \sum_{I=1}^{K} \max_{I \neq J} \{\frac{S_I + S_J}{M_{i,J}}\}$$

A lower non-negative value of DB means the better clustering results.

The *Calinski-Harabasz index* is a measure of how similar a data sample is to its own cluster (cohesion) compared to other clusters (separation) [39]. The cohesion is estimated based on the distance from the objective in a cluster to its cluster centroid and the separation is based on the distance of the cluster centroids from the global centroid. This cohesion can be defined as:

$$CO = \frac{\sum_{I=1}^{K} |C_I| dis(A_I - A)}{K - 1}$$

where A is the global centroid of the whole dataset.

The inter-cluster dispersion can be defined as:

$$SE = \frac{\sum_{I=1}^{K} \sum_{i}^{|C_I|} dis(i, A_I)}{N - K}$$

where N is the number of data samples of the whole dataset.

Therefore, the Calinski-Harabasz index is:

$$CH = CO/SE$$

A higher value of the CH index means the clusters are dense and well separated.

2.5. Statistical analyses

In this study, the Mann–Whitney U test was applied to examine the differences in demographic characteristics. Based on the optimal clusters, the receiver operating characteristic (ROC) curve analysis [40] was used to suggest the optimal CSI cut-off value (shown in Fig. 1(e).). The area under the ROC curve (AUC) represents the harmonic ratio of sensitivity and specificity. The Youden index was calculated to evaluate the performance of the accuracy of a diagnostic test. Positive predictive values (PPV) and negative predictive values (NPV) serve as valuable indicators of diagnostic accuracy, reflecting the proportion of true positives (corresponding to high level of HACS) and true negatives (corresponding with low level of HACS) among all positive and negative findings, respectively. These metrics contribute to a comprehensive understanding of the diagnostic performance. In addition to PPV and NPV, likelihood ratios are employed as critical statistical measures to assess the diagnostic efficacy of tests. The positive likelihood ratio (PLR) is calculated by dividing the true positive rate by the false positive rate. Similarly, the negative likelihood ratio (NLR) is determined by dividing the false negative rate by the true negative rate. These ratios provide insights into the ability of diagnostic tests to discriminate between individuals with different CS levels.

As a whole, the combined utilization of AUC, Youden-index, sensitivity, specificity, predictive values, and likelihood ratios was used to determine the optimal CSI cut-off values. For clinical use, the cut-off value should have a sensitivity plus specificity of at least 1.5, which is halfway between 1 (useless) and 2 (perfect) [41].

To ensure transparency and reproducibility of the findings, we have made the project repository publicly accessible at https://github.com/xzheng93/CSI_cutoff_establishment.

3. Results

In this study, 296 subjects were included, while 139 subjects were excluded due to the incomplete questionnaires data, 6 subjects in the HC group were excluded since they reported moderate pain. Therefore, 151 subjects (63 HC and 88 CLBP) were included in the data analysis. Table 1 shows the characteristics. The HC and CLBP groups were age-matched and were significantly different in BMI, but not in height and weight. The HC group reported less pain, better physical functioning, better psychological status, and lower CSI values. In terms of gender, females and males had matched BMI. Females were younger and smaller, and reported more pain, more disability, worse psychological status (e.g., depression, anxiety, distress, and pain catastrophising), and higher CSI values compared to male.

Table 1. Demography of participants

	HC (n=63)	CLBP (n = 88)	p-value	F(n=74)	M(n=77)	p-value
Gender	23F/40M	51F/37M	-	-	-	-
Age, years	40.9 ± 13.5	41.4 ± 12.3	=.988	37.5 ± 13.7	44.7 ± 10.7	<.001
Height, cm	160.5 ± 56.7	175.2 ± 10.1	=.085	158.6 ± 42.7	179.1 ± 29.9	<.001
Weight, kg	83.6 ± 17.0	86.6 ± 16.9	=.131	79.1 ± 16.3	91.3 ± 15.4	<.001
BMI, kg/m^2	25.8 ± 3.9	28.3 ± 5.6	=.003	27.5 ± 6.2	27.0 ± 3.8	=.64
VAS (0–10)	0.5 ± 0.7	4.2 ± 2.3	<.001	3.3 ± 2.8	2.1 ± 2.3	=.008
PDI (0–70)	5.1 ± 7.4	29.8 ± 14.4	<.001	22.8 ± 16.1	16.3 ± 17.4	=.009
WAS (0–10)	8.5 ± 1.3	4.9 ± 2.4	<.001	5.9 ± 2.9	6.8 ± 2.3	=.088
Rand36-PF (0–100)	28.4 ± 2.4	56.2 ± 21.4	<.001	45.3 ± 20.4	44.0 ± 22.3	=.39
PCS (0–52)	4.3 ± 4.8	15.5 ± 10.5	<.001	13.6 ± 9.8	8.2 ± 9.9	<.001
IEQ (0–48)	3.4 ± 5.5	14.1 ± 8.2	<.001	12.5 ± 9.1	7.0 ± 7.8	<.001
BSI (t-score)	32.5 ± 5.3	36.6 ± 9.4	<.001	36.4 ± 8.4	33.5 ± 7.7	=.004
CSI (0–100)	22.1 ± 9.7	38.1 ± 12.5	<.001	34.5 ± 12.6	28.5 ± 14.4	=.004

HC: Healthy Controls; CLBP: Chronic Low Back Pain; F: Female; M: Male; VAS: Visual Analogue Scale. BMI: Body mass index. PDI: Pain Disability Index. WAS: Work Ability Score. Rand36-PF: Rand 36-Physical Functioning subscale. PCS: Pain Catastrophizing Scale. IEQ: Injustice Experience Questionnaire. BSI: Brief Symptom Inventory, CSI: Central Sensitization Inventory.

After conducting 4 clustering approaches, their performance was compared and summarized in Table 2. With respect to internal indicators, hierarchical clustering, K-Means, and SOM demonstrated similar optimal values for Silhouette, Calinski-Harabasz, and Davies-Bouldin. In terms of external indicators, DBSCAN clustered all the HC subjects (n=63) in the same cluster, but incorrectly classified 6 CLBP subjects within this cluster. Hierarchical clustering yielded a more balanced outcome, clustering 62 HC subjects in the same cluster while misclassifying 3 CLBP subjects in the same cluster. Thus, hierarchical clustering may be considered the most suitable trade-off approach. Therefore, the results obtained from hierarchical clustering will be further analyzed to determine the optimal cut-off values for the CSI.

Fig. 2 graphically demonstrates the clustering results of hierarchical clustering. On the y-axis, it represents the distances between distinct subjects and clusters. Meanwhile, the various colour blocks along the x-axis signify individuals from different groups, with red representing HC and blue representing CLBP. This figure distinctly demonstrates the clear separation between HC and CLBP. Within the CLBP cluster, two primary subgroups are distinguishable, represented by the colours grey and green. Consequently, this dendrogram implies the existence of three main clusters.

The CSI values for the three clusters identified by hierarchical clustering are depicted in Fig. 3 using a box plot. In this figure, red dots correspond to subjects from the HC group, while blue dots denote subjects from CLBP group. Females are represented by dots, while males are indicated by stars. Additionally, green triangles are employed to indicate the mean CSI values

of each box (cluster), and orange lines are used to indicate the median values. The boxes encapsulate the CSI values ranging from the first quartile to the third quartile, while the whiskers extend to show the minimum and maximum CSI values for each cluster. The box plot figures for other clustering approaches can be found in Appendix A. Fig. 1, 2, and 3.

Table 2. Clustering performance evaluation

Indicators	Approaches	Hierarchical Clustering	K-Means	DBSCAN	SOM
Internal	Silhouette	0.47	0.48	0.34	0.47
	Calinski-Harabasz	145.66	154.44	62.46	153.44
	Davies-Bouldin	0.91	0.89	3.89	0.90
External	True HC/Predicted HC	62/65	60/65	63/69	60/63

DBSCAN: Density-based Spatial Clustering of Applications with Noise; SOM: Self-Organizing Map; HC: healthy controls.

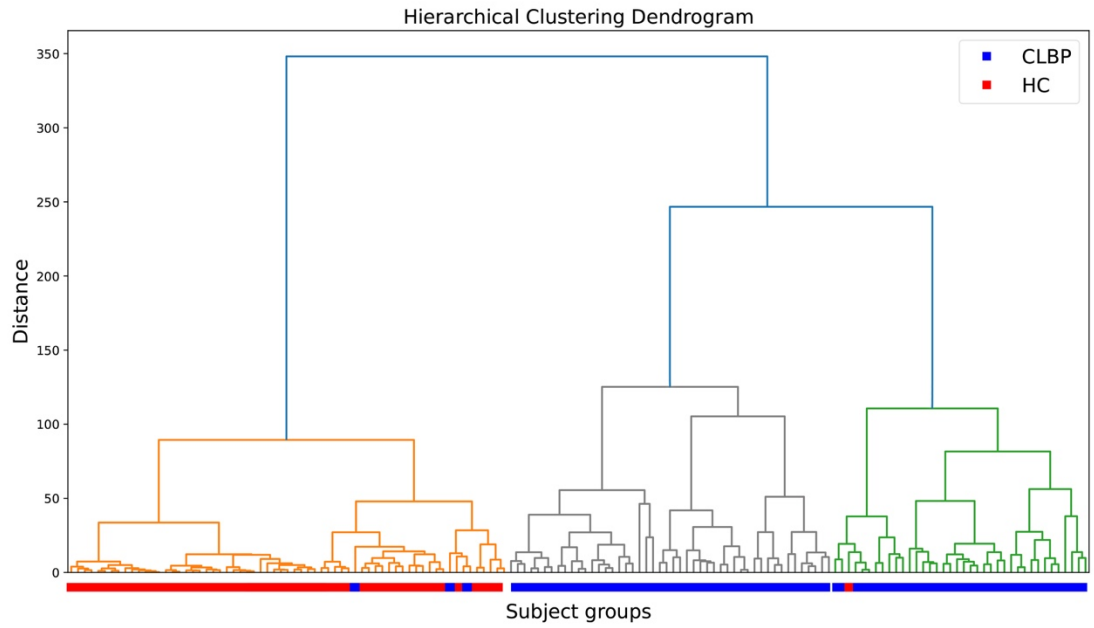


Figure 2. Dendrogram result of hierarchical clustering: red and blue blocks mean subjects from HC and CLBP groups, trees coloured in orange, grey, and green indicate the presence of three primary clusters.

Fig. 3 shows that cluster A predominantly comprises most of the red dots (HC, N=62 out of 63) and a small number of blue dots (CLBP, N=3), indicating that cluster A may represent the healthy group. To understand the meaning of clusters B and C, the demographic characteristics of the hierarchical clustering results are displayed in Table 3. In comparison to cluster B, cluster C exhibited significantly higher levels of pain (VAS) and disability (PDI), lower work ability (WAS) and physical functioning (Rand36-PF), higher pain catastrophizing (PCS), injustice (IEQ), distress (BSI), and CSI values. Consequently, cluster C is characterized as

patients with high HACS level, while cluster B represents patients with low HACS level. Hence, cluster A, B, and C represent the healthy, patients with low HACS levels, and patients with high CS levels groups, respectively.

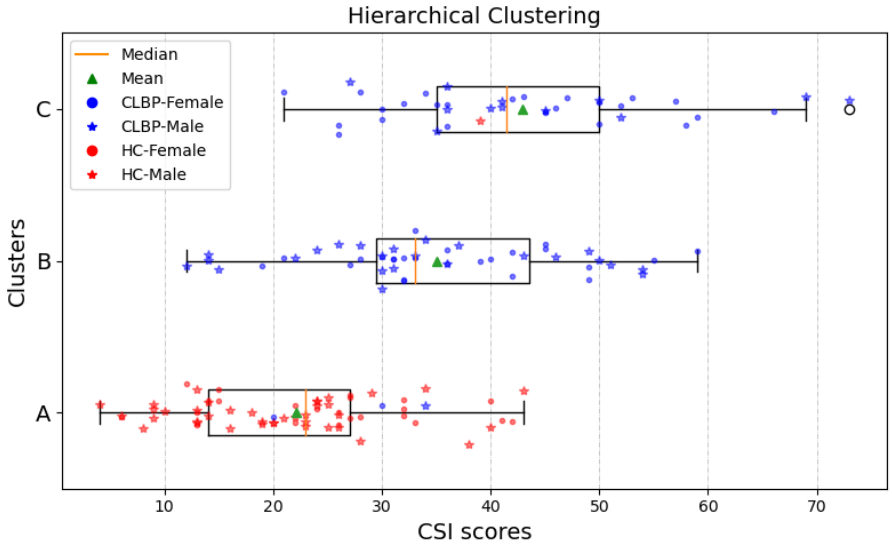


Figure 3. CSI clustering results of Hierarchical clustering. HC: Healthy Controls; CLBP: Chronic Low Back Pain; CSI: Central Sensitization Inventory.

Table 3. Demography of different clustering groups

	Cluster A	Cluster B	p-value of A & B	Cluster C	p-value of A & C	p-value of B & C
Gender	25F/40M	24F/24M	-	25F/13M	-	-
Age, years	41.2±13.6	39.4±11.8	=.436	43.2±12.3	=.77	=.155
Height, cm	160.7±55.8	175.5±8.9	=.232	175.4±11.6	=.278	=.807
Weight, kg	84.5±17.9	82.3±13.1	=.9	90.5±18.6	=.038	=.028
BMI, kg/m^2	26.3±4.8	26.8±4.1	=.216	29.5±6.0	=.002	=.023
VAS (0–10)	0.5±0.9	3.3±1.9	<.001	5.4±2.4	<.001	<.001
PDI (0–70)	5.3±7.2	22.9±13.2	<.001	39.3±10.1	<.001	<.001
WAS (0–10)	8.4±1.4	6.0±2.0	<.001	3.4±2.0	<.001	<.001
Rand36-PF (0–100)	28.7±2.6	71.1±14.4	<.001	38.4±13.7	<.001	<.001
PCS (0–52)	4.5±4.8	11.0±8.1	<.001	21.5±10.6	<.001	<.001
IEQ (0–48)	3.5±5.4	11.1±6.4	<.001	18.3±8.5	<.001	<.001
BSI (t-score)	32.4±5.2	33.9±9.6	=.038	40.5±8.0	<.001	=.001
CSI (0–100)	22.1±9.5	35.0±11.5	<.001	42.9±12.2	<.001	=.007

F: Female; M: Male; VAS: Visual Analogue Scale. BMI: Body mass index. PDI: Pain Disability Index. WAS: Work Ability Score. Rand36-PF: Rand 36-Physical Functioning subscale. PCS: Pain Catastrophizing Scale. IEQ: Injustice Experience Questionnaire. BSI: Brief Symptom Inventory, CSI: Central Sensitization Inventory.

To establish the cut-off value for CSI to distinguish low and high levels of HACS, cluster A and B were combined to represent the low HACS levels population and cluster C was used to represent a high HACS levels population. The demographic comparison of the overall low HACS and high HACS groups is presented in Table 4. Furthermore, to establish the cut-off

values for females and males respectively, the females and males in the low HACS and high HACS groups were extracted for analysis, and the corresponding demographics are provided in Appendix B. Tables 1 and 2.

Table 4. Demography of low and high HACS samples

	Low HACS	High HACS	p-value
Gender	49F/64M	25F/13M	-
Age, years	40.5±12.9	43.2±12.3	=.357
Height, cm	167.0±43.3	175.4±11.6	=.422
Weight, kg	83.6±16.1	90.5±18.6	=.017
BMI, kg/m^2	26.5±4.5	29.5±6.0	=.002
VAS (0–10)	1.7±2.0	5.4±2.4	<.001
PDI (0–70)	12.8±13.4	39.3±10.1	<.001
WAS (0–10)	7.4±2.0	3.4±2.0	<.001
Rand36-PF (0–100)	46.7±23.1	38.4±13.7	=.301
PCS (0–52)	7.3±7.2	21.5±10.6	<.001
IEQ (0–48)	6.8±6.9	18.3±8.5	<.001
BSI (t-score)	33.0±7.4	40.5±8.0	<.001
CSI (0–100)	27.6±12.2	42.9±12.2	<.001

HACS: Human assumed central sensitization; F: Female; M: Male; VAS: Visual Analogue Scale. BMI: Body mass index. PDI: Pain Disability Index. WAS: Work Ability Score. Rand36-PF: Rand 36-Physical Functioning subscale. PCS: Pain Catastrophizing Scale. IEQ: Injustice Experience Questionnaire. BSI: Brief Symptom Inventory, CSI: Central Sensitization Inventory.

Based on the low and high levels HACS groups of overall, females, and males, ROC analysis was performed respectively, and the cut-off values for CSI, along with corresponding AUC, Youden Index, sensitivity, specificity, predictive values and likelihood ratio are presented in Table 5. In the table, the darker red colour represents better performance. By taking all the metrics into consideration, especially AUC and YI, the optimal cut-off value for the overall. group is 35. Although cut-off values 34 and 35 for overall yielded the same AUC (=0.76) and YI (=0.52), the cut-off value of 35 showed a more balanced sensitivity and specificity. Therefore, the optimal cut-off value for the overall group is 35, with a AUC of 0.76, Youden Index of 0.52, sensitivity of 0.76, specificity of 0.76, PPV of 0.52, NPV of 0.91, PLR of 3.19, and NLR of 0.31. For females, the cut-off value remains 34 (AUC=0.71, Youden index=0.41, sensitivity=0.72, specificity=0.69, PPV=0.55, NPV=0.83, PLR=2.35, and NLR=0.4), while for males, the cut-off is 35 (AUC=0.87, Youden index=0.74, sensitivity=0.92, specificity=0.81, PPV=0.5, NPV=0.98, PLR=4.92, and NLR=0.09).

Table 5. CSI cut-off values

	Overall								Females								Males							
CF	AUC	YI	Sen.	Spe.	PPV	NPV	PLR	NLR	AUC	YI	Sen.	Spe.	PPV	NPV	PLR	NLR	AUC	YI	Sen.	Spe.	PPV	NPV	PLR	NLR
20	0.63	0.27	1	0.27	0.31	1	1.36	0	0.58	0.16	1	0.16	0.38	1	1.2	0	0.67	0.34	1	0.34	0.24	1	1.52	0
21	0.65	0.29	1	0.29	0.32	1	1.41	0	0.59	0.18	1	0.18	0.38	1	1.22	0	0.69	0.38	1	0.38	0.25	1	1.6	0
22	0.64	0.28	0.97	0.31	0.32	0.97	1.41	0.08	0.58	0.16	0.96	0.2	0.38	0.91	1.21	0.2	0.7	0.39	1	0.39	0.25	1	1.64	0
23	0.66	0.32	0.97	0.35	0.33	0.98	1.49	0.08	0.6	0.2	0.96	0.24	0.39	0.92	1.27	0.16	0.71	0.42	1	0.42	0.26	1	1.73	0
24	0.67	0.35	0.97	0.37	0.34	0.98	1.55	0.07	0.6	0.2	0.96	0.24	0.39	0.92	1.27	0.16	0.73	0.47	1	0.47	0.28	1	1.88	0
25	0.69	0.39	0.97	0.42	0.36	0.98	1.67	0.06	0.61	0.23	0.96	0.27	0.4	0.93	1.31	0.15	0.77	0.53	1	0.53	0.3	1	2.13	0
26	0.71	0.42	0.97	0.44	0.37	0.98	1.75	0.06	0.61	0.23	0.96	0.27	0.4	0.93	1.31	0.15	0.79	0.58	1	0.58	0.32	1	2.37	0
27	0.7	0.41	0.92	0.49	0.38	0.95	1.79	0.16	0.59	0.19	0.88	0.31	0.39	0.83	1.27	0.39	0.81	0.62	1	0.62	0.35	1	2.67	0
28	0.71	0.42	0.89	0.52	0.39	0.94	1.87	0.2	0.63	0.27	0.88	0.39	0.42	0.86	1.44	0.31	0.77	0.55	0.92	0.62	0.33	0.98	2.46	0.12
29	0.71	0.43	0.87	0.56	0.4	0.93	1.96	0.24	0.63	0.27	0.84	0.43	0.43	0.84	1.47	0.37	0.79	0.58	0.92	0.66	0.35	0.98	2.69	0.12
30	0.72	0.43	0.87	0.57	0.4	0.93	2	0.23	0.63	0.27	0.84	0.43	0.43	0.84	1.47	0.37	0.8	0.59	0.92	0.67	0.36	0.98	2.81	0.11
31	0.71	0.43	0.82	0.61	0.41	0.91	2.1	0.3	0.61	0.23	0.76	0.47	0.42	0.79	1.43	0.51	0.82	0.64	0.92	0.72	0.4	0.98	3.28	0.11
32	0.73	0.46	0.82	0.65	0.44	0.91	2.3	0.29	0.64	0.27	0.76	0.51	0.44	0.81	1.55	0.47	0.84	0.67	0.92	0.75	0.43	0.98	3.69	0.1
33	0.74	0.49	0.79	0.7	0.47	0.91	2.62	0.3	0.68	0.35	0.72	0.63	0.5	0.82	1.96	0.44	0.84	0.67	0.92	0.75	0.43	0.98	3.69	0.1
34	0.76	0.52	0.79	0.73	0.5	0.91	2.97	0.29	0.71	0.41	0.72	0.69	0.55	0.83	2.35	0.4	0.84	0.69	0.92	0.77	0.44	0.98	3.94	0.1
35	0.76	0.52	0.76	0.76	0.52	0.91	3.19	0.31	0.69	0.37	0.68	0.69	0.53	0.81	2.22	0.46	0.87	0.74	0.92	0.81	0.5	0.98	4.92	0.09
36	0.74	0.47	0.71	0.76	0.5	0.89	2.97	0.38	0.67	0.33	0.64	0.69	0.52	0.79	2.09	0.52	0.83	0.66	0.85	0.81	0.48	0.96	4.51	0.19
37	0.7	0.39	0.61	0.79	0.49	0.86	2.85	0.5	0.65	0.29	0.56	0.73	0.52	0.77	2.11	0.6	0.76	0.52	0.69	0.83	0.45	0.93	4.03	0.37
38	0.7	0.4	0.61	0.8	0.5	0.86	2.97	0.5	0.65	0.29	0.56	0.73	0.52	0.77	2.11	0.6	0.77	0.54	0.69	0.84	0.47	0.93	4.43	0.36
39	0.71	0.41	0.61	0.81	0.51	0.86	3.11	0.49	0.65	0.29	0.56	0.73	0.52	0.77	2.11	0.6	0.78	0.55	0.69	0.86	0.5	0.93	4.92	0.36
40	0.7	0.39	0.58	0.81	0.51	0.85	3.12	0.52	0.66	0.32	0.56	0.76	0.54	0.77	2.29	0.58	0.74	0.47	0.62	0.86	0.47	0.92	4.38	0.45
41	0.7	0.39	0.55	0.84	0.54	0.85	3.47	0.53	0.68	0.36	0.56	0.8	0.58	0.78	2.74	0.55	0.71	0.41	0.54	0.88	0.47	0.9	4.31	0.53
42	0.67	0.35	0.5	0.85	0.53	0.83	3.32	0.59	0.69	0.38	0.56	0.82	0.61	0.78	3.05	0.54	0.63	0.26	0.38	0.88	0.38	0.88	3.08	0.7
43	0.67	0.35	0.47	0.88	0.56	0.83	3.82	0.6	0.7	0.4	0.52	0.88	0.68	0.78	4.25	0.55	0.63	0.26	0.38	0.88	0.38	0.88	3.08	0.7
44	0.67	0.34	0.45	0.89	0.59	0.83	4.21	0.62	0.68	0.36	0.48	0.88	0.67	0.77	3.92	0.59	0.65	0.29	0.38	0.91	0.45	0.88	4.1	0.68
45	0.67	0.34	0.45	0.89	0.59	0.83	4.21	0.62	0.68	0.36	0.48	0.88	0.67	0.77	3.92	0.59	0.65	0.29	0.38	0.91	0.45	0.88	4.1	0.68

CF: cut-off values; AUC: area under the curve; YI: Youden index; Spe: specificity; Sen: sensitivity; PPV: positive predictive values; NPV: negative predictive values; PLR: positive likelihood ratio; NLR: negative likelihood ratio. The optimal cut-off values are underlined. The darker red colour represents better performance.

4. Discussion

The aim of the present study was to explore HACS-related subgroups via unsupervised clustering approaches based on questionnaires data and to establish CSI cut-off values for the Dutch-speaking population with CLBP. The clustering results showed three distinct clusters: a healthy group, patients with low HACS levels, and patients with high HACS levels. These clusters exhibited variations in pain intensity, disability levels, and psychological status. By comparing the low HACS level individuals (including the healthy group), ROC analysis indicated optimal cut-off value of 35 for the total group and for males, and 34 for females.

To evaluate the performance of the clustering algorithms, both external and internal metrics were utilized in the present study. The clustering outcomes demonstrated that, across all methods, the majority of HC subjects as grouped into the same cluster. This may suggest that the proposed clustering approaches were accurate to a certain degree. Internal metrics evaluate the separation and cohesion of clusters; and, the values of these indicators may not support that the result clusters were well-separated. This might be attributed to the inherent nature of HACS, which is not strictly binary in its essence. Rather, HACS likely exists along a continuum, spanning from absent to more pronounced degrees [42]. The demographics of the clustering results reveal that cluster C (high HACS level group) exhibited the most severe pain, greatest disability, and poorest psychological status (e.g., depression, anxiety, distress, and pain catastrophising). In contrast, cluster B (patients with CLBP and low HACS level) occupies the middle of the spectrum, while cluster A (healthy group) is situated at the opposite end.

The cutoff value for CSI in our study for Dutch-speaking patients with CLBP is established at 35, with an AUC of 0.76, Youden Index of 0.52, sensitivity of 0.76, and specificity of 0.76. Three other studies have established CSI cutoff values for Dutch-speaking patients with chronic pain [18, 43, 44]. Initially, the CSI was translated into Dutch, and the Dutch version demonstrated sufficient test-retest reliability (ICC=0.88) internal consistency (Cronbach's alpha=0.91), and appropriate structural validity [44]. This study recommended employing a cutoff value of 40 for identifying patients (not solely CLBP) at risk of exhibiting signs of HACS based on earlier research [12]. The determination of this cutoff value of 40 was based solely on patients with chronic pain and CSS, as well as HCs, while patients without CSS were excluded [12]. It yielded a sensitivity of 0.81 and specificity of 0.75. A recent study established a cut-off value of 30 for Dutch-speaking patients with chronic pain and at least one CSS, compared to HCs, reporting high sensitivity (0.85) and specificity (0.92) [18]. However, because patients with chronic pain and without CSS were excluded, the sensitivity and specificity scores of the cut-off values may not accurately reflect the discriminative ability between patients with or without CSS. A follow-up study [45] employed a cutoff value of 40 to distinguish between patients with chronic pain with or without CSS, showing similar sensitivity (0.83), but a notably decreased

specificity of 0.55. As demonstrated in our study, there are distinct differences in pain intensity, disability levels, and psychological status between HCs and patients with chronic pain and high levels of HACS, whereas patients with chronic pain and low levels of HACS fall in the middle. In our study we conducted a sensitivity analysis by comparing cluster C (comprising patients with high HACS levels) exclusively with the HC group, as elaborated in Appendix C. Remarkably, the optimal cutoff values remained consistent, and, simultaneously, all metrics exhibited substantial improvements (AUC=0.83, Youden Index=0.65, sensitivity=0.76, and specificity=0.89). These results may suggest the need to include patients with low levels of HACS when determining and evaluating the cutoff values. Another research, which established cutoff values for Dutch-speaking patients with chronic pain, identified four clinically relevant categories: low (0–26), mild (27–39), moderate (40–52), and high (53+) [43]. However, this study only utilized the CSI value distribution of patients with chronic pain while excluding HCs. The distribution may change and vary based on the assessed population. Apart from this, this approach may lead to suboptimal cutoff values since it does not allow for discrimination between patients with HACS, as well as HCs. Therefore, in our study, the optimal cut-off values of CSI were determined based on high HACS level and low HACS level (with HC) groups.

Because no gold standard exists for assessing HACS, previous studies have employed various methods to indirectly determine the presence of HACS. Some studies assume that the presence of HACS can be indicated by one or several CSS [12, 15, 18]. The presence of CSS was assessed by a physician based on symptom complaints or thorough physical examination (e.g., tender-point evaluations fibromyalgia) [46] or self-reported questions (such as section B of the CSI which asks participants if they have been diagnosed with CSS) [12, 18]. However, due to the lack of a gold standard, expert judgment may vary across clinicians, potentially introducing bias into the cut-off value [12]. Apart from this, HACS can exist even when a CSS is absent [6]. Some studies use quantitative sensory testing (QST) to evaluate the dynamic modulation of nociceptive signals [47] as an indicator for the presence of HACS based on pressure pain threshold, temporal summation, conditioned pain modulation, and thermal QST [14, 48, 49]. However, QST is time-consuming and requires specialized equipment and trained personnel, and does not take the physical functioning and psychosocial issues into consideration while HACS is also related to these factors [50, 51]. Moreover, there is an absence of established cut-off values for QST for the assessment of HACS [2]. Our study employed data-driven clustering approaches to automatically uncover potential patterns in individuals based on pain, physical functioning, and psychological factors. Through the examination of the interrelationships among these factors, the clustering results indicated the division of CLBP group and HC group into three primary clusters (cluster A, B and C). Their demographic results may support the notion that clusters B and C are associated with HACS, as they exhibit significant differences in pain, physical functioning, and psychological status

compared with healthy group (cluster A). In addition to this, our study, through the association of CSI values and psychological states across the three clusters, may corroborate the finding that CSI is associated with psychological constructs [19]. However, since no biological measures were included in our study, it is not possible to determine if CSI is exclusively associated with psychological constructs.

Literature indicates significant differences in pain perception between genders [52-54]. Females generally exhibit a higher prevalence of clinical pain disorders and lower pain thresholds compared to males [52-54] . Accordingly, it was expected that females would demonstrate higher cut-off values for CSI, indicating higher sensitivity to nociceptive stimuli. Several studies have provided evidence in support of this hypothesis [45, 55, 56]. In our study, females reported higher levels of pain, disability, CSI values, and worse psychological status. However, this could be confounded by the higher number of females in CLBP group. Contrary to previous research [55, 56], our study did not find higher cut-off value for females compared to males (34 vs. 35) , and the sum of our sensitivity and specificity for the females group was 1.41, below 1.5 [41]. Given the unequal distribution of male and female participants in both CLBP and HC groups, the distributions of their CSI values are also uneven. Hence, the gender-related cut-off values derived from this study should be interpreted cautiously.

5. Strengths and Limitations

The utilization of clustering approaches offers a flexible and adaptable methodology that can accommodate diverse data types, thereby providing valuable insights into the complex and multidimensional characteristics of HACS. As knowledge of HACS continues to expand, the methodology employed in this study can be applied to identify patterns associated with HACS, incorporating increasingly precise factors that accurately capture the essence of HACS. Furthermore, these clustering approaches can be implemented within the electronic health record system. As the system expands, the scalability of the clustering approaches allows for learning from each case, leading to the generation of increasingly robust cut-off values. This contributes to the advancement of “Data Driven Health Care”.

There were several limitations to the current study. Firstly, the data for our study were obtained from a larger study with different objectives. As a result, some critical information, such as part B of the CSI and objective measurements like QST, was absent. These missing details could provide insights into changes in pain thresholds and the presence of widespread pain, which could suggest alterations in central pain processing mechanisms [57] which could possibly be indicative of HACS. However, once we acquired these additional pieces of information, the flexibility and adaptability of the proposed approach enabled us to redo the analysis easily, thereby identifying patterns related to HACS and determining more accurate CSI cutoff values. Secondly, in our study, levels of HACS were clustered using unsupervised

machine learning based on questionnaire results and gender. While cluster C exhibited patterns associated with high levels of HACS, characterized by increased pain, poorer physical and psychological conditions, and higher CSI values, it is important to note that individuals in cluster C should not be directly classified as patients with HACS, and individuals in cluster B should not be assumed to be without HACS. Instead, the clustering results provide insights into the severity of HACS levels among participants [32]. It is worth acknowledging that utilizing CSI as an input for identifying optimal clustering results and determining the best cut-off values for CSI introduces the potential risk of circular reasoning. However, the results of the feature importance analysis using PCA (further details provided in Appendix D. Fig. 1) indicate that the clustering process primarily relied on information such as gender, Rand-36, and BSI, with CSI ranking 7 out of 9 variables. Consequently, the associated risk is minimal. Thirdly, our study did not exclude other pain conditions due to the complex nature of chronic pain and the lack of comprehensive documentation. It is plausible that HACS may not be solely attributed to CLBP, and consequently, the CSI cut-off values established by our study may apply to patients with CLBP in combination with other pain conditions. Lastly, the cut-off value for females may not be valid enough since sensitivity plus specificity was lower than 1.5. Apart from this, the distribution in gender was imbalanced between CLBP and HC groups. Consequently, the proposed cut-off values for the genders separately should be interpreted with caution.

6. Conclusion

This study explored the HACS-related clusters based on pain intensity, disability levels, psychological status, and gender. Three distinct clusters were found by the data-driven approaches. Patients with high HACS levels group and healthy group were at 2 ends, while patients with low HACS levels group were at the middle. A cut-off value of 35 on the CSI for the Dutch-speaking population with CLBP was established, aiming to differentiate between low and high levels of HACS.

The methodology employed in this study offers a data-driven means to identify subgroups, establish optimal diagnostic thresholds, and enriches the comprehension of this intricate HACS. Ultimately, this methodology empowers researchers and clinicians to craft more personalized and effective approaches for assessing and managing conditions associated with HACS and other diseases.

Reference

[1] J. D. Loeser, and R.-D. Treede, “The Kyoto protocol of IASP basic pain Terminology☆,” *Pain,* vol. 137, no. 3, pp. 473-477, 2008.

[2] I. Schuttert, H. Timmerman, K. K. Petersen, M. E. McPhee, L. Arendt-Nielsen, M. F. Reneman, and A. P. Wolff, “The Definition, Assessment, and Prevalence of (Human

Assumed) Central Sensitisation in Patients with Chronic Low Back Pain: A Systematic Review," *Journal of Clinical Medicine,* vol. 10, no. 24, pp. 5931, 2021.

[3] C. J. Woolf, "Central sensitization: implications for the diagnosis and treatment of pain," *pain,* vol. 152, no. 3, pp. S2-S15, 2011.

[4] D. Hoy, L. March, P. Brooks, F. Blyth, A. Woolf, C. Bain, G. Williams, E. Smith, T. Vos, and J. Barendregt, "The global burden of low back pain: estimates from the Global Burden of Disease 2010 study," *Annals of the rheumatic diseases,* vol. 73, no. 6, pp. 968-974, 2014.

[5] Y. Cruz-Almeida, and R. B. Fillingim, "Can quantitative sensory testing move us closer to mechanism-based pain management?," *Pain medicine,* vol. 15, no. 1, pp. 61-72, 2014.

[6] J. Nijs, S. Z. George, D. J. Clauw, C. Fernández-de-las-Peñas, E. Kosek, K. Ickmans, J. Fernández-Carnero, A. Polli, E. Kapreli, and E. Huysmans, "Central sensitisation in chronic pain conditions: latest discoveries and their potential for precision medicine," *The Lancet Rheumatology,* vol. 3, no. 5, pp. e383-e392, 2021.

[7] R. J. Gatchel, Y. B. Peng, M. L. Peters, P. N. Fuchs, and D. C. Turk, "The biopsychosocial approach to chronic pain: scientific advances and future directions," *Psychological bulletin,* vol. 133, no. 4, pp. 581, 2007.

[8] S. H. Kim, K. B. Yoon, D. M. Yoon, J. H. Yoo, and K. R. Ahn, "Influence of centrally mediated symptoms on postoperative pain in osteoarthritis patients undergoing total knee arthroplasty: a prospective observational evaluation," *Pain Practice,* vol. 15, no. 6, pp. E46-E53, 2015.

[9] E. E. Bennett, K. M. Walsh, N. R. Thompson, and A. A. Krishnaney, "Central sensitization inventory as a predictor of worse quality of life measures and increased length of stay following spinal fusion," *World neurosurgery,* vol. 104, pp. 594-600, 2017.

[10] T. G. Mayer, R. Neblett, H. Cohen, K. J. Howard, Y. H. Choi, M. J. Williams, Y. Perez, and R. J. Gatchel, "The development and psychometric validation of the central sensitization inventory," *Pain Practice,* vol. 12, no. 4, pp. 276-285, 2012.

[11] T. Scerbo, J. Colasurdo, S. Dunn, J. Unger, J. Nijs, and C. Cook, "Measurement properties of the central sensitization inventory: a systematic review," *Pain Practice,* vol. 18, no. 4, pp. 544-554, 2018.

[12] R. Neblett, H. Cohen, Y. Choi, M. M. Hartzell, M. Williams, T. G. Mayer, and R. J. Gatchel, "The Central Sensitization Inventory (CSI): establishing clinically significant values for identifying central sensitivity syndromes in an outpatient chronic pain sample," *J Pain,* vol. 14, no. 5, pp. 438-45, May, 2013.

[13] A. Mibu, T. Nishigami, K. Tanaka, M. Manfuku, and S. Yono, "Difference in the impact of central sensitization on pain-related symptoms between patients with chronic low back pain and knee osteoarthritis," *Journal of Pain Research,* vol. 12, pp. 1757, 2019.

[14] J. Zafereo, S. Wang-Price, and E. Kandil, “Quantitative sensory testing discriminates central sensitization inventory scores in participants with chronic musculoskeletal pain: an exploratory study,” *Pain Practice,* vol. 21, no. 5, pp. 547-556, 2021.

[15] N. L. Orr, K. J. Wahl, M. Lisonek, A. Joannou, H. Noga, A. Albert, M. A. Bedaiwy, C. Williams, C. Allaire, and P. J. Yong, “Central sensitization inventory in endometriosis,” *Pain,* vol. 163, no. 2, pp. e234-e245, 2022.

[16] R. Neblett, “The central sensitization inventory: a user’s manual,” *Journal of Applied Biobehavioral Research,* vol. 23, no. 2, pp. e12123, 2018.

[17] R. E. Sorge, and S. K. Totsch, “Sex differences in pain,” *Journal of neuroscience research,* vol. 95, no. 6, pp. 1271-1281, 2017.

[18] I. Schuttert, A. P. Wolff, R. H. Schiphorst Preuper, A. G. Malmberg, M. F. Reneman, and H. Timmerman, “Validity of the Central Sensitization Inventory to Address Human Assumed Central Sensitization: Newly Proposed Clinically Relevant Values and Associations,” *Journal of Clinical Medicine,* vol. 12, no. 14, pp. 4849, 2023.

[19] G. R. Adams, W. Gandhi, R. Harrison, C. M. Van Reekum, I. Gilron, and T. V. Salomons, “Do “central sensitization” questionnaires reflect measures of nociceptive sensitization or psychological constructs? Protocol for a systematic review,” *Pain reports,* vol. 6, no. 4, 2021.

[20] J. Kregel, C. Schumacher, M. Dolphens, A. Malfliet, D. Goubert, D. Lenoir, B. Cagnie, M. Meeus, and I. Coppieters, “Convergent validity of the Dutch central sensitization inventory: associations with psychophysical pain measures, quality of life, disability, and pain cognitions in patients with chronic spinal pain,” *Pain practice,* vol. 18, no. 6, pp. 777-787, 2018.

[21] D. Xu, and Y. Tian, “A comprehensive survey of clustering algorithms,” *Annals of Data Science,* vol. 2, pp. 165-193, 2015.

[22] R. Xu, and D. Wunsch, “Survey of clustering algorithms,” *IEEE Transactions on neural networks,* vol. 16, no. 3, pp. 645-678, 2005.

[23] G. Gan, C. Ma, and J. Wu, *Data clustering: theory, algorithms, and applications*: SIAM, 2020.

[24] J. A. Echeita, H. R. S. Preuper, R. Dekker, I. Stuive, H. Timmerman, A. P. Wolff, and M. F. Reneman, “Central Sensitisation and functioning in patients with chronic low back pain: protocol for a cross-sectional and cohort study,” *Bmj Open,* vol. 10, no. 3, Mar, 2020.

[25] M. Nicholas, J. W. Vlaeyen, W. Rief, A. Barke, Q. Aziz, R. Benoliel, M. Cohen, S. Evers, M. A. Giamberardino, and A. Goebel, “The IASP classification of chronic pain for ICD-11: chronic primary pain,” *Pain,* vol. 160, no. 1, pp. 28-37, 2019.

[26] M. P. Jensen, C. Chen, and A. M. Brugger, “Interpretation of visual analog scale ratings and change scores: a reanalysis of two clinical trials of postoperative pain,” *The Journal of pain,* vol. 4, no. 7, pp. 407-414, 2003.

[27] C. A. Pollard, “Preliminary validity study of the pain disability index,” *Perceptual and motor skills,* vol. 59, no. 3, pp. 974-974, 1984.

[28] R. D. Hays, C. D. Sherbourne, and R. M. Mazel, “The rand 36-item health survey 1.0,” *Health economics,* vol. 2, no. 3, pp. 217-227, 1993.

[29] M. El Fassi, V. Bocquet, N. Majery, M. L. Lair, S. Couffignal, and P. Mairiaux, “Work ability assessment in a worker population: comparison and determinants of Work Ability Index and Work Ability score,” *BMC public health,* vol. 13, no. 1, pp. 1-10, 2013.

[30] M. J. Sullivan, S. R. Bishop, and J. Pivik, “The pain catastrophizing scale: development and validation,” *Psychological assessment,* vol. 7, no. 4, pp. 524, 1995.

[31] R. M. Bults, M. Reneman, C. van Wilgen, and H. S. Preuper, “Test-retest reliability and construct validity of the dutch injustice experience questionnaire in patients with chronic pain,” *Psychological Injury and Law,* vol. 13, pp. 316-325, 2020.

[32] R. Neblett, M. M. Hartzell, T. G. Mayer, H. Cohen, and R. J. Gatchel, “Establishing clinically relevant severity levels for the central sensitization inventory,” *Pain Practice,* vol. 17, no. 2, pp. 166-175, 2017.

[33] G. Hamerly, and C. Elkan, “Learning the k in k-means,” *Advances in neural information processing systems,* vol. 16, 2003.

[34] F. Murtagh, and P. Contreras, “Algorithms for hierarchical clustering: an overview,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* vol. 2, no. 1, pp. 86-97, 2012.

[35] T. Kohonen, “The self-organizing map,” *Proceedings of the IEEE,* vol. 78, no. 9, pp. 1464-1480, 1990.

[36] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, “DBSCAN revisited, revisited: why and how you should (still) use DBSCAN,” *ACM Transactions on Database Systems (TODS),* vol. 42, no. 3, pp. 1-21, 2017.

[37] S. Aranganayagi, and K. Thangavel, "Clustering categorical data using silhouette coefficient as a relocating measure." pp. 13-17.

[38] S. Petrovic, "A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters." pp. 53-64.

[39] X. Wang, and Y. Xu, "An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index." p. 052024.

[40] M. H. Zweig, and G. Campbell, “Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine,” *Clinical chemistry,* vol. 39, no. 4, pp. 561-577, 1993.

[41] M. Power, G. Fell, and M. Wright, “Principles for high-quality, high-value testing,” *BMJ Evidence-Based Medicine,* vol. 18, no. 1, pp. 5-10, 2013.

[42] E. Lluch, J. Nijs, C. A. Courtney, T. Rebbeck, V. Wylde, I. Baert, T. H. Wideman, N. Howells, and S. T. Skou, “Clinical descriptors for the recognition of central sensitization

pain in patients with knee osteoarthritis," *Disability and rehabilitation,* vol. 40, no. 23, pp. 2836-2845, 2018.

[43] R. van der Noord, D. Paap, and C. P. van Wilgen, "Convergent validity and clinically relevant categories for the Dutch Central Sensitization Inventory in patients with chronic pain," *Journal of Applied Biobehavioral Research,* vol. 23, no. 2, pp. e12119, 2018.

[44] J. Kregel, P. J. Vuijk, F. Descheemaeker, D. Keizer, R. van der Noord, J. Nijs, B. Cagnie, M. Meeus, and P. van Wilgen, "The Dutch Central Sensitization Inventory (CSI): factor analysis, discriminative power, and test-retest reliability," *The Clinical journal of pain,* vol. 32, no. 7, pp. 624-630, 2016.

[45] R. Neblett, M. M. Hartzell, H. Cohen, T. G. Mayer, M. Williams, Y. Choi, and R. J. Gatchel, "Ability of the central sensitization inventory to identify central sensitivity syndromes in an outpatient chronic pain sample," *The Clinical journal of pain,* vol. 31, no. 4, pp. 323-332, 2015.

[46] K. C. Fleming, and M. M. Volcheck, "Central sensitization syndrome and the initial evaluation of a patient with fibromyalgia: a review," *Rambam Maimonides medical journal,* vol. 6, no. 2, 2015.

[47] M. Mucke, H. Cuhls, L. Radbruch, R. Baron, C. Maier, T. Tolle, R. Treede, and R. Rolke, "Quantitative sensory testing (QST). English version," *Schmerz,* vol. 35, pp. 153-160, 2016.

[48] C. Falling, S. Stebbings, G. D. Baxter, C. A. Siegel, R. B. Gearry, J. Nijs, and R. Mani, "Symptoms of central sensitization in patients with inflammatory bowel diseases: a case-control study examining the role of musculoskeletal pain and psychological factors," *Scandinavian journal of pain,* vol. 21, no. 2, pp. 283-295, 2021.

[49] J. Gervais-Hupé, J. Pollice, J. Sadi, and L. C. Carlesso, "Validity of the central sensitization inventory with measures of sensitization in people with knee osteoarthritis," *Clinical rheumatology,* vol. 37, pp. 3125-3132, 2018.

[50] M. B. Yunus, "Fibromyalgia and overlapping disorders: the unifying concept of central sensitivity syndromes." pp. 339-356.

[51] M. Meeus, and J. Nijs, "Central sensitization: a biopsychosocial explanation for chronic widespread pain in patients with fibromyalgia and chronic fatigue syndrome," *Clinical rheumatology,* vol. 26, no. 4, pp. 465-473, 2007.

[52] E. J. Bartley, and R. B. Fillingim, "Sex differences in pain: a brief review of clinical and experimental findings," *British journal of anaesthesia,* vol. 111, no. 1, pp. 52-58, 2013.

[53] N. Geva, S. Golan, L. Pinchas, and R. Defrin, "Sex effects in the interaction of acute stress and pain perception," *Pain*, pp. 10.1097, 2022.

[54] E. Keogh, "Sex and gender differences in pain: past, present, and future," *Pain,* vol. 163, no. S1, pp. S108-S116, 2022.

[55] C. Roldán-Jiménez, D. Pérez-Cruzado, R. Neblett, R. Gatchel, and A. Cuesta-Vargas, "Central sensitization in chronic musculoskeletal pain disorders in different populations: A cross-sectional study," *Pain Medicine,* vol. 21, no. 11, pp. 2958-2963, 2020.

[56] S. Sharma, J. Jha, A. Pathak, and R. Neblett, "Translation, cross-cultural adaptation, and measurement properties of the Nepali version of the central sensitization inventory (CSI)," *BMC neurology,* vol. 20, no. 1, pp. 1-10, 2020.

[57] S. E. Harte, R. E. Harris, and D. J. Clauw, "The neurobiology of central sensitization," *Journal of Applied Biobehavioral Research,* vol. 23, no. 2, pp. e12137, 2018.

Appendix A.

The CSI clustering results of K-means, self-organizing map, and DBSCAN are shown in Appendix A. Fig. 1, Fig. 2, and Fig. 3.

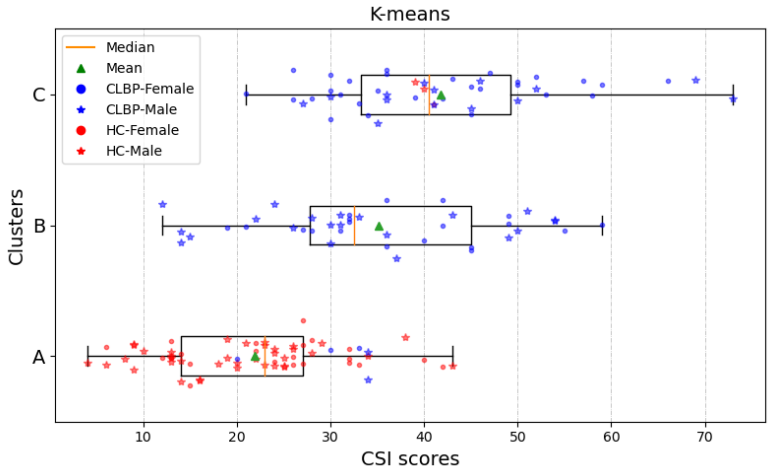


Figure 1. CSI clustering results of K-means. HC: Healthy Controls; CLBP: Chronic Low Back Pain; CSI: Central Sensitization Inventory.

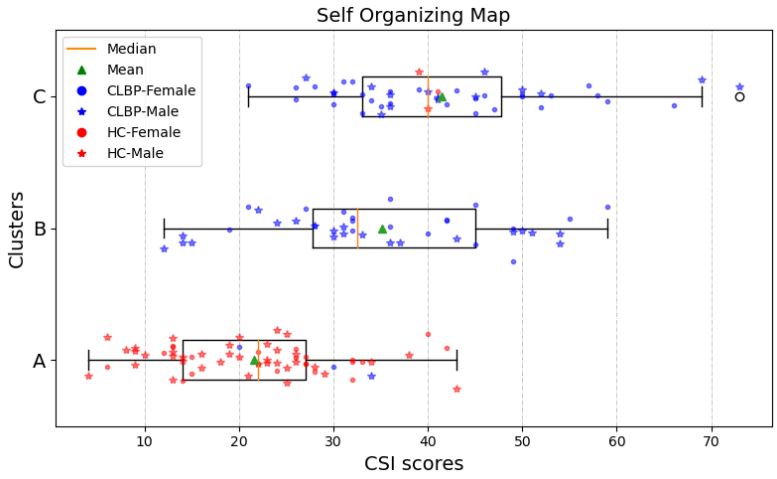


Figure 2. CSI clustering results of self-organizing. HC: Healthy Controls; CLBP: Chronic Low Back Pain; CSI: Central Sensitization Inventory.

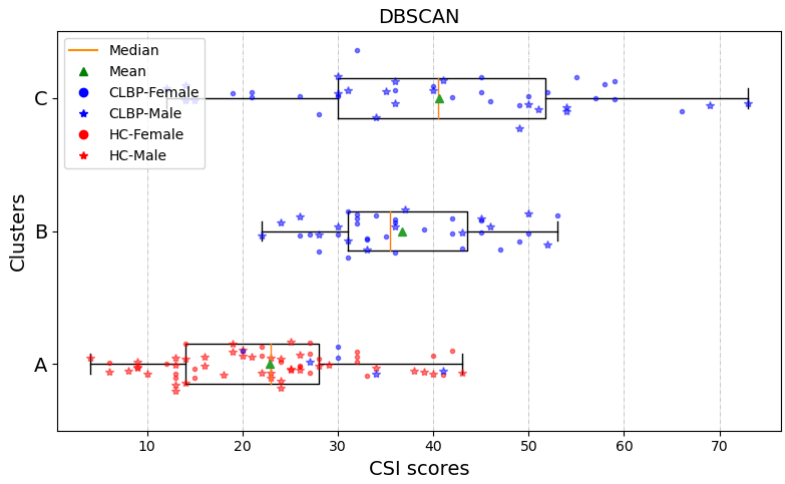


Figure 3. CSI clustering results of DBSCAN. HC: Healthy Controls; CLBP: Chronic Low Back Pain; CSI: Central Sensitization Inventory; DBSCAN: Density-based spatial clustering of applications with noise.

Appendix B.

The Demography of females and males with low and high HACS are shown in Table 1 and Table 2. The cut-off values for females and males were established based on their corresponding low and high HACS groups.

Table 1. Demography of females with low and high HACS

	Low HACS	High HACS	p-value
Gender	49F/0M	25F/0M	-
Age, years	35.5±13.5	41.4±13.3	=.107
Height, cm	153.4±51.4	168.9±6.8	=.909
Weight, kg	76.2±15.0	84.8±17.3	=.008
BMI, kg/m^2	26.2±5.4	29.9±6.9	=.024
VAS (0–10)	1.9±2.0	5.9±2.3	<.001
PDI (0–70)	14.8±12.9	38.4±8.4	<.001
WAS (0–10)	7.3±2.2	3.1±2.0	<.001
Rand36-PF (0–100)	48.5±22.0	39.0±14.8	=.163
PCS (0–52)	8.6±6.3	23.4±7.8	<.001
IEQ (0–48)	8.2±6.8	20.8±6.9	<.001
BSI (t-score)	34.2±7.4	40.7±8.7	=.003
CSI (0–100)	30.7±11.3	41.9±11.8	=.001

HACS: Human assumed central sensitization; F: Female; M: Male; VAS: Visual Analogue Scale. BMI: Body mass index. PDI: Pain Disability Index. WAS: Work Ability Score. Rand36-PF: Rand 36-Physical Functioning subscale. PCS: Pain Catastrophizing Scale. IEQ: Injustice Experience Questionnaire. BSI: Brief Symptom Inventory, CSI: Central Sensitization Inventory.

Table 2. Demography of males with low and high HACS

	Low HACS	High HACS	p-value
Gender	0F/64M	0F/13M	-
Age, years	44.3±11.0	46.8±9.1	=.624
Height, cm	177.3±32.3	188.0±7.7	=.031
Weight, kg	89.2±14.5	101.5±15.8	=.001
BMI, kg/m^2	26.7±3.7	28.7±3.6	=.072
VAS (0–10)	1.6±1.9	4.5±2.1	<.001
PDI (0–70)	11.3±13.5	41.2±12.6	<.001
WAS (0–10)	7.4±1.9	3.9±2.0	<.001
Rand36-PF (0–100)	45.3±23.7	37.4±11.1	=.989
PCS (0–52)	6.2±7.6	17.8±13.7	=.001
IEQ (0–48)	5.6±6.8	13.5±9.1	=.002
BSI (t-score)	32.2±7.3	40.1±6.6	<.001
CSI (0–100)	25.2±12.2	44.9±12.7	<.001

HACS: Human assumed central sensitization; F: Female; M: Male; VAS: Visual Analogue Scale. BMI: Body mass index. PDI: Pain Disability Index. WAS: Work Ability Score. Rand36-PF: Rand 36-Physical Functioning subscale. PCS: Pain Catastrophizing Scale. IEQ: Injustice Experience Questionnaire. BSI: Brief Symptom Inventory, CSI: Central Sensitization Inventory.

Appendix C.

The HC group was used as the standard to compare with cluster C (high HACS level group) to determine the optimal cut-off values for CSI. The results are shown in Table 1.

Table 1. CSI cut-off values

	Overall								Females								Males							
CF	AUC	YI	Sen.	Spe.	PPV	NPV	PLR	NLR	AUC	YI	Sen.	Spe.	PPV	NPV	PLR	NLR	AUC	YI	Sen.	Spe.	PPV	NPV	PLR	NLR
20	0.69	0.38	1	0.38	0.47	1	1.61	0	0.64	0.28	1	0.28	0.58	1	1.39	0	0.72	0.44	1	0.44	0.34	1	1.78	0
21	0.71	0.42	1	0.42	0.49	1	1.74	0	0.66	0.32	1	0.32	0.6	1	1.47	0	0.74	0.49	1	0.49	0.36	1	1.95	0
22	0.71	0.41	0.97	0.44	0.49	0.97	1.74	0.06	0.64	0.28	0.96	0.32	0.59	0.89	1.41	0.13	0.76	0.51	1	0.51	0.38	1	2.05	0
23	0.73	0.46	0.97	0.48	0.51	0.97	1.89	0.06	0.68	0.36	0.96	0.4	0.62	0.91	1.6	0.1	0.77	0.54	1	0.54	0.39	1	2.16	0
24	0.75	0.5	0.97	0.53	0.54	0.97	2.07	0.05	0.68	0.36	0.96	0.4	0.62	0.91	1.6	0.1	0.8	0.61	1	0.61	0.43	1	2.56	0
25	0.78	0.56	0.97	0.59	0.57	0.98	2.38	0.05	0.7	0.4	0.96	0.44	0.63	0.92	1.71	0.09	0.84	0.68	1	0.68	0.48	1	3.15	0
26	0.8	0.61	0.97	0.64	0.6	0.98	2.68	0.04	0.7	0.4	0.96	0.44	0.63	0.92	1.71	0.09	0.88	0.76	1	0.76	0.55	1	4.1	0
27	0.81	0.62	0.92	0.7	0.63	0.94	3.03	0.12	0.7	0.4	0.88	0.52	0.65	0.81	1.83	0.23	0.9	0.8	1	0.8	0.6	1	5.13	0
28	0.82	0.63	0.89	0.74	0.66	0.92	3.46	0.15	0.76	0.52	0.88	0.64	0.71	0.84	2.44	0.19	0.86	0.72	0.92	0.8	0.58	0.97	4.7	0.1
29	0.82	0.64	0.86	0.77	0.68	0.91	3.81	0.17	0.76	0.52	0.84	0.68	0.72	0.81	2.63	0.24	0.87	0.75	0.92	0.83	0.61	0.97	5.37	0.1
30	0.83	0.65	0.86	0.79	0.7	0.91	4.08	0.17	0.76	0.52	0.84	0.68	0.72	0.81	2.63	0.24	0.89	0.77	0.92	0.85	0.65	0.97	6.26	0.1
31	0.81	0.61	0.81	0.8	0.7	0.88	4.12	0.24	0.74	0.48	0.76	0.72	0.73	0.75	2.71	0.33	0.89	0.77	0.92	0.85	0.65	0.97	6.26	0.1
32	0.81	0.61	0.81	0.8	0.7	0.88	4.12	0.24	0.74	0.48	0.76	0.72	0.73	0.75	2.71	0.33	0.89	0.77	0.92	0.85	0.65	0.97	6.26	0.1
33	0.82	0.63	0.78	0.85	0.74	0.88	5.17	0.25	0.78	0.56	0.72	0.84	0.82	0.75	4.5	0.33	0.89	0.77	0.92	0.85	0.65	0.97	6.26	0.1
34	0.82	0.65	0.78	0.86	0.76	0.88	5.75	0.25	0.8	0.6	0.72	0.88	0.86	0.76	6	0.32	0.89	0.77	0.92	0.85	0.65	0.97	6.26	0.1
35	0.83	0.65	0.76	0.89	0.8	0.87	7.14	0.27	0.78	0.56	0.68	0.88	0.85	0.73	5.67	0.36	0.91	0.82	0.92	0.9	0.73	0.97	9.4	0.09
36	0.8	0.6	0.7	0.89	0.79	0.84	6.63	0.33	0.76	0.52	0.64	0.88	0.84	0.71	5.33	0.41	0.87	0.74	0.83	0.9	0.71	0.95	8.54	0.18
37	0.74	0.49	0.59	0.89	0.76	0.8	5.61	0.45	0.72	0.44	0.56	0.88	0.82	0.67	4.67	0.5	0.78	0.57	0.67	0.9	0.67	0.9	6.83	0.37
38	0.74	0.49	0.59	0.89	0.76	0.8	5.61	0.45	0.72	0.44	0.56	0.88	0.82	0.67	4.67	0.5	0.78	0.57	0.67	0.9	0.67	0.9	6.83	0.37
39	0.75	0.5	0.59	0.91	0.79	0.8	6.54	0.45	0.72	0.44	0.56	0.88	0.82	0.67	4.67	0.5	0.8	0.59	0.67	0.93	0.73	0.9	9.11	0.36
40	0.76	0.52	0.59	0.92	0.81	0.8	7.85	0.44	0.72	0.44	0.56	0.88	0.82	0.67	4.67	0.5	0.81	0.62	0.67	0.95	0.8	0.91	13.7	0.35
41	0.76	0.52	0.57	0.95	0.88	0.8	12.5	0.45	0.74	0.48	0.56	0.92	0.88	0.68	7	0.48	0.78	0.56	0.58	0.98	0.88	0.89	23.9	0.43
42	0.74	0.48	0.51	0.97	0.9	0.78	16.9	0.5	0.76	0.52	0.56	0.96	0.93	0.69	14	0.46	0.7	0.39	0.42	0.98	0.83	0.85	17.1	0.6
43	0.74	0.47	0.49	0.98	0.95	0.77	32.1	0.52	0.76	0.52	0.52	1	1	0.68	-	0.48	0.7	0.39	0.42	0.98	0.83	0.85	17.1	0.6
44	0.73	0.46	0.46	1	1	0.77	-	0.54	0.74	0.48	0.48	1	1	0.66	-	0.52	0.71	0.42	0.42	1	1	0.85	-	0.58
45	0.73	0.46	0.46	1	1	0.77	-	0.54	0.74	0.48	0.48	1	1	0.66	-	0.52	0.71	0.42	0.42	1	1	0.85	-	0.58

CF: cut-off values; AUC: area under the curve; YI: Youden index; Spe: specificity; Sen: sensitivity; PPV: positive predictive values; NPV: negative predictive values; PLR: positive likelihood ratio; NLR: negative likelihood ratio. The optimal cut-off values are underlined. The darker red colour represents better performance

Appendix D.

To identify the features that contributed to the clustering process and reveal the important factors for CS, feature importance was determined using PCA. The results are shown in figure 1.

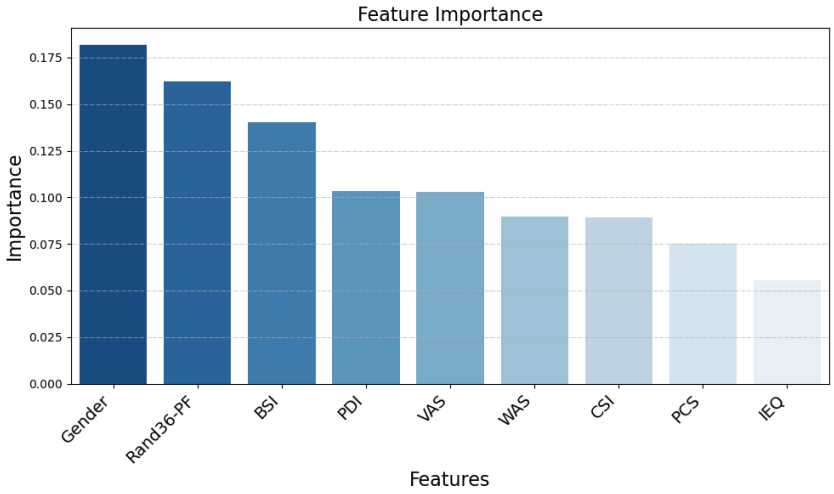


Figure 1. Feature importance

Chapter 7

General Discussion

Main Findings

The structure of this thesis and the findings of each chapter are briefly summarized in Fig. 1. The analysis begins with focusing on the gait patterns of healthy older adults. **Chapter 2** showed that deep learning (DL) outperformed conventional machine learning (CML) in classifying adults and older adults. Building on these results, **Chapter 3** utilized Explainable Artificial Intelligence (XAI) to investigate what DL models had learned from accelerometer signal data. This chapter found that DL models primarily used data surrounding heel contact for classification, indicating potential differences in acceleration and deceleration patterns between adults and older adults during walking. These gait analysis insights and the XAI methodologies could further aid in analyzing the movement of patients with chronic low back pain (CLBP). In **Chapter 4**, it was aimed to use CML and XAI to examine how human assumed central sensitization (HACS) was associated with the gait of patients with CLBP. The findings showed different gait patterns in patients with low or high levels of HACS (CLBP- and CLBP+), possibly associated with different motor control strategies characterized as “loose” or “tight”. Going beyond gait analysis, **Chapter 5** used unsupervised CML (with ante-hoc explainability) to explore the relationships between physical activity intensity (PAI), CLBP, and HACS. The findings indicate distinct PAI patterns, suggesting that patients in the CLBP+ group may exhibit an endurance pain response pattern. Given that severe levels of HACS in Chapters 4 and 5 were determined by the central sensitization inventory (CSI) questionnaire with a predetermined cut-off value of 40, **Chapter 6** used unsupervised CML (with ante-hoc explainability) to define a more accurate cut-off value for the population with CLBP. This analysis found two distinct subgroups within CLBP, possibly correlated with low and high levels of HACS, leading to the establishment of a new cutoff value of 35.

Gait Performance in Healthy Aging

The rapidly increasing aging population raises concerns about maintaining the quality of life in older adults, especially their mobility [1]. Aging is a continuous process that often involves the loss of muscle mass, reduced bone density, and declining nerve function [2]. These age-related alterations can lead to changes in gait patterns, such as lower walking speed, shorter stride/step length, increased gait variability, and increased gait instability compared to young controls [3, 4]. Given the limited availability of medical resources, automatic gait analysis plays a crucial role in monitoring the mobility of older adults [5] and contributed to a better understanding of aging.

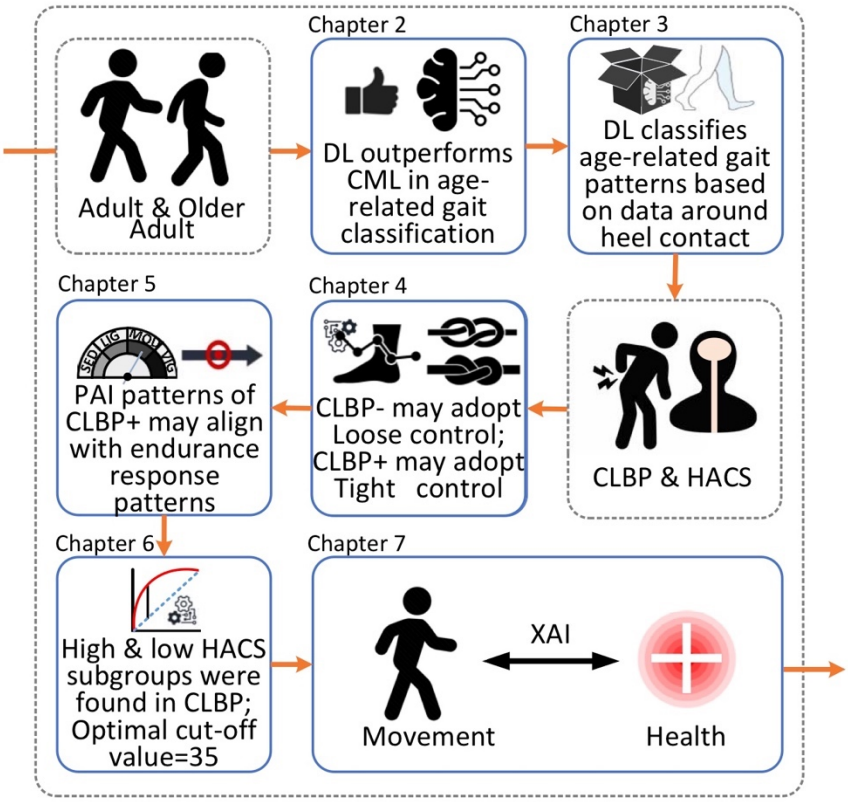


Figure 1. The summaries of findings this thesis. DL: deep learning; CML: conventional machine learning; CLBP: chronic low back pain; HACS: human assumed central sensitization; CLBP+: patients with CLBP and high levels of HACS; CLBP-: patients with CLBP and low levels of HACS; PAI: physical activity intensity; XAI: Explainable artificial intelligence.

CML has been widely used to classify age-related gait patterns based on gait outcomes [6, 7]. However, these gait outcomes are handcrafted, requiring extensive labor for design and selection [8]. Furthermore, the offline computation of gait outcomes may limit their use in real-time applications [9], such as monitoring gait of older adults in daily living environments. To overcome these challenges, DL has been introduced to gait analysis, utilizing accelerometer signal data as input, rather than gait outcomes (as discussed in **Chapter 2**). The results indicated that DL outperformed CML, achieving a high classification performance with an area under the curve (AUC) exceeding 0.94, compared to the highest AUC of 0.83 for the best CML models. These results not only highlighted the potential of DL in classifying age-related gait patterns but also suggested that DL had learned specific features that can efficiently capture the age-related changes in gait. However, the black-box nature of DL obscures the understanding of what the models have learned from the data.

To improve the transparency of DL, XAI was utilized in **Chapter 3**. The results showed that the accelerometer signal data spanning from the terminal swing to the loading response phase contributed most to the classification process. This implies that there may be notable differences in acceleration and deceleration patterns between adults and older adults during

walking. The study also found that these distinct patterns within a single stride were sufficient for a convolutional neural network (CNN) model to accurately differentiate between adult and older adult groups, achieving an AUC of 0.89. Additionally, employing a recurrent neural network (RNN) model allowed for an examination of gait evolution by analyzing acceleration and deceleration patterns across strides, revealing subtle differences and relationships that might indicate postural control ability [10]. RNN achieved a high level of classification accuracy, with an AUC of 0.94. These distinct patterns could be associated with age-related changes, such as declining muscular capabilities, but our study lacked adequate data to confirm these links. To gain a more comprehensive understanding of these patterns, incorporating kinematics and kinetics [11], as well as electromyography (EMG) data [12], would be beneficial. For example, ground reaction forces (GRF) can provide insights into the forces involved in body weight absorption and push-off during walking, range of motion could track alterations at specific joints, and EMG can reveal muscle activation patterns.

Declining balance is associated with aging [13] and it is often observed that older adults exhibit less stable gait patterns [14]. **Chapter 3,** however, showed that DL mostly employed accelerometer signals in the anteroposterior (AP) and vertical (V) directions to distinguish older adults from adults, rather than relying on mediolateral (ML) direction data. To further explore gait stability differences between adults and older adults, additional steps could be taken. First, using long-term data (such as 24-hour recordings) from daily living environments may better reflect actual gait performance during daily living [15]. The walking task in our study, a simple 3-minute walk, lacked complexity and did not include perturbations encountered in daily life, such as navigating around obstacles or moving through doorways, which impose higher demands and often require multitasking. Second, this study used a single accelerometer to record walking data, which may not fully capture all aspects of balance-related postural control. Therefore, using more sensors, like additional accelerometers or gyroscopes, could yield a better understanding of the balance during gait of older adults.

Chapters 2 and 3 highlighted the superior performance of DL for classifying age-related gait patterns, and the use of XAI to enhance understanding of gait changes in aging. The findings indicated that distinctions in walking patterns between adults and older adults can be characterized by acceleration and deceleration patterns within a single stride or across multiple strides. This knowledge enhances the comprehension of how gait changes as individuals age.

Movement Characteristics of Patients with Chronic Low Back Pain

The methodologies of AI-driven gait analysis and XAI in Chapters 2 and 3, may shed light not only on exploring the relationship between gait and aging but also on movement and other health-related conditions, such as back pain.

Low back pain (LBP) is highly prevalent, with up to 84% of individuals experiencing LBP at least once in their lifetime [16]. In approximately 90% of patients, the cause of LBP is nonspecific [17], and around 20% of patients continue to report persistent back pain one year after the onset of acute LBP [18]. When LBP persists beyond three months, it is classified as chronic low back pain (CLBP). CLBP poses significant socioeconomic burdens and causes great individual suffering. Physical exercise is often recommended to manage CLBP but the effect size is modest [19].

CLBP is a heterogeneous condition [20]. Gait analysis reports inconsistent evidence when comparing patients with CLBP to healthy controls, including walking speed, step width, and stride variability [21-25]. Similarly, studies on PAI in patients with CLBP versus healthy controls have yielded mixed findings, with some indicating reduced daily PAI in patients with CLBP [26, 27] and others showing no significant difference [28, 29]. This inconsistent evidence highlights the heterogeneity and complexity of CLBP.

The presence of central sensitization (CS) in CLBP may be one of the key factors contributing to this heterogeneity, as CS is associated with long-lasting chronic pain [30]. CS is defined as an increased responsiveness of nociceptive neurons in the central nervous system to their normal or subthreshold afferent input [31]. However, due to the current inability to directly measure CS mechanisms in individual humans, the term human assumed central sensitization (HACS) is used [32]. Since movement may be changed due to pain, it was hypothesized that different levels of HACS might be associated with gait patterns (Chapter 4) and PAI patterns (Chapter 5) in patients with CLBP.

Chapter 4 used CML to classify gait patterns in patients with CLBP and with low or high levels of HACS (CLBP- and CLBP+, respectively). A satisfactory performance (an accuracy rate of 84.4%) was achieved, indicating distinct gait patterns between CLBP- and CLBP+ groups. XAI was employed to explain the classification process, revealing that CLBP- exhibited higher gait smoothness and stability, whereas CLBP+ showed a more regular, less variable, and more predictable gait pattern. These findings suggested different motor control strategies: “loose control” in CLBP- characterized by reduced trunk muscle excitability leading to reduced control over trunk movements [33], and “tight control” in CLBP+ marked by increased activation and co-contraction of trunk muscles for enhanced movement control [33]. These results could inform the categorization of CLBP patients into different treatment groups, but further research is needed for more direct validation of these motor control strategies. For example, the increased activation and co-contraction of trunk muscles might be observable through EMG [34]. Moreover, gait perturbations can be used to examine the presence of tight control. Tight control might result in increased trunk stiffness to counterbalance anticipated

perturbations and hence will show larger trunk displacement due to the unanticipated perturbations [35].

In **Chapter 5**, unsupervised CML with ante-hoc explainability was employed to explore and clarify PAI patterns in CLBP- and CLBP+ groups. The findings showed that patients in the CLBP- group tended to break tasks into smaller bouts of activity, taking frequent and short rests in between, while patients in the CLBP+ group engaged in prolonged periods of activity with extended rest intervals. These patterns might be interpreted by the avoidance-endurance model (AEM) [36]. AEM postulates that a subgroup of patients shows a pattern of fear-avoidance responses caused by high fear of pain, leading to avoidance behaviour; another subgroup of patients shows a pattern of endurance responses with overuse and overload of physical structures, despite having pain. The endurance responses seemed to correspond with CLBP+ group, while the behaviour of CLBP- group might not necessarily align with fear-avoidance responses, as there was no evidence to support the presence of fear beliefs. These findings also need to be examined by future studies, since this chapter did not directly assess the fear or endurance belief. To assess avoidance and endurance beliefs more directly, the avoidance-endurance questionnaire [37] could be utilized. Alternatively, fear beliefs could be evaluated through responses to perturbations during gait, where individuals fearful of falling might demonstrate anticipatory postural adjustments prior to the perturbation [38, 39]

Chapters 4 and 5 provide insights into the movement characteristics of CLBP- and CLBP+ groups within their daily living environments. However, the relatively small sample size in these studies (n=42) is a limitation. Apart from this, further longitudinal research is required to explore potential causal links between movement changes and HACS. It is possible that HACS is a consequence of the observed movement patterns. Tight control presumably leads to increased muscle activation and co-contraction, resulting in higher spinal loading. This continuous muscle co-contraction, even at rest [40], can produce long-lasting peripheral noxious stimuli, potentially contributing to the development and/or persistence of HACS [41]. Additionally, patients with an endurance response pattern may subject their muscles to ongoing stress and repetitive strain during prolonged activities, leading to laxity and inflammation [42]. This persistent nociceptive input could also play a role in HACS development [36]. Conversely, HACS might also be a cause of the observed movement differences. Patients in CLBP+ group show higher levels of HACS. Considering HACS's mechanism, the relationship between movement and pain may become irrelevant, as pain can occur without tissue loading [43]. Thus, these patients might manage their trunk by increasing spinal loading, allowing them to continue activities for longer periods. Therefore, HACS could be either a result or a cause of changes in movement, or perhaps both.

In Chapters 4 and 5, HACS levels of CBLP- and CLBP+ were determined using a cut-off value of 40 from the CSI questionnaire [44] which serves as an indirect method for assessing HACS. It has been reported that the cut-off value of 40 for the CSI may vary depending on different types of musculoskeletal pain [45, 46], as well as across various cultural and national contexts [47]. Currently, the gold standard for diagnosing HACS is unavailable and hence, **Chapter 6** of this thesis discussed the exploration of HACS-related subgroups using unsupervised CML with ante-hoc explainability based on clinical outcomes (questionnaire data) to establish a more accurate cut-off value for the CSI. The findings found two distinct HACS-related groups and suggested adopting 35 as the new cut-off value for the Dutch-speaking population with CLBP.

The methodologies of AI, particularly XAI, as employed in chapters 4 to 6, provide a data-driven approach to enhance our understanding of complex issues related to HACS, CLBP, and movement characteristics. Ultimately, these methodologies enable researchers and clinicians to develop more personalized and effective strategies for assessing and managing conditions associated with HACS and CLBP.

Selection of AI Models for HMS

This thesis has demonstrated that movement analysis can benefit from the application of AI. The integration of AI into movement analysis begins with the selection of appropriate AI models.

Fig. 3 graphically illustrates the hierarchy of commonly used terms related to AI. AI is a broad concept, with subsets of machine learning, XAI, and others. Within machine learning, a spectrum of methodologies includes CML, DL, and others. Depending on the learning paradigm, AI models can be categorized into supervised learning models [48], unsupervised learning models [49], and others such as semi-supervised learning models. For XAI, it contains two main categories: ante-hoc explainability and post-hoc explainability [50]. Some supervised and unsupervised CML models are also categorized under ante-hoc explainability, e.g., K-nearest neighbors and K-means [51].

a. Supervised Learning or Unsupervised Learning

Supervised learning is tasked with recognizing patterns from labeled datasets. For instance, **Chapter 2** employed five supervised DL models to learn the gait patterns of adults and older adults. These models were trained using accelerometer data as input features, along with their respective labels ("adults" or "older adults"). Once trained, these models are capable of predicting labels for new accelerometer data inputs. The underlying assumption is that if there are distinct patterns that correlate with the labels, then the models should classify them with a high degree of accuracy. For example, if older adults and adults have different gait patterns, DL models are expected to classify them accurately, and vice versa.

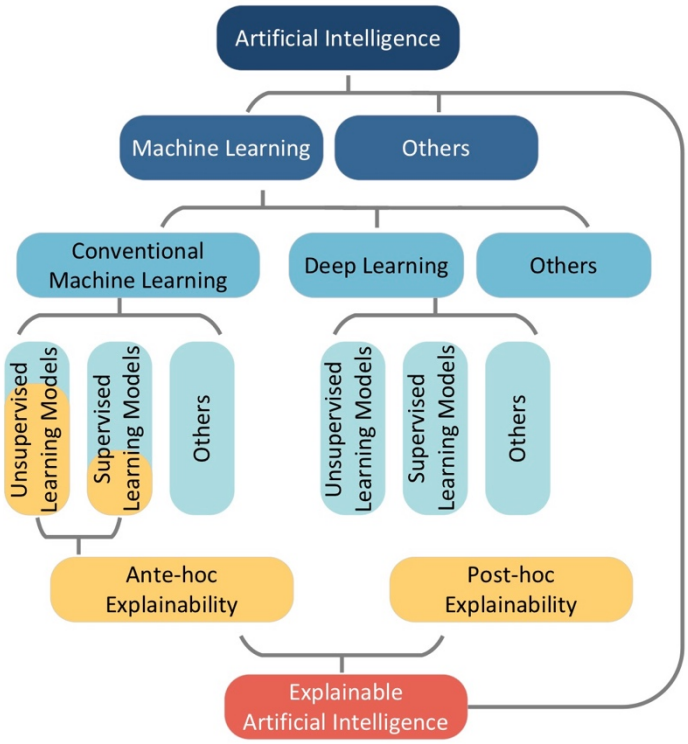


Figure 3. Hierarchy of artificial intelligence related terms.

Unsupervised learning, on the other hand, learns patterns within unlabeled data and automatically clusters them. In **Chapter 6,** four unsupervised CML models were employed to explore potential HACS-related subgroups. Clinical outcomes data from questionnaires were used as input, and the models determined the clusters automatically. The rationale for unsupervised learning is that within a dataset, if a subset of data samples is in close proximity to each other and distinct from others, they form a cluster.

In summary, the selection of a supervised or unsupervised model depends on the objectives of the analysis. If the goal is to recognize/predict specific gait patterns (**Chapters 2**, and **3**) or examine distinct gait patterns between groups (**Chapter 4**), supervised learning models are appropriate. Conversely, to uncover latent patterns (**Chapter 5**) or potential subgroups within the targeted population (**Chapter 6**), unsupervised learning models are preferable.

b. CML or DL

Once the decision is made between supervised or unsupervised learning, the next step is to select the specific models within CML or DL.

CML and DL each present different advantages and disadvantages. CML can be performing well even for small datasets. It is a critical consideration in medical fields where acquiring extensive gait data from patients can be challenging [52]. DL, on the other hand, has been

shown to surpass CML in classifying gait patterns [53, 54], although this enhanced performance comes at the cost of increased model complexity, which can impede the interpretability of the models [51]. Consequently, **Chapter 2** utilized supervised DL models for accurate classification of age-related gait patterns, whereas **Chapter 4** employed supervised CML models to examine the relationships between CLBP, HACS, and gait outcomes, prioritizing explainability. Beyond objective constraints, such as dataset size, computing power, and etc., choosing between DL and CML often reflects a trade-off between explainability and performance [55]. Hence, for a better explainability, only unsupervised CML with ante-hoc explainability were used in this thesis (**Chapters 5**, and **6**).

c. Selection of Specific Models

This thesis evaluates the performance of various supervised CML models in **Chapters 2** and **4,** including support vector machines (SVM), random forests (RF), artificial neural networks, k-nearest neighbors, and naive bayes. These CML models have been extensively compared in other research, and RF and SVM are frequently utilized classifiers in gait analysis [6, 56]. Despite well-tuned RF and SVM potentially achieving similar levels of performance, as demonstrated in **Chapter 2**, RF is recommended over SVM for gait analysis, as suggested in **Chapter 4**. The performance of SVM is highly influenced by the choice of kernel function, which should align with the data's underlying nature (e.g., linear or non-linear) [57]. As the nature of the data is often not know, determining the optimal kernel function for SVM requires specialized knowledge and additional analysis. Conversely, RF is more user-friendly.

Regarding the supervised DL models, five supervised DL models within three categories were discussed in **Chapter 2**, including CNN, RNN (gate recurrent unit, long short-term memory, and bi-directional long short-term memory), and hybrid neural networks (HNN; convolutional long short-term memory). These supervised DL models are able to use time-series signal data as input. The time-dependent information within time-series data is critical for certain gait analyses and should be considered when selecting models. For instance, CNN excels at extracting local spatial-temporal features [58] and may capture independent gait outcomes (e.g., step length). Consequently, CNN could be effective at recognizing gait patterns observable within one or two strides, like asymmetric gait patterns [59]. RNN is specifically designed to learn both short- and long-term features [58] and may capture temporal-dependent gait outcomes (e.g., gait regularity). Therefore, RNN might benefit from analyzing extended time-series data to detect gradual changes in gait patterns, such as fatigue-related gait patterns [60]. HNN contains the structures of CNN and RNN. It is expected to benefit from both structures. However, **Chapter 2** did not support this idea. The unique characteristics of these models may lead to different intention of applications and may have different data processing requirements. For instance, CNN might achieve satisfactory performance with only one or two strides of data, while RNN and HNN might require longer data, potentially

exceeding 10 seconds (8 strides), as the minimum for optimal performance (as noted in **Chapter 2**).

Limitations of XAI in Human Movement Sciences

This thesis showcases the practical implementation of XAI to enhance the transparency and explainability of human movement analysis, thereby making it more trustworthy and facilitating the extraction of scientific knowledge from the data. However, several challenges remain.

A fundamental challenge in XAI is the absence of a clear metric for evaluating the "interpretability" or "explainability" [61]. Due to the lack of a ground truth, the classic metrics (such as accuracy, sensitivity, and etc.) are not available. Hence, it is difficult to compare the performance among XAI approaches. In **Chapter 4**, the XAI approach, SHapley Additive exPlanations (SHAP) [62], was employed to explain the CML model (RF), instead of using the conventional Gini impurity [63]. This choice was made based on the authors' experience, as SHAP has a stronger theoretical foundation [64]. Although we suggested that SHAP will offer superior explanations, it may be challenging to provide definitive proof of this. Additionally, due to the lack of metrics, it becomes a complex task to assess the trustworthiness of the explanations. The explanations provided by XAI may not align with domain expertise. These discrepancies may provide fresh insights when the explanations are trustworthy. However, these discrepancies could also be raised by biases or unknown errors. Due to the lack of metrics, caution is advised when using explanations provided by XAI methods. Recent efforts have been made in evaluating the effectiveness of XAI [65, 66]. Measurement techniques, such as the goodness checklist and the explanation satisfaction scale, seem to be a good step in the direction of evaluating XAI.

The explanatory capabilities of XAI are still limited. In **Chapter 3**, XAI was used to explain the classification process for age-related gait patterns based on accelerometer time-series. XAI provided explanations by highlighting the importance of specific segments of accelerometer data. However, the authors need to explain these XAI-generated explanations by manually linking these segments with the specific gait events and guessing why these gait events are important when classifying age-related gait patterns. The current explanations may not provide a complete picture of what happened during these gait events. XAI should not leave the majority explanation generation to users, as their diverse backgrounds and knowledge may lead to different interpretations and explanations [67]. Additionally, XAI generates explanations based on correlations and associations within the data, which may not be sufficient to unveil cause-effect relationships [68]. In **Chapter 4**, XAI identified the top 10 critical gait outcomes that have the most influence on the classification process. However, it

remains challenging to determine whether changes in these gait outcomes are the causes or the consequences of HACS.

XAI is still in its early stages, especially in the context of human movement analysis. To effectively apply XAI in this field, collaborations with interdisciplinary fields like human-computer interaction and data sciences are essential. However, these collaborations can be challenging to establish, potentially hindering the application of XAI in human movement analysis. Therefore, the development of automated XAI solutions becomes crucial [69]. These services should be designed to offer substantial assistance to a broad spectrum of end-users, including non-technical experts. In line with this vision, steps have been taken to contribute. In **Chapters 2**, **3**, and **6** of this thesis, the authors have made the code openly accessible to fellow researchers, thereby facilitating the reuse and expansion of this work, and fostering collaborative efforts in the application of XAI to human movement analysis.

Future of XAI in Human Movement Sciences

In the domain of human movement analysis, numerous AI-driven systems, such as AI-driven gait analysis, have exhibited impressive levels of accuracy and performance. Looking forward, XAI is poised to clarify the opaque “black-box” nature of AI, laying the groundwork for the wider implementation of AI-driven solutions, not just in gait analysis but throughout various healthcare domains.

The widespread use of wearable smart devices (such as smart watches) has generated massive amounts of individual movement data and provided the computational capacity necessary for deploying AI-driven gait analysis. Trustworthy AI-driven gait analysis will facilitate personalized health monitoring, early disease prediction, and the assessment of treatment effectiveness, among other benefits. Leveraging both historical clinical data and the burgeoning influx of newly collected data, AI-generated insights have the potential to rapidly advance our understanding of health and movement. This, in turn, assists clinicians and computer scientists in further refining AI-driven gait analysis. These innovations are expected to not only improve gait analysis but also transform a broad spectrum of healthcare practices, making the processes of prognosis, diagnosis, treatment, patient follow-up, and clinical decision-making more streamlined, precise, and efficient.

AI in human movement analysis is becoming integral to clinical settings and everyday life. Rather than supplanting clinical experts, AI will support them in treating diseases and monitoring the health in daily live environments more effectively. It will not disrupt patient-clinician connections but instead help patients better understand their individual health. AI is not an isolated knowledge generator; it is a catalyst for deepening the understanding of

movement scientists in health issues related to movement. Ultimately, AI contributes to a deeper comprehension of movement and paves the way for more effective treatments.

Conclusion

This thesis further highlighted the importance of AI in movement analysis and demonstrated the potential of XAI in enhancing our understanding of gait patterns in healthy older adults and movement characteristics in patients with back pain. It provided guidance on selecting appropriate AI models for movement (especially gait) analysis through a comparison of various AI models. Based on the insights from XAI, it revealed that differences in gait patterns between adults and older adults can be characterized by acceleration and deceleration patterns. In the context of CLBP, the findings indicated different motor control strategies and pain response patterns among patients. These findings highlighted the possibility of personalized treatment approaches in managing CLBP and HACS. The methodologies used in this thesis, including CML, DL, and XAI, bring us closer to understanding the complex interplay between movement, and aging or back pain. Looking ahead, further steps are necessary to developing more transparent and reliable AI tools in healthcare.

Reference

[1] U. Bechtold, N. Stauder, and M. Fieder, “Let’s walk it: Mobility and the perceived quality of life in older adults,” *International journal of environmental research and public health,* vol. 18, no. 21, pp. 11515, 2021.

[2] M. Intriago, G. Maldonado, R. Guerrero, O. Messina, and C. Rios, “Bone mass loss and Sarcopenia in Ecuadorian patients,” *Journal of Aging Research,* vol. 2020, 2020.

[3] A. Aboutorabi, M. Arazpour, M. Bahramizadeh, S. W. Hutchins, and R. Fadayevatan, “The effect of aging on gait parameters in able-bodied older subjects: a literature review,” *Aging clinical and experimental research,* vol. 28, no. 3, pp. 393-405, 2016.

[4] P. Terrier, and F. Reynard, “Effect of age on the variability and stability of gait: a cross-sectional treadmill study in healthy individuals between 20 and 69 years of age,” *Gait & posture,* vol. 41, no. 1, pp. 170-174, 2015.

[5] A. Hanley, C. Silke, and J. Murphy, “Community-based health efforts for the prevention of falls in the elderly,” *Clinical interventions in aging*, pp. 19-25, 2011.

[6] Y. Zhou, R. Romijnders, C. Hansen, J. van Campen, W. Maetzler, T. Hortobagyi, and C. J. C. Lamoth, “The detection of age groups by dynamic gait outcomes using machine learning approaches,” *Scientific Reports,* vol. 10, no. 1, Mar, 2020.

[7] C. Prakash, R. Kumar, and N. Mittal, “Recent developments in human gait research: parameters, approaches, applications, machine learning techniques, datasets and challenges,” *Artificial Intelligence Review,* vol. 49, pp. 1-40, 2018.

[8] F. Gu, M.-H. Chung, M. Chignell, S. Valaee, B. Zhou, and X. Liu, “A survey on deep learning for human activity recognition,” *ACM Computing Surveys (CSUR),* vol. 54, no. 8, pp. 1-34, 2021.

[9] Y. Hutabarat, D. Owaki, and M. Hayashibe, “Recent advances in quantitative gait analysis using wearable sensors: a review,” *IEEE Sensors Journal*, 2021.

[10] D. Borah, S. Wadhwa, U. Singh, S. L. Yadav, M. Bhattacharjee, and V. Sindhu, “Age related changes in postural stability,” *Indian J Physiol Pharmacol,* vol. 51, no. 4, pp. 395-404, 2007.

[11] K. A. Boyer, R. T. Johnson, J. J. Banks, C. Jewell, and J. F. Hafer, “Systematic review and meta-analysis of gait mechanics in young and older adults,” *Experimental gerontology,* vol. 95, pp. 63-70, 2017.

[12] A. Schmitz, A. Silder, B. Heiderscheit, J. Mahoney, and D. G. Thelen, “Differences in lower-extremity muscular activation during walking between healthy older and young adults,” *Journal of electromyography and kinesiology,* vol. 19, no. 6, pp. 1085-1091, 2009.

[13] R. W. Bohannon, P. A. Larkin, A. C. Cook, J. Gear, and J. Singer, “Decrease in timed balance test scores with aging,” *Physical therapy,* vol. 64, no. 7, pp. 1067-1070, 1984.

[14] S. M. Bruijn, O. Meijer, P. Beek, and J. H. van Dieen, “Assessing the stability of human locomotion: a review of current measures,” *Journal of the Royal Society Interface,* vol. 10, no. 83, pp. 20120999, 2013.

[15] I. Hillel, E. Gazit, A. Nieuwboer, L. Avanzino, L. Rochester, A. Cereatti, U. D. Croce, M. O. Rikkert, B. R. Bloem, and E. Pelosin, “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, no. 1, pp. 1-12, 2019.

[16] B. F. Walker, “The prevalence of low back pain: a systematic review of the literature from 1966 to 1998,” *Clinical Spine Surgery,* vol. 13, no. 3, pp. 205-217, 2000.

[17] J. Hartvigsen, M. J. Hancock, A. Kongsted, Q. Louw, M. L. Ferreira, S. Genevay, D. Hoy, J. Karppinen, G. Pransky, J. Sieper, R. J. Smeets, M. Underwood, and W. Lancet Low Back Pain Series, “What low back pain is and why we need to pay attention,” *Lancet,* vol. 391, no. 10137, pp. 2356-2367, Jun 9, 2018.

[18] M. Von Korff, and K. Saunders, “The course of back pain in primary care,” *Spine,* vol. 21, no. 24, pp. 2833-2837, 1996.

[19] N. E. Foster, J. R. Anema, D. Cherkin, R. Chou, S. P. Cohen, D. P. Gross, P. H. Ferreira, J. M. Fritz, B. W. Koes, W. Peul, J. A. Turner, C. G. Maher, and W. Lancet Low Back Pain Series, “Prevention and treatment of low back pain: evidence, challenges, and promising directions,” *Lancet,* vol. 391, no. 10137, pp. 2368-2383, Jun 9, 2018.

[20] D. R. Fourney, G. Andersson, P. M. Arnold, J. Dettori, A. Cahana, M. G. Fehlings, D. Norvell, D. Samartzis, and J. R. Chapman, “Chronic low back pain: a heterogeneous

condition with challenges for an evidence-based approach," *Spine,* vol. 36, pp. S1-S9, 2011.

[21] C. J. Lamoth, O. G. Meijer, A. Daffertshofer, P. I. Wuisman, and P. J. Beek, "Effects of chronic low back pain on trunk coordination and back muscle activity during walking: changes in motor control," *European Spine Journal,* vol. 15, pp. 23-40, 2006.

[22] C. J. Lamoth, J. F. Stins, M. Pont, F. Kerckhoff, and P. J. Beek, "Effects of attention on the control of locomotion in individuals with chronic low back pain," *Journal of neuroengineering and rehabilitation,* vol. 5, no. 1, pp. 1-8, 2008.

[23] G. Christe, F. Kade, B. M. Jolles, and J. Favre, "Chronic low back pain patients walk with locally altered spinal kinematics," *Journal of biomechanics,* vol. 60, pp. 211-218, 2017.

[24] S. Ebrahimi, F. Kamali, M. Razeghi, and S. A. Haghpanah, "Comparison of the trunk-pelvis and lower extremities sagittal plane inter-segmental coordination and variability during walking in persons with and without chronic low back pain," *Human movement science,* vol. 52, pp. 55-66, 2017.

[25] S. P. Gombatto, T. Brock, A. DeLork, G. Jones, E. Madden, and C. Rinere, "Lumbar spine kinematics during walking in people with and people without low back pain," *Gait & posture,* vol. 42, no. 4, pp. 539-544, 2015.

[26] M. Soysal, B. Kara, and M. N. Arda, "Assessment of physical activity in patients with chronic low back or neck pain," *Turk Neurosurg,* vol. 23, no. 1, pp. 75-80, 2013.

[27] C. G. Ryan, P. M. Grant, P. M. Dall, H. Gray, M. Newton, and M. H. Granat, "Individuals with chronic low back pain have a lower level, and an altered pattern, of physical activity compared with matched controls: an observational study," *Australian Journal of Physiotherapy,* vol. 55, no. 1, pp. 53-58, 2009.

[28] J. A. Verbunt, K. R. Westerterp, G. J. van der Heijden, H. A. Seelen, J. W. Vlaeyen, and J. A. Knottnerus, "Physical activity in daily life in patients with chronic low back pain," *Arch Phys Med Rehabil,* vol. 82, no. 6, pp. 726-30, Jun, 2001.

[29] M. G. van Weering, M. M. Vollenbroek-Hutten, T. M. Tonis, and H. J. Hermens, "Daily physical activities in chronic lower back pain patients assessed with accelerometry," *Eur J Pain,* vol. 13, no. 6, pp. 649-54, Jul, 2009.

[30] J. A. Echeita, H. R. S. Preuper, R. Dekker, I. Stuive, H. Timmerman, A. P. Wolff, and M. F. Reneman, "Central Sensitisation and functioning in patients with chronic low back pain: protocol for a cross-sectional and cohort study," *Bmj Open,* vol. 10, no. 3, Mar, 2020.

[31] J. D. Loeser, and R.-D. Treede, "The Kyoto protocol of IASP basic pain Terminology☆," *Pain,* vol. 137, no. 3, pp. 473-477, 2008.

[32] I. Schuttert, H. Timmerman, K. K. Petersen, M. E. McPhee, L. Arendt-Nielsen, M. F. Reneman, and A. P. Wolff, "The Definition, Assessment, and Prevalence of (Human Assumed) Central Sensitisation in Patients with Chronic Low Back Pain: A Systematic Review," *Journal of Clinical Medicine,* vol. 10, no. 24, pp. 5931, 2021.

[33] J. H. van Dieen, N. P. Reeves, G. Kawchuk, L. R. van Dillen, and P. W. Hodges, "Motor Control Changes in Low Back Pain: Divergence in Presentations and Mechanisms," *Journal of Orthopaedic & Sports Physical Therapy,* vol. 49, no. 6, pp. 370-379, Jun, 2019.

[34] C. T. Candotti, J. F. Loss, A. M. S. Pressi, F. A. de Souza Castro, M. La Torre, M. de Oliveira Melo, L. D. Araújo, and M. Pasini, "Electromyography for assessment of pain in low back muscles," *Physical therapy,* vol. 88, no. 9, pp. 1061-1067, 2008.

[35] N. W. Mok, S. G. Brauer, and P. W. Hodges, "Failure to use movement in postural strategies leads to increased spinal displacement in low back pain," *Spine,* vol. 32, no. 19, pp. E537-E543, 2007.

[36] M. I. Hasenbring, and J. A. Verbunt, "Fear-avoidance and Endurance-related Responses to Pain: New Models of Behavior and Their Consequences for Clinical Practice," *Clinical Journal of Pain,* vol. 26, no. 9, pp. 747-753, Nov-Dec, 2010.

[37] M. I. Hasenbring, D. Hallner, and A. C. Rusu, "Fear-avoidance-and endurance-related responses to pain: development and validation of the Avoidance-Endurance Questionnaire (AEQ)," *European Journal of Pain,* vol. 13, no. 6, pp. 620-628, 2009.

[38] S. B. Swart, R. den Otter, and C. J. Lamoth, "Anticipatory control of human gait following simulated slip exposure," *Scientific Reports,* vol. 10, no. 1, pp. 9599, 2020.

[39] T. J. Ellmers, A. Maslivec, and W. R. Young, "Fear of falling alters anticipatory postural control during cued gait initiation," *Neuroscience,* vol. 438, pp. 41-49, 2020.

[40] A. Schinkel-Ivy, B. C. Nairn, and J. D. Drake, "Investigation of trunk muscle co-contraction and its association with low back pain development during prolonged sitting," *Journal of Electromyography and Kinesiology,* vol. 23, no. 4, pp. 778-786, 2013.

[41] J. Nijs, S. Z. George, D. J. Clauw, C. Fernández-de-las-Peñas, E. Kosek, K. Ickmans, J. Fernández-Carnero, A. Polli, E. Kapreli, and E. Huysmans, "Central sensitisation in chronic pain conditions: latest discoveries and their potential for precision medicine," *The Lancet Rheumatology,* vol. 3, no. 5, pp. e383-e392, 2021.

[42] M. I. Hasenbring, N. E. Andrews, and G. Ebenbichler, "Overactivity in Chronic Pain, the Role of Pain-related Endurance and Neuromuscular Activity An Interdisciplinary, Narrative Review," *Clinical Journal of Pain,* vol. 36, no. 3, pp. 162-171, Mar, 2020.

[43] J. Nijs, A. Lahousse, E. Kapreli, P. Bilika, İ. Saraçoğlu, A. Malfliet, I. Coppieters, L. De Baets, L. Leysen, and E. Roose, "Nociplastic pain criteria or recognition of central sensitization? Pain phenotyping in the past, present and future," *Journal of clinical medicine,* vol. 10, no. 15, pp. 3203, 2021.

[44] E. E. Bennett, K. M. Walsh, N. R. Thompson, and A. A. Krishnaney, "Central sensitization inventory as a predictor of worse quality of life measures and increased length of stay following spinal fusion," *World neurosurgery,* vol. 104, pp. 594-600, 2017.

[45] R. Neblett, H. Cohen, Y. Choi, M. M. Hartzell, M. Williams, T. G. Mayer, and R. J. Gatchel, "The Central Sensitization Inventory (CSI): establishing clinically significant values for

identifying central sensitivity syndromes in an outpatient chronic pain sample," *J Pain,* vol. 14, no. 5, pp. 438-45, May, 2013.

[46] A. Mibu, T. Nishigami, K. Tanaka, M. Manfuku, and S. Yono, "Difference in the impact of central sensitization on pain-related symptoms between patients with chronic low back pain and knee osteoarthritis," *Journal of Pain Research,* vol. 12, pp. 1757, 2019.

[47] R. Neblett, "The central sensitization inventory: A user's manual," *Journal of Applied Biobehavioral Research,* vol. 23, no. 2, Jun, 2018.

[48] P. Cunningham, M. Cord, and S. J. Delany, "Supervised learning," *Machine learning techniques for multimedia: case studies on organization and retrieval*, pp. 21-49: Springer, 2008.

[49] H. B. Barlow, "Unsupervised learning," *Neural computation,* vol. 1, no. 3, pp. 295-311, 1989.

[50] T. Speith, "A review of taxonomies of explainable artificial intelligence (XAI) methods." pp. 2239-2250.

[51] P. Lisboa, S. Saralajew, A. Vellido, R. Fernández-Domenech, and T. Villmann, "The coming of age of interpretable and explainable machine learning models," *Neurocomputing,* vol. 535, pp. 25-39, 2023.

[52] P. Khera, and N. Kumar, "Role of machine learning in gait analysis: a review," *Journal of Medical Engineering & Technology,* vol. 44, no. 8, pp. 441-467, 2020.

[53] C. Tunca, G. Salur, and C. Ersoy, "Deep learning for fall risk assessment with inertial sensors: Utilizing domain knowledge in spatio-temporal gait parameters," *IEEE journal of biomedical and health informatics,* vol. 24, no. 7, pp. 1994-2005, 2019.

[54] R. Kaur, J. Levy, R. W. Motl, R. Sowers, and M. E. Hernandez, "Deep Learning for Multiple Sclerosis Differentiation Using Multi-Stride Dynamics in Gait," *IEEE Transactions on Biomedical Engineering*, 2023.

[55] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning." pp. 80-89.

[56] E. Balaji, D. Brindha, and R. Balakrishnan, "Supervised machine learning based gait classification system for early detection and stage classification of Parkinson's disease," *Applied Soft Computing,* vol. 94, pp. 106494, 2020.

[57] R. Zhang, and W. Wang, "Facilitating the applications of support vector machine by using a new kernel," *Expert systems with applications,* vol. 38, no. 11, pp. 14225-14230, 2011.

[58] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Deep learning for time series classification: a review," *Data mining and knowledge discovery,* vol. 33, no. 4, pp. 917-963, 2019.

[59] S. Viteckova, P. Kutilek, Z. Svoboda, R. Krupicka, J. Kauler, and Z. Szabo, "Gait symmetry measures: A review of current and prospective methods," *Biomedical Signal Processing and Control,* vol. 42, pp. 89-100, 2018.

[60] F. A. Barbieri, P. C. R. Dos Santos, E. Lirani-Silva, R. Vitório, L. T. B. Gobbi, and J. H. Van Diëen, “Systematic review of the effects of fatigue on spatiotemporal gait parameters,” *Journal of back and musculoskeletal rehabilitation,* vol. 26, no. 2, pp. 125-131, 2013.

[61] L. Arras, A. Osman, and W. Samek, “CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations,” *Information Fusion,* vol. 81, pp. 14-40, 2022.

[62] S. M. Lundberg, and S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in neural information processing systems,* vol. 30, 2017.

[63] C. Kelly, and K. Okada, "Variable interaction measures with random forest classifiers." pp. 154-157.

[64] G. Owen, *Game theory*: Emerald Group Publishing, 2013.

[65] S. Mohseni, N. Zarei, and E. D. Ragan, “A multidisciplinary survey and framework for design and evaluation of explainable AI systems,” *ACM Transactions on Interactive Intelligent Systems (TiiS),* vol. 11, no. 3-4, pp. 1-45, 2021.

[66] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, “Metrics for explainable AI: Challenges and prospects,” *arXiv preprint arXiv:1812.04608*, 2018.

[67] D. Doran, S. Schulz, and T. R. Besold, “What does explainable AI really mean? A new conceptualization of perspectives,” *arXiv preprint arXiv:1710.00794*, 2017.

[68] C. Molnar, G. Casalicchio, and B. Bischl, "Interpretable machine learning–a brief history, state-of-the-art and challenges." pp. 417-431.

[69] C. Molnar, *Interpretable machine learning*: Lulu. com, 2020.

Appendix A

Summary

Walking is one of the most common daily physical activities. Gait serves as a window into health by reflecting the intricate coordination of the musculoskeletal, cardiorespiratory, and nervous systems. The advent of wearable sensor technologies, such as accelerometers, enables the recording of gait performance in daily living environments. Artificial Intelligence (AI), applied in gait analysis, has shown success in monitoring the gait of older adults and analyzing complex conditions, such as back pain. AI-driven gait analysis exhibits accuracy levels that are comparable with trained clinicians. However, these achievements are accompanied by increasing the complexity of AI models which are lack of explainability. The black-box nature makes clinical experts and patients hard to trust the AI-driven gait analysis, and it may fail to meet the law requirements and ethical standards. Thus, explainable AI (XAI) has emerged to address these challenges by enhancing the transparency and interpretability of AI models. It may help to establish trust between AI-driven gait analysis and their users. Additionally, XAI could offer valuable insights into movement expertise for human movement scientists and clinical experts by explaining the knowledge extracted from data by these black-box AI models. Therefore, the aim of this thesis was to enhance the understanding of movement, especially gait, in healthy older adults and patients with back pain by leveraging insights from XAI (**Chapter 1**).

Conventional machine learning (CML) is commonly used for classifying various gait patterns based on gait outcomes. However, these gait outcomes are handcrafted, and the process of their design and selection are labor-intensive. Additionally, the computation of gait outcomes is usually offline, implying that such approaches may not be suitable for real-time applications such as monitoring the gait of older adults in daily living environments. Hence, the study described in **Chapter 2** aimed to eliminate the need for handcrafted gait outcomes and improve the performance of age-related gait pattern classification by employing deep learning (DL) based on accelerometer data collected from 3-minute walking. This chapter compared the performance of five DL models (using accelerometer signal data as input) with four CML models (using handcrafted gait outcomes as input). The results showed that all DL models surpassed CML models, achieving an area under the curve (AUC) greater than 0.94, compared to the highest AUC of 0.83 achieved by the best CML model. These findings not only highlighted the superiority of DL, but also suggested that DL has learned valuable gait outcomes reflecting age-related changes in gait, which had been overlooked by CML. In addition, this chapter presented an investigation into the effects of different window sizes on classification performance, as varying window sizes result in different numbers of consecutive gait cycles being used as DL input. It was observed that convolutional neural network (CNN) was capable of using single stride data to differentiate between adults and older adults, while recurrent neural network (RNN) might depend on the distinctions and connections among various gait cycles for classification.

Building on the insights from Chapter 2, the study described in **Chapter 3** aimed to investigate what DL models had learned from data in classifying adults and older adults, utilizing XAI. The findings indicated that accelerometer signal data spanning from the terminal swing to the loading response phase, especially data around heel contact, contributed most to the classification process. These findings suggested that DL captured different acceleration and deceleration patterns to differentiate the gait of older adults from adults. Additionally, the study revealed that variations in acceleration and deceleration within a single stride were adequate for CNN to classify (achieving an AUC of 0.89). RNN classified based on subtle differences and relationships in acceleration and deceleration patterns across multiple strides, attaining an AUC of 0.94. These results implied that RNN considered the postural control ability of older adults. Notably, XAI revealed that DL primarily utilized data in the vertical and anterior-posterior directions for classification, rather than data in the mediolateral direction, which is more closely related to balance.

The insights and methodologies from these two chapters could shed light on the gait analysis of other conditions, such as back pain. Gait analysis in patients with chronic low back pain (CLBP) reported conflicting evidence, supporting the idea that CLBP is a heterogeneous condition. The presence of human assumed central sensitization (HACS) in CLBP might contribute to this heterogeneity. In **Chapter 4**, it was hypothesized that different levels of HACS (low or high) could be related to the gait patterns of patients with CLBP, and these differences could be effectively classified using CML. Additionally, XAI could be used to interpret these differences in gait patterns. By analyzing accelerometer data collected from daily living environments for about one week, the results of this chapter confirmed that patients with CLBP and with low or high levels of HACS (CLBP- and CLBP+, respectively), could be effectively classified by CML (e.g., Random Forest), achieving an accuracy of 84.4%. XAI revealed that patients in CLBP- group exhibited a higher smoothness and better stability in gait, whereas patients in CLBP+ group showed a more regular, less variable, and predictable gait pattern. These findings suggested that CLBP- and CLBP+ patients might adopt different motor control strategies, namely “loose control” and “tight control”. The loose control strategy, characterized by reduced trunk movement control, could explain the gait patterns in CLBP-, while the tight control strategy, with enhanced trunk movement control, could explain those in CLBP+. These findings emphasize the need for personalized treatment approaches.

Building on the findings from Chapter 4, the study described in **Chapter 5** delved into the physical activity intensity (PAI) patterns in patients with CLBP, where inconsistent evidence was also observed, based on the accelerometer data collected from daily living environments over a period of approximately one week. This study employed unsupervised CML with XAI

(Hidden semi-Markov Model, HSMM) to explore PAI patterns in CLBP- and CLBP+ groups. While traditional methods using preset cut-points failed to detect statistical differences in overall PAI between CLBP- and CLBP+ groups, HSMM can learn the PAI patterns from data. HSMM identified distinct PAI patterns in these two groups. CLBP- group tended to segment tasks into smaller bouts, interspersed with frequent and short rests. CLBP+ group exhibited prolonged periods of activity and rest. PAI patterns in CLBP+ group could be explained by the endurance response pattern, characterized by overuse and overload of physical structures, despite experiencing pain. PAI patterns in CLBP- might not align with the fear-avoidance response pattern, as there was no evidence indicating fear belief in this group. The insights from Chapters 4 and 5 contributed to a better understanding of CLBP, movement, and HACS, paving the way for personalized treatment strategies in the future.

In Chapters 4 and 5, the low and high levels of HACS were determined using the central sensitization inventory (CSI) questionnaire, based on a predefined cut-off value of 40. However, it has been reported that the cut-off value for CSI may vary according to different pain conditions, as well as different cultural and national contexts. It is notable that there is no universally accepted gold standard for assessing HACS. Based on the data patterns and structure within gender and clinical outcomes (questionnaires reflecting pain, physical status, and psychological status), the objective of **Chapter 6** was to utilize unsupervised CML with XAI to investigate subgroups related to HACS among patients with CLBP. The results identified two distinct subgroups within the CLBP population. One subgroup, characterized by higher pain, greatest disability, worse psychological status, and higher CSI values, was assigned to higher levels of HACS, while the other represented lower levels of HACS. Based on these subgroups, a new cut-off value of 35 was established for the Dutch-speaking population with CLBP. The methodology used in this chapter provided new understanding in identifying HACS-related patterns and contributes to setting more accurate cut-off values.

Chapter 7 provided a summary of the main findings in this thesis and discussed potential future directions. Using the thesis as a case study, it provided guidelines for selecting appropriate AI models for movement analysis, elaborating on their advantages and disadvantages. By underscoring the potential of XAI in human movement analysis, this chapter highlighted the need for ongoing improvements to make XAI more reliable, understandable, and user-friendly. Despite acknowledged limitations, this chapter presented the view that the potential of AI, particularly XAI, in movement analysis, holds promise for AI-driven human movement analysis in the future. Furthermore, this advancement will aid clinical experts and movement scientists in deciphering knowledge extracted from data by AI models.

Samenvatting

Lopen is een van de meest voorkomende dagelijkse fysieke activiteiten. Het loopgedrag geeft inzicht in de gezondheid doordat het de coördinatie van het musculoskeletale-, het cardiorespiratoire systeem en het zenuwstelsel weerspiegelt. Met de komst van draagbare sensortechnologieën, zoals accelerometers, is het mogelijk geworden om het lopen in de dagelijkse leefomgevingen te registreren. Met gebruik van data analysemethoden en artificiële intelligentie (AI) kan uit de accelerometer signalen informatie worden gehaald, waarmee inzicht wordt gekregen over veranderingen in het lopen als gevolg van het ouder worden of door lage rugklachten. AI-gestuurde loopanalyse heeft een nauwkeurigheidsniveaus die vergelijkbaar is met die van getrainde clinici. De prestaties van de voorspelling en nauwkeurigheid gaat echter gepaard met een toenemende complexiteit van de AI-modellen waarbij het lastig is te beschrijven op grond waarvan het model tot zijn uitkomst komt. Deze zogenoemde “black-box” maakt de toepassing er van in de kliniek lastig. Klinische experts en patiënten zullen moeite hebben om de AI-gestuurde loopanalyse te vertrouwen omndat niet altijd duidelijk is hoe de uiteindelijke resultaten tot stand zijn gekomen. Daarnaast is wetgeving omtrent het gebruik van AI in de zorg nog in ontwikkeling. Explainable AI (XAI), een recente ontwikkeling beoogt deze uitdagingen aan te pakken en de transparantie en interpreteerbaarheid van AI-modellen te verbeteren. Het kan bijdragen aan het creëren van meer inzicht in de achtergrond van de uitkomst van de modellen en bijdragen aan het vertrouwen tussen AI-gestuurde loopanalyses en de gebruikers. Daarnaast kan XAI waardevolle inzichten en kennis genereren over het begrijpen van veranderend bewegen als gevolg van leeftijd en aandoeningen. Daarom was het doel van dit proefschrift om het begrip van beweging, met name het loopgedrag, bij gezonde oudere volwassenen en patiënten met rugpijn te verbeteren, door gebruik te maken van inzichten uit XAI (**hoofdstuk 1**).

Conventionele Machine Learning (CML) wordt vaak gebruikt voor het classificeren van verschillende looppatronen op basis van een grote verscheidenheid aan loopparameters, zoals spatio-temporele parameters (stap lengte, stap tijd, snelheid) en dynamische parameters (stabiliteit, vloeiendheid, frequentie) van het lopen. Deze loopparameters worden afzonderlijk berekend vanuit accelerometer signalen, en dit is doorgaans een arbeidsintensief proces met veel voorbewerkingen. Bovendien wordt de berekening van de loopparameters meestal offline gedaan, waardoor real-time toepassingen zoals het monitoren van het looppatronen van oudere volwassenen in dagelijkse leefomgevingen, met directe terugkoppeling niet mogelijk is. **Hoofdstuk 2** heeft als doel om te onderzoeken of met Deep Learning (DL) modellen, op basis van accelerometer signalen van drie minuten lopen, de noodzaak van het vooraf berekenen van aparte loopvariabelen voor classificatiemodellen kan worden vermeden. De prestatie van modellen voor de classificatie van looppatronen van mensen met verschillende leeftijd wordt in dit hoofdstuk vergeleken. De prestaties van vijf DL-modellen (met het ‘ruwe’ accelerometer signaal als invoer) is vergeleken met vier CML-

modellen (met afzonderlijke loopvariabelen). De resultaten toonden aan dat alle DL-modellen de CML-modellen overtroffen, met een area under the curve (AUC) van meer dan 0,94, vergeleken met de hoogste AUC van 0,83 van het beste CML-model. Deze bevindingen benadrukten niet alleen de superioriteit van DL, maar suggereerden ook dat DL waardevolle loopuitkomsten heeft geleerd die leeftijdsgerelateerde veranderingen in het lopen weerspiegelen, welke door de CML niet werden geïdentificeerd. Daarnaast beschrijft het hoofdstuk de effecten van verschillende venstergroottes van data selectie op de classificatieprestaties. Verschillende venstergroottes resulteren in verschillende aantallen van opeenvolgende loopcycli die gebruikt worden als input voor het DL-model. Er werd geconstateerd dat Convolutionele Neurale Netwerken (CNN) in staat waren om op basis van een enkele loopcycli onderscheid te maken tussen volwassenen en oudere volwassenen, terwijl Recurrent Neural Network (RNN) de relaties over de tijd tussen loopcycli meeneemt in de classificatie.

Voortbouwend op de inzichten uit hoofdstuk 2, heeft het onderzoek dat in **hoofdstuk 3** wordt beschreven als doel om te onderzoeken wat DL-modellen hadden geleerd van de data bij het classificeren van volwassenen en oudere volwassenen, met behulp van XAI. De resultaten lieten zien dat het accelerometer signaal tijdens de late zwaaifase tot aan het neerzetten van de voet binnen een loopcyclus, met name het moment rond het hielcontact, het meest bijdroeg aan het classificatieproces. Deze bevindingen suggereren dat DL verschillende versnellings- en vertragingspatronen identificeert die het looppatroon van oudere volwassenen onderscheidt van dat van volwassenen. Daarnaast toont het onderzoek aan dat variaties in versnelling en vertraging binnen enkele schrede voldoende is voor CNN om te classificeren (met een AUC van 0,89). Recurrent Neural Network (RNN) methoden, classificeert daarentegen op basis van relaties in versnellings- en vertragingspatronen over verschillende loopcycli, waarbij een AUC van 0,94 werd bereikt. De resultaten impliceerden dat RNN rekening houdt met dynamische veranderingen over de tijd, die gerelateerd zijn aan de houdingscontrole. Echter, opmerkelijk was dat XAI liet zien dat DL voornamelijk datagegevens in de verticale en anterolaterale richtingen gebruikte voor de classificatie in plaats van datagegevens in de mediolaterale richting, die sterker gerelateerd zijn aan balans.

De inzichten en methodologieën uit deze twee hoofdstukken kunnen gebruikt worden om inzicht te krijgen in het looppatroon van mensen met aandoeningen, zoals rugklachten. Studies naar het looppatroon bij patiënten met chronische lage rugpijn (CLBP) laten tegenstrijdige resultaten zien. Dit komt overeen met het beeld dat er is van van CLBP als een heterogene aandoening. De aanwezigheid van ‘human assumed central sensitization’ (HACS) in CLBP zou kunnen bijdragen aan deze heterogeniteit. In **hoofdstuk 4** wordt de hypothese getest dat verschillende niveaus van HACS (laag of hoog) gerelateerd zijn aan de looppatronen van patiënten met CLBP, en dat deze verschillen effectief geclassificeerd kunnen worden met

behulp van CML. Daarnaast zou XAI gebruikt kunnen worden om deze verschillen in looppatronen te interpreteren. Door de accelerometer signalen van één week uit de dagelijkse leefomgeving te analyseren, bevestigden de resultaten, gepresenteerd in dit hoofdstuk, dat patiënten met CLBP en met lage of hoge niveaus van HACS (respectievelijk CLBP- en CLBP+) effectief geclassificeerd konden worden door CML (bijvoorbeeld met Random Forest), met een nauwkeurigheid van 84,4%. XAI liet zien dat patiënten in de CLBP-groep een grotere soepelheid en een betere stabiliteit in loopgedrag vertoonden, terwijl patiënten in de CLBP+ groep een regelmatiger, minder variabel en voorspelbaarder looppatroon vertoonden. Deze bevindingen suggereren dat CLBP- en CLBP+-patiënten mogelijk verschillende motorische controle strategieën aannemen, namelijk "losse controle" en "strakke controle". De "losse controle" strategie, gekenmerkt door verminderde controle over de rompbewegingen, zou de looppatronen in CLBP- kunnen verklaren, terwijl de "strakke controle" strategie, met verbeterde controle over de rompbewegingen, de looppatronen in CLBP+ zou kunnen verklaren. Deze bevindingen benadrukken de noodzaak van een gepersonaliseerde behandeling.

Voortbouwend op de bevindingen uit hoofdstuk 4, onderzocht het in **hoofdstuk 5** beschreven onderzoek de intensiteit van fysieke activiteit patronen bij patiënten met CLBP, gemeten in de dagelijkse omgeving met een draagbare accelerometer gedurende ongeveer één week. Dit onderzoek maakte gebruik van CML zonder gebruik te maken van XAI (Hidden semi-Markov Model, HSMM) om patronen van fysieke activiteit van patiënten met CLBP- en CLBP+- te onderzoeken. Terwijl traditionele methoden die gebruik maken van vooraf ingestelde afkapwaarde van het acceleratiesignaal er niet in slaagden om verschillen in algemene fysieke activiteit intensiteit tussen CLBP- en CLBP+ -groepen statistisch significant vast te stellen, kan HSMMM de PAI-patronen vanuit de datagegevens leren. HSMM identificeerde zo verschillende PAI-patronen in deze twee groepen. CLBP- groepen hadden de neiging om taken op te splitsen in kleinere periodes, afgewisseld met frequente en korte rustpauzes. De CLBP+ groep vertoonde langere periodes van activiteit en rust. Fysieke activiteit patronen in de CLBP+ groep kunnen verklaard worden door een persisterende respons patroon, die gekenmerkt wordt door overmatig gebruik van fysieke structuren, ondanks het ervaren van pijn. De fysieke activiteitenpatronen in de CLBP-groep komen mogelijk niet overeen met het angst-vermijdingsrespons patroon, aangezien er geen aanwijzingen waren voor angstovertuiging in deze groep. De inzichten uit hoofdstuk 4 en 5 droegen bij aan een beter begrip van CLBP, beweging en HACS, en maken de weg vrij voor gepersonaliseerde behandelstrategieën in de toekomst.

In de hoofdstukken 4 en 5 werden de lage en hoge niveaus van HACS bepaald met behulp van de Central Sensitization Inventory (CSI) vragenlijst, op basis van een vooraf gedefinieerde afkapwaarde van 40. Er is echter gerapporteerd dat de afkapwaarde voor CSI kan variëren

door zowel verschil in pijnconditie als verschil in culturele en nationale context. Het is opmerkelijk dat er geen universeel geaccepteerde gouden standaard is voor het beoordelen van HACS. Gebaseerd op de patronen en structuren in de datagegevens van mensen met hetzelfde geslacht en dezelfde klinische uitkomsten (vragenlijsten over pijn, fysieke status en psychologische status), was het doel van het in **hoofdstuk 6** beschreven onderzoek om CML zonder toezicht van XAI te gebruiken voor het onderzoeken van HACS-gerelateerde subgroepen binnen patiënten met CLBP. Uit de resultaten werden twee verschillende subgroepen binnen de CLBP-populatie geïdentificeerd. De ene subgroep, die gekenmerkt werd door meer pijn, de grootste invaliditeit, slechtere psychologische status en hogere CSI-waarden, werd toegewezen aan hogere HACS-niveaus, terwijl de andere subgroep lagere HACS-niveaus vertegenwoordigde. Op basis van deze subgroepen werd een nieuwe afkapwaarde van 35 vastgesteld voor de Nederlandstalige populatie met CLBP. De methodologie die in dit hoofdstuk gebruikt werd, gaf nieuw inzicht in het identificeren van HACS-gerelateerde patronen en draagt bij aan het vaststellen van nauwkeurigere afkapwaarden.

Hoofdstuk 7 geeft een samenvatting van de belangrijkste bevindingen in dit proefschrift en bespreekt mogelijke richtingen voor toekomstig onderzoek. Door het proefschrift als casestudy te gebruiken, worden richtlijnen gegeven voor het selecteren van geschikte AI modellen voor bewegingsanalyse, waarbij de voor- en nadelen worden besproken. Door de potentie van XAI voor de bewegingswetenschappen te onderstrepen, benadrukte dit hoofdstuk de noodzaak van voortdurende verbeteringen om XAI betrouwbaarder, begrijpelijker en gebruiksvriendelijker te maken. Ondanks de erkende beperkingen, toont het hoofdstuk wat de potentie is van AI, met name XAI, voor de (klinische) bewegingsanalyse van verschillende patiëntengroepen. Deze ontwikkeling zal klinische experts en bewegingswetenschappers kunnen ondersteunen in het inzicht krijgen in veranderingen in het bewegingen tijdens dagelijkse activiteiten in verschillende patiëntengroepen.

Acknowledgements

As my Ph.D. journey nears its conclusion, upon reflection, I am pleasantly surprised to realize that I have gained far more than I ever could have imagined. Although this four-year journey has been relatively short, its impact on my life will be enduring. I would like to take this opportunity to express my deepest gratitude to all those who have supported and accompanied me throughout this remarkable journey, from my supervisors and colleagues to my friends and family. Your presence has served as significant milestones in this beautiful journey of my life, and it will forever be etched into my memories.

Dear Prof. **Claudine Lamoth**, you have been an exceptional supervisor to me, and so much more. I am deeply grateful for your guidance, invaluable support, and endless patience throughout this journey. In my early steps in research, your extensive knowledge and expertise provided me with opportunities to reshape my thinking, transitioning from a data science mindset to a more human movement science thinking. In the middle stages of our research, you gave me the freedom to explore, while always being there to help and keep me on track whenever I needed it. As we reached the final stages of our projects, your insightful and critical scientific input immensely benefited our projects. I genuinely appreciate the suggestions and comments you provided to improve our projects and our papers, even though you often mentioned that maybe you are too critical. In my view, it reminds me of a quote, "Mom is the only one who will tell you the cold and hard truth about yourself". Your critical thinking has played a crucial role in my growth as a researcher. You are a remarkable example in the pursuit of science and in nurturing students. If I am fortunate enough to become a teacher one day, you will undoubtedly be my role model. Additionally, I would like to express my gratitude for your care in my daily life. These experiences not only provided an energy boost but also fostered a sense of unity that made my journey more meaningful. I still cherish the memories of our time spent in your garden, having drinks, enjoying BBQ, or sharing hometown cuisines with colleagues. The fragrance of flowers, the sensation of grass, the warmth of sunshine, and our conversations still linger in my mind. I have learned a lot from you, both in academic and non-academic realms, directly and indirectly. Thank you for being an integral part of my journey as a supervisor and as a mentor. I sincerely hope that our collaboration does not end here.

Dear Prof. **Michiel Reneman**, I would like to sincerely thank you for your intelligence and valuable support throughout my Ph.D. journey. Your sense of humor and friendly nature make working with you enjoyable. Thank you for giving me the opportunity to work under your supervision. I vividly recall a particular incident when you and Judith invited my wife and me to visit your home. Unfortunately, there was a mix-up with the timing, and we arrived one hour early. As we were cycling along the countryside road, immersing ourselves in the picturesque scenery, we unexpectedly stumbled upon each other. What a coincidence!

Despite the wrong timing, it was clear that you were the right person, just as we were when we first made contact. The first email I received from you mentioned that my CV had caught your attention. Although my background did not directly align with Human Movement Sciences, after reading the project background you provided in the email, I made the right decision, which ultimately led to this thesis. My Ph.D. journey has been memorable, much like that day we spent together. Sitting in the sunshine in your garden, enjoying your cooking prowess, exploring the presence of wild deer and foxes in the fields, and admiring the sunset from a tower—these memories, as well as my Ph.D. journey, hold a special place in my heart. I am truly honored to have the privilege to learn under your guidance.

Dear Prof. **Bert Otten**, I am particularly grateful for your valuable interactions and discussions in guiding me through the project. Thank you for imparting essential skills that I learned from you, like storytelling skills. Whenever I encounter questions, you effortlessly provide suitable and straightforward examples, presenting them in a storytelling manner to address my questions. Your unwavering passion for research is truly admirable. Even after your retirement, you continue to devote yourself to research. Apart from this, I am amazed by your diverse range of interests, encompassing roles as a scientist in NeuroMechanics, a Software Developer, a Photographer, a Cycling Champion, and so much more. Thank you for your guidance and the knowledge you have imparted to me. Your influence will undoubtedly have a lasting impact on my academic and non-academic journey.

I would like to sincerely thank my thesis reading committee Prof. **Jaap van Dieeen**, Prof. **Natasha Maurtis**, and Prof. **Nils Strodthoff**. Thank you very much for investing time in evaluating and giving advice to improve my thesis. I extend my heartfelt thanks to Dr. **Yijian Yang**, Dr. **Juha Heijmans**, and Dr. **Alessio Murgia** for being my defense opponents.

I would like to extend my gratitude to the pioneers who have made significant contributions to the field in the past. Your excellent work laid the groundwork and built a solid foundation upon which this thesis stands. Your research and findings have provided important references and inspiration for my thesis.

Additionally, I am very grateful to **Jorine Schoenmaker** and **Jelmer Braaksma** for being my paranymphs and walking me through this stressful time! Jorine, I want to express my gratitude for sharing the office space with me. Our coffee breaks, lunch breaks, and after-work drinks together were truly enjoyable. Jelmer, every time hang out with you was a pleasure because of your incredible sense of humor. Thank you for organizing numerous lively activities that added joy to my Ph.D. journey. I would like to express my sincere gratitude to **Channah BellinK** for translating the summary into Dutch. Your assistance was invaluable in helping me complete this thesis.

I thank all my colleagues, especially my Chinese colleagues, and friends for their company and support. Although I am not mentioning their names here, I wish everyone a fulfilling academic and personal journey ahead.

Last but not least, I would like to express my gratitude to my family for their unwavering support throughout my life. I extend my heartfelt thanks to my **parents** for their endless love and care in educating and preparing me for the future.

I reserve my deepest appreciation for **Xia Wu**, my wife. Thank you for being my constant companion and supporting me. I am grateful for your willingness to listen to my research ideas as my first audience, even though these ideas may always not be attractive. Your significant contributions as the co-author of my life and research mean more to me than words can express.

About the Author

Xiaoping Zheng (郑潇平) was born on April 8, 1993, in Shantou, China. From 2012 to 2016, he pursued his Bachelor's degree in Information Engineering at the Chinese University of Geosciences in Wuhan, over 1,000 km away from his hometown. He continued at the same university for his Master's in Software Engineering, completing it in 2019. This education laid a foundation in computer science for him. In October 2019, Xiaoping embarked on a new journey over 10,000 km away from hometown, to study a new subject, Human Movement Science, at the University of Groningen. His 4-year PhD program was supported by the China Scholarship Council-UG Joint Scholarship (Grant No. 201906410084).

Xiaoping's PhD research revolved around exploring artificial intelligence in studying human movement characteristics. Working within a multidisciplinary team, he collaborated with data scientists to apply advanced techniques in extracting insights from complex data collected from movements. He cooperated closely with movement scientists, aiding in the interpretation of data-driven findings and their implications in human movement. Furthermore, he liaised with clinicians and rehabilitation experts to translate these findings into clinically relevant applications. In this research, he contributed as a bridge, bridging the gap between data science and practical applications, making complex data understandable and useful across various disciplines.

Currently, Xiaoping continues to explore the relationship between movement and health by utilizing artificial intelligence at the Chinese University of Hong Kong. After his PhD, he will officially commence his Postdoctoral research at the same institution.

Scientific Output

Journal Publications

Zheng, X., Reneman, M., Echeita, J., Schiphorst Preuper, R., Kruitbosch, H., Ottem, E., and Lamoth, C., Association between central sensitization and gait in chronic low back pain: Insights from a machine learning approach. Computers in biology and medicine, 144 (2022): 105329. doi.org/10.1016/j.compbiomed.2022.105329

Zheng, X., Reneman, M., Schiphorst Preuper, R., Ottem, E., and Lamoth, C., Relationship between physical activity and central sensitization in chronic low back pain: Insights from machine learning. Computer Methods and Programs in Biomedicine, 232 (2023): 107432. doi.org/10.1016/j.cmpb.2023.107432

Submitted for Publication

Zheng, X., Wilhelm, E., Reneman, M., Ottem, E., and Lamoth, C., Age-related gait patterns classification using deep learning based on time-series data from one accelerometer. doi.org/10.36227/techrxiv.22643314.v1

Zheng, X., Reneman, M., Ottem, E., and Lamoth, C., Explaining deep learning models for age-related gait classification based on acceleration time series. doi.org/10.48550/arXiv.2311.12089

Zheng, X., Lamoth, C., Timmerman, H., Ottem, E., and Reneman, M., Establishing central sensitization inventory cut-off Values in patients with chronic low back pain by unsupervised machine learning. doi.org/10.48550/arXiv.2311.11862

Conference Contributions

Oral Presentations

Zheng, X., Reneman, M., Ottem, E., and Lamoth, C., Explaining deep learning models for age-related gait classification based on time-series acceleration. ISB 2023, Fukuoka, Japan.

Zheng, X., Wilhelm, E., Reneman, M., Ottem, E., and Lamoth, C., Deep learning for age-related gait patterns classification based on raw accelerometer signal from 3 minutes walking. ISPGR 2023, Brisbane, Australia.

Zheng, X., Reneman, M., Schiphorst Preuper, R., Ottem, E., and Lamoth, C., Physical activity patterns of patients with chronic low back pain and central sensitization: insights from a machine learning method. BME 2023, Egmond aan Zee, The Netherlands.

Zheng, X., Reneman, M., Schiphorst Preuper, R., Ottem, E., and Lamoth, C., Effects of level of central sensitization on physical activity patterns in chronic low back pain: insights from a machine learning approach. ISPGR 2022, Online.

Zheng, X., Reneman, M., Echeita, J., Schiphorst Preuper, R., Kruitbosch, H., Ottem, E., and Lamoth, C., Exploring effects of central sensitization on gait in chronic low back pain by using machine learning approach. ICAMPAM 2022, Online.

Zheng, X., Reneman, M., Echeita, J., Schiphorst Preuper, R., Kruitbosch, H., Ottem, E., and Lamoth, C., Classification of patients with chronic low back pain and high or low central sensitization by gait outcomes using machine learning methods. BME 2021, Online.

Zheng, X., Lamoth, C., Timmerman, H., Ottem, E., and Reneman, M., Establishing central sensitization inventory cutoff-scores in chronic low back pain population by deep learning. PA!N Congres 2021, Online.

Zheng, X., Reneman, M., Echeita, J., Schiphorst Preuper, R., Kruitbosch, H., Ottem, E., and Lamoth, C., Exploring effects of central sensitization on gait in chronic low back pain by machine learning. PA!N Congres 2021, Online.

Zheng, X., Reneman, M., Schiphorst Preuper, R., Ottem, E., and Lamoth, C., Physical activity levels of patients with chronic low back pain and central sensitization: insights from a machine learning method. ICAMPAM 2021, Online.

Research Institute SHARE

This thesis is published within the **Research Institute SHARE** (Science in Healthy Ageing and healthcaRE) of the University Medical Center Groningen / University of Groningen.
Further information regarding the institute and its research can be obtained from our internet site: https://umcgresearch.org/w/share

More recent theses can be found in the list below (supervisors are between brackets).

2024

Oostrum I van
Survival extrapolation models' impact on cost-effectiveness assessments for reimbursement of oncolytics
(Prof E Buskens, Prof MJ Postma)

Bouma SE
Lifestyle-related treatment modalities in hip and knee osteoarthritis care: A mixed-methods investigation of implementation determinants and strategies among healthcare professionals
(Dr M Stevens, Dr I van den Akker-Scheek, Prof RL Diercks, Prof LHV van der Woude)

Pardoel ZE
The contribution of community-based programmes to health
(Prof MJ Postma, Prof SA Reijneveld, Prof BW Lensink, Dr JA Landsman-Dijkstra)

Mylius CF
Physical Fitness & Activity: Reference values and a clinical application in major abdominal surgery
(Prof CP van der Schans, Dr WP Krijnen, Dr T Takken)

Tuin S van der
Zooming in on the development of psychotic experiences: Insights from daily diaries
(Prof AJ Oldehinkel, Dr Wardenaar-Wigman, Dr SH Booij)

Jima GH
Promoting contraceptive uptake to reduce the unmet need for family planning during the postpartum period in Ethiopia
(Prof J Stekelenburg, Dr RG Biesma-Blanco, Dr Tegbar Yigzaw)

Dijkhuis TB
Physical performance in daily life and sports: bridging the data analytics gap
(Prof KAPM Lemmink, Prof M Aiello, Dr H Velthuijsen)

Zhang X
Novel Methods in Preference-based Health Outcome Measurement: Development, Validation, Application
(Dr PFM Krabbe, Dr KM Vermeulen)

Minaeva O
Unraveling the Rhythm of Depression: Exploring Physical Activity, Sleep, and Circadian Markers for Depression Detection and Prediction
(Dr H Riese, Prof MC Wichers, Dr SH Booij)

Kamp T
Return to work after total hip or total knee arthroplasty
(Prof S Brouwer, Dr M Stevens)

Libutzki B
Comorbidities and medical costs of Attention-Deficit/Hyperactivity Disorder
(Dr CA Hartman, Prof A Reif, Prof B Neukirch)

2023

Hashim MSMM
Exploring clinical and economic value of novel therapies in oncology; contemporary challenges and applications
(Prof MJ Postma, Dr B Heeg)

Silva Gurgel do Amaral M
Can you help me take care of my health? Exploring the role of patients' health literacy to improve the prevention and management of chronic kidney disease
(Dr AF de Winter, Prof SA Reijneveld, Prof GJ Navis)

Bao M
Early Life Exposures and Offspring Health: From Animal Models to Human Studies
(Prof T Plösch, Dr E Corpeleijn)

Zeevat F
Assessment of vaccination policies from a health economic perspective: opportunities and emerging foci
(Prof MJ Postma, Prof C Boersma)

Burger J
The Future of Case Formulation in Clinical Psychology: Advancements in Network Modeling and Simulation-based Science
(Dr H Riese, Prof RA Schoevers)

For earlier theses visit the website: Find Research outputs — the University of Groningen research portal (rug.nl)