TU WIEN Informatics

Human Gait Analysis

Machine Learning-Based Classification of Gait Disorders

DISSERTATION

zur Erlangung des akademischen Grades

Doktor der Technischen Wissenschaften

eingereicht von

Dipl.-Ing. Djordje Slijepčević, BSc
Matrikelnummer 00925240

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.-Prof. Dipl.-Ing. Dr.techn. Christian Breiteneder
Zweitbetreuung: FH-Prof. Priv.-Doz. Dipl.-Ing. Mag. Dr. Matthias Zeppelzauer

Diese Dissertation haben begutachtet:

Neil Cronin

Morgan Sangeux

Wien, 21. Mai 2024

Djordje Slijepčević

Informatics

Human Gait Analysis

Machine Learning-Based Classification of Gait Disorders

DISSERTATION

submitted in partial fulfillment of the requirements for the degree of

Doktor der Technischen Wissenschaften

by

Dipl.-Ing. Djordje Slijepčević, BSc
Registration Number 00925240

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.-Prof. Dipl.-Ing. Dr.techn. Christian Breiteneder
Second advisor: FH-Prof. Priv.-Doz. Dipl.-Ing. Mag. Dr. Matthias Zeppelzauer

The dissertation has been reviewed by:

Neil Cronin

Morgan Sangeux

Vienna, 21st May, 2024

Djordje Slijepčević

Erklärung zur Verfassung der Arbeit

Dipl.-Ing. Djordje Slijepčević, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 21. Mai 2024

Djordje Slijepčević

Acknowledgements

I am deeply grateful for the guidance, support, and expertise of my first supervisor, Prof. Christian Breiteneder, whose insights and direction were invaluable throughout this journey. His profound knowledge has been a driving force and inspiration behind this dissertation.

A heartfelt thanks to Matthias Zeppelzauer, my second supervisor and daily mentor at the university. Matthias, your involvement in the day-to-day aspects of my academic work and your mentorship have been fundamental in shaping the quality and direction of this research.

My gratitude extends to Brian Horsak, my mentor in the field of human gait analysis. Brian, your leadership in the gait analysis projects at the university and your insightful feedback have been essential in my understanding and exploration of this complex subject.

A special thanks to Fabian Horst for the enriching collaboration over the last years. Your moral and content-wise support, especially during the crucial stages of this thesis, have been of immense value.

I would like to express my appreciation to all the co-authors of the publications related to this thesis. Your contributions and collaborative spirit have been essential in achieving the quality of the presented work.

A special thanks to Jürgen Pannosch and Lukas Daniel Klausner for their meticulous proofreading and feedback.

To my parents, Nataša and Radomir, and my brother Dimitrije, your unwavering support and belief in me have been the foundation for this achievement. This journey would not have been possible without your love, encouragement, and sacrifice. Your motivating words from the Master's thesis period, "Piši, piši ponekad dva-tri reda" have motivated me throughout this endeavor as well.

Last but not least, to my wife Birgit and my two children, Lora and Iva, your patience, understanding, and unending support have been my greatest strength. Balancing family life with academic endeavors is a challenge, one that would have been impossible to overcome without your constant love and support (e.g., such as during the late-night proofreading sessions).

Kurzfassung

Die klinische Ganganalyse ermöglicht die Bewertung des menschlichen Gangbildes. Sie liefert die Grundlage für KlinikerInnen, um präzise Diagnosen zu stellen und effektive Behandlungspläne zu entwickeln. Die klinische dreidimensionale Ganganalyse, die als Goldstandard in der klinischen Praxis gilt, umfasst verschiedene Datenmodalitäten wie z.B. kinematische Daten (z.B. Gelenkwinkel), die mithilfe optischer Bewegungserfassungssysteme berechnet werden, und kinetische Daten (z.B. Bodenreaktionskräfte), die über Kraftmessplatten aufgezeichnet werden. Diese Daten sind multivariate hochdimensionale Zeitreihen, die zeitliche Abhängigkeiten und nichtlineare Beziehungen zueinander aufweisen. Die Komplexität dieser Daten und der entsprechenden klinischen Aufgabenstellungen, insbesondere bei der Identifikation spezifischer pathologischer Gangmuster, hat zur Anwendung von Methoden des maschinellen Lernens (ML) geführt. Der Einsatz von ML-Methoden zielt darauf ab, die Effizienz der klinischen Ganganalyse zu erhöhen und zu einer besser informierten Entscheidungsfindung beizutragen. Durch den Einsatz von ML-Methoden können ForscherInnen und KlinikerInnen große Mengen von Gangdaten analysieren, um neue Erkenntnisse zu gewinnen, die mit konventionellen Methoden schwer zu erlangen wären. Viele dieser Ansätze sind jedoch mit Einschränkungen verbunden, wie zum Beispiel die Verwendung von kleinen Datensätzen oder vereinfachten Aufgabenstellungen mit wenigen Klassen.

In dieser Dissertation werden bestehende Limitationen in der klinischen Ganganalyse behandelt, und es wird ein methodischer Beitrag dazu geleistet, komplexe Mehrklassen-Klassifikationsaufgaben mit der Entwicklung erklärbarer ML-Ansätzen zu bewältigen. Zu diesem Zweck werden traditionelle ML- und Deep-Learning-Ansätze entwickelt, und ihre Anwendbarkeit auf Gangdaten und entsprechende Klassifikationsaufgaben untersucht. In dieser Dissertation werden erstmals Erklärungsansätze für ML-Methoden (einschließlich Deep-Learning-Methoden) für klinische Gangdaten vorgestellt, die es ermöglichen, Entscheidungen für KlinikerInnen nachvollziehbar zu machen. Darüber hinaus wird die Nützlichkeit von Erklärbarkeitsmethoden bei der Identifizierung von Verzerrungen innerhalb der Daten und der trainierten ML-Modelle aufgezeigt. Neben einer systematischen Evaluierung von Datenaufbereitungsstrategien in Bezug auf die Skalierung und Extraktion von Merkmalen sowie Unausgewogenheit der Daten wird auch die diskriminative Fähigkeit von Bodenreaktionskräften und kinematischen Daten untersucht.

Die vorliegende Dissertation leistet einen bedeutenden Beitrag, indem sie einen großen realen Datensatz namens GaitRec einführt. Dieser Datensatz soll als Benchmark-Datensatz dienen und bildet eine entscheidende Grundlage für die standardisierte Bewertung der Leistung von ML-Ansätzen. In dieser Arbeit werden zwei Anwendungsfälle mit unterschiedlich komplexen Klassifikationsaufgaben untersucht, die große Mengen klinischer Daten nutzen. Der erste Anwendungsfall verwendet den GaitRec Datensatz und beinhaltet Bodenreaktionskraftdaten sowohl von gesunden Personen als auch von PatientInnen mit funktionellen Gangstörungen. Der zweite Anwendungsfall umfasst kinematische Daten (z.B. Gelenkwinkel) und Bodenreaktionskraftdaten von PatientInnen mit Zerebralparese.

Abschließend werden in dieser Arbeit zukünftige Forschungsrichtungen aufgezeigt, die das Potenzial haben, den Bereich der automatisierten Klassifizierung von klinischen Gangdaten voranzubringen.

Abstract

Clinical gait analysis is a central approach for assessing human gait, which forms the foundation for clinicians to make accurate diagnoses and to develop effective treatment plans. Clinical three-dimensional gait analysis, considered as gold standard in clinical practice, involves various data modalities such as kinematic data (e.g., joint angles) calculated using optical motion capture systems and kinetic data (e.g., ground reaction forces) recorded via force plates. These data represent multivariate high-dimensional time series signals that exhibit temporal dependencies and non-linear relationships. The complexity of these data and corresponding clinical tasks, particularly in the identification of specific pathological gait patterns, has motivated researchers to investigate the suitability of machine learning (ML) methods to solve gait analysis tasks. The use of ML methods aims to improve the efficiency of clinical gait analysis and to contribute to better informed decision-making. By using ML, researchers and clinical experts can analyze large amounts of gait data to gain new insights, which would be difficult with conventional methods. However, many of these approaches are also subject to limitations, such as using small datasets for training or addressing simplified tasks with only a few classes.

The present thesis addresses existing gaps and limitations and makes a significant methodological contribution to explainable ML approaches for complex multi-class gait classification tasks. For this purpose, traditional ML and deep learning approaches are developed, and their suitability for gait data and corresponding classification tasks is investigated. This thesis proposes for the first time explainability approaches for ML methods (including deep learning methods) for clinical gait data that enable clinicians to trace decisions. Additionally, it demonstrates the usefulness of explainability methods in identifying biases within ML pipelines and gait data. In addition to a systematic evaluation of data handling strategies concerning feature scaling, feature extraction, and data imbalances, this thesis investigates the discriminative power of ground reaction force and joint angle data.

A significant contribution of the current thesis lies in the publication of a large-scale real-world dataset named GaitRec. This dataset serves as a benchmark, providing a crucial foundation for assessing the performance of ML approaches in a standardized way. In this work, two use cases with complex binary and multi-class classification tasks are investigated, utilizing large-scale clinical datasets. The first use case utilizes the

GaitRec dataset and involves ground reaction force data from both healthy individuals and patients with functional gait disorders. The second use case encompasses kinematic (i.e., joint angles) and ground reaction force data from patients with cerebral palsy.

Finally, the present thesis identifies future research directions that have the potential to advance the field of automated classification of clinical gait data.

Contents

Preface

The present thesis follows a cumulative approach, encompassing a collection of selected publications derived from my contributions to machine learning in the field of clinical gait analysis. The following five journal publications constitute the main body of this thesis:

- Brian Horsak, Djordje Slijepcevic, Anna-Maria Raberger, Caterine Schwab, Marianne Worisch, and Matthias Zeppelzauer. **GaitRec, a Large-Scale Ground Reaction Force Dataset of Healthy and Impaired Gait**. *Scientific Data*, 7(1):143, 2020. DOI: 10.1038/s41597-020-0481-z

- Djordje Slijepcevic, Matthias Zeppelzauer, Anna-Maria Gorgas, Caterine Schwab, Michael Schüller, Arnold Baca, Christian Breiteneder, and Brian Horsak. **Automatic Classification of Functional Gait Disorders**. *IEEE Journal of Biomedical and Health Informatics*, 22(5):1653–1661, 2017. DOI: 10.1109/JBHI.2017.2785682

- Djordje Slijepcevic, Matthias Zeppelzauer, Caterine Schwab, Anna-Maria Raberger, Christian Breiteneder, and Brian Horsak. **Input Representations and Classification Strategies for Automated Human Gait Analysis**. *Gait & Posture*, 76:198–203, 2020. DOI: 10.1016/j.gaitpost.2019.10.021

- Djordje Slijepcevic, Fabian Horst, Sebastian Lapuschkin, Brian Horsak, Anna-Maria Raberger, Andreas Kranzl, Wojciech Samek, Christian Breiteneder, Wolfgang Immanuel Schöllhorn, and Matthias Zeppelzauer. **Explaining Machine Learning Models for Clinical Gait Analysis**. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(2):1–27, 2021. DOI: 10.1145/3474121

- Djordje Slijepcevic, Matthias Zeppelzauer, Fabian Unglaube, Andreas Kranzl, Christian Breiteneder, and Brian Horsak. **Explainable Machine Learning in Human Gait Analysis: A Study on Children With Cerebral Palsy**. *IEEE Access*, 11:65906–65923, 2023. DOI: 10.1109/ACCESS.2023.3289986

Chapter 1 motivates the undertaking of the thesis (Section 1.1), outlines the aims (Section 1.2), offers a synopsis and summary of the publications (Section 1.3), details the methodology employed (Section 1.4), discusses the results (Section 1.5), and addresses both limitations and future research directions (Section 1.6). Section 1.1 provides a general overview of the field of clinical gait analysis and motivates the use of machine learning for the automated analysis of clinical gait data. Section 1.1 concludes with the current gaps and limitations of machine learning application in clinical gait analysis that serve as the foundation for motivating the aims of the thesis. The aims and corresponding research questions of the thesis are presented in Section 1.2. Section 1.3 provides concise summaries of the publications contributing to this thesis, along with individual contributions classified according to the contributor roles taxonomy (CRedIT) [1]. Section 1.4 provides an overview of the methodology (i.e., investigated datasets, machine learning methods, and explainability methods) that was utilized to address the research questions and to achieve the goals of this thesis. Section 1.4 concludes with a general overview of the field of explainable artificial intelligence and matches the utilized methods to an established taxonomy. Section 1.5 provides a comprehensive discussion of the research findings from the perspective of the defined goals and the corresponding research questions. Section 1.6 offers an extensive exploration of limitations in the addressed research field while identifying potential directions for future research. Section 1.7 summarizes the scientific contributions of the thesis.

Chapter 2 contains the publications that constitute the main body of this thesis in their original form as they were published. The order in which the publications are presented is selected to ensure logical coherence for the reader and alignment with the research questions (independent of the chronological order of publication dates).

CHAPTER 1

Introduction

1.1 Motivation

Diseases or injuries of the musculoskeletal locomotor system as well as neurological disorders can affect people of any age (although they are more prevalent in older populations), regardless of gender and social status, and are one of the main causes of pathological impairments of human motor function. Various factors, including, e.g., infections, inflammations, degenerative processes, traumatic events, neoplastic and vascular diseases, as well as neurological conditions such as cerebral palsy, Parkinson's disease, multiple sclerosis, and stroke, can cause impairments in human motor function [2]. Affected people may lose the ability to interact with their environment and participate fully in social activities or the labor market as a result of these disabilities.

According to the Global Burden of Disease Study 2019, musculoskeletal disorders were identified as one of the leading factors contributing to the growing burden on health systems across 202 countries [3]. In Austria, similar trends can be observed, with diseases related to the musculoskeletal system and connective tissues accounting for approximately 21.9% of the causes of sick leave in 2021 [4]. Furthermore, among the population aged 15 and above in Austria, 9.1% experienced challenges when walking longer distances, while 11.3% encountered difficulties while climbing stairs [5]. The aforementioned statistics, their implications, and the burden they impose on national health care systems provide strong motivation for conducting extensive research on causes and symptoms associated with diseases related to the musculoskeletal system. Various factors can be linked to diseases and injuries related to the musculoskeletal system, such as physical activity, diet, obesity, and smoking. Extensive studies have been dedicated to human gait, due to its role as an indicator of both physical activity and overall quality of life. To better understand gait impairments and allow for an optimal patient treatment, an accurate assessment of underlying movement mechanisms is essential. Different gait analysis approaches of varying complexity have been developed for this purpose.

1.1.1 Clinical Gait Analysis

Clinical gait analysis serves as a tool for the evaluation of human gait, with the primary aim to identify impairments that affect a patient's gait pattern [6]. Clinical gait analysis supports clinicians in making accurate diagnoses and developing individualized treatment plans for their patients. For this reason, clinical gait analysis has become a crucial assessment tool in hospitals and rehabilitation centers. There are different gait analysis approaches with varying levels of complexity and accuracy, as well as different requirements for equipment and personnel. These approaches range from observational gait analysis [7] to more quantitative measures such as kinematic (e.g., joint angles) and kinetic (e.g., joint moments) data derived from instrumented three-dimensional gait analysis (3DGA) [8].

Clinical 3DGA is well-established in clinical practice and regarded as the gold standard for the quantification of a patient's gait performance due to the high accuracy and quality of derived information [6]. This approach relies primarily on motion capture techniques in which retro-reflective markers are placed at specific anatomical landmarks on the human body. Using the 3D trajectories of these markers in conjunction with geometric biomechanical models, kinematic data such as joint angles can be accurately calculated [6]. Alternatively, recent approaches use inertial measurement units (IMUs) to extract kinematic information outside the gait laboratory [9, 10, 11]. In addition to kinematic data, assessments often include the measurement of muscle activation via electromyography as well as ground reaction forces via force plates [6]. The ground reaction force (GRF) corresponds to the force generated by the ground as a reaction force equal to the (averaged) force applied by the human body to the ground (i.e. body weight) [8]. By utilizing these different data modalities, a comprehensive understanding of the patient's walking behavior can be obtained as they capture complementary information. The general drawbacks of 3DGA are the need for highly trained staff as well as high acquisition and maintenance costs. In addition, a major drawback is the time-consuming and complex setup process, which includes the system calibration and the attachment of markers or sensors to specific landmarks on the patient's body. For certain groups of patients, this time investment might not be feasible, leading to the recording of only GRF data.

1.1.2 Automated Classification of Gait Data

The prevalent approach in the current clinical setting involves the manual analysis of gait data obtained via 3DGA by representing it in the form of line plots during the assessment and diagnostic process (see Figure 1.1). However, this approach is susceptible to subjectivity, time-consuming, and can be costly. Furthermore, experienced and qualified domain experts are necessary for the manual analysis of gait data due to the high-dimensional nature and the presence of temporal dependencies, strong variability, non-linear relationships, and inter-correlations within the different signals [12].

Clinical gait data and medical history are stored in databases that implicitly contain a vast amount of valuable clinical knowledge, which is, however, currently hardly accessible.

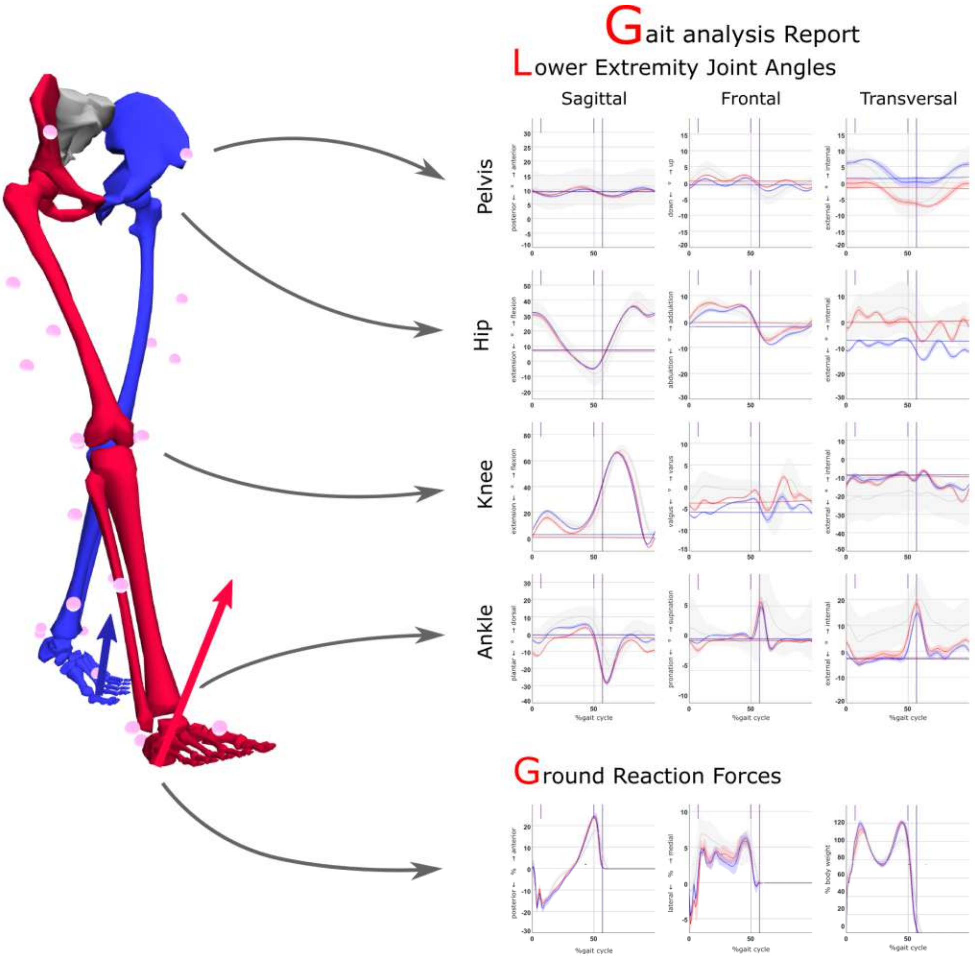


Figure 1.1: Data typically presented in a 3D gait analysis report used in clinical practice. Retro-reflective markers (depicted as pink spheres) are attached to specific anatomical landmarks on the human body, enabling the quantification of human locomotion through a 3D motion capture system. The 3D trajectories of these markers combined with geometrical biomechanical models are utilized to determine joint angles. In addition, ground reaction forces are determined via force plates. In clinical practice, these data are typically used to inform medical decision-making. The data from clinical gait analysis reports are typically presented as simple line plots. Blue and red colors encode the right and left body sides, respectively. Deriving a diagnosis from these abstract line plots is a challenging task that requires the expertise of trained medical professionals.

Automated data analysis methods that utilized machine learning (ML) bear the potential to exploit this implicit knowledge and provide an efficient and data-driven way for the automated detection of pathological gait patterns. The application of automated data analysis methods can assist clinicians by providing efficient insights into gait data without the need for extensive manual analysis of the complex data. The aim of developing automated analysis approaches is not to replace clinicians, but rather to enhance their capabilities and provide them with valuable tools for faster and more accurate assessment. Clinicians could benefit from immediate insights and data-driven support that enable them to make better-informed decisions and create personalized treatment plans. Furthermore, accelerating the diagnosis and decision-making process through ML-based assistance systems would also save time and thus healthcare costs.

In recent years, ML has made significant contributions to healthcare applications. For example, in the medical domain, ML methods have already been able to detect skin and breast cancer more efficiently and accurately than clinicians [13, 14, 15]. However, the field of clinical gait analysis still lags behind despite having accumulated a wealth of data through different measurement methods over the past decades. The demand for rapid and accurate decision-making, coupled with the complexity of gait data, has driven research efforts to leverage ML [16]. However, existing literature addressing ML approaches in clinical gait analysis exhibits limitations, which are outlined in Section 1.1.4. To overcome the limitations of existing ML approaches based on handcrafted domain-specific features and linear compression using principal component analysis (PCA), there has been a strong motivation to explore non-linear representations [17]. Utilizing deep learning enables the autonomous learning of non-linear representations through a data-driven approach. The motivation for the application of deep learning builds on the idea that human motor actions consist of elementary building blocks, so-called motion primitives, at different levels (e.g., neural or kinematic) [18, 19, 20]. Different transformations and combinations of these motion primitives to more complex modules form increasingly complex motor actions. Thus, the human gait as such a complex motor action is also based on redundant modules on every level of the motor hierarchy. This structural property makes gait data especially interesting for feature learning [18]. For hierarchically structured data, e.g., images, music, or speech, deep learning methods have shown to be particularly suitable to learn hierarchical representations that combine basic building blocks to complex and abstract concepts [18]. The basic assumption for biomechanical gait data is that specific pathologies are associated with distinct motion primitives that compose the gait pattern of a patient. Deep learning-based approaches represent a promising candidate to learn meaningful gait representations from the data. Furthermore, the application of explainability methods could serve as a valuable tool in identifying the location and pattern of motion primitives associated with the investigated pathologies.

1.1.3 Significance of Explainability

While ML approaches show promising outcomes in terms of classification performance, they often suffer from a significant drawback, which is their black-box nature [21]. This implies that even if we understand the underlying mathematical principles of these methods, their decision-making process is often incomprehensible and their predictions are hard to trace. Thus, the problem in the context of clinical gait analysis is that it remains unclear to clinical experts whether predictions are based on clinically relevant patterns or if they are influenced by spurious correlations or biases in the data that are not causally related to the targeted pathologies. The inability to validate the functioning of complex ML models and the challenge of understanding the learned patterns and rules are currently restricting the application of ML-based decision-support systems in clinical practice. However, clinicians require full transparency of decisions [22, 23]. The absence of transparency in ML approaches poses a significant challenge in offering justifications for their predictions. These justifications are essential for compliance with regulations such as the General Data Protection Regulation (GDPR, EU 2016/679) [24] and the recently proposed Artificial Intelligence Act [25] by the European Commission.

For simpler ML models that are inherently explainable, such as decision trees, generating decision and model explanations can be relatively straightforward (e.g., by utilizing feature importance). However, to identify patterns within the input data that contribute to the predictions of complex ML models, methods from the field of *explainable artificial intelligence* (XAI) are necessary. In general, these explainability methods aim to reveal the workings of complex non-linear ML models and the way they produce their predictions.

1.1.4 Current Gaps and Limitations

In the context of clinical gait analysis and the automated analysis of gait data, various limitations become evident. These limitations manifest across different levels, and the subsequent listing is not intended to present a comprehensive compilation but a summary of limitations that serve as motivation for the present thesis.

Data and annotation availability. Existing literature on ML approaches in clinical gait analysis has primarily focused on simple use cases and small-scale datasets. Furthermore, in the field of clinical gait analysis, there is a lack of comprehensive publicly available datasets containing data from patients and healthy controls. An important constraint for gait data is also the absence of annotations, which are particularly crucial in clinical scenarios (e.g., pathological gait patterns) as the annotation process typically involves a subjective and time-consuming evaluation by clinical experts. Each laboratory collects the data independently, and the use of different laboratory settings further complicates the merging of different data modalities. However, it is critical to consider incorporating heterogeneous and large datasets to train and validate robust ML models. This approach is central to ensuring the applicability of these models in diverse populations and to improve their generalizability.

Differently expressive data modalities. As the gold standard for assessing human gait, 3DGA considers the kinematic and kinetic aspects of movement. However, in everyday clinical practice, clinicians and therapists face challenges due to the necessity to examine a large number of patients. There is a trade-off between the accuracy and time efficiency of 3DGA. Additionally, motion capture systems utilized for 3DGA are expensive and the operation of such systems requires trained personnel, which further complicates their integration into clinical practice. Thus, an alternative that is sometimes used involves recording only GRFs using force plates. Considering the time-efficient process of collecting only GRF data, as opposed to 3DGA, the accumulation of datasets suitable for automated analysis becomes more feasible. Moreover, the availability of GRF data is higher, as they can be obtained from regular 3DGA, as well as from gait laboratories without motion capture systems and can include historic data from periods when such complex recording systems were not utilized. In addition to simplified data collection, GRF data also offers advantages such as easier integration of datasets from different gait laboratories. This facilitates the setup of multi-center studies, while the integration of kinematic data from different gait laboratories is more complicated due to differences in marker setups and biomechanical models across gait laboratories. As a result, numerous studies in the literature have utilized GRF data and demonstrated high classification performance. However, these studies primarily focused on distinguishing between one or two specific pathological gait patterns and healthy controls (physiological gait) [16]. These studies investigated pathological gait patterns associated with conditions such as Parkinson's disease [26, 27, 28], cerebral palsy [29], multiple sclerosis [29], osteoarthritis [30], transfemoral amputation [31], and lower limb fracture [32]. The main drawback of utilizing only GRF data is that the view on the biomechanical processes of the lower extremities is narrowed compared to data derived from 3DGA, as kinematic processes are not explicitly represented. This is also the reason why GRF data have often only been utilized for binary classification tasks, e.g., to distinguish between healthy controls and a single pathological gait pattern. The quantitative assessment of the discriminative power of GRF data for multi-class classification tasks, in comparison to 3DGA data, remains unexplored in the literature.

Lack of systematic evaluation of data handling strategies. In the existing literature on automated gait classification, evaluating the impact of different data processing strategies on performance has not yet been thoroughly addressed for complex multi-class classification tasks. In particular, there is a gap in the study of how factors such as feature scaling and feature extraction, data imbalance, and dealing with various trials per individual affect the performance of ML approaches. Understanding the impact of these data factors on complex, clinically relevant datasets is crucial for optimizing the performance and robustness of such approaches.

Limitations in systematically evaluating traditional ML and deep learning approaches. The main difference between deep learning and traditional ML relates to the concept of feature extraction. In deep learning, there is no need for explicit feature extraction since the model inherently learns the features directly from the data (i.e., feature learning). This capability is enabled by the architecture of deep neural networks. These models are composed of multiple stacked layers that facilitate the learning of higher-level, more abstract features from the raw input data. These high-level features enable deep learning models to handle complex multi-class classification tasks [17]. There has been an increasing trend towards the use of deep learning methods for gait data in the literature in recent years [17]. These studies are often subject to limitations such as very small datasets or simplified classification tasks with few classes. Therefore, uncertainties remain regarding the suitability of deep learning for complex multi-class gait classification tasks and how deep learning methods compare to traditional ML methods.

Lack of explainability. Explainability methods (see Section 1.4.3) have been successfully used to explain ML models in a variety of domains and their application in the medical field has also received considerable attention [33]. The motivation behind this is to increase transparency and thereby trust in ML models among medical professionals [34]. However, the use of explainability methods in the context of clinical gait analysis still needs to be explored. This is particularly interesting because most explainability methods have been developed for image data and structured data and evaluating explanations becomes particularly challenging when dealing with more abstract data such as multivariate time series. The suitability and usefulness of explainability methods for gait analysis and for clinical practice in general is currently an open question.

1.2 Aims of the Thesis

The primary aim of this thesis is to address and overcome the aforementioned gaps and limitations through the development and investigation of novel ML approaches. These approaches are intended for the automated analysis of measurement data obtained from clinical gait analysis with the overall aim of supporting clinical decision-making.

The present thesis focuses on developing and evaluating the performance of explainable ML and deep learning methods in modeling motion primitives at both kinematic (i.e., joint angle) and kinetic (i.e., GRF) levels while addressing complex classification tasks and larger datasets compared to the current state of the art. For the development and evaluation of these methods, two use cases with binary and multi-class classification tasks will be explored: i) the use case on functional gait disorders (**UC: functional gait disorders**), which includes GRF data from healthy controls and four classes with functional gait disorders related to the hip, knee, ankle, or calcaneus, and ii) the use case on cerebral palsy (**UC: cerebral palsy**) that utilizes a dataset containing GRF and joint angle data from patients with cerebral palsy with four distinct pathological gait patterns. The **UC: cerebral palsy** aims to enable a quantitative assessment of the discriminative power of both GRF and joint angle data for classifying multiple pathological gait patterns.

Overall, this thesis proposes a set of methodologies (published in high-ranked peer-reviewed journals) designed and implemented to achieve the following research goals:

- **Goal 1 – Creation of high-quality dataset**: Creation and publication of a high-quality (from a biomechanical point of view) gait dataset that contains clinically relevant annotations and GRF data and is comprehensive concerning the quantity of participants and number of trials per participant, as well as the diversity of pathological gait patterns.
- **Goal 2 – Evaluation of discriminative power of 3DGA modalities**: Evaluation of the discriminative power of different 3DGA data modalities, i.e., GRF and joint angle data, for automated gait classification.
- **Goal 3 – Evaluation of data handling strategies**: Evaluation of the impact of various data handling strategies (including feature scaling, feature extraction, data imbalance, and different aggregation strategies) on the performance of automated gait classification.
- **Goal 4 – Comparison of traditional ML and deep learning**: Development and comparison of traditional ML models and deep neural networks in terms of the classification performance.
- **Goal 5 – Evaluation of explainability approaches**: Development and evaluation of explainability approaches for traditional ML models and deep neural networks and assessing their ability to utilize clinically relevant input features for gait classification.

In the context of the above-mentioned challenges and goals, the main research questions (RQs) addressed in the present thesis are the following.

Research question related to **Goal 1**:

- **RQ1.1: Which steps should a preprocessing pipeline for GRF data include to facilitate the collaborative use of data gathered from different gait laboratories?**

  A common challenge in using ML for gait analysis is the limited availability of large datasets. Typically, ML models are trained and evaluated on small datasets from a single gait laboratory. The absence of comprehensive benchmark datasets makes it challenging to provide clear guidance on appropriate data preprocessing and classification methods for specific classification tasks. Regarding data preprocessing in the field of gait analysis, it is important to introduce and assess domain-specific as well as ML-related preprocessing procedures. These procedures, including data thresholding, data filtering, outlier detection, and data normalization, have been evaluated on a large real-world dataset. The outcome is a standardized preprocessing pipeline for GRF data that can be applied across various gait laboratory settings, enabling the collaborative use of these data from different research laboratories.

Research questions related to **Goal 2**:

- **RQ2.1: What level of classification performance can be achieved using only GRF data for automated gait classification?**

  In the related literature, GRF data are commonly utilized for binary classification tasks to distinguish between physiological and pathological gait. Multi-class classification tasks using GRF data are less common and are usually employed to identify patient groups exhibiting large differences in their gait patterns. To evaluate the discriminative power of GRF data, two complex multi-class datasets were employed. For the **UC: functional gait disorders** the classification performance was evaluated in a binary classification task by merging all pathological classes and distinguishing them from physiological gait. Subsequently, the classification performance was evaluated on the more complex multi-class classification task (as originally defined in the GaitRec dataset). The **UC: cerebral palsy** examined the discriminative power of GRF data for different gait patterns within the cerebral palsy population. The outcome is a quantitative comparison of classification performance, based exclusively on the use of GRF data, for the two given use cases.

- **RQ2.2: What is the advantage in classification performance when using kinematic data compared to GRF data for automated gait classification, and is there an improved classification performance when using both inputs together as opposed to using them separately?**

  Recording only GRF data is a more time- and resource-efficient approach compared to 3DGA. However, the exclusive use of GRF data represents a significant limitation for understanding the biomechanical processes of the human body. The amount of relevant information is significantly limited compared to the data obtained from 3DGA, as the exclusive use of GRF data excludes the explicit representation of gait kinematics. This drawback emphasizes the importance of incorporating 3DGA data, in particular for multi-class classification tasks. The experiments to assess the classification performance of the different data modalities were carried out on the complex multi-class classification task within the **UC: cerebral palsy** (see Section 1.4.1). The outcome is a quantitative comparison of classification performance, evaluating the effectiveness of each individual data modality separately and in combination.

- **RQ2.3: To what degree do the signals from the affected and unaffected sides differ in terms of their discriminative power for automated gait classification?**

  Conditions affecting the musculoskeletal locomotor system or neurological disorders have consequences not only for the (more) affected leg but also for the unaffected (less affected) leg. Individuals experiencing these conditions tend to develop compensatory strategies in the unaffected side, primarily influenced by the increased use of this side [35]. Leveraging these additional compensatory strategies encoded in the data from the unaffected side could potentially enhance classification performance. This research question was addressed by evaluating the classification performance on the classification tasks defined in the **UC: functional gait disorders**. The outcome is a quantitative comparison of classification performance, evaluating the discriminative power of data from the affected and unaffected sides, both separately and in combination.

Research questions related to **Goal 3**:

- **RQ3.1: To what extent do different feature scaling and feature extraction techniques impact the performance of automated gait classification?**

  In ML practice, it is well established that feature scaling and feature extraction techniques can greatly aid in the training process of ML models. Feature scaling is a necessary step before applying ML models to ensure uniform numerical ranges across different input features and signals. Thus, feature scaling prevents that signals with larger numeric ranges (amplitude) dominate those with smaller dynamic ranges. The primary focus in this thesis is on different feature scaling techniques (e.g., min-max normalization and z-standardization). Furthermore, various methods for feature extraction were evaluated, aiming to obtain diverse data parameterizations. The investigated parameterizations encompass handcrafted domain-specific parameters as well as PCA-derived representations of the raw data and handcrafted parameters. The assessment of these feature scaling variants and representations involved evaluating the classification performance on the tasks defined in the **UC: functional gait disorders**. The outcome is a quantitative comparison of classification performance, evaluating the different feature scaling and feature extraction approaches.

- **RQ3.2: What is the impact of data imbalance on the performance of automated gait classification?**

  One of the most significant factors influencing the classification performance of ML models is class imbalance [36]. Imbalanced data refers to a scenario with an unequal distribution of samples among different classes. Imbalanced data can result in the training of biased ML models, which in turn can lead to lower classification performance especially for the minority classes. Real-world datasets in the field of human gait analysis exhibit various imbalances. Certain pathological classes are inherently much rarer than others. Furthermore, in some conditions, such as those involving long therapy processes, subjects may have significantly more sessions recorded than in other cases (in which only a low number of sessions are recorded). In a single session, the number of recorded trials can also vary, influenced by factors such as the patient's condition. In this thesis, two causes of imbalance, i.e., variations in the number of patients and sessions per patient, were investigated both individually and in combination in the **UC: functional gait disorders**. To this end, the classification performance was evaluated on subsets that are balanced with respect to these two causes of imbalance. The outcome is a quantitative comparison of achievable classification performance in balanced and imbalanced scenarios.

- **RQ3.3: To what extent do different data aggregation methods impact the performance of automated gait classification?**

  In clinical practice, multiple trials are often recorded during a recording session. Clinicians usually analyze these trials by averaging them to achieve more robust representations. With multiple trials per recording session in the datasets, the question arises whether these trials can be combined or aggregated to enhance prediction robustness of ML models. The baseline approach involved using all available trials from a session without aggregation to train ML models. Furthermore, different early fusion techniques were investigated, such as aggregating (i.e., averaging) and subselecting (i.e., using the median or the most representative trial) trials from a session prior to training the ML model. Additionally, a late fusion strategy was evaluated, which aggregated the predictions of the ML model trained on all trials (i.e., baseline approach) using majority voting. The evaluation of these aggregation methods involved assessing the classification performance on the classification tasks within the **UC: functional gait disorders**. The outcome is a quantitative comparison of classification performance for the three early fusion approaches, the late fusion approach, and the baseline approach.

Research question related to **Goal 4**:

- **RQ4.1: How do traditional ML models compare to deep neural networks for the automated gait classification in terms of performance?**

  Deep learning and traditional ML methods differ in their learning paradigms, as elaborated in Section 1.1.4 and Section 1.4.2. To assess the classification performance of these methods, systematic comparisons using the datasets of the two use cases were performed. Within the **UC: functional gait disorders**, the classification performance of convolutional neural networks (CNNs), multi-layer perceptrons (MLPs), and support vector machines (SVMs) was evaluated across six tasks (comprising four binary and two multi-class tasks). In the **UC: cerebral palsy**, the performance of convolutional neural networks, self-normalizing neural networks, random forests, decision trees, support vector machines, and gradient boosting classifiers was evaluated. The outcome is a quantitative comparison of classification performance that should provide information regarding the strengths and limitations of the ML methods when utilized in specific gait classification tasks.

Research questions related to **Goal 5**:

- **RQ5.1: To what extent can explainability approaches be employed to determine the input features on which ML models base their decisions for automated gait classification, and are these relevant input features statistically justified and in line with clinical assessment?**

  To develop decision-support systems for clinical practice using ML, it is essential to integrate explainability approaches, which can be implemented at following levels (see Section1.4.3): i) at the data level (i.e., data exploration), ii) at the decision level (i.e., explanation of a specific prediction), and iii) at the model level (i.e., explanation of class-specific and model-specific patterns and learning strategies). Different explainability approaches were implemented and investigated explanations on the three levels using the datasets of the two use cases. The evaluation of explainability at the data level was performed through the use of linear discriminant analysis (LDA). The evaluation of explainability at decision and model level was conducted using two state-of-the-art methods, i.e., layer-wise relevance propagation (LRP) [37] and gradient-weighted class activation mapping (Grad-CAM) [38]. For the **UC: functional gait disorders**, LRP [37] was utilized to explain convolutional neural networks, multi-layer perceptrons, and support vector machines across the binary tasks. For the **UC: cerebral palsy**, Grad-CAM [38] was employed to generate explanations for convolutional neural networks and self-normalizing neural networks. Additionally, for random forests and decision trees, the feature importance based on Gini impurity served as model explanation. The outcomes are i) a quantitative evaluation from a statistical perspective using statistical parametric mapping (SPM) [39] to assess whether relevant input features exhibit also statistical differences between the classes, and ii) a qualitative examination of the explainability results and the differences in these results among the different ML methods, conducted via a series of focus group interviews with clinical experts.

- **RQ5.2: How effective are explainability approaches in detecting bias in ML models used for automated gait classification?**

  In practice, ML approaches are prone to biases. These biases are often present in the training data and originate from factors such as imbalanced data distributions, differences in walking speeds among different populations (e.g., physiological vs. pathological classes), or inadequate data preprocessing (e.g., unequal data scaling). An explainability approach was utilized to identify biases in ML models caused by the absence of feature scaling and variations in walking speed between healthy controls and patients. The outcome is a qualitative evaluation of the explainability results via focus group interviews with clinical experts to identify specific biases, followed by experiments designed to address and mitigate the underlying causes of these biases.

1.3 Synopsis and Publications

In the following, the reader will find a brief summary of the publications that contribute to this thesis. The emphasis lies on five journal publications encompassing methodological advancements beyond the respective state of the art. Table 1.1 illustrates the relationship between the publications and the research goals and questions defined in Section 1.2. Table 1.2 states the personal contributions (indicated with ✓) for each paper according to the contributor roles taxonomy (CRedIT) [1].

Publications	**Goal 1**	**Goal 2**			**Goal 3**			**Goal 4**	**Goal 5**	
	RQ1.1	RQ2.1	RQ2.2	RQ2.3	RQ3.1	RQ3.2	RQ3.3	RQ4.1	RQ5.1	RQ5.2
Horsak & Slijepcevic et al. (2020)*	✓									
Slijepcevic et al. (2017)		✓			✓	✓			✓	
Slijepcevic et al. (2020)		✓		✓			✓			
Slijepcevic & Horst et al. (2022)*		✓		✓				✓	✓	✓
Slijepcevic et al. (2023)		✓	✓					✓	✓	

Table 1.1: Correspondence (indicated with ✓) between the publications comprising the present thesis and the research goals and question (RQ) addressed within the thesis. The asterisk (*) indicates publications with co-shared first authorship.

The research question **RQ1.1** related to **Goal 1** is addressed in Horsak et al. (2020), a publication that was released along a real-world dataset containing clinical gait data. This dataset forms also the fundamental basis for three of the other publications.

Several RQs have been explored in multiple publications, with **RQ2.1** being the most frequently addressed and covered in all of the publications. In Slijepcevic et al. (2017) [40], Slijepcevic et al. (2020) [41], Slijepcevic et al. (2022) [42], and Slijepcevic et al. (2023) [43], we utilized GRF data from the **UC: functional gait disorders** and examined various classification tasks, offering a comprehensive evaluation related to **RQ2.1**. **RQ2.2** was addressed in Slijepcevic et al. (2023) [43] due to the availability of both GRF and joint angle data in the **UC: cerebral palsy**. The assessment of the discriminative power between the affected and unaffected side (**RQ2.3**) was conducted in Slijepcevic et al. (2020) [41] and Slijepcevic et al. (2022) [42].

Aspects related to **RQ3.1**, such as exploring different feature extraction methods (i.e., handcrafted domain-specific parameters as well as PCA-derived representations of the data), and investigating the influence of feature scaling techniques on classification performance were examined in Slijepcevic et al. (2017) [40]. In Slijepcevic et al. (2017) [40], we examined also the impact of data imbalance (**RQ3.2**), which guided our approach to utilize balanced datasets in subsequent publications. Slijepcevic et al. (2020) [41] evaluated the influence of different data aggregation approaches, i.e., early and late fusion strategies, on the classification performance in scenarios with multiple trials per person (**RQ3.3**).

The research question **RQ4.1** related to **Goal 4** was predominantly addressed in Slijepcevic et al. (2022) [42] and Slijepcevic et al. (2023) [43]. These two publications explored the comparison between traditional ML models and deep neural networks concerning classification performance.

Slijepcevic et al. (2017) [40], Slijepcevic et al. (2022) [42], and Slijepcevic et al. (2023) [43], examined **RQ5.1** from different perspectives. In Slijepcevic et al. (2017) [40], explainability was investigated on the data level by utilizing linear discriminant analysis. This approach allowed the assessment of the discriminative power of handcrafted domain-specific features and PCA representations. In Slijepcevic et al. (2022) [42] and Slijepcevic et al. (2023) [43], various explainability approaches were proposed to obtain explanations at the prediction, class, and model level. Finally, the investigation of how explainability methods enable the identification of bias related to walking speed differences and data scaling (**RQ5.2**) was addressed in Slijepcevic et al. (2022) [42].

Publications	Conceptualization	Data Curation	Formal Analysis	Funding Acquisition	Investigation	Methodology	Project Administration	Resources	Software	Supervision	Validation	Visualization	Writing – Original Draft	Writing – Review & Editing
Horsak & Slijepcevic et al. (2020)*	✓	✓				✓		✓	✓		✓	✓		✓
Slijepcevic et al. (2017)	✓	✓	✓		✓	✓		✓	✓		✓	✓	✓	✓
Slijepcevic et al. (2020)	✓	✓	✓		✓	✓		✓	✓		✓	✓	✓	✓
Slijepcevic & Horst et al. (2022)*	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓
Slijepcevic et al. (2023)	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓

Table 1.2: Correspondence (indicated with ✓) between the publications comprising the present thesis and the personal contributions based on the contributor roles taxonomy (CRedIT) [1]. The asterisk (*) indicates publications with co-shared first authorship.

The following subsections contain a brief summary of each publication included in the present thesis. For more detailed information, please refer to the corresponding publication in Chapter 2.

1.3.1 GaitRec, a Large-Scale Ground Reaction Force Dataset of Healthy and Impaired Gait

Brian Horsak, Djordje Slijepcevic, Anna-Maria Raberger, Caterine Schwab, Marianne Worisch, and Matthias Zeppelzauer. **GaitRec, a Large-Scale Ground Reaction Force Dataset of Healthy and Impaired Gait**. *Scientific Data*, 7(1):143, 2020. DOI: 10.1038/s41597-020-0481-z

The GaitRec dataset is derived from a clinical gait database maintained by an Austrian rehabilitation center. The dataset contains anonymized GRF measurements from 2,085 patients with various musculoskeletal impairments and data from 211 healthy controls, along with accompanying metadata such as age, sex, footwear, and walking speed. The dataset covers the entire rehabilitation progress of a patient during the patient's stay. The labels provided in the dataset indicate the anatomical joint level of orthopedic impairment, i.e., hip, knee, ankle, and calcaneus. During data collection, patients and healthy controls were asked to walk unassisted at a self-selected walking speed on a walkway equipped with two centrally embedded force plates that recorded bilateral GRF data. The dataset contains multiple left and right foot contacts of one person from one session. In addition to the unprocessed GRF data, the dataset also contains preprocessed data that are ready for immediate use.

We developed and published a preprocessing pipeline (including data filtering and thresholding) with the aim of standardizing GRF datasets from various gait laboratories. This pipeline facilitates the consolidation of GRF data from diverse laboratories, allowing for their collaborative utilization.

1.3.2 Automatic Classification of Functional Gait Disorders

Djordje Slijepcevic, Matthias Zeppelzauer, Anna-Maria Gorgas, Caterine Schwab, Michael Schüller, Arnold Baca, Christian Breiteneder, and Brian Horsak. **Automatic Classification of Functional Gait Disorders**. *IEEE Journal of Biomedical and Health Informatics*, 22(5):1653–1661, 2017. DOI: 10.1109/JBHI.2017.2785682

This publication presents a comprehensive investigation on the **UC: functional gait disorders** (i.e., automated classification of functional gait disorders using GRF data). The main objective of the study was to assess the effectiveness of i) handcrafted domain-specific GRF parameters and ii) PCA-based representations of GRF data for distinguishing functional gait disorders, as well as to establish a performance baseline for the automated classification of functional gait disorders using a large-scale dataset. This study was one of the first to examine such a comprehensive dataset of domain-specific gait features for automated gait classification.

We utilized a subset of the GaitRec datasets that included measurements from 279 patients with gait disorders and data from 161 healthy controls and resulted in a total of 9,496 gait measurements. The study included two classification experiments: i) a binary task that distinguishes between healthy and impaired gait (healthy controls vs. gait disorder patients) and ii) multi-class classification between healthy gait and all four gait disorder classes. Various data parameterization methods were examined, including domain-specific handcrafted GRF parameters, PCA-based representations, and a combined representation using PCA on these handcrafted GRF parameters. Boxplots were generated for each parameter and each class, allowing for an initial assessment of both intra- and inter-class variation. These boxplots offered valuable insights into the potential of the parameters to distinguish between the different classes. A more comprehensive assessment of the discriminative power of each parameterization was conducted using linear discriminant analysis.

The results of the experiments showed promising outcomes, but also highlighted the impact of factors such as data imbalance (i.e., differences in class sizes and varying numbers of measurements per patient) and feature scaling (i.e., min-max normalization and z-standardization) on the classification performance. The overall results showed that: i) for the multi-class classification task, the accuracy was 54.3%, and for the binary classification task, it was 90.8%; ii) when considering balanced data with an equal number of persons and sessions, the accuracies were 59.2% for multi-class and 85.4% for binary classification (it should be noted that the accuracy was significantly higher than the random baseline in this case compared to the unbalanced setting); iii) the linear support vector machine outperformed the radial basis function kernel in terms of classification performance; and iv) the application of PCA-based parameterization of the raw GRF data yielded better results compared to using handcrafted domain-specific GRF features, with a difference of 7.5% in classification accuracy.

1.3.3 Input Representations and Classification Strategies for Automated Human Gait Analysis

Djordje Slijepcevic, Matthias Zeppelzauer, Caterine Schwab, Anna-Maria Raberger, Christian Breiteneder, and Brian Horsak. **Input Representations and Classification Strategies for Automated Human Gait Analysis**. *Gait & Posture*, 76:198–203, 2020. DOI: 10.1016/j.gaitpost.2019.10.021

In this study, we compared two data aggregation methods, i.e., early fusion and late fusion, within the **UC: functional gait disorders**. In the gait classification literature, prior approaches employed either an early fusion method, which involved averaging multiple recorded trials of a subject into a single waveform, or a classification approach without data aggregation was performed, in which all available trials were used to train the ML models. In addition to these two methods, we further investigated an early fusion approach which determined the most representative trial based on a statistical method. Additionally, we explored a late fusion approach where the model was trained on all trials, but during inference, a majority voting scheme was used to combine the decisions from individual trials.

Subsequently, we explored the optimal input representations and combinations thereof for automated gait classification. This involved various options, such as using raw gait waveforms, relative changes within these waveforms, or signal differences between the affected and unaffected side.

We utilized a subset of the GaitRec dataset, which included measurements from 728 patients with gait disorders and data from 182 healthy controls. The dataset was balanced in terms of the number of persons per class, recorded sessions per person, and trials per person. The multi-class classification task focused on distinguishing between healthy controls and each gait disorder class associated with the hip, knee, ankle, and calcaneus. In line with the results from Slijepcevic et al. (2017) [40], we employed an ML pipeline that involved PCA, z-standardization, and support vector machines as classifier.

The results demonstrated the advantage of aggregating multiple trials from a single subject, especially when using late fusion or the mean waveform approach. In addition, the results suggested that the inclusion of both the original signals and their derived representations increased the informativeness of the data in feature extraction and classification. Even when certain input signals or representations contain redundancies, the combination of these signals, such as the GRF and center of pressure components with derived representations, improved classification performance in this study. Thus, the main finding from these experiments is that using a larger number of input signals and representations, even when redundancies exist, can lead to better results. This observation is especially true when combining GRF and center of pressure data and using derivatives from both the affected and unaffected sides. In addition, the inclusion of both the affected and unaffected side, whether explicitly or implicitly, seems to be beneficial.

1.3.4 Explaining Machine Learning Models for Clinical Gait Analysis

Djordje Slijepcevic, Fabian Horst, Sebastian Lapuschkin, Brian Horsak, Anna-Maria Raberger, Andreas Kranzl, Wojciech Samek, Christian Breiteneder, Wolfgang Immanuel Schöllhorn, and Matthias Zeppelzauer. **Explaining Machine Learning Models for Clinical Gait Analysis**. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(2):1–27, 2021. DOI: 10.1145/3474121

This publication investigated explainability methods to enhance transparency in automated gait classification within the **UC: functional gait disorders**. The main goal was to investigate and explain the class-specific characteristics learned by ML models from these data. To this end, various classification models, i.e., convolutional neural networks, multi-layer perceptrons, and support vector machines, were trained for different gait classification tasks, and prediction explanations were derived using a popular explainability method for the image domain, i.e., layer-wise relevance propagation (LRP). In addition, we proposed also two types of model explanations using the individual prediction explanations: The initial approach involved averaging relevance scores across all samples within a specific class. However, to conduct a more comprehensive analysis capable of identifying different learning strategies employed by the ML models, we adapted spectral relevance analysis (SpRAy) [44] for GRF data. This approach clustered the relevance scores obtained from various samples and classes and allowed to conduct a detailed examination of the resulting clusters and subclusters.

The evaluation of the obtained explanations followed a two-step approach. First, a statistical analysis was conducted using statistical parametric mapping (SPM) [39] to assess whether relevant input features exhibit also statistical differences between the classes. Second, two clinical experts interpreted the explainability results from a clinical perspective to assess whether the explanations align with clinical practice. Additionally, the investigation explored various aspects that could influence classification performance and explainability. These aspects included the impact of different classification methods, feature scaling techniques, and the role of various input signal components (i.e., horizontal forces and measurements of the affected and unaffected side).

The study utilized a subset of the GaitRec dataset, comprising GRF measurements during barefoot walking from 132 patients with lower-body gait disorders and data from 62 healthy controls with varying physical composition and gender. The dataset comprised three classes of orthopedic gait disorders related to the hip, knee, and ankle, in addition to a class representing healthy controls.

The results emphasize that ML models used in various clinical gait classification tasks base their predictions mostly on meaningful features from GRF data. These features have been validated through statistical and clinical evaluation. Within the scope of the analysis, several significant observations were made. First, highly relevant regions were identified in both the affected and unaffected sides, suggesting that the unaffected side contains complementary information that is relevant for the classification. Second, statistical parametric mapping proved to be a suitable statistical reference for the explainability results. Regions identified as highly relevant by the explainability method were generally found to be significantly different according to statistical parametric mapping and aligned with clinical evaluation. Furthermore, our results showed that not only the vertical GRF force but also the other force components exhibit highly relevant regions. This observation is consistent with the existing literature on clinical gait analysis. The results suggest that ML models tend to learn an over-complete set of features that may contain redundant information. This finding potentially explains why certain changes, such as occluding certain force components and using different input normalization methods, had negligible influence on the classification performance. Furthermore, ML models for gait classification exhibited the capability to learn different strategies for individual persons and patient groups, reflecting the capability to adapt to different patterns in the data. Finally, the implementation of the proposed explainability approaches allowed clinical experts to identify a bias related to the walking speed in ML models and accurately assess their functionality. This aspect is crucial for clinicians, as it is the only way to strengthen their trust in the predictions generated by these models.

1.3.5 Explainable Machine Learning in Human Gait Analysis: A Study on Children With Cerebral Palsy

Djordje Slijepcevic, Matthias Zeppelzauer, Fabian Unglaube, Andreas Kranzl, Christian Breiteneder, and Brian Horsak. **Explainable Machine Learning in Human Gait Analysis: A Study on Children With Cerebral Palsy**. *IEEE Access*, 11:65906–65923, 2023. DOI: 10.1109/ACCESS.2023.3289986

The main objective of this publication was to explore the effectiveness of various ML methods for the **UC: cerebral palsy**. Similar to the work presented in Slijepcevic et al. (2022) [42], this research also aimed to develop explainability approaches to assess the clinical relevance of the features learned by these models. In our study, we conducted a comparison between various traditional ML methods, such as random forests, decision trees, and gradient boosting classifiers, and deep learning methods, including convolutional neural networks and self-normalizing neural networks. For decision trees and random forests, Gini impurity-based feature importance served as the basis for the model explanation. For the deep neural networks, individual prediction explanations determined via the gradient-weighted class activation mapping (Grad-CAM) [38] method, were aggregated at different levels to provide insights at the decision, class, and model levels.

The study investigated the discriminative power of two different data modalities recorded during 3DGA, i.e., joint angle and GRF data, for classifying gait patterns associated with cerebral palsy. We conducted experiments using a 3DGA dataset comprising 302 patients with cerebral palsy exhibiting four distinct gait patterns associated with this condition.

The results indicate that joint angle data (peak performance of 93.4%) significantly outperforms GRF data (peak performance of 47.2%) for this classification task. Moreover, traditional ML approaches like random forests and decision trees achieved better results and focused on clinically relevant regions more effectively than deep neural networks. The best configuration, utilizing sagittal knee and ankle angles with a random forest, achieved a classification accuracy of 93.4%. Deep neural networks employed both clinically relevant features but also additional features for their predictions. These additional features could offer novel insights into the data and raise new research questions. Overall, this publication highlights the significance of explainability in fostering understanding of ML models for clinical practice.

1.4 Methodology

This section outlines the methodology that was utilized to address the research questions and achieve the goals of this thesis. More detailed specification of the employed data, approaches, and methods can be found in the publications presented in Chapter 2.

For each RQ, a systematic approach is followed, beginning with the experimental design, followed by the selection of a suitable dataset or subset, then the ML pipelines are implemented and finally an evaluation of the results is conducted. Almost all RQs were evaluated quantitatively by evaluating standard performance metrics (e.g., classification accuracy, precision, recall, and F_1 score) of the ML methods either through a dedicated train/validation/test split or a k-fold cross-validation approach. In all four publications, when evaluating classification accuracy, a comparison was made with the zero-rule baseline (i.e., representing the theoretical accuracy obtained by assigning class labels based on selecting the most frequent class in the dataset). In the case of Slijepcevic et al. (2020) [41], the zero-rule baseline equals the random baseline due to the perfectly balanced nature of the utilized dataset. RQs associated with explainability were evaluated qualitatively in a series of focus group interviews with the clinical experts. In addition to the qualitative assessment, this thesis proposes an additional assessment of the explainability results based on statistical analysis of the underlying data.

1.4.1 Clinical Use Cases and Datasets

To develop methods which are capable of learning higher-level and non-linear features, certain prerequisites regarding the quality and size of the utilized datasets have to be fulfilled. Primarily, the dataset should contain informative gait data and clinically relevant annotations determined by clinical experts. Additionally, the dataset has to exhibit substantial variability in the metadata that is specific to the respective population. These metadata include anthropometric properties, such as body weight and height, as well as other factors that influence gait, such as age, sex, and walking speed. Furthermore, a subset of clinically relevant pathological classes has to be identified for the purpose of automated analysis and classification. The literature demonstrates a broad range of use cases, ranging from orthopedic issues related to post-joint replacement surgery, ligament ruptures, and osteoarthritis, to complex diseases that cause neuromuscular mobility impairments in the lower extremities, including e.g., cerebral palsy, Parkinson's disease, and Alzheimer's disease. The dataset should contain data from several hundred people, to allow for the modeling of inter-individual variability within a specific pathological group. Generally, clinical gait analysis involves the recording of several trials (i.e., steps) in order to account for intra-individual step variability. Thus, a representative sample size per subject has to be ensured [45]. The imbalance of a dataset represents an additional challenge. This imbalance can naturally occur when the number of patients with different pathologies varies significantly. In order to model also healthy gait, a comprehensive dataset should comprise not only data from pathological gait, but also data from healthy controls.

The proposed thesis investigates two clinical use cases employing two comprehensive real-world datasets: i) a dataset comprising GRF data from patients with different functional deficits associated with a patient's condition after joint replacement surgery, fractures, ligament ruptures, and osteoarthritis, and ii) a dataset containing joint angle data and GRF data obtained via 3DGA from patients with cerebral palsy.

For the **UC: functional gait disorders**, different subsets of the GaitRec dataset [46] served as the basis for addressing and exploring research questions **RQ1.1**, **RQ2.1**, **RQ2.3**, **RQ3.1**, **RQ3.2**, **RQ3.3**, **RQ4.1**, **RQ5.1**, and **RQ5.2**. This dataset is one of the largest publicly available collections of clinical gait data and it was published within the scope of this thesis. The GaitRec dataset comprises anonymized GRF and center of pressure data from an existing clinical gait database maintained by a rehabilitation center of the Austrian Workers' Compensation Board (Allgemeine Unfallversicherungsanstalt, AUVA). Kinematic data was not recorded during the gait analyses. The entire dataset comprises GRF measurements from 2,085 patients with gait disorders and data from 211 healthy controls, both of various physical composition and sex. Data were manually classified into four classes – hip, knee, ankle, and calcaneus – by a physical therapist, based on the available medical diagnosis of each patient. The individual pathological gait patterns are related to joint replacement surgery, fractures, ligament ruptures, and related disorders associated with the hip, knee, ankle, or calcaneus. Participants walked unassisted at a self-selected walking speed on an approximately 10 m long walkway with two force plates. In healthy controls, measurements were also conducted at different walking speeds, which differ from the habitual speed. Each participant performed one or several measurement sessions. In each session, at least eight valid recordings for two consecutive steps were performed, leading to a total of 75,732 bilateral individual measurements for the entire dataset. The preprocessed GRF data, which includes the vertical, anterior-posterior, and medio-lateral force components, along with the center of pressure data, were normalized to 100% stance phase and to the body weight.

The **UC: cerebral palsy** focuses on a dataset comprising anonymized 3DGA data from an existing database created and maintained by the Laboratory of Gait and Human Movement of the Orthopaedic Hospital Vienna-Speising (Austria). The dataset includes 3DGA measurements, which consist of simultaneously recorded kinematic and GRF data. The dataset comprises anonymized data from 302 patients with cerebral palsy. Furthermore, the dataset included anthropometric data, along with annotations of four pathological gait patterns associated with cerebral palsy: true equinus, jump gait, apparent equinus, and crouch gait. This dataset served as the basis for addressing and exploring research questions **RQ2.1**, **RQ2.2**, **RQ4.1**, and **RQ5.1**. The 3DGA was conducted on a 12 m walkway using a motion capture system consisting of a minimum of 14 infrared cameras and three force plates. Patients walked without a walking aid and at a self-selected walking speed until a minimum of five valid recordings had been collected. Kinematic data in terms of joint angles were computed using the raw marker trajectories. Additionally, the kinematic data were time-normalized to 100% of the corresponding gait cycle (or stance phase in the case of GRFs). Time normalization of the gait data to

100% of the gait cycle or stance phase ensures that the input data has a uniform length and therefore no additional padding is required. To obtain more robust data, average curves were computed for each joint angle (for the pelvis, hip, knee, and ankle) and GRF component by aggregating data from all gait cycles within one recording session.

1.4.2 Representation Learning and Classification Methods

The initial classification baseline was established on the **UC: functional gait disorders** by applying traditional ML methods, such as k-nearest neighbor (k-NN) classifier, multi-layer perceptron, and support vector machine. These methods were trained on handcrafted domain-specific features derived from raw gait signals (e.g., local minima and maxima of the waveforms, as well as spatio-temporal gait parameters such as cadence, walking speed, and step length), which clinicians commonly use in clinical practice. For this purpose, Slijepcevic et al. (2017) [40] examined a comprehensive set of handcrafted domain-specific features.

In the automated analysis of gait data, another state-of-the-art approach employs PCA as a feature extraction method on the raw data and combines it with traditional ML methods [16]. Despite providing a linear feature representation, PCA has shown great suitability for biomechanical gait data, resulting in higher classification performances in the literature compared to handcrafted features [16]. Slijepcevic et al. (2017) [40] assessed the suitability of using PCA and kernel PCA (with a polynomial kernel) [47] as a feature extraction method for the **UC: functional gait disorders**. According to **RQ3.1**, the suitability of various feature extraction techniques and the resulting data parameterizations for gait analysis data is investigated in Slijepcevic et al. (2017) [40].

The ML methods most frequently used in the literature are support vector machines with different kernel functions [48, 49, 50, 30, 51, 28].This served as motivation for employing support vector machines as either the main classifier [40, 41, 42] or as a baseline method [43] in the publications of this thesis. In addition, this thesis applied the following traditional ML methods to handcrafted gait features, PCA-based representations, or the raw gait data: k-nearest neighbor [40], multi-layer perceptron [40, 42], random forest [43], decision tree [43], and gradient boosting classifier [43].

In order to compensate for the limitations of existing representations for clinical gait data, higher-level and non-linear feature representations are investigated within the scope of this thesis. Deep learning methods inherently learn feature representations directly from the input data (i.e., feature learning) and do not demand specific feature engineering. Therefore, no domain-specific knowledge is needed to derive specific input features. The architecture of deep neural networks incorporates multiple stacked layers and enables the learning of higher-level hierarchically related features that are employed by the top-most (classification) layers to tackle complex tasks.

Recently, various deep learning approaches have been employed for the analysis of human gait data [52, 17]. Matsushita et al. [17] identified convolutional neural networks, recurrent neural networks, and auto-encoders as the most commonly employed deep learning methods within the existing literature for gait analysis. To this end, the present thesis investigates the suitability of convolutional neural networks [42, 43] and self-normalizing neural networks [43] for analyzing GRF and kinematic data. Additionally, within the scope of this thesis, bi-directional long short-term memory (LSTM) networks have been explored. However, despite their intended design to capturing temporal dependencies in time series data, this recurrent network architecture yielded poorer results when compared to other methods.

The increasing use of deep learning has raised questions concerning the suitability and efficiency of deep learning versus traditional ML methods for gait analysis (see **RQ4.1**). The comparison of the classification performance between traditional ML and deep learning methods in Slijepcevic et al. (2023) [43] addresses this question.

1.4.3 Explainability Approaches

The lack of transparency in complex ML models has led to significant progress in the development of explainability methods. These methods are specifically designed to provide explanations for automated predictions and model behavior, aiding clinical experts in understanding the patterns and rules behind specific predictions. The present thesis involved the development and investigation of explainability approaches to tackle the research questions **RQ4.1** and **RQ5.1**.

Explainability methods can be classified based on the type of explanation they offer. According to the taxonomy proposed by Arya et al. [53], these approaches can be categorized into three coarse types: i) data exploration approaches, ii) decision explanations (also known as local model explanations), and iii) global model explanations. These different types of explanations complement each other.

Data exploration approaches do not provide explanations for an ML model, but instead focus on the data that were used to train the model. These approaches aim to visualize and transform the data, enabling domain experts to uncover significant structures and patterns within the data with the final goal of generating novel insights from the data. In the context of this thesis, various data exploration methods were employed, with a focus on visualizing the various distributions in the data (Figure 1.2: *static* → *data* → *distributions*). Slijepcevic et al. (2017) [40] employed boxplots to evaluate each manually crafted gait parameter. The examination of boxplots for each parameter and class allowed an assessment of both intra-class and inter-class variability, providing insights into the parameters' ability to distinguish between different classes. Subsequently, linear discriminant analysis was applied to the individual parameters and their combinations, as well as to the higher-dimensional PCA-based representations. The purpose of this analysis was to quantify the discriminative power of the studied representations and assess their suitability for the classification task. Furthermore, we employed one-dimensional statistical parametric mapping [39], a method that allows for the statistical analysis of time series data, to identify statistically significant differences in clinical gait data among various patient groups [42].

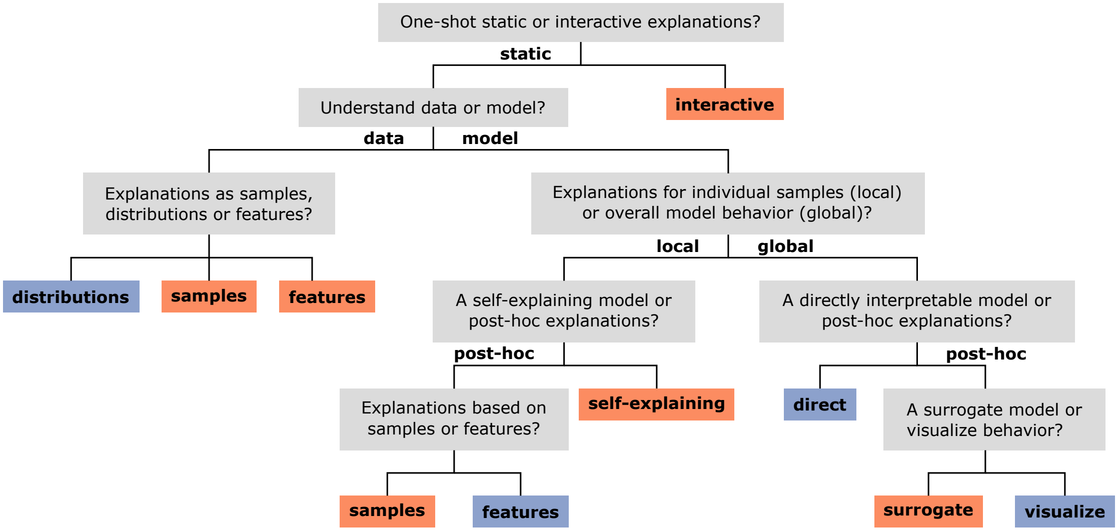

One-shot static or interactive explanations?
static
interactive
Understand data or model?
data
model
Explanations as samples, distributions or features?
distributions
samples
features
Explanations for individual samples (local) or overall model behavior (global)?
local
global
A self-explaining model or post-hoc explanations?
post-hoc
self-explaining
Explanations based on samples or features?
samples
features
A directly interpretable model or post-hoc explanations?
post-hoc
direct
A surrogate model or visualize behavior?
surrogate
visualize

Figure 1.2: A taxonomy introduced by Arya et al. [53] classifies explanations based on the following criteria: what is being explained (e.g., data or the model), the way in which the explanation is determined/provided (e.g., direct or post-hoc explanations; static or interactive explanations), and the level of explanation (whether it is local or global). The color of the leaves indicates whether the explainability approach has been implemented and evaluated within the context of this thesis. Blue leaves indicate the explainability approaches that have been implemented for clinical gait data. Adopted from [53].

Decision explanation methods explain the *local* behavior of ML models. Thus, such explanations can reveal the contributing regions of the input data responsible for the prediction of a particular data sample. Most decision explanation methods are *post-hoc* approaches that provide a certain flexibility as they can be directly applied to already trained ML models [53]. These methods typically produce saliency maps, which highlight the input features that are most relevant for a specific prediction [38]. When applied to gait data, these methods have the ability to detect distinctive regions in the input data that the ML model associates with a particular gait disorder [42]. Within the scope of this thesis, two decision explanation methods were applied, with a specific emphasis on post-hoc explanations of the features utilized by ML models (Figure 1.2: *static* → *model* → *local* → *post-hoc* → *features*). At first, we implemented layer-wise relevance propagation (LRP) [37] for clinical gait data [42]. This method propagates relevance scores from the output layer to the input layer throughout the entire network. The final relevance scores at the input layer can be mapped back to the original signals, thereby highlighting the input features that contributed to the prediction. For the final publication [43], we implemented gradient-weighted class activation mapping (Grad-CAM) [38] for clinical gait data. Grad-CAM is a method that provides explanations based on abstract features learned in the last convolutional layer. Unlike propagating gradients (or relevance scores) back to the input space, Grad-CAM propagates the gradients with respect to the class to be explained back to the last convolutional layer in a convolutional neural network. Subsequently, the activation map of the last convolutional layer is weighted with these gradients and then averaged over all channels of the layer. This results in an activation map that captures more abstract patterns used for the prediction. For the final decision explanation, the activation pattern is upscaled and mapped to the input signal. We recently developed gaitXplorer [54], a visual analytics approach for classifying gait patterns associated with cerebral palsy, which utilizes Grad-CAM to provide explanations for the predictions made by convolutional neural networks.

Both of the aforementioned methods are regarded as propagation-based approaches because they identify the impact of input features on the model's prediction by (partially) back-propagating either the gradient or relevance scores from the output to the input of the model. In this thesis, the focus has been on propagation-based methods instead of perturbation-based methods, mainly due to the computational efficiency of the former and the well-documented issues with reliability and consistency of the latter [55]. Additionally, perturbation-based methods are highly depended on the choice of hyperparameters, such as the number of perturbations.

Model explanation methods aim to explain which learning strategies and patterns a trained ML model has learned at a *global* level. Model explanations enable the assessment of whether an ML model has been trained correctly and whether the modeled classes rely on meaningful patterns. As a result, model explanations facilitate the identification of ambiguous features and biases that the model has learned, while also enabling the detection of overlaps in learning strategies between different classes.

In the context of this thesis, multiple model explanation approaches have been developed that rely on aggregating individual decision explanations (Figure 1.2: *static* → *model* → *global* → *post-hoc* → *visualize*). In Slijepcevic et al. (2022) [42], we averaged the individual decision explanations for each class, allowing to derive common patterns that ML models use to predict a specific class. Building upon this approach, we developed an explanation by incorporating the median, which proved to be more robust for Grad-CAM explanations compared to the mean [43]. Additionally, we introduced a visualization of individual decision explanations to allow visual evaluation of the distribution rather than relying only on the median/mean relevances. Furthermore, we adopted a model explanation approach based on SpRAy [44], which clusters individual decision explanations, enabling the identification of learning strategies for subgroups in the data utilized by the ML models [42]. The aforementioned approaches have been explored to explain sex- and age-dependent gait patterns utilized by ML models [56, 57].

For inherently explainable models like decision trees, we employed feature importance based on Gini impurity (Figure 1.2: *static* → *model* → *global* → *direct*) [43]. For more complex tree-based models like random forests, we adopted a similar approach.

In the area of clinical gait analysis, there has been a lack of usage of explainable methods to unveil the inner workings of black-box models and facilitate their application in clinical settings. Our efforts in this domain have played an important role in introducing and promoting explainability approaches specifically tailored for clinical gait analysis.

1.5 Results and Discussion

The five goals and the corresponding research questions of the thesis, outlined in Section 1.2, are utilized to present and discuss the obtained results.

1.5.1 Goal 1 – Creation of High-Quality Dataset

RQ1.1: Which steps should a preprocessing pipeline for GRF data include to facilitate the collaborative use of data gathered from different gait laboratories?

Despite the existence of publicly available gait datasets, access to fully annotated, comprehensive datasets with patient data remains quite limited. The survey conducted by Matsushita et al. [17] indicates that many of the publicly available gait datasets involve only a limited number of subjects, typically in the range of a few dozen. Exceptions to this trend are the dataset provided by Hausdorff via PhysioNet [58], which includes insole force data from 93 patients with Parkinson's disease and 73 healthy controls, as well as the dataset by Ferrari et al. [59], which contains kinematic data (i.e., marker trajectories) from 178 patients with cerebral palsy. The GaitRec dataset [46] is currently among the largest publicly accessible gait datasets, containing GRF and center of pressure data from 211 healthy controls and 2,085 patients with various musculoskeletal impairments. The dataset exhibits a remarkable degree of diversity due to several factors, including the number of subjects, multiple sessions, and trials within each session. This diversity also extends to different orthopedic conditions as well as the heterogeneous conditions under which the data were collected. To this end, each healthy control subject made walking trials at three different walking speed conditions (i.e., slow, self-selected, and fast), both with and without footwear and patients walked also under different conditions, such as barefoot, with orthopedic or normal shoes, and with or without orthopedic insoles.

In combination with the dataset, we introduced a universal preprocessing pipeline suitable for GRF data from different gait laboratories. Clinical experts were consulted to validate this preprocessing pipeline, which was designed to address **RQ1.1**. This pipeline includes several steps: i) ensuring uniform orientation of the medio-lateral and anterior-posterior signals (independent of the walking direction in the gait laboratory); ii) applying a threshold of 25 N to remove noise at the signal edges; iii) noise reduction using a second-order low-pass Butterworth filter with a cutoff frequency of 20 Hz; iv) time normalization to 100% stance; and v) normalization based on body weight. For the preprocessing of center of pressure signals, we applied a threshold of 80 N with respect to the vertical GRF component, aiming to reduce inaccuracies in calculation of the center of pressure during lower force values. Additionally, the medio-lateral and anterior-posterior center of pressure components were mean-centered and zero-centered, respectively. Furthermore, to ensure a high level of data quality, we applied an outlier detection algorithm proposed by Sangeux and Polak [60] to the data of a single session per individual.

In the course of this thesis, the same pipeline was also employed to preprocess and publish the Gutenberg Gait Database [61]. This dataset is one of the largest publicly available GRF dataset containing data from healthy controls. Additionally, the same pipeline was applied to the publicly available AIST Gait Database [62]. The combination of these three datasets opened up unprecedented data dimensions and led us to the exploration of research questions related to the uniqueness of gait data in person re-identification [63] and the identification of sex-related [64] and age-related [65] walking patterns. Furthermore, several research papers proposed ML approaches based on the GaitRec dataset [66, 67, 68, 69, 70]. Additionally, the dataset has been employed for transfer learning for an auditory feedback system based on GRF data [71, 72], and in the context of biomechanical analysis [73]. This demonstrates that the proposed standardization of GRF data from different gait laboratories already offers the possibility to investigate innovative aspects in the field of gait analysis.

1.5.2 Goal 2 – Evaluation of Discriminative Power of 3DGA Modalities

RQ2.1: What level of classification performance can be achieved using only GRF data for automated gait classification?

The results indicate that bilateral GRF data (especially when combining all three force components) can be effectively utilized for a binary task of distinguishing physiological gait from pathological gait. This was demonstrated using various subsets of the GaitRec dataset. In the task of distinguishing the healthy control class from a combined gait disorder class (i.e., encompassing all pathological patterns), we achieved peak classification accuracies of 89.5% (90.8% in combination with center of pressure data) [40] and 88.8%[42] using GRF data. Furthermore, in the tasks of distinguishing the healthy control class from individual pathological classes (e.g., healthy control class versus hip class), we also achieved high accuracies ranging from 86.5% to 88.8% [42]. In multi-class tasks, GRF data did not yield the desired results in either of the use cases. For the **UC: functional gait disorders** based on the GaitRec data, accuracies of 51.6% (54.3% in combination with center of pressure data) [40], and a maximum of 60% (62.0% in combination with center of pressure data) for a balanced setting [41] were achieved in the task with five classes. When the calcaneus class was removed and only the barefoot condition was selected, comparable results were achieved with an accuracy of 59.5% [42]. It is noteworthy that lower classification results with a peak accuracy of 51.8% were obtained when attempting to classify the hip, knee, and ankle classes [42]. This observation implies that GRF data may have limited capability in capturing distinct patterns among pathological classes in this specific context. This observation was also confirmed in experiments conducted within the **UC: cerebral palsy**. In this multi-class task consisting of four classes, a peak performance of 47.2% was achieved [43]. With respect to **RQ2.1**, it can be concluded that GRF data have the potential to classify physiological and pathological gait patterns (as a binary classification task). However, GRF data lack sufficient discriminative power for multi-class classification tasks.

RQ2.2: What is the advantage in classification performance when using kinematic data compared to GRF data for automated gait classification, and is there an improved classification performance when using both inputs together as opposed to using them separately?

The **UC: cerebral palsy** revealed that kinematic data are significantly more discriminative than GRF data. In the multi-class task involving four pathological gait patterns, the kinematic data achieved a peak performance of 93.4%, marking a substantial difference of 46.2% compared to GRF data [43]. From a clinical perspective, this outcome may not be surprising, as kinematic data alone often contain sufficient information for the analysis and diagnosis. However, it is important to assess the discriminative power to identify potential use cases where utilizing only GRF data might be sufficient. Furthermore, the classification results showed that combining kinematic and GRF data does not yield any advantage [43]. Moreover, for almost all classification methods the use of GRF data resulted in a slight decrease in performance. This observation suggests that GRF data do not offer complementary information compared to the kinematic data for the task at hand. However, it is important to consider the potential benefits of including both types of data for a more comprehensive analysis, even though the combination did not yield benefits in this use case. By integrating kinematic and GRF data, ML models could gain a more holistic picture of gait patterns.

RQ2.3: To what degree do the signals from the affected and unaffected sides differ in terms of their discriminative power for automated gait classification?

In relation to **RQ2.3**, several studies within the **UC: functional gait disorders** revealed that leveraging data from both the affected and unaffected side provides a slight advantage [74, 41]. Including both the affected and unaffected sides, either explicitly or implicitly through calculating the sample-wise difference between them, prove beneficial for certain input scenarios. The explainability results presented in Slijepcevic et al. (2022) [42] provide further support for this finding. In all classification tasks, relevant regions are evident not only in the GRF data of the affected side but also in the unaffected side, although to a slightly lesser degree. This observation suggests that the unaffected side contains complementary information for the classification task.

Furthermore, it is essential to note that using only the GRF data from the unaffected side resulted in significantly poorer classification results compared to utilizing only the GRF data from the affected side [74, 41]. This observation contradicts the findings of Williams et al. [75], who obtained higher classification performance for the less affected side in classifying six pathological gait patterns associated with traumatic brain injury.

1.5.3 Goal 3 – Evaluation of Data Handling Strategies

RQ3.1: To what extent do different feature scaling and feature extraction techniques impact the performance of automated gait classification?

With respect to **RQ3.1**, we explored various parameterizations for clinical gait data, including handcrafted domain-specific parameters, PCA-based representations of raw gait data, and a combined representation using PCA on GRF parameters [40]. The first parameterization involved 52 handcrafted parameters extracted from the GRF and center of pressure data. To address the significant variation in parameter value ranges, feature scaling was crucial. We evaluated both min-max normalization and z-standardization, with z-standardization showing slightly better results. The second parameterization relied on PCA of raw GRF data. PCA representations obtained from only the three force components performed better than the handcrafted GRF parameters. Incorporating PCA representations of the center of pressure further improved results for both tasks. Normalization of PCA-based representations proved to be vital, as performance significantly dropped without it. The third parameterization applied PCA on the normalized handcrafted GRF parameters. However, the results did not exhibit an improvement compared to using handcrafted GRF parameters without PCA. Based on these findings, PCA-representations of raw GRF data are recommended as input instead of relying only on handcrafted GRF parameters. These results are consistent with a study by Burdack et al. [76], where the highest performance was also obtained by employing PCA on raw GRF data along with support vector machines for the task of person re-identification in healthy controls.

When taking into account additional aspects such as the explainability of the utilized ML methods, it is advisable to employ raw input data. In the publications in which we focused on explainability [42, 43], we intentionally avoided using PCA, as it introduces an additional abstract feature space prior to the application of ML methods. Our main goal was to provide explanations at the input level, which is crucial because it is the domain where clinical experts analyze the data.

RQ3.2: What is the impact of data imbalance on the performance of automated gait classification?

In Slijepcevic et al. (2017) [40], three experiments were conducted for both tasks of **UC: functional gait disorders** to investigate the impact of imbalanced data on the classification results, specifically addressing **RQ3.2**. The classification results were compared to those of the unbalanced setting, which served as the baseline. *Balanced number of sessions:* The dataset was balanced by randomly selecting only one session per person (while the number of individuals per class remained unbalanced) to assess the effect of balanced numbers of recorded sessions per individual. *Balanced number of persons:* The dataset was balanced by randomly subselecting individuals per class to match the size of the smallest class (while including all sessions from these individuals) to examine the effect of balanced numbers of individuals per class. *Balanced number of persons and sessions:* A fully balanced dataset was created, containing only one session per person and equal numbers of persons per class (i.e., with respect to the smallest class), to explore the combined effect of balancing the number of individuals and sessions.

In all three experimental settings, and especially in the last case, balancing the dataset led to significant improvements in terms of the deviation from the random baseline compared to the results without balancing. These findings emphasize the importance of considering intra-patient variability and data imbalance when conducting automated analysis of clinical gait data. Moreover, these findings demonstrate that using balanced datasets, in terms of both the number of sessions per person and the number of persons per class, can lead to considerable improvements in classification performance (with respect to the random baseline). This observation served as motivation to predominantly use balanced datasets in subsequent publications.

RQ3.3: To what extent do different data aggregation methods impact the performance of automated gait classification?

To address **RQ3.3**, we explored the effectiveness of different aggregation methods for classifying gait analysis data. The baseline approach involved using all available trials from a session without aggregation. We explored various early fusion approaches, which involved aggregating or subselecting data samples from an individual before training the ML model. Additionally, we considered a late fusion approach that aggregated the predictions of the ML model, trained on all trials from an individual, using majority voting. The median waveform and most representative trial approaches failed to surpass the baseline performance. In contrast, among the early fusion approaches, the mean waveform method showed the most significant improvement. The late fusion approach demonstrated better results compared to early fusion methods, suggesting that introducing an abstraction layer to the classifier's outputs could enhance robustness.

1.5.4 Goal 4 – Comparison of Traditional ML and Deep Learning

RQ4.1: How do traditional ML models compare to deep neural networks for the automated gait classification in terms of performance?

Regarding **RQ4.1**, we observed somewhat diverse outcomes when comparing deep learning and traditional ML methods in the two examined use cases. In the **UC: functional gait disorder**, convolutional neural networks, support vector machines, and multi-layer perceptrons were examined across six classification tasks [42]. The classification results revealed that there were no significant performance differences among the ML methods.

In the **UC: cerebral palsy**, we investigated the performance and explainability of various ML models, including convolutional neural networks, self-normalizing neural networks, random forests, and decision trees. For performance comparison, support vector machines and gradient boosting classifiers served as baseline models. The results revealed that random forests outperformed all other ML methods achieving consistent results across diverse input scenarios with kinematic data (peak performance of 93.4%). Gradient boosting exhibited slightly lower performance (peak performance of 92.0%), while decision trees ranked third in most input scenarios (peak performance of 89.7%). Both convolutional neural networks and self-normalizing neural networks achieved peak performances of 86.6% and 85.7%, respectively, which were slightly inferior to the performance of decision trees. Surprisingly, support vector machines achieved the lowest overall performance, reaching only a peak performance of 78.8%. The higher performance of tree-based ML models can be attributed to their robust generalization ability with limited training data, a characteristic not shared by convolutional neural networks and self-normalizing neural networks. Deep learning methods tend to overfit when trained on smaller datasets, which may have contributed to their comparatively lower performance.

1.5.5 Goal 5 – Evaluation of Explainability Approaches

RQ5.1: To what extent can explainability approaches be employed to determine the input features on which ML models base their decisions for automated gait classification, and are these relevant input features statistically justified and in line with clinical assessment?

The evaluation of **RQ5.1** was conducted on different levels in three publications. The study presented in Slijepcevic et al. (2017) [40] evaluated explainability on the data level by employing linear discriminant analysis to assess the discriminative power of handcrafted domain-specific features and PCA representations. The publications focusing on explainability [42, 43] proposed various explainability approaches to provide explanations at the prediction, class, and model level.

Utilizing linear discriminant analysis and the visual assessment of boxplots of the handcrafted domain-specific features revealed that discrete parameters identified at the local minima and maxima within the GRF signals, as well as spatio-temporal parameters, showed the highest discriminative properties. These results are in line with clinical research, as this subset of domain-specific features is frequently employed to evaluate the progress of therapy in clinical practice [77].

For the **UC: functional gait disorders**, the model explanations for the three investigated ML methods, i.e., convolutional neural networks, multi-layer perceptrons, support vector machines, exhibited a high degree of overlap, particularly regarding the location of relevant regions in the input data [42]. In certain signal regions, there were only slight differences in the amplitude of relevance scores. Furthermore, in Slijepcevic et al. (2022) [42], we proposed the use of statistical parametric mapping for the statistical assessment of input data. By employing this method, we were able to identify regions in the input data that exhibit significant statistical differences between the classes. This analysis played a crucial role in evaluating the explainability results from a statistical perspective. The results demonstrate that in the majority of cases, statistical parametric mapping reveals statistically significant differences in regions that are highly relevant according to the explainability method. Furthermore, according to clinical experts, relevant regions are strongly linked to the existing clinical literature and are considered clinically plausible.

Similarly, for the **UC: cerebral palsy**, model explanations demonstrated the highest relevance in the two clinically most relevant signals, i.e., sagittal knee and ankle angles [43]. This observation aligns with clinical expectations and is consistent with findings from other studies, which have also identified these signals as the most promising for distinguishing crouch gait, apparent equinus, jump gait, and true equinus. The explainability results revealed that deep neural networks showed a tendency to learn patterns from a wide range of input signals, including clinically relevant regions but also on less relevant and potentially unrelated regions. In contrast, random forests and decision trees focused specifically on the clinically relevant regions. Interestingly, for the deep neural networks some of the relevant regions outside the sagittal knee and ankle angles were also considered clinically meaningful by clinicians, such as the sagittal hip angle. On the other hand, some regions were not considered clinically meaningful. These regions can be either attributed to a bias in the data or might not have been considered in clinical practice because they exhibit subtle differences that haven't been recognized as clinically relevant yet. These findings highlight the potential of explainability approaches not only to assist in evaluating the behavior of ML models but also to gain novel clinical insights into the underlying data.

RQ5.2: How effective are explainability approaches in detecting bias in ML models used for automated gait classification?

Two experiments presented in Slijepcevic et al. (2022) [42] investigated the suitability of explainability approaches for detecting biases related to differences in walking speed between healthy controls and patients and the absence of feature scaling in ML models.

During the evaluation of the explainability results in the **UC: functional gait disorders**, clinicians identified relevant regions in the unaffected side that they believed were not directly linked to the specific gait disorders. The clinicians hypothesized that these regions might be influenced by differences in walking speed between healthy controls and patients (rather than compensatory strategies of the unaffected side in patients' data), suggesting a potential bias in the trained ML model. We were able to confirm this hypothesis in an experiment using a subset of the data in which walking speed was not statistically significantly different between the two groups. We trained the same model architecture on this subset and observed that relevant regions remained consistent between the two models, except for the regions previously identified by the clinicians. In this case, the explainability results provided the necessary information that led to a deeper understanding of the ML model and the underlying data. This allowed clinicians to identify the bias related to differences in walking speeds between the healthy controls and patients with functional gait disorders.

To investigate the effect of feature scaling on ML models, we conducted experiments with and without min-max normalization of the input data for the **UC: cerebral palsy**. For the classification of non-normalized data, the most relevant input features were found in the vertical GRF component. The absence of relevant regions in the horizontal forces suggests that the ML models might not effectively utilize them, as a result of their small value range. On the other hand, explainability results for min-max normalized input data revealed highly relevant regions in the vertical and horizontal forces. The normalization process expanded the value range of the horizontal forces, allowing them to contribute at a level comparable to the vertical component. Despite the slightly better classification results achieved with non-normalized data for the multi-class tasks, the explainability results suggest that normalization is crucial for obtaining unbiased predictions. These results underscore the effectiveness of explainability approaches in identifying biases introduced by the absence of feature scaling techniques.

1.6 Limitations and Future Work

This section discusses limitations observed in the research related to this thesis and identifies future research directions that hold the potential to advance the field of automated classification of clinical gait data.

1.6.1 Performance Considerations

The results obtained in this thesis have demonstrated that the classification performance is highly dependent on the data modality and the classification task at hand. For example, when utilizing only GRF data, the multi-class classification yields relatively moderate results, with the highest accuracy reaching 62.0% in the **UC: functional gait disorders** [41] and 47.2% in the **UC: cerebral palsy** [43]. However, in the case of employing GRF data for a binary task, such as distinguishing between physiological and pathological gait, classification accuracy can reach up to 90.8% [40]. Compared to GRF data, joint angles offer a more detailed representation of the kinematics during walking. Consequently, the utilization of joint angles results in a significant enhancement of performance in the multi-class classification task for the **UC: cerebral palsy**, achieving a classification accuracy of 93.4% (i.e., a difference of 46.2%) [43]. In comparison to previous studies addressing the same classification task (i.e., classification of four gait patterns associated with cerebral palsy as defined by Rodda et al. [78]), our results achieved a similar level of classification performance. Reported performances in the literature ranged from 93.5% in the study by Zhang and Ma [79], which was based on a dataset comprising 200 samples, to 94.0% as reported by Darbandi et al. [80], utilizing a dataset of 60 samples. In comparison, our study comprises a much larger dataset consisting of 302 children and shows that this performance level can be achieved even for large-scale data [43].

Generally, the observed results leave room for improvement and may still not meet clinical requirements. However, the assessment of whether this level of performance is adequate and clinically suitable predominantly relies on comprehending the human baseline. *A promising direction for future research involves establishing a human baseline for a range of classification tasks in the field of human gait analysis.* To define this baseline performance, an analysis of evaluations and annotations from multiple clinical experts from different gait laboratories is essential. In the course of this thesis, we conducted an evaluation of a human baseline for the multi-class classification task within the **UC: functional gait disorders**. Interestingly, the performance of the human baseline was significantly lower than the ML performance. One possible reason for this outcome is that clinical experts had to assess only the GRF data, which deviated from their typical clinical practice, without access to any contextual information (e.g., observing the patients while walking). Moreover, it is important to acknowledge the presence of uncertainties and class overlaps in annotations, resulting in classification outcomes that may not be perfect.

1.6.2 Unexplored ML Methodologies

Multi-label learning. Through the interviews with the clinical experts conducted during the evaluation of the explainability results, we identified that the classes in the **UC: cerebral palsy** are not always mutually exclusive in practice. These classes can exhibit overlaps (e.g., different trends in the patterns of the knee and ankle) but clinical experts inherently assess which pattern is more pronounced and use this for the annotation. These uncertainties and class overlaps can introduce bias into the annotation process. *Within this context, a potential future research direction involves exploring the appropriateness of a multi-label classification approach for gait pattern classification.* With suitable ML approaches, dependencies between variables that are relevant to different classes can be modeled with greater accuracy and flexibility. A multi-label approach might be closer to the real-world setting and therefore more suitable for clinical practice.

Few-shot and zero-shot learning. A related topic is the challenge of modeling out-of-distribution samples, which include patterns that deviate from predefined categories, as well as handling "unseen" classes, referring to patterns or conditions not encountered during the training process. This is frequently the case in situations where data collection is limited, especially when encountering rare or unusual pathological gait patterns. *These challenges underscore the significance of exploring few-shot and zero-shot learning in future research.* Few-shot learning allows to model effectively even sparsely sampled classes, by extracting information from only a few training samples per class. Zero-shot learning, an extreme case of few-shot learning, represents a learning paradigm that enables the detection of classes that were not part of the initial training data at all. Few-shot and zero-shot learning approaches have been rarely investigated in the context of gait data analysis [81, 82, 83]. However, addressing the aforementioned challenges is crucial for developing more adaptable and clinically relevant decision-support systems.

Multi-modal learning. An aspect we realized while determining the aforementioned human baseline is that clinical experts utilize far more than just raw gait data during the assessment process of patients. Clinical experts utilize also contextual information, e.g., they can visually observe individuals walking and estimate anthropometric data. This implies that clinicians employ a multi-modal approach when assessing gait patterns. *A promising future direction is multi-modal learning for human gait data.* A preliminary step in this direction was taken in this thesis by combining joint angles and GRF data. The additional inclusion of GRF data did not impact the results in this case, but it may yield different outcomes in other classification tasks. *Promising modalities for modeling gait patterns might encompass not only joint angles and GRF data, but also subject-specific metadata (sex, age, walking speed, and anthropometric data), muscle activation determined via electromyography, data from inertial measurement units, and video recordings of patients.*

Physics-informed ML approaches. Nowadays, ML approaches typically learn gait representations in a completely data-driven way, with the consequence of neglecting the biomechanical context and constraints in the data modeling process. Data-driven approaches have demonstrated limitations in effectively capturing gait primitives with respect to biomechanical constraints. *A promising direction for future research is to incorporate kinematic and kinetic constraints directly in the ML process via physics-informed ML methods* (e.g., [84]). The loss function of these methods can be constrained to follow biomechanical principles. This enables a more accurate modeling of the underlying physics of biomechanical data and motion primitives. Consequently, the ML models could demonstrate greater generalizability.

1.6.3 Effects of Influencing Factors on Gait Data

Human gait data exhibit a high level of inter-subject [63] and intra-subject [85] variability. Furthermore, pathological and physiological gait patterns are strongly influenced by numerous interacting factors, including sex, age, body height, body weight, walking speed, and the use of footwear and prostheses. Hence, when evaluating a pathology using gait data during walking, it is crucial to consider that these data can be affected not only by the presence of an underlying pathology but also by the aforementioned influencing factors. Preliminary investigations of some of these influencing factors, i.e., sex [56] and age [57], were conducted within the scope of this thesis. Figure 1.3 illustrates the outcomes with respect to the influencing factor of sex. In the subfigures B) and C), the color coding represents relevance scores (obtained via layer-wise relevance propagation), highlighting relevant input feature for the distinction between male and female healthy controls. The results show a certain agreement of relevant features (according to the explainability method), the gait literature, and statistical assessment. However, there are also discrepancies among these three approaches. This motivates future research regarding sex differences on larger datasets. *Future research should conduct similar investigations to explore the effects of all types of influencing factors on human gait as well as their interactions.* These investigations can provide insights into the actual extent of these influences, opening up follow-up questions on how to incorporate these factors into modeling and how to make automated gait analysis robust to their effects.

1.6.4 Data Availability

The limited size of gait datasets could be the reason why deep learning methods fail to meet expectations in terms of outperforming traditional ML methods. Therefore, this thesis emphasizes the importance of considering heterogeneous and large-scale benchmark datasets to train and evaluate robust ML models and ensure their generalizability across different populations and gait laboratories. *Another important aspect of large-scale benchmark datasets would be the potential to train foundational models for human gait data (e.g., for various gait data modalities), which can then be adapted to specific gait use cases using a transfer learning approach* [92]. In addition to data quantity, benchmark datasets should be controlled for various influencing factors, such as age, sex, body height, body mass, and speed differences (i.e., to ensure they represent a wide range of population variability and are balanced), enabling unbiased training and evaluation of ML models. Moreover, benchmark datasets should incorporate also data obtained from

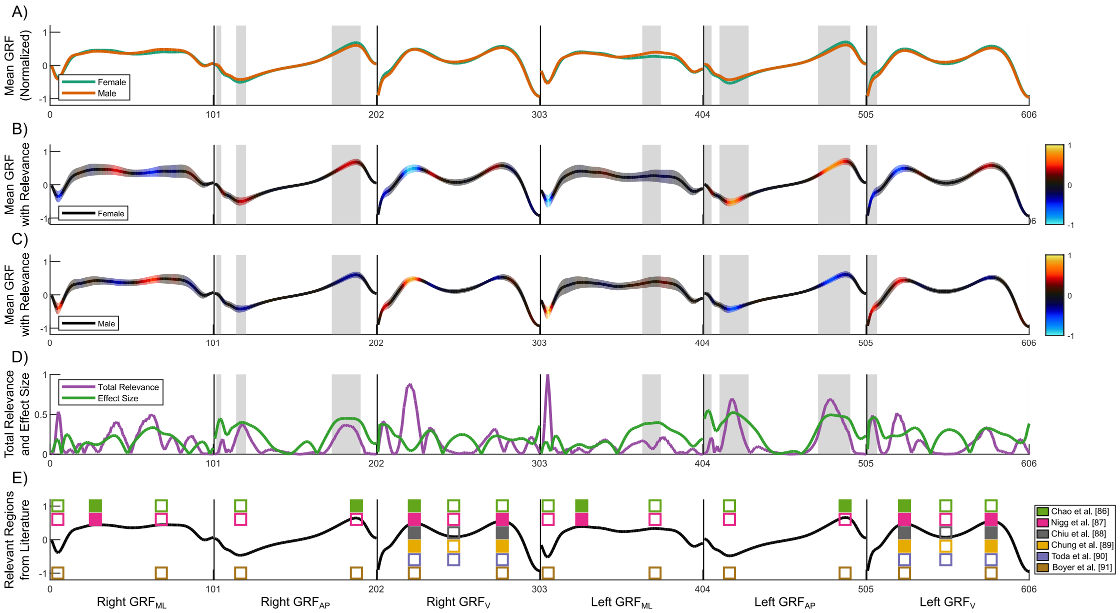


Figure 1.3: Explainability results for sex classification (adapted from [56]). A) Averaged GRF signals for both classes. The first three signals represent the three GRF components of the right side and are followed by the three GRF components of the left side. The shaded areas highlight the input features where statistical parametric mapping (two-sample t-test ($p < 0.05$)) indicated a statistically significant difference between both classes. B)–C) Averaged GRF signals for female/male class, with a band of one standard deviation, color-coded via relevance scores. D) Effect size obtained from statistical parametric mapping and total relevance (absolute sum of input relevance scores of both classes). The total relevance indicates the common relevance of the input signal for the classification task. E) Significant (filled boxes) and non-significant (empty boxes) handcrafted GRF parameters according to the literature [86, 87, 88, 89, 90, 91].

different walking surfaces (i.e., including indoor and outdoor environments), as well as various footwear conditions. This inclusion is important to enable future ML models to capture the inter- and intra-subject variability observed in real-world biomechanical data.

To reach this goal, collaboration with various research institutes and health care facilities is essential to gain access to a broader range of clinical gait data. In this regard, we have made an initial step by merging the GRF data from the GaitRec [46] dataset and the Gutenberg Gait Database [61] in a consistent and directly comparable data format. *In the future, it is crucial to expand such data sharing initiatives to encompass also further data modalities (e.g., joint angles, joint moments, and muscle activation).*

The imbalance within a dataset poses an additional challenge when working with real-world data. Inequalities in class cardinality may arise naturally, as certain pathologies may be less common than others, or healthy controls may be measured less frequently in gait laboratories than pathological cases. There are various strategies for dealing with data imbalances, which can be divided into two main groups, i.e., data-centered and algorithm-centered approaches [93]. In the present thesis, to address **RQ3.2**, the most commonly employed data-centered approach of data subsampling has been explored. Another frequently used data-centered approach involves upsampling the data using either new measurements, data augmentation techniques, or the generation of synthetic data. In the course of the present thesis, various data augmentation techniques for time series data, as presented in the survey by Iwana and Uchita [94] (i.e., jittering, magnitude warping, scaling, window slicing, window warping, guided warping, and time aligned averaging) have been experimented with for the data in **UC: cerebral palsy**. However, none of the augmentation approaches yielded improvements in the results for random forests and convolutional neural networks. Promising algorithm-centered approaches include the use of cost-sensitive methods, where upweighting is utilized for the samples of the minority class (i.e., assigning more weight to those samples in terms of cost). For gait analysis data, for example, Chia et al. [95] investigated the weighted Brier score as a cost function for the classification of musculoskeletal impairments in cerebral palsy, while Dumphart et al. [96] utilized the weighted cross-entropy loss to address the high imbalance in gait event classification. *Future work should investigate various data-centered and algorithm-centered approaches to address data imbalances in real-world gait datasets.*

In addition to the recorded data collected in laboratories, *synthetic* data can be employed to expand the volume of training data and compensate for data imbalances in the datasets. For this purpose, generative adversarial networks (GANs) [97] have been employed for other domains [98, 99]. In generative adversarial networks two models are trained simultaneously, a generative model that learns the distribution of the training data and generates synthetic data and a discriminative model which decides if a sample originates from the training data or was generated artificially. The use of generative methods could be particularly valuable, especially for imbalanced datasets and pathological gait patterns that are rare. Alternatively, auxiliary classifier generative adversarial networks (AC-GANs) [100] could be utilized to simultaneously generate synthetic data and model the classification task. For this purpose, the discriminator is trained not only to discriminate between real and artificial data, but also to classify the input data according to the task at hand. *Future research should explore whether generative methods are appropriate for human gait data and the available datasets and whether the synthetic data they generate can enhance the learning process of ML models.*

1.6.5 From Explainability to Trustworthiness

The present thesis in particular underscores the significance of explainability in automated gait classification. The proposed explainability approaches enable the identification and comparison of learning strategies across various classification methods. They effectively highlight the signal regions on which predictions of specific classes are grounded. However, approaches that provide saliency maps provide explanations of which features are relevant to a certain prediction and to what extent, but they fall short in exploring the underlying reasons for this relevance or the specific patterns and concepts involved. This circumstance sometimes complicates the interpretation of explainability results and consequently limits trustworthiness. Therefore, there is a need for the development of human-centered interactive explanation methods (see Figure 1.2: *interactive*) that would enable clinicians to manipulate the input data, create counterexamples, and observe the behavior of ML models in near real-time. Another approach that could be combined with interactive methods to further increase trust in ML models is the development of self-explaining (deep) learning methods (see Figure 1.2: *static* → *model* → *local* → *self-explaining*) that are inherently explainable by nature. Baumhauer et al. [101] introduced an appropriate explainability method for this purpose, known as bounded logit attention. This approach introduces a trainable explanation module that can be integrated into a deep neural network (typically a one- or two-dimensional convolutional neural network), whether it is pretrained or not. By training this module or the entire network, it serves as a feature extractor at the final convolutional layer, inherently providing the features used for classification as an explanation. *A promising direction for future research is the development of self-explaining and interactive explainability approaches as these would provide a deeper understanding of the ML models' decision-making process and aid clinicians in gaining valuable insights from ML predictions.* Developing models with transparent decision-making processes and providing insightful explanations for predictions would not only improve trust among clinicians but also pave the way for wider adoption of such methods in clinical practice.

1.7 Conclusion

The present thesis has made significant scientific contributions to the field of clinical gait analysis, addressing key challenges and gaps in the automated analysis of human gait. The publications of this thesis have accumulated a total of 159 citations according to Google Scholar until May 2024 (Horsak et al. (2020) [46]: 42 citations, Slijepcevic et al. (2017) [40]: 57 citations, Slijepcevic et al. (2020) [41]: 17 citations, Slijepcevic et al. (2022) [42]: 39 citations, and Slijepcevic et al. (2023) [43]: 4 citation).

For more than two decades, machine learning has been applied to clinical gait data with the aim to enhance the efficiency of clinical gait analysis and contribute to better informed decision-making. However, many of the ML approaches proposed prior to the present thesis fail to satisfy the prerequisites for clinical practice. These limitations include reliance on small datasets for training, tackling simplified tasks, and utilizing non-transparent ML approaches. The present thesis addressed the existing gaps and limitations. The main conclusions from the thesis are summarized in the following:

- The creation of a high-quality publicly available dataset establishes a solid foundation for further research, providing a valuable resource for the development and validation of gait analysis algorithms. The evaluation of the discriminative power of 3DGA modalities demonstrated significant quantitative benefits in favor of kinematic data over GRF data, especially in complex multi-class classification tasks. However, employing GRF data could be advantageous in efficiently differentiating between physiological and pathological gait patterns (i.e., binary classification) or longitudinally monitoring the gait pattern of individuals, for instance, on a population basis utilizing wearable pressure insoles.
- The evaluation of data handling strategies offers effective solutions to manage the complexities inherent in large-scale gait datasets. These complexities encompass, e.g., variations in value ranges of gait signals, imbalances in the dataset, and the requirement to handle multiple trials for each individual. The findings can serve as valuable guidelines for data preprocessing and data aggregation in the domain of automated gait analysis.
- This thesis involves the development and evaluation of various classification approaches, ranging from traditional ML to deep learning methods. Interestingly, the performance of deep learning approaches did not meet expectations, as they either performed at a comparable level to or were surpassed by traditional ML methods. To assess further the full potential of deep learning for gait analysis, it is essential to support data sharing initiatives and conduct experiments on datasets even larger than those utilized in the present thesis.

- Furthermore, the development and evaluation of explainability approaches for the utilized classification models addresses a crucial aspect of translating automated gait classification approaches and findings into real-world applications. By offering techniques for data exploration and presenting methods for both decision and model explanations, this thesis lays the groundwork for clinicians to establish trust in automated gait classification.

In conclusion, the goals and outcomes attained in this thesis create a fertile foundation for significant progress in patient care, elevating diagnostic standards and contributing to the development of more efficient treatment plans in the future.

Bibliography

[1] A. Brand, L. Allen, M. Altman, M. Hlava, and J. Scott, “Beyond Authorship: Attribution, Contribution, Collaboration, and Credit,” *Learned Publishing*, vol. 28, no. 2, pp. 151–155, 2015.

[2] National Academies of Sciences, Engineering, and Medicine, *Selected Health Conditions and Likelihood of Improvement with Treatment.* National Academies Press, 2020.

[3] T. Vos, S. S. Lim, C. Abbafati, K. M. Abbas, M. Abbasi, M. Abbasifard, M. Abbasi-Kangevari, H. Abbastabar, F. Abd-Allah, A. Abdelalim, *et al.*, “Global Burden of 369 Diseases and Injuries in 204 Countries and Territories, 1990—2019: A Systematic Analysis for the Global Burden of Disease Study 2019,” *The Lancet*, vol. 396, no. 10258, pp. 1204–1222, 2020.

[4] C. Mayrhuber, B. Bittschi, *et al.*, “Fehlzeitenreport 2022. Krankheits-und unfallbedingte Fehlzeiten in Österreich,” *WIFO Studies*, 2022.

[5] J. Klimont, “Österreichische Gesundheitsbefragung 2019: Hauptergebnisse des Austrian Health Interview Survey (ATHIS) und methodische Dokumentation,” 2020.

[6] R. Baker, *Measuring Walking: A Handbook of Clinical Gait Analysis.* Mac Keith Press, 2013.

[7] B. Toro, C. Nester, and P. Farren, “A Review of Observational Gait Assessment in Clinical Practice,” *Physiotherapy Theory and Practice*, vol. 19, no. 3, pp. 137–149, 2003.

[8] C. Kirtley, *Clinical Gait Analysis: Theory and Practice.* Elsevier Health Sciences, 2006.

[9] E. Dorschky, M. Nitschke, A.-K. Seifer, A. J. van den Bogert, and B. M. Eskofier, “Estimation of Gait Kinematics and Kinetics From Inertial Sensor Data Using Optimal Control of Musculoskeletal Models,” *Journal of Biomechanics*, vol. 95, p. 109278, 2019.

[10] M. Mundt, W. R. Johnson, W. Potthast, B. Markert, A. Mian, and J. Alderson, "A Comparison of Three Neural Network Approaches for Estimating Joint Angles and Moments From Inertial Measurement Units," *Sensors*, vol. 21, no. 13, p. 4535, 2021.

[11] R. Caldas, M. Mundt, W. Potthast, F. B. de Lima Neto, and B. Markert, "A Systematic Review of Gait Analysis Methods Based on Inertial Sensors and Adaptive Algorithms," *Gait & Posture*, vol. 57, pp. 204–210, 2017.

[12] 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.

[13] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-Level Classification of Skin Cancer With Deep Neural Networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[14] H. A. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, A. Kalloo, A. B. H. Hassen, L. Thomas, A. Enk, *et al.*, "Man Against Machine: Diagnostic Performance of a Deep Learning Convolutional Neural Network for Dermoscopic Melanoma Recognition in Comparison to 58 Dermatologists," *Annals of Oncology*, vol. 29, no. 8, pp. 1836–1842, 2018.

[15] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. S. Corrado, A. Darzi, *et al.*, "International Evaluation of an AI System for Breast Cancer Screening," *Nature*, vol. 577, no. 7788, pp. 89–94, 2020.

[16] J. Figueiredo, C. P. Santos, and J. C. Moreno, "Automatic Recognition of Gait Patterns in Human Motor Disorders Using Machine Learning: A Review," *Medical Engineering & Physics*, 2018.

[17] 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.

[18] M. Längkvist, L. Karlsson, and A. Loutfi, "A Review of Unsupervised Feature Learning and Deep Learning for Time-Series Modeling," *Pattern Recognition Letters*, vol. 42, pp. 11–24, 2014.

[19] T. Flash and B. Hochner, "Motor Primitives in Vertebrates and Invertebrates," *Current Opinion in Neurobiology*, vol. 15, no. 6, pp. 660–666, 2005.

[20] J. Guerra, J. Uddin, D. Nilsen, J. Mclnerney, A. Fadoo, I. B. Omofuma, S. Hughes, S. Agrawal, P. Allen, and H. M. Schambra, "Capture, Learning, and Classification of Upper Extremity Movement Primitives in Healthy Controls and Stroke Patients," in *2017 International Conference on Rehabilitation Robotics (ICORR)*, pp. 547–554, IEEE, 2017.

[21] 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.

[22] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, “What Do We Need to Build Explainable AI Systems for the Medical Domain?,” *CoRR*, vol. abs/1712.09923, 2017.

[23] W. Samek, T. Wiegand, and K.-R. Müller, “Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models,” *ITU Journal: ICT Discoveries*, vol. 1, no. 1, pp. 39–48, 2017.

[24] European Union, “Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons With Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation),” *Official Journal of the European Union*, vol. L 119, pp. 1–88, 2016.

[25] European Commission, “Proposal for a Regulation of the European Parliament and the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts,” *EUR-Lex-52021PC0206*, 2021.

[26] I. El Maachi, G.-A. Bilodeau, and W. Bouachir, “Deep 1D-Convnet for Accurate Parkinson Disease Detection and Severity Prediction From Gait,” *Expert Systems with Applications*, vol. 143, p. 113075, 2020.

[27] W. Zeng, F. Liu, Q. Wang, Y. Wang, L. Ma, and Y. Zhang, “Parkinson’s Disease Classification Using Gait Analysis via Deterministic Learning,” *Neuroscience Letters*, vol. 633, pp. 268–278, 2016.

[28] 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.

[29] M. Alaqtash, T. Sarkodie-Gyan, H. Yu, O. Fuentes, R. Brower, and A. Abdelgawad, “Automatic Classification of Pathological Gait Patterns Using Ground Reaction Forces and Machine Learning Algorithms,” in *Engineering in Medicine and Biology Society, 2011 Annual International Conference of the IEEE*, pp. 453–457, IEEE, 2011.

[30] C. Nüesch, V. Valderrabano, C. Huber, V. von Tscharner, and G. Pagenstert, “Gait Patterns of Asymmetric Ankle Osteoarthritis Patients,” *Clinical Biomechanics*, vol. 27, no. 6, pp. 613–618, 2012.

[31] D. Soares, M. de Castro, E. Mendes, and L. Machado, “Principal Component Analysis in Ground Reaction Forces and Center of Pressure Gait Waveforms of People With Transfemoral Amputation,” *Prosthetics and Orthotics International*, vol. 40, no. 6, pp. 729–738, 2016.

[32] A. Muniz and J. Nadal, "Application of Principal Component Analysis in Vertical Ground Reaction Force to Discriminate Normal and Abnormal Gait," *Gait & Posture*, vol. 29, no. 1, pp. 31–35, 2009.

[33] E. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, 2020.

[34] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and Explainability of Artificial Intelligence in Medicine," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 4, p. e1312, 2019.

[35] C. Beyaert, R. Vasa, and G. E. Frykberg, "Gait Post-stroke: Pathophysiology and Rehabilitation Strategies," *Neurophysiologie Clinique/Clinical Neurophysiology*, vol. 45, no. 4-5, pp. 335–355, 2015.

[36] W. Zheng and M. Jin, "The Effects of Class Imbalance and Training Data Size on Classifier Learning: An Empirical Study," *SN Computer Science*, vol. 1, pp. 1–13, 2020.

[37] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On Pixel-Wise Explanations for Non-linear Classifier Decisions by Layer-Wise Relevance Propagation," *PLoS One*, vol. 10, no. 7, p. e0130140, 2015.

[38] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

[39] T. C. Pataky, "Generalized n-dimensional Biomechanical Field Analysis Using Statistical Parametric Mapping," *Journal of Biomechanics*, vol. 43, no. 10, pp. 1976–1982, 2010.

[40] D. Slijepcevic, M. Zeppelzauer, A.-M. Gorgas, C. Schwab, M. Schüller, A. Baca, C. Breiteneder, and B. Horsak, "Automatic Classification of Functional Gait Disorders," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1653–1661, 2017.

[41] D. Slijepcevic, M. Zeppelzauer, C. Schwab, A.-M. Raberger, C. Breiteneder, and B. Horsak, "Input Representations and Classification Strategies for Automated Human Gait Analysis," *Gait & Posture*, vol. 76, pp. 198–203, 2020.

[42] D. Slijepcevic, F. Horst, S. Lapuschkin, B. Horsak, A.-M. Raberger, A. Kranzl, W. Samek, C. Breiteneder, W. I. Schöllhorn, and M. Zeppelzauer, "Explaining Machine Learning Models for Clinical Gait Analysis," *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 2, pp. 1–27, 2021.

[43] D. Slijepcevic, M. Zeppelzauer, F. Unglaube, A. Kranzl, C. Breiteneder, and B. Horsak, "Explainable Machine Learning in Human Gait Analysis: A Study on Children With Cerebral Palsy," *IEEE Access*, vol. 11, pp. 65906–65923, 2023.

[44] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking Clever Hans Predictors and Assessing What Machines Really Learn," *Nature Communications*, vol. 10, no. 1, p. 1096, 2019.

[45] G. Giakas and V. Baltzopoulos, "Time and Frequency Domain Analysis of Ground Reaction Forces During Walking: An Investigation of Variability and Symmetry," *Gait & Posture*, vol. 5, no. 3, pp. 189–197, 1997.

[46] B. Horsak, D. Slijepcevic, A.-M. Raberger, C. Schwab, M. Worisch, and M. Zeppelzauer, "GaitRec, a Large-Scale Ground Reaction Force Dataset of Healthy and Impaired Gait," *Scientific Data*, vol. 7, no. 1, p. 143, 2020.

[47] D. Slijepcevic, B. Horsak, C. Schwab, A. Raberger, M. Schüller, A. Baca, C. Breiteneder, and M. Zeppelzauer, "Ground Reaction Force Measurements for Gait Classification Tasks: Effects of Different PCA-Based Representations," *Gait & Posture*, vol. 57, pp. 4–5, 2017.

[48] D. Janssen, W. I. Schöllhorn, K. M. Newell, J. M. Jäger, F. Rost, and K. Vehof, "Diagnosing Fatigue in Gait Patterns by Support Vector Machines and Self-Organizing Maps," *Human Movement Science*, vol. 30, no. 5, pp. 966–975, 2011.

[49] B. M. Eskofier, P. Federolf, P. F. Kugler, and B. M. Nigg, "Marker-Based Classification of Young–Elderly Gait Pattern Differences via Direct PCA Feature Extraction and SVMs," *Computer Methods in Biomechanics and Biomedical Engineering*, vol. 16, no. 4, pp. 435–442, 2013.

[50] J. Christian, J. Kröll, G. Strutzenberger, N. Alexander, M. Ofner, and H. Schwameder, "Computer Aided Analysis of Gait Patterns in Patients With Acute Anterior Cruciate Ligament Injury," *Clinical Biomechanics*, vol. 33, pp. 55–60, 2016.

[51] R. Altilio, M. Paoloni, and M. Panella, "Selection of Clinical Features for Pattern Recognition Applied to Gait Analysis," *Medical & Biological Engineering & Computing*, vol. 55, no. 4, pp. 685–695, 2017.

[52] E. J. Harris, I.-H. Khoo, and E. Demircan, "A Survey of Human Gait-Based Artificial Intelligence Applications," *Frontiers in Robotics and AI*, vol. 8, p. 749274, 2022.

[53] V. Arya, R. K. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović, *et al.*, "One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques," *arXiv preprint arXiv:1909.03012*, 2019.

[54] A. Rind, D. Slijepčević, M. Zeppelzauer, F. Unglaube, A. Kranzl, and B. Horsak, "Trustworthy Visual Analytics in Clinical Gait Analysis: A Case Study for Patients with Cerebral Palsy," in *2022 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX)*, pp. 8–15, IEEE, 2022.

[55] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 180–186, 2020.

[56] F. Horst, D. Slijepcevic, M. Zeppelzauer, A. Raberger, S. Lapuschkin, W. Samek, W. Schöllhorn, C. Breiteneder, and B. Horsak, "Explaining Automated Gender Classification of Human Gait," *Gait & Posture*, vol. 81, pp. 159–160, 2020.

[57] D. Slijepcevic, F. Horst, M. Simak, S. Lapuschkin, A.-M. Raberger, W. Samek, C. Breiteneder, W. I. Schöllhorn, M. Zeppelzauer, and B. Horsak, "Explaining Machine Learning Models for Age Classification in Human Gait Analysis," *Gait & Posture*, vol. 97, pp. S252–S253, 2022.

[58] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[59] A. Ferrari, L. Bergamini, G. Guerzoni, S. Calderara, N. Bicocchi, G. Vitetta, C. Borghi, R. Neviani, and A. Ferrari, "Gait-Based Diplegia Classification Using LSMT Networks," *Journal of Healthcare Engineering*, vol. 2019, 2019.

[60] M. Sangeux and J. Polak, "A Simple Method to Choose the Most Representative Stride and Detect Outliers," *Gait & Posture*, vol. 41, no. 2, pp. 726–730, 2015.

[61] F. Horst, D. Slijepcevic, M. Simak, and W. I. Schöllhorn, "Gutenberg Gait Database, a Ground Reaction Force Database of Level Overground Walking in Healthy Individuals," *Scientific Data*, vol. 8, no. 1, p. 232, 2021.

[62] Y. Kobayashi, N. Hida, K. Nakajima, M. Fujimoto, and M. Mochimaru, "AIST Gait Database 2019." `https://unit.aist.go.jp/harc/ExPART/GDB2019_e.html`, 2019. Online; accessed 28 July 2023.

[63] F. Horst, D. Slijepcevic, M. Simak, B. Horsak, W. I. Schöllhorn, and M. Zeppelzauer, "Modeling Biological Individuality Using Machine Learning: A Study on Human Gait," *Computational and Structural Biotechnology Journal*, 2023.

[64] F. Horst, D. Slijepcevic, M. Zeppelzauer, A. Raberger, S. Lapuschkin, W. Samek, W. Schöllhorn, C. Breiteneder, and B. Horsak, "Explaining Automated Gender Classification of Human Gait," *Gait & Posture*, vol. 81, pp. 159–160, 2020. ESMAC 2020 Abstracts.

[65] D. Slijepcevic, F. Horst, M. Simak, S. Lapuschkin, A. Raberger, W. Samek, C. Breiteneder, W. Schöllhorn, M. Zeppelzauer, and B. Horsak, “Explaining Machine Learning Models for Age Classification in Human Gait Analysis,” *Gait & Posture*, vol. 97, pp. S252–S253, 2022. ESMAC 2022 Abstracts.

[66] M. N. I. Shuzan, M. E. Chowdhury, M. B. I. Reaz, A. Khandakar, F. F. Abir, M. A. A. Faisal, S. H. M. Ali, A. A. A. Bakar, M. H. Chowdhury, Z. B. Mahbub, *et al.*, “Machine Learning-Based Classification of Healthy and Impaired Gaits Using 3D-GRF Signals,” *Biomedical Signal Processing and Control*, vol. 81, p. 104448, 2023.

[67] D. Jani, V. Varadarajan, R. Parmar, M. H. Bohara, D. Garg, A. Ganatra, and K. Kotecha, “An Efficient Gait Abnormality Detection Method Based on Classification,” *Journal of Sensor and Actuator Networks*, vol. 11, no. 3, p. 31, 2022.

[68] C. Pandey, D. S. Roy, R. C. Poonia, A. Altameem, S. R. Nayak, A. Verma, and A. K. J. Saudagar, “GaitRec-Net: A Deep Neural Network for Gait Disorder Detection Using Ground Reaction Force,” *PPAR Research*, vol. 2022, 2022.

[69] R. Yun, M. Salama, and L. Elrefaei, “An Exploratory Study on the Effect of Applying Various Artificial Neural Networks to the Classification of Lower Limb Injury,” *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 31, no. 2, pp. 448–461, 2023.

[70] J. Chakraborty, S. Upadhyay, and A. Nandy, “Musculoskeletal Injury Recovery Assessment Using Gait Analysis With Ground Reaction Force Sensor,” *Medical Engineering & Physics*, vol. 103, p. 103788, 2022.

[71] M. Iber, B. Dumphart, V.-A. de Jesus Oliveira, S. Ferstl, J. M. Reis, D. Slijepčević, M. Heller, A.-M. Raberger, and B. Horsak, “Mind the Steps: Towards Auditory Feedback in Tele-Rehabilitation Based on Automated Gait Classification,” in *Proceedings of the 16th International Audio Mostly Conference*, pp. 139–146, 2021.

[72] V. A. de Jesus Oliveira, D. Slijepčević, B. Dumphart, S. Ferstl, J. Reis, A.-M. Raberger, M. Heller, B. Horsak, and M. Iber, “Auditory Feedback in Tele-Rehabilitation Based on Automated Gait Classification,” *Personal and Ubiquitous Computing*, pp. 1–14, 2023.

[73] J. E. Deffeyes and D. M. Peters, “Time-Integrated Propulsive and Braking Impulses Do Not Depend on Walking Speed,” *Gait & Posture*, vol. 88, pp. 258–263, 2021.

[74] D. Slijepcevic, M. Zeppelzauer, C. Schwab, A.-M. Raberger, B. Dumphart, A. Baca, C. Breiteneder, and B. Horsak, “P 011–Towards an Optimal Combination of Input Signals and Derived Representations for Gait Classification Based on Ground Reaction Force Measurements,” *Gait & Posture*, vol. 65, pp. 249–250, 2018.

[75] G. Williams, D. Lai, A. Schache, and M. Morris, “Classification of Gait Disorders Following Traumatic Brain Injury,” *The Journal of Head Trauma Rehabilitation*, vol. 30, no. 2, pp. E13–E23, 2015.

[76] J. Burdack, F. Horst, S. Giesselbach, I. Hassan, S. Daffner, and W. I. Schöllhorn, “Systematic Comparison of the Influence of Different Data Preprocessing Methods on the Performance of Gait Classifications Using Machine Learning,” *Frontiers in Bioengineering and Biotechnology*, vol. 8, p. 260, 2020.

[77] S. Winiarski and A. Rutkowska-Kucharska, “Estimated Ground Reaction Force in Normal and Pathological Gait,” *Acta of Bioengineering & Biomechanics*, vol. 11, no. 1, 2009.

[78] J. Rodda and H. Graham, “Classification of Gait Patterns in Spastic Hemiplegia and Spastic Diplegia: A Basis for a Management Algorithm,” *European Journal of Neurology*, vol. 8, pp. 98–108, 2001.

[79] Y. Zhang and Y. Ma, “Application of Supervised Machine Learning Algorithms in the Classification of Sagittal Gait Patterns of Cerebral Palsy Children With Spastic Diplegia,” *Computers in Biology and Medicine*, vol. 106, pp. 33–39, 2019.

[80] H. Darbandi, M. Baniasad, S. Baghdadi, A. Khandan, A. Vafaee, and F. Farahmand, “Automatic Classification of Gait Patterns in Children With Cerebral Palsy Using Fuzzy Clustering Method,” *Clinical Biomechanics*, vol. 73, pp. 189–194, 2020.

[81] C. Dindorf, J. Konradi, C. Wolf, B. Taetz, G. Bleser, J. Huthwelker, F. Werthmann, P. Drees, M. Fröhlich, and U. Betz, “Machine Learning Techniques Demonstrating Individual Movement Patterns of the Vertebral Column: The Fingerprint of Spinal Motion,” *Computer Methods in Biomechanics and Biomedical Engineering*, vol. 25, no. 7, pp. 821–831, 2022.

[82] P. Krondorfer, D. Slijepčević, F. Unglaube, A. Kranzl, C. Breiteneder, M. Zeppelzauer, and B. Horsak, “Deep Learning-Based Similarity Retrieval in Clinical 3D Gait Analysis,” *Gait & Posture*, vol. 90, pp. 127–128, 2021.

[83] K. A. Duncanson, S. Thwaites, D. Booth, G. Hanly, W. S. Robertson, E. Abbasnejad, and D. Thewlis, “Deep Metric Learning for Scalable Gait-Based Person Re-identification Using Force Platform Data,” *Sensors*, vol. 23, no. 7, p. 3392, 2023.

[84] J. Zhang, Y. Zhao, F. Shone, Z. Li, A. F. Frangi, S. Q. Xie, and Z.-Q. Zhang, “Physics-Informed Deep Learning for Musculoskeletal Modeling: Predicting Muscle Forces and Joint Kinematics From Surface EMG,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 484–493, 2022.

[85] F. Horst, F. Kramer, B. Schäfer, A. Eekhoff, P. Hegen, B. Nigg, and W. Schöllhorn, “Daily Changes of Individual Gait Patterns Identified by Means of Support Vector Machines,” *Gait & Posture*, vol. 49, pp. 309–314, 2016.

[86] E. Chao, R. Laughman, E. Schneider, and R. Stauffer, “Normative Data of Knee Joint Motion and Ground Reaction Forces in Adult Level Walking,” *Journal of Biomechanics*, vol. 16, no. 3, pp. 219–233, 1983.

[87] B. Nigg, V. Fisher, and J. Ronsky, “Gait Characteristics as a Function of Age and Gender,” *Gait & Posture*, vol. 2, no. 4, pp. 213–220, 1994.

[88] M.-C. Chiu and M.-J. Wang, “The Effect of Gait Speed and Gender on Perceived Exertion, Muscle Activity, Joint Motion of Lower Extremity, Ground Reaction Force and Heart Rate During Normal Walking,” *Gait & Posture*, vol. 25, no. 3, pp. 385–392, 2007.

[89] M.-J. Chung and M.-J. J. Wang, “The Change of Gait Parameters During Walking at Different Percentage of Preferred Walking Speed for Healthy Adults Aged 20-–60 Years,” *Gait & Posture*, vol. 31, no. 1, pp. 131–135, 2010.

[90] H. Toda, A. Nagano, and Z. Luo, “Age and Gender Differences in the Control of Vertical Ground Reaction Force by the Hip, Knee and Ankle Joints,” *Journal of Physical Therapy Science*, vol. 27, no. 6, pp. 1833–1838, 2015.

[91] K. A. Boyer, G. S. Beaupre, and T. P. Andriacchi, “Gender Differences Exist in the Hip Joint Moments of Healthy Older Walkers,” *Journal of Biomechanics*, vol. 41, no. 16, pp. 3360–3365, 2008.

[92] N. J. Cronin, “Using deep neural networks for kinematic analysis: Challenges and opportunities,” *Journal of Biomechanics*, vol. 123, p. 110460, 2021.

[93] B. Krawczyk, “Learning from imbalanced data: open challenges and future directions,” *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016.

[94] B. K. Iwana and S. Uchida, “An empirical survey of data augmentation for time series classification with neural networks,” *PLoS One*, vol. 16, no. 7, p. e0254841, 2021.

[95] K. Chia, I. Fischer, P. Thomason, K. Graham, and M. Sangeux, “Is It Feasible to Use an Automated System to Identify Gait Impairments?,” *Gait & Posture*, vol. 57, pp. 167–168, 2017.

[96] B. Dumphart, D. Slijepcevic, M. Zeppelzauer, A. Kranzl, F. Unglaube, A. Baca, and B. Horsak, “Robust deep learning-based gait event detection across various pathologies,” *PLoS One*, vol. 18, no. 8, p. e0288555, 2023.

[97] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative Adversarial Nets,” in *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

[98] E. Brophy, Z. Wang, Q. She, and T. Ward, “Generative Adversarial Networks in Time Series: A Systematic Literature Review,” *ACM Computing Surveys*, vol. 55, no. 10, pp. 1–31, 2023.

[99] S. Kazeminia, C. Baur, A. Kuijper, B. van Ginneken, N. Navab, S. Albarqouni, and A. Mukhopadhyay, “GANs for Medical Image Analysis,” *Artificial Intelligence in Medicine*, vol. 109, p. 101938, 2020.

[100] A. Odena, C. Olah, and J. Shlens, “Conditional Image Synthesis With Auxiliary Classifier GANs,” in *International Conference on Machine Learning*, pp. 2642–2651, PMLR, 2017.

[101] T. Baumhauer, D. Slijepcevic, and M. Zeppelzauer, “Bounded Logit Attention: Learning to Explain Image Classifiers,” in *NeurIPS’22 Workshop on All Things Attention: Bridging Different Perspectives on Attention*, 2022.

CHAPTER 2

Publications

2.1 GaitRec, a Large-Scale Ground Reaction Force Dataset of Healthy and Impaired Gait

Brian Horsak, Djordje Slijepcevic, Anna-Maria Raberger, Caterine Schwab, Marianne Worisch, and Matthias Zeppelzauer. **GaitRec, a Large-Scale Ground Reaction Force Dataset of Healthy and Impaired Gait**. *Scientific Data*, 7(1):143, 2020. DOI: 10.1038/s41597-020-0481-z

The final version of this publication is available at: `https://doi.org/10.1038/s41597-020-0481-z`. Permission for reprint granted, © 2020 Slijepcevic

www.nature.com/scientificdata

SCIENTIFIC DATA

Check for updates

OPEN

DATA DESCRIPTOR

GaitRec, a large-scale ground reaction force dataset of healthy and impaired gait

Brian Horsak[1,4]✉, Djordje Slijepcevic[2,4], Anna-Maria Raberger[1], Caterine Schwab[1], Marianne Worisch[3] & Matthias Zeppelzauer[2]

The quantification of ground reaction forces (GRF) is a standard tool for clinicians to quantify and analyze human locomotion. Such recordings produce a vast amount of complex data and variables which are difficult to comprehend. This makes data interpretation challenging. Machine learning approaches seem to be promising tools to support clinicians in identifying and categorizing specific gait patterns. However, the quality of such approaches strongly depends on the amount of available annotated data to train the underlying models. Therefore, we present GaitRec, a comprehensive and completely annotated large-scale dataset containing bi-lateral GRF walking trials of 2,084 patients with various musculoskeletal impairments and data from 211 healthy controls. The dataset comprises data of patients after joint replacement, fractures, ligament ruptures, and related disorders at the hip, knee, ankle or calcaneus during their entire stay(s) at a rehabilitation center. The data sum up to a total of 75,732 bi-lateral walking trials and enable researchers to classify gait patterns at a large-scale as well as to analyze the entire recovery process of patients.

Background & Summary

The quantification of ground reaction forces (GRF) is a standard tool for clinicians to objectively measure human locomotion and to describe and analyze a patient's gait performance in detail. The primary aim of instrumented gait analysis, regardless of which technology used, is to identify impairments that affect a patient's gait pattern and to describe those quantitatively[1]. Recordings obtained during clinical gait analyses produce a vast amount of data which are difficult to comprehend and analyze due to their high-dimensionality, temporal dependencies, strong variability, non-linear relationships and correlations within the data[2]. This makes data interpretation challenging and requires an experienced clinician to draw valid conclusions. Therefore, there is a constantly growing interest in applying machine learning techniques to clinical gait analysis data for the purpose of pattern identification and automated classification. Such systems might bear potential to assist clinicians in identifying and categorizing specific gait patterns into clinically relevant categories[2,3]. Machine learning methods employed in this context comprise, but are not limited to, neural networks[4–6], support vector machines[7–9], nearest neighbor classifiers[10,11], and different clustering approaches[12].

Our research group is collaborating with a local Austrian rehabilitation center of the Austrian Workers' Compensation Board (AUVA). The AUVA is the social insurance for occupational risks for more than 3.3 million employees and 1.4 million pupils and students in Austria. They have been using GRF assessments during walking to diagnose, plan and evaluate therapy outcomes for more than two decades. Our main research goal within this collaboration was to develop automatic classification algorithms which support clinicians during data inspection and interpretation. To this end, we have developed a machine learning framework for gait classification and have performed comprehensive experiments[13–16]. One conclusion of our experiments is that the performance of automatic classification methods strongly depends on the amount of available training data. One reason for this is that state-of-the-art classifiers such as deep neural networks[17] are extremely data hungry and require large-scale data to learn meaningful and generalizable patterns from the data. The training process, however, requires each walking-trial in the dataset to be annotated and categorized exactly. Even though there are datasets available

[1]St. Pölten University of Applied Sciences, Institute of Health Sciences, St. Pölten, Austria. [2]St. Pölten University of Applied Sciences, Institute of Creative Media Technologies, St. Pölten, Austria. [3]Rehabilitation Center Weißer Hof, Austrian Workers' Compensation Board (AUVA), Klosterneuburg, Austria. [4]These authors contributed equally: Brian Horsak, Djordje Slijepcevic. ✉e-mail: brian.horsak@fhstp.ac.at

relevant to instrumented gait analysis, e.g.[18], the availability of completely annotated large-scale datasets is very scarce. Our collaboration with the AUVA and their gait laboratory gave us the unique opportunity to process and manually annotate thousands of walking GRF trials from several years of clinical practice. These data have been used in our previous research and show a large potential for further research in gait analysis (see section usageNotes) to achieve the long-term goal to put assistive machine learning techniques into clinical gait analysis practice. For this purpose, we make these data available to the public as the GaitRec dataset.

Methods

Data recording & testing protocol. The presented dataset is part of an existing clinical gait database maintained by a local Austrian rehabilitation center, which offers care to patients across entire Austria. Prior to the experiments involved and the publication of the dataset, approval was obtained from the local Ethics Committee of Lower Austria (GS1-EK-4/299-2014). Data were recorded during clinical practice between 2007 and 2018. Bi-lateral GRF were recorded by asking patients and healthy controls to walk unassisted and without a walking aid at self-selected walking speed on an approximately 10 m walkway with two centrally embedded force plates (Kistler, Type 9281B12, Winterthur, CH). The force plates were placed in a consecutive order and flush with the ground. Both plates were covered with the same walkway surface material, so that targeting was not an issue. During one session, subjects walked until a minimum number of (usually) ten valid recordings were available. These recordings were defined as valid by the assessor when the participant walked naturally (e.g. with respect to targeting) and there was a clean foot strike on each force plate. Left and right foot contacts for each force plate were identified and set by visual inspection by the assessor during each recording. Patients were asked to walk at their self-selected walking speed. Healthy controls walked at three different walking speeds (mean and standard deviation, m/s): slow 0.98 (0.14), self-selected 1.27 (0.13), and fast 1.55 (0.15). In accordance with the internal rehabilitation center's standards, patients walked either barefoot, with their orthopedic or normal shoes, and with or without orthopedic insoles. Healthy controls walked either barefoot or with their normal shoes. Prior to the gait analysis session, each participant underwent rigorous physical examination by a physician. The three analog GRF signals (vertical, anterior-posterior and medio-lateral force components) as well as the center of pressure (COP) were converted to digital signals using a sampling rate of 2000 Hz and a 12-bit analog-digital converter (DT3010, Data Translation Incorporation, Marlboro, MA, USA) with a signal input range of ±10 V. COP and GRF were recorded in the local force plate coordinate system (reaction-orientated). For easier usage the orientation of the medio-lateral and anterior-posterior signals for all data were uniformed, so that medial and anterior forces are always represented as positive values. Due to the center's internal standards raw signals were only available down-sampled to 250 Hz. To avoid noise and signal peaks at the beginning and end of the signals, a threshold of 25 N was applied to all force data and the COP was calculated afterwards. These data are referred to as unprocessed (raw) GRF signals. Additionally, we have generated processed "ready to use" data. For this purpose the COP was only calculated when the vertical force reached 80 N to avoid inaccuracies in COP calculation at small force values. Additionally, the medio-lateral COP coordinates were mean-centered and anterior-posterior coordinates zero-centered. This was in line with the internal standards of the rehabilitation center. The processed force signals were then filtered using a 2nd order low-pass butterworth filter with a cut-off frequency of 20 Hz to reduce noise and were time-normalized to 100% stance (i.e. 101 points). The choice of appropriate cut-off frequency ranges widely in the literature, 20 Hz seems as a good trade-off between reducing noise and attaining as much physiological frequency content as possible[19]. The interested reader may also refer to [ref. [20], p.49]. Amplitude values of the three force components were expressed as a multiple of body weight (*BW*) by dividing the force by the product of body mass times acceleration due to gravity (g). Amplitude and time normalization are both necessary operations to reduce effects of covariates (such as anthropometry) on the signals and to reduce temporal differences which make comparisons of different steps difficult, e.g.[21,22]. Note that the processed and amplitude normalized data show small variations at the first and last frame of each signal. This might affect machine learning outcomes and therefore needs to be recognized. Sessions with less than three bi-lateral trials per participant were not included in the dataset. Additionally, we have used an algorithm proposed by Sangeux and Polak to eliminate any outliers before they were included in the GaitRec dataset[23]. This algorithm is based on the notion of depth, where the deepest signal is the equivalent to the median for univariate data and is sensitive to both shape and position of the signals. As suggested by Sangeux and Polak we have used a score of three to run their algorithm. All processing steps were performed in Matlab 2019a (The MathWorks Inc., Natick, MA, USA).

Dataset & annotation. The presented dataset comprises completely anonymized GRF measurements from 2085 patients with different musculoskeletal impairments ("gait disorders", GD) and data from 211 healthy controls (HC) including additional metadata such as age, sex, shod condition, walking speed condition, etc. For details see Table 1. Note that there is a considerable large gender imbalance in all GD classes. Healthy controls were recruited in the geographical region around the clinic's by public posting and considered eligible if they were free of pain and complaints at the lower extremity and spine and did not have any orthotics or orthopedic insoles. Exclusion criteria were any history of surgery or trauma at the spine or lower extremities. This was assessed by an experienced therapist. A typical stay of a patient at the rehabilitation center ranged from a few days to several weeks and depends on factors such as diagnosis, administered therapy/surgery, and progress in recovery. During that time a patient is usually administered once a week to the gait analysis. At the beginning of a patient's stay, therapy outcomes are mutually defined between the therapist and the patient. After reaching these goals in whole or in part, patients are usually discharged. However, they can be readmitted if necessary. The present dataset contains the data gathered during the entire stay(s) of each patient and covers a patient's entire rehabilitation progress. Different types of analyses can thus be performed on the data set: an *inter-participant analysis* based on the initial assessment (first measurement session), e.g. for gait pattern classification, an *intra-participant analysis*, e.g. for the assessment of rehabilitation progress, or combinations.

Class	N	Age (yrs.) Mean (SD)	Body mass (kg) Mean (SD)	Sex (m/f)	Bi-lateral Trials
Healthy C.	211	34.7 (13.9)	73.9 (15.6)	104/107	7,755
Hip	450	42.6 (12.8)	82.4 (15.6)	373/77	12,748
Knee	625	41.6 (12.0)	84.3 (18.6)	426/199	19,873
Ankle	627	41.6 (11.4)	87.0 (18.0)	498/129	21,386
Calcaneus	382	43.5 (10.4)	84.0 (14.5)	339/43	13,970
Total	**2,295**	**41.5 (12.1)**	**83.6 (17.3)**	**1,740/555**	**75,732**

Table 1. Demographic overview of the dataset and the pre-defined classes.

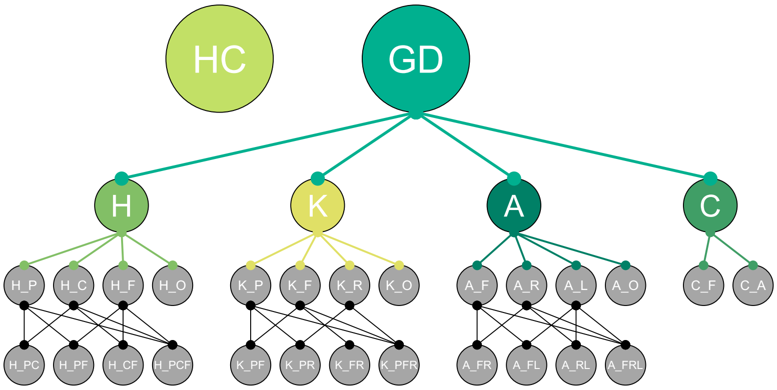


Fig. 1 Class taxonomy. The class structure and the dependencies between the classes of the GaitRec dataset: Healthy Controls (HC), Gait Disorders (GD), Hip (H), Knee (K), Ankle (A), and Calcaneus (C). Details of the subclasses are described in Section Dataset & Annotation.

Regarding annotation, the dataset was manually labeled by a well-experienced physical therapist (with more than a decade of clinical experience) based on the available medical diagnosis of each patient. The annotation labels are formed by two strings concatenated with an underscore "*X_xxx*", where "*X*" denotes the general anatomical joint level at which the orthopedic impairment was located, i.e. at the hip "*H*", knee "*K*", ankle "*A*", or calcaneus "*C*". The second string ("*xxx*") gives a more detailed localization and is joint dependent, see the following paragraphs for details. An overview of the class structure is shown in Fig. 1.

- **Hip class (*H_xxx*):** The most common injuries present in the hip class are fractures of the pelvis and thigh as well as luxation of the hip joint, coxarthrosis, and total hip replacement. The second string "*xxx*" refers to the following specific anatomical regions: pelvis (H_P), coxa (H_C), the femur (H_F), and their combinations when two or more anatomical areas are affected (*H_PC, H_PF, H_CF, H_PCF*), as well as one class for other diagnoses (*H_O*).
- **Knee class (*K_xxx*):** The knee class comprises patients after patella, femur or tibia fractures, ruptures of the cruciate or collateral ligaments or the meniscus, and total knee replacements. The second string "*xxx*" refers to the following specific anatomical regions or diagnosis: patella (K_P), a fracture near the knee joint of the femur or the tibia (K_F), rupture of ligaments or the menisci (K_R), and their combinations (*K_PF, K_PR, K_FR, K_PFR*, as well as one class for other diagnoses (*K_O*).
- **Ankle class (*A_xxx*):** The ankle class includes patients after fractures of the malleoli, talus, tibia, or lower leg, and ruptures of ligaments or the Achilles tendon. The second string "*xxx*" refers to the following specific anatomical regions or diagnosis: fracture of the tibia, fibula or talus near the ankle joint (A_F), rupture of ligaments or the Achilles tendon (A_R), lower leg shaft fracture (A_L), and their combinations (*A_FR, A_FL, A_RL, A_FRL*, as well as one class for other diagnoses (*A_O*).
- **Calcaneus class (*C_xxx*):** The calcaneus class comprises patients after calcaneus fractures or ankle fusion surgery. The second string "*xxx*" refers to the following specific anatomical regions or diagnosis: fracture (C_F) or arthrodesis (C_A).

The hierarchical multi-level categorization allows for grouping the data into a dataset with four GD classes (H ∪ K ∪ A ∪ C) and one healthy controls (HC) class, but also holds more details if needed. Figure 1 and Table 1 give a brief overview of the dataset. Although the metadata includes a structured labelling of musculoskeletal impairments for each subject, there is no information available about the history of similar or other types of musculoskeletal injuries for both, the patient and the healthy controls. This limiting factors needs to be recognized when using GaitRec.

Variables	Associated file	Format	Dimension	Unit	Description
Vertical GRF	`GRF_F_V-RAW_*.csv`	double	1 × n	Newton	Raw vertical ground reaction force
Anterior-posterior GRF	`GRF_F_AP-RAW_*.csv`	double	1 × n	Newton	Raw breaking and propulsive shear force
Medio-lateral GRF	`GRF_F_ML_RAW_*.csv`	double	1 × n	Newton	Raw medio-lateral shear force
COP anterior-posterior	`GRF_COP_AP_RAW_*.csv`	double	1 × n	Centimeter	Raw COP coordinate in walking direction
COP medio-lateral	`GRF_COP_ML_RAW_*.csv`	double	1 × n	Centimeter	Raw COP coordinate in medio-lateral direction
Vertical GRF	`GRF-F_V_PRO_*.csv`	double	1 × n	Multiple of body weight	Post-processed vertical ground reaction force
Anterior-posterior GRF	`GRF_F_AP_PRO_*.csv`	double	1 × n	Multiple of body weight	Post-processed breaking and propulsive shear force
Medio-lateral GRF	`GRF-F_ML_PRO_*.csv`	double	1 × n	Multiple of body weight	Post-processed medio-lateral shear force
COP anterior-posterior	`GRF_COP_AP_PRO_*.csv`	double	1 × n	% stance	Post-processed COP coordinate in walking direction
COP medio-lateral	`GRF_COP_ML_PRO_*.csv`	double	1 × n	% stance	Post-processed COP coordinate in medio-lateral direction

Table 2. Description of the data stored in the "`GRF_*.csv" files`. "*" for the associated file name is a placeholder for "right" and "left". n is either the number of frames during one step across the force plate for the unprocessed data ("`RAW`") or a time-normalized vector of 101 points for the post-processed ("`PRO`") data. Note that the first three columns of each file hold the `SUBJECT_ID`, `SESSION_ID`, and `TRIAL_ID`.

Categories/Variables	Format	Unit	Description
Identifiers			
`SUBJECT_ID`	integer	—	Unique identifier of a subject
`SESSION_ID`	integer	—	Unique identifier of a session
Labels			
`CLASS_LABEL`	string	—	Annotated class labels
`CLASS_LABEL_DETAILED`	string	—	Annotated class labels for subclasses
Subject Metadata			
`SEX`	binary	—	female = 0, male = 1
`AGE`	integer	years	Age at recording date
`HEIGHT`	integer	centimeter	Body height in centimeters
`BODY_WEIGHT`	double	$\frac{kg\,m}{s^2}$	Body weight in Newton
`BODY_MASS`	double	kg	Body mass
`SHOE_SIZE`	double	EU	Shoe size in the Continental European System
`AFFECTED_SIDE`	integer	—	left = 0, right = 1, both = 2
Trial Metadata			
`SHOD_CONDITION`	integer	—	barefoot & socks = 0, normal shoe = 1, orthopedic shoe = 2
`ORTHOPEDIC_INSOLE`	binary	—	without insole = 0, with insole = 1
`SPEED`	integer	—	slow = 1, self-selected = 2, fast = 3 walking speed
`READMISSION`	integer	—	indicates the number of re-admission = 0 … n
`SESSION_TYPE`	integer	—	initial measurement = 1, control measurement = 2, initial measurement after readmission = 3
`SESSION_DATE`	string	—	date of recording session in the format "DD-MM-YYYY"
Train-Test Split Information			
`TRAIN`	binary	—	is part (=1) or is not part (=0) of `TRAIN`
`TRAIN_BALANCED`	binary	—	is part (=1) or is not part (=0) of `TRAIN_BALANCED`
`TEST`	binary	—	is part (=1) or is not part (=0) of `TEST`

Table 3. Description of the information stored in the metadata file.

Data Records

All published data are fully anonymized. The data records are available online from figshare[24]. The dataset consists of twenty files holding the GRF data (see Table 2) and one file holding the metadata, including the annotations and additional subjects' information, e.g. category label, sex, body mass, etc. All files are available as comma-separated value files (CSV). The twenty GRF data files are organized according to the following naming convention: "*GRF-type-processing-side*.csv". The *type* denotes, whether the file holds the vertical ("`F_V`"), anterior-posterior ("`F_AP`"), medio-lateral ("`F_ML`") or the anterior-posterior or medio-lateral COP ("`COP_AP`", "`COP_ML`")

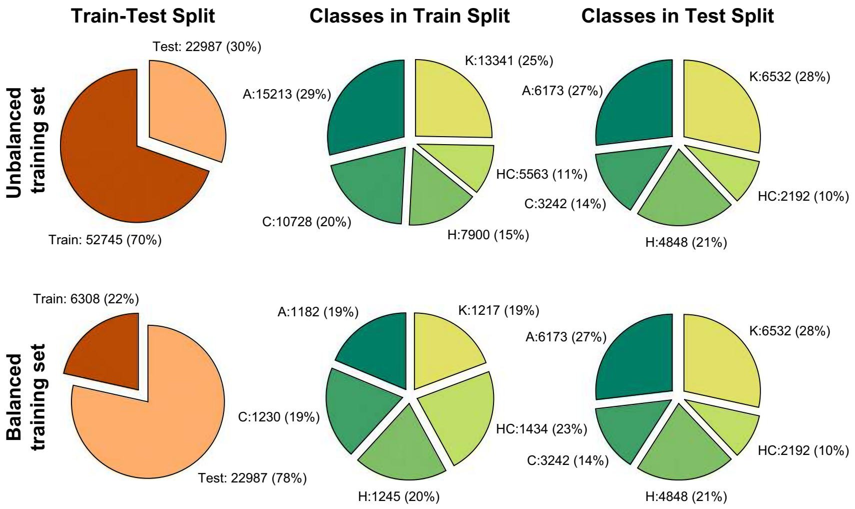


Fig. 2 Dataset composition. Configuration of the balanced and unbalanced train/test splits of the GaitRec dataset. The pie-charts show the amount of trials populated (in total amount and percentage) within each class and split.

time-series. *Processing* denotes, if the files hold the unprocessed raw data ("RAW") or the post-processed data ("PRO"). The *side* denotes, if the data are from the "left" or "right" body side. The common prefix for all files is "GRF-". An example filename is thus: "GRF_F_V_RAW_left.csv".

Each of the "*GRF-type-processing-side*.csv" files is structured as a matrix with *N* rows × *M* columns. Each row holds the data of one subject and trial. The first column identifies each subject ("SUBJECT_ID"), the second column each recording session ("SESSION_ID"), and the third column each single trial within a recording session ("TRIAL_ID"). Note that due to the non-normalized nature of the data and the resulting different vector lengths in the "RAW" files, non-available numbers have been replaced by "NaN" to maintain a constant matrix-dimension.

The metadata file, which contains annotations and additional subject-related information is available in "GRF-metadata.csv". It is structured as a matrix with *N* rows × *M* columns (see Table 3). Here, the first two columns hold the SUBJECT_ID and SESSION_ID, the other columns hold information such as class labels, sex, body mass, age, shod-condition, see Table 3 for details. Note that this information is available in all records. Missing values are identified as "NaN". A particularly notable field is "AFFECTED_SIDE", which indicates which leg is affected by a certain impairment (e.g. left knee) or if both sides are affected.

To foster comparability of classification results derived from the GaitRec dataset, we included a predefined randomized partitioning of the dataset into three subsets for training and testing. This information is stored in the metadata file. The GaitRec dataset is split into an unbalanced training set (TRAIN) and a test set (TEST). The first can be used for training and optimization of the machine learning models (e.g. by cross-validation) and the latter for the final evaluation. However, unbalanced classes might negatively affect the optimization of machine learning models, therefore we have created a balanced subset of TRAIN, referred to as TRAIN_BALANCED. The TRAIN_BALANCED subset comprises only data from initial assessments (first measurement session), which at least hold five trials for each body side per session. This is also the reason why the balanced splits populated sightly different amounts of trials. The data allocation to the different subsets was always performed on a random basis. Details of the train/test split configuration are depicted in Fig. 2.

Technical Validation

The provided data are available in raw format and post-processed with well-established de-noising and normalization procedures. This allows future researchers to either use the raw data and post-process them as desired (e.g., filtering, thresholding, normalization, etc.) or to employ the ready-to use post-processed data. The accuracy of the force plates was not specifically assessed during the data capturing period. However, the force plates and the measurement equipment has been checked and serviced regularly during clinical practice. To get a picture of the data integrity, the post-processed data are plotted in Fig. 3.

Usage Notes

The data records are stored in *.csv files and can be easily imported into any desired software package for further data analysis. The dataset also contains two scripts which allow easy data import for Matlab (The MathWorks, Inc., Natick, Massachusetts, United States, 2019a) and Python (Python Software Foundation, 3.7). Benchmarks for automatically classifying the presented data based on the first annotation level into five classes, i.e. *H vs. K vs. A vs. C*

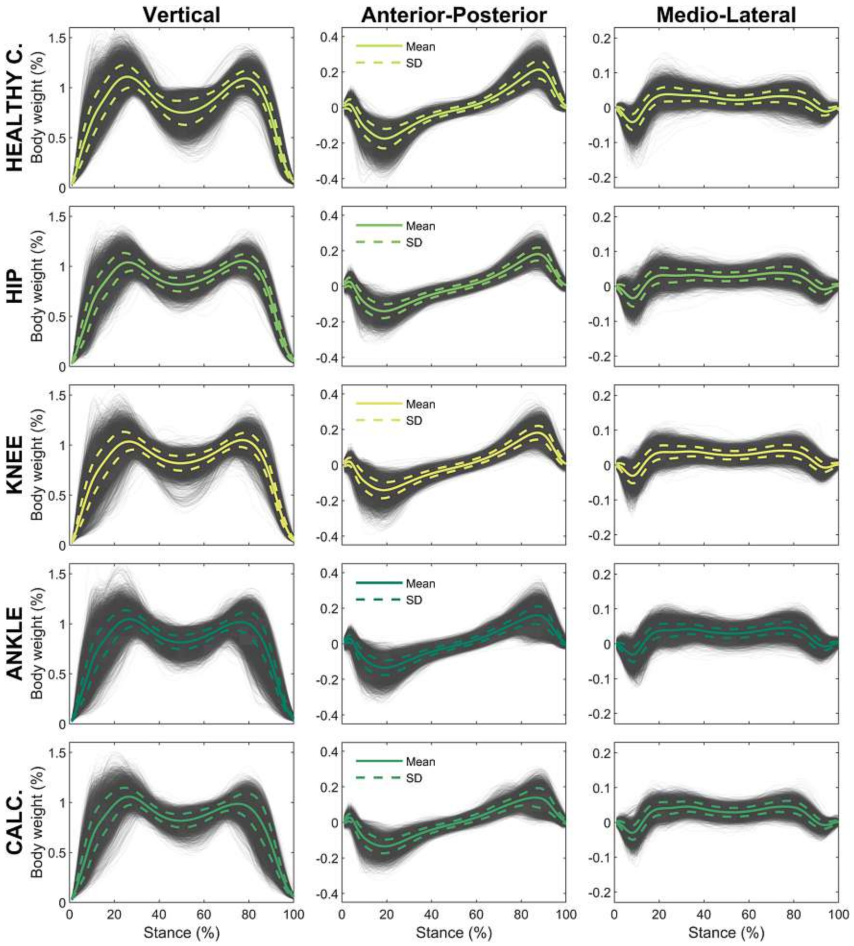


Fig. 3 Data overview. Visualization of all body-weight normalized vertical, anterior-posterior, and medio-lateral GRF signals of the affected side available per subject and class. For healthy controls all available recordings are visualized. The plots also show the mean (solid line) and its one-fold standard deviation (dotted line). Note that for easier usage the orientation of the medio-lateral and anterior-posterior signals were uniformed, so that medial and anterior forces are always represented as positive values.

vs. HC, can be found in our earlier work[13–15]. These works also provide a baseline approach that employs a signal representation based on Principal Component Analysis (PCA) combined with a Support Vector Machine (SVM) as a classifier for orientation and comparison. Note, however, that the presented dataset is an extended version of the dataset used in these studies and that results may thus slightly deviate from those of our previous studies. The studies further elaborate on the optimization of post-processing of GRF data for the purpose of gait classification.

Future work with the GaitRec dataset might focus on one of the research questions stated below. However, one should be aware that depending on the research question not all subsets of our dataset might be perfectly applicable due to their reduced sample size (i.e. for the balanced subsamples).

- Classifying healthy vs. pathological gait
- Build statistical models of normative walking
- Classify gait disorders
- Evaluation and prediction of therapy progress
- Gait data-record retrieval and similarity retrieval of trials
- Identification of subject-specific gait patterns
- Modeling dependencies between anthropometric/demographic data and the GRF signals

For the purpose of comparability of derived results from the GaitRec dataset, we highly recommend performing model optimization (e.g. by cross-validation) on the training set only and to keep the test set untouched until the final evaluation. However, it has to be noted that the train/test split does not coincide exactly with the splits in our baseline experiments because both are larger now[13–15].

Received: 20 December 2019; Accepted: 6 April 2020;
Published: 12 May 2020

References

1. Baker, R. *Measuring Walking: A Handbook of Clinical Gait Analysis* (Mac Keith Press, London, 2013).
2. Chau, T. A review of analytical techniques for gait data. Part 1: fuzzy, statistical and fractal methods. *Gait Posture* **13**, 49–66 (2001).
3. Chau, T. A review of analytical techniques for gait data. Part 2: neural network and wavelet methods. *Gait Posture* **13**, 102–120 (2001).
4. Lozano-Ortiz, C. A., Muniz, A. M. S. & Nadal, J. Human gait classification after lower limb fracture using Artificial Neural Networks and principal component analysis. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2010**, 1413–1416 (2010).
5. Zeng, W. *et al.* Parkinson's disease classification using gait analysis via deterministic learning. *Neurosci. Lett.* **633**, 268–278 (2016).
6. Vieira, A. *et al.* Software for human gait analysis and classification. In *2015 IEEE 4th Portuguese Meeting on Bioengineering (ENBENG)*, 1–1 (2015).
7. Wu, J., Wang, J. & Liu, L. Feature extraction via KPCA for classification of gait patterns. *Hum. Movement Sci.* **26**, 393–411 (2007).
8. Wu, J. & Wang, J. PCA-based SVM for automatic recognition of gait patterns. *J. Appl. Biomech.* **24**, 83–87 (2008).
9. Levinger, P., Lai, D., Begg, R. K., Webster, K. E. & Feller, J. A. The application of support vector machines for detecting recovery from knee replacement surgery using spatio-temporal gait parameters. *Gait Posture* **29**, 91–96 (2009).
10. Mezghani, N. *et al.* Automatic classification of asymptomatic and osteoarthritis knee gait patterns using kinematic data features and the nearest neighbor classifier. *IEEE T. Bio-Med. Eng.* **55**, 1230–1232 (2008).
11. Alaqtash, M. *et al.* Automatic classification of pathological gait patterns using ground reaction forces and machine learning algorithms. In *Conf. Proc. IEEE. Eng. Med. Biol. Soc.*, 453–457 (2011).
12. Ferrarin, M. *et al.* Gait pattern classification in children with Charcot–Marie–Tooth disease type 1a. *Gait Posture* **35**, 131–137 (2012).
13. Slijepcevic, D. *et al.* Automatic classification of functional gait disorders. *IEEE J. Biomed. Health* **22**, 1653–1661 (2017).
14. Slijepcevic, D. *et al.* Ground reaction force measurements for gait classification tasks: Effects of different PCA-based representations. *Gait Posture* **57**, 4–5 (2017).
15. Slijepcevic, D. *et al.* P 011–Towards an optimal combination of input signals and derived representations for gait classification based on ground reaction force measurements. *Gait Posture* **65**, 249–250 (2018).
16. Slijepcevic, D. *et al.* Input representations and classification strategies for automated human gait analysis. *Gait Posture* **76**, 198–203 (2020).
17. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
18. Brantley, J., Luu, T., Nakagome, S., Zhu, F. & Contreras-Vidal, J. Full body mobile brain-body imaging data during unconstrained locomotion on stairs, ramps, and level ground. *Sci. Data* **5**, 180133 (2018).
19. Mai, P. & Willwacher, S. Effects of low-pass filter combinations on lower extremity joint moments in distance running. *J. Biomech.* **95**, 109311 (2019).
20. Winter, D. A. *Biomechanics and Motor Control of Human Movement* (Wiley, Hoboken, NJ, 2009), 4 edn.
21. Mullineaux, D. R., Milner, C. E., Davis, I. S. & Hamill, J. Normalization of ground reaction forces. *J. Appl. Biomech.* **22**, 230–233 (2006).
22. Helwig, N. E., Hong, S., Hsiao-Wecksler, E. T. & Polk, J. D. Methods to temporally align gait cycle data. *J. Biomech.* **44**, 561–566 (2011).
23. Sangeux, M. & Polak, J. A simple method to choose the most representative stride and detect outliers. *Gait Posture* **41**, 726–730 (2015).
24. Horsak, B. *et al.* GaitRec, a large-scale ground reaction force dataset of healthy and impaired gait. *figshare*, https://doi.org/10.6084/m9.figshare.c.4788012 (2020).

Acknowledgements

This work was partly funded by the NFB - Lower Austrian Research and Education Company (NFB) and the Provincial Government of Lower Austria, Department of Science and Research (LSC14-005 and FTI17-014). We want to thank Marianne Worisch, Szava Zoltán, and Theresa Fischer for their great assistance in data preparation and their great support in clinical and technical questions.

Author contributions

B.H. and M.Z. developed the research agenda behind this work and raised the funding for this research. Both supervised the team during the entire project. B.H. and A.M.R. drafted the first manuscript of this article and coordinated the manuscript with all co-authors. MW was responsible for dataset annotation. D.S. was responsible for data cleaning, dataset construction and in creating the final files. D.J. was supported by C.S., M.W., and B.H. D.S. (post-)processed the GRF data, verified their validity in classification experiments and created the main data record files. D.S. implemented the data import scripts. All authors contributed to the writing of the manuscript and to proof-reading.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to B.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

2.2 Automatic Classification of Functional Gait Disorders

Djordje Slijepcevic, Matthias Zeppelzauer, Anna-Maria Gorgas, Caterine Schwab, Michael Schüller, Arnold Baca, Christian Breiteneder, and Brian Horsak. **Automatic Classification of Functional Gait Disorders**. *IEEE Journal of Biomedical and Health Informatics*, 22(5):1653–1661, 2017. DOI: 10.1109/JBHI.2017.2785682

The final version of this publication is available at: `https://doi.org/10.1109/JBHI.2017.2785682`. Permission for reprint granted, © 2018 IEEE

IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 22, NO. 5, SEPTEMBER 2018 1653

Automatic Classification of Functional Gait Disorders

Djordje Slijepcevic, Matthias Zeppelzauer, Anna-Maria Gorgas, Caterine Schwab, Michael Schüller, Arnold Baca, Christian Breiteneder, and Brian Horsak

***Abstract*—This paper proposes a comprehensive investigation of the automatic classification of functional gait disorders (GDs) based solely on ground reaction force (GRF) measurements. The aim of this study is twofold: first, to investigate the suitability of the state-of-the-art GRF parameterization techniques (representations) for the discrimination of functional GDs; and second, to provide a first performance baseline for the automated classification of functional GDs for a large-scale dataset. The utilized database comprises GRF measurements from 279 patients with GDs and data from 161 healthy controls (*N*). Patients were manually classified into four classes with different functional impairments associated with the "hip", "knee", "ankle", and "calcaneus". Different parameterizations are investigated: GRF parameters, global principal component analysis (PCA) based representations, and a combined representation applying PCA on GRF parameters. The discriminative power of each parameterization for different classes is investigated by linear discriminant analysis. Based on this analysis, two classification experiments are pursued: distinction between healthy and impaired gait (*N* versus GD) and multiclass classification between healthy gait and all four GD classes. Experiments show promising results and reveal among others that several factors, such as imbalanced class cardinalities and varying numbers of measurement sessions per patient, have a strong impact on the classification accuracy and therefore need to be taken into account. The results represent a promising first step toward the automated classification of GDs and a first performance baseline for future developments in this direction.**

***Index Terms*—Ground reaction force (GRF), gait classification, principal component analysis (PCA), gait parameters, machine learning.**

I. Introduction

GAIT analysis is a tool for clinicians to objectively quantify human locomotion and to describe and analyze a patient‘s gait performance. The primary aim is to identify impairments that affect a patient's gait pattern [1].

Recordings obtained during clinical gait analyses produce a vast amount of data which are difficult to comprehend and analyze due to their high-dimensionality, temporal dependences, strong variability, non-linear relationships and correlations within the data [2]. This makes data interpretation challenging and requires an experienced clinician to draw valid conclusions. Several automatic analysis approaches based on machine learning have been published in recent years to tackle these problems and to support clinicians in identifying and categorizing specific gait patterns into clinically relevant categories [2], [3]. Machine learning methods employed in this context comprise neural networks [4]–[6], support vector machines (SVMs) [7]–[9], nearest neighbor classifiers [10], [11], and different clustering approaches (hierarchical, k-means, etc.) [12]. The performance of such methods strongly depends on the input data representation [13]. Frequently used representations in gait analysis comprise discrete kinematic gait parameters (e.g. local minima and maxima of gait signals and time-distance parameters) [11], [14], [15]. Additionally, previous research has shown that global signal representations obtained by principal component analysis (PCA) [16], [17], kernel-based PCA (KPCA) [18], [19] and discrete wavelet transformation (DWT) [10], [11] are suitable for subsequent classification [10], [16].

Typical use cases for automatic gait analysis described in the literature show a moderate to high accuracy in distinguishing between different pathologies or patient groups [4], [7]–[9], [11], [16], [17]. However, most of the existing literature investigated rather simple cases such as the differentiation between the affected/non-affected limb in hemiplegic patients [20], and the distinction of healthy gait from people with neurological disorders [5], [11], transfemoral amputation [16], and lower limb fractures [4], [17]. A more complex study is presented in [21], where several disorders associated with traumatic brain injuries are classified. The majority of published articles employed kinematic and kinetic data derived from three-dimensional gait

Manuscript received June 28, 2017; revised October 30, 2017; accepted December 11, 2017. Date of publication December 20, 2017; date of current version August 31, 2018. This work was supported by the NFB Lower Austrian Research and Education Company and the Provincial Government of Lower Austria, Department of Science and Research (LSC14-005). *(Corresponding author: Djordje Slijepcevic.)*

D. Slijepcevic and M. Zeppelzauer are with the ICMT Institute of Creative Media Technologies, St. Pölten University of Applied Sciences, St. Pölten 3100, Lower Austria, Austria (e-mail: djordje.slijepcevic@fhstp.ac.at; matthias.zeppelzauer@fhstp.ac.at).

A.-M. Gorgas, C. Schwab, and B. Horsak are with the St. Pölten University of Applied Sciences, St. Pölten 3100, Austria (e-mail: Anna-Maria.Gorgas@fhstp.ac.at; Caterine.Schwab@fhstp.ac.at; brian.horsak@fhstp.ac.at).

M. Schüller and A. Baca are with the University of Vienna, Vienna 1010, Austria (e-mail: schueller.michael@gmx.at; arnold.baca@univie.ac.at).

C. Breitender is with the TU Wien, Vienna 1040, Austria (e-mail: christian.breiteneder@tuwien.ac.at).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JBHI.2017.2785682

analysis (3DGA), which provide a vast amount of kinematic and kinetic information for multiple joints. Drawbacks of such 3DGA measurement systems are the relatively time-consuming data recording, the need for highly trained staff as well as high acquisition and maintenance costs. Therefore, such analysis tools are often not suitable for daily use in clinical practice.

To manage the high patient throughput in rehabilitation centers, a frequently used approach is to combine simple visual inspection or 2D video recordings with the quantification of ground reaction forces (GRF) by force platforms, as changes in the morphology of the GRF waveforms reflect pathological gait [11], [17]. One major drawback of this approach is the loss of clinically relevant and quantifiable information (e.g. gait kinematics), causing a potential decrease in classification accuracy [22]. However, such simple approaches are common in clinical practice as they overcome the before-mentioned limitations of 3DGA. To date, few attempts have been published that use only GRF data for automated gait pattern classification [16], [23]. Most of these gait classification approaches show promising results. However, the majority of previous works employed relatively small datasets. Alaqtash *et al.* [11], for example, compared the data of 12 healthy adults to those of patients with cerebral palsy and multiple sclerosis (4 patients each), Muniz and Nadal [17] used data from 38 healthy controls and 13 patients with lower limb fracture, and Soares *et al.* [16] classified GRF data of 20 able-bodied and 12 patients with transfemoral amputation. Such small datasets make it difficult to train robust and reliable classifiers that are applicable in complex real-world scenarios. Furthermore, a majority of studies [10], [17], [23] relies solely on the vertical ground reaction force for classification purposes, rather than considering all available GRF components, including the center of pressure (COP), for a more conclusive picture of the underlying gait pattern. Previous classification attempts mainly focused on the differentiation between specific diseases rather than drawing a distinction between functional gait disorders. The work of Köhle and Merkl [24], [25], who clustered and classified GRF measurements into deficits of different body regions, represents an exception in this regard. Their dataset was about half of the size of the one presented in this article and their work also focused on patients walking with a prosthesis. In this article we define a functional gait disorder as the cause of a gait impairment, which is reflected by the individual gait patterns. These may be associated with a patient's condition after joint replacement surgery, fractures, ligament ruptures, osteoarthritis or related disorders. The classification of functional gait disorders is of particular interest in clinical examinations, as it may play a key role in detecting arthropathies or diseases at an early stage. In addition, such a classification may also indicate secondary disorders that otherwise might be easily overlooked during clinical examination.

The aim of this article is to present a detailed investigation of the automated classification of several functional gait disorders solely based on GRF data. The presented approach builds upon the aforementioned studies, e.g. [16], [17], [23], investigates the suitability of frequently used state-of-the-art GRF parameterization techniques for gait classification and analyzes their discriminative power. In the experiments we evaluate the individual representations on a large-scale and real-world dataset for different classification tasks. This paper therefore presents a first performance baseline for the automatic classification of different gait disorders in a real-world setting.

II. Material and Methods

A. Patients and Dataset

The presented retrospective study was approved by the local Ethics Committee of Lower Austria (GS1-EK-4/299-2014). The anonymized data used in this study are part of an existing clinical gait database maintained by a rehabilitation center of the Austrian Workers' Compensation Board (AUVA). The AUVA is the social insurance for occupational risks for more than 3.3 million employees and 1.4 million pupils and students in Austria. The utilized database comprises GRF measurements from 279 patients with gait disorders (GD) and data from 161 healthy controls (N), both of various physical composition and gender (see Table I for details on the dataset). Patients were manually classified into four classes - calcaneus "C" (n = 82), ankle "A" (n = 62), knee "K" (n = 69), and hip "H" (n = 66) - by a physical therapist, based on the available medical diagnosis of each patient. Thus, GD refers to $C \cup A \cup K \cup H$. The individual GD classes include patients after joint replacement surgery, fractures, ligament ruptures, and related disorders associated with the above-mentioned anatomical areas. The most common injuries present in the hip class are fractures of the pelvis and thigh as well as luxation of the hip joint, coxarthrosis, and total hip replacement. The knee class comprises patients after patella, femur or tibia fractures, ruptures of the cruciate or collateral ligaments or the meniscus and total knee replacements. The ankle class includes patients after fractures of the calcaneus, malleoli, talus, tibia or lower leg, and ruptures of ligaments or the achilles tendon. The calcaneus class comprises patients after calcaneus fractures or ankle fusion surgery. All of the above-mentioned injuries may occur individually or in combination within each class.

Each patient performed one or several measurement sessions. In each session, eight recordings for two consecutive steps were performed. Each bilateral recording is referred to as one trial in this paper. Thus, the utilized dataset contains 1,187 sessions comprising 9,496 individual trials (see Table I for details).

B. Data Recording and Preprocessing

Gait analysis was performed on a 10 m walkway with two centrally embedded force plates (Kistler, Type 9281B12). The force plates were placed in a consecutive order, allowing a person to walk across by placing one foot on each plate. Both plates were flush with the ground and covered with the same walkway surface material, so that targeting was not an issue. During a session, participants walked unassisted and without a walking aid at a self-selected walking speed until a minimum of eight valid recordings were available. These recordings were defined as valid by a supervisor when the participant walked naturally and there was a clean foot strike on the force plate. Prior to

TABLE I
DETAILS OF THE DATASET AND CLASSES

Class	Amount	Age (yrs.) Mean ± SD	Body Mass (kg) Mean ± SD	Sex (m/f)	Num. Sessions	Num. Trials
Healthy Control (N)	161	32.4 ± 13.6	74.1 ± 16.2	84/77	161	1,288
Calcaneus (C)	82	42.4 ± 9.9	84.5 ± 12.1	74/8	320	2,560
Ankle (A)	62	40.0 ± 11.5	88.3 ± 16.9	56/6	259	2,072
Knee (K)	69	41.5 ± 11.4	83.7 ± 19.6	44/25	258	2,064
Hip (H)	66	43.6 ± 14.7	81.6 ± 18.3	53/13	189	1,512
SUM	**440**	**38.4** ± **13.3**	**80.7** ± **17.3**	**311/129**	**1,187**	**9,496**

the gait analysis session, each participant underwent rigorous physical examination by a physician.

All processing steps and subsequent analyses were performed in Matlab 2016a (The MathWorks Inc., Natick, MA, USA). The three analog GRF signals as well as the two COP signals were converted to digital signals using a sampling rate of 2000 Hz and a 12-bit analog-digital converter (DT3010, Data Translation Incorporation, Marlboro, MA, USA) with a signal input range of ± 10 V. A threshold of 10 N was used for step detection and 30 N for COP calculation. Raw signals were filtered using a 2nd order low-pass butterworth filter with a cut-off frequency of 20 Hz. All gait measurements were temporally aligned so that they all started with the initial contact and ended with toe-off. They were further time-normalized to 100% stance by re-sampling the data to 1000 points. The processed signals are referred to as waveforms in this article. Amplitude values of the three force components, i.e. vertical (V), medio-lateral (ML), and anterior-posterior (AP), were expressed as a multiple of body weight (BW) by dividing the force by the product of body mass times acceleration due to gravity (g). The COP waveforms from each trial were normalized by the foot length (FL) determined during each session, expressed as a multiple of foot length.

C. Signal Representation

The representations employed in our investigation comprise (1) discrete GRF parameters (DP) in combination with time-distance parameters (TDP) [11], [14], [15]; (2) PCA-based parameterizations of the entire GRF waveforms [4], [8], [16] and (3) a combination of the first two approaches, i.e. PCA applied to DPs and TDPs [7]. In the following, all three approaches are described in detail.

DPs were calculated for the affected limb and extracted from all three force components, $F_V(t)$ (vertical), $F_{AP}(t)$ (anterior-posterior), and $F_{ML}(t)$ (medio-lateral), as well as from the COP displacement in the anterior-posterior (walking) direction $COP_{AP}(t)$ and in the medio-lateral direction $COP_{ML}(t)$. An example of the GRF and corresponding COP waveforms is presented in Fig. 1. Furthermore, a more detailed visualization of the mean GRF waveforms over each class is illustrated in Fig. S1 (supplementary material). DPs include a set of predefined (most prominent) local minima and maxima of the waveforms, which were extracted by peak detection in a fully automatic way from each trial. Furthermore, impulses were calculated over different segments of the waveform by multiplying the average force (in N) by the time this force is active. To account for differences in body mass between participants [26], all impulses were divided by the product of body mass times acceleration due to gravity (g) and then multiplied by 100 (%BW·s). TDPs such as cadence (CAD), double support time (DS), gait velocity (GV), step length $(STEPLEN)$, and stance time (ST) were calculated from two consecutive steps (affected and unaffected limb) and averaged over the eight valid trials. Table II lists all 52 extracted parameters.

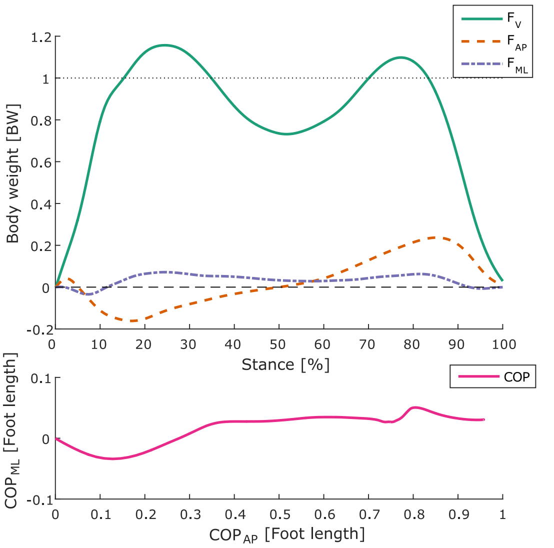


Fig. 1. (Top) The characteristic shape of the three components of the GRF: the vertical force (F_V), the anterior-posterior shear (F_{AP}), and the medio-lateral shear (F_{ML}). (bottom) The corresponding COP path for one step. Note that x and y axes are scaled slightly differently for better visualization.

In contrast to the GRF parameters (DPs and TDPs), the PCA takes the entire waveforms[1] of the affected limb into account and provides a holistic representation of the data. Complementary information to the parameters is thus captured. The main goal of PCA is to reduce the dimensionality of a dataset by transforming the data into a set of uncorrelated variables, i.e. the principal components (PCs) [27]. Each PC points in (and thus explains) one orthogonal direction of variance in the data.

[1] For the purpose of the present study, every third sample was used in order to reduce redundancy in the data, thereby improving the robustness of the decomposition.

TABLE II
Discrete and Time-Distance Parameters, Description, Type of Normalization and Physical Unit

Abbreviation	Description	Normalization	Unit
ST	Stance time is the duration of the stance phase of one foot	–	s
F_{V1}	Maximum value of F_V within the breaking phase of stance	Body weight	BW
T_{V1}	Time of F_{V1}	Stance time	%ST
F_{V2}	Minimum value of F_V between T_{V1} and T_{V3}	Body weight	BW
T_{V2}	Time of F_{V2}	Stance time	%ST
F_{V3}	Maximum value of F_V within the propulsive phase of stance	Body weight	BW
T_{V3}	Time of F_{V3}	Stance time	%ST
F_{AP1}	Maximum value of F_{AP} between initial contact and T_{AP2}	Body weight	BW
T_{AP1}	Time of F_{AP1}	Stance time	%ST
F_{AP2}	Minimum value of F_{AP} within the breaking phase of stance	Body weight	BW
T_{AP2}	Time of F_{AP2}	Stance time	%ST
F_{AP3}	Maximum value of F_{AP} within the propulsive phase of stance	Body weight	BW
T_{AP3}	Time of F_{AP3}	Stance time	%ST
F_{ML1}	Minimum value of F_{ML} within the breaking phase of stance	Body weight	BW
T_{ML1}	Time of F_{ML1}	Stance time	%ST
F_{ML2}	Maximum value of F_{ML} within the breaking phase of stance	Body weight	BW
T_{ML2}	Time of F_{ML2}	Stance time	%ST
F_{ML3}	Maximum value of F_{ML} within the propulsive phase of stance	Body weight	BW
T_{ML3}	Time of F_{ML3}	Stance time	%ST
F_{VAVG}	Mean value of F_V	Body weight	BW
F_{APAVG}	Mean value of F_{AP}	Body weight	BW
F_{MLAVG}	Mean value of F_{ML}	Body weight	BW
IF_V	Impulse of F_V during stance	Body weight	%BW·s
IF_{AP}	Impulse of F_{AP} during stance	Body weight	%BW·s
IF_{ML}	Impulse of F_{ML} during stance	Body weight	%BW·s
IF_{V1}	Impulse of F_V between initial contact and T_{V1}	Body weight	%BW·s
IF_{V2}	Impulse of F_V between initial contact and T_{V2}	Body weight	%BW·s
IF_{V3}	Impulse of F_V between initial contact and T_{V3}	Body weight	%BW·s
IF_{APDEC}	Impulse of F_{AP} during the breaking phase	Body weight	%BW·s
IF_{APACC}	Impulse of F_{AP} during the propulsive phase	Body weight	%BW·s
IF_{LAT}	Impulse of the lateral component of F_{ML}	Body weight	%BW·s
IF_{MED}	Impulse of the medial component of F_{ML}	Body weight	%BW·s
$COPANG$	COP angle is the horizontal angle between the COP linear regression line and the x-axes ($\neq$ foot rotation)	–	deg
$COPDEV$	COP deviation is the root mean square error of the COP linear regression	Foot length	FL
COP_{AP}	COP range is the range in the anterior-posterior direction during stance phase	Foot length	FL
$COPV$	COP velocity is calculated as the ratio of foot length and stance time	Foot length	FL/s
COP_{ML}	COP range is the range in the medio-lateral direction during stance phase	Foot length	FL
$DECT$	Deceleration time (breaking phase) is the duration of F_{AP} being negative	–	s
$ACCT$	Acceleration time (propulsive phase) is the duration of F_{AP} being positive	–	s
$LR0080$	Loading rate represented as the slope of F_V from the initial contact to 80% of F_{V1}	Body weight	N/s
$LR2080$	Loading rate represented as the slope of F_V from 20% to 80% of F_{V1}	Body weight	N/s
$UR8000$	Unloading rate represented as the slope of F_V from 80% of F_{V3} to the toe-off	Body weight	N/s
$UR8020$	Unloading rate represented as the slope of F_V from 80% to 20% of F_{V3}	Body weight	N/s
DS	Double support time during one stride	–	s
$STEPLEN$	Step length is the distance of the COP position from initial contact to following contralateral initial contact	–	m
$STEPV$	Step velocity is calculated as the ratio of step length and step time	–	km/h
$STRIDET$	Stride time is the duration from initial contact to initial contact of the ipsilateral foot	–	s
BF	Basic frequency is the mean number of strides per second ($1/STRIDET$)	–	Hz
CAD	Cadence is the number of steps per minute	–	1/min
$STEPWD$	Step width is the medio-lateral distance of the mean COP between both feet	–	m
$STRLEN$	Stride length is the distance of the COP position from initial contact to following ipsilateral initial contact	–	m
GV	Gait velocity is calculated as the mean step velocity of both feet	–	km/h

Body weight (BW): product of body mass and acceleration due to gravity;
%ST: percentage of stance time; %BW: percentage of body weight; FL: multiple of foot length.

The main intention is to obtain a lower-dimensional representation of our time- and weight-normalized waveforms similar to [4], [8], [16] by projecting the data onto those PCs which explain most variance in the data. This dimensionality reduction fosters subsequent machine learning [3]. We performed PCA on each of the five signals separately and concatenated the resulting PCs to obtain a feature vector for classification. This approach proved to be superior to other PCA-based representations in a preliminary study [28]. The final dimensionality of the obtained representations is specified by the amount of variance preserved in a particular projection, i.e. 98%, 95%, and 90%. An exemplary visualization of the different PCA representations is presented in Fig. S2 (supplementary material). A preliminary evaluation indicated that preserving 98% of the variance results in a good trade-off between data reduction and classification performance. Thus, all results presented in the following are based on the approach in which 98% of the variance is preserved (the number of resulting PCs is waveform-specific and ranges from four to twelve, i.e. for all five signals there are 39 PCs in total).

As a third representation, PCA was applied to the previously extracted DPs and TDPs (a vector comprising of 52 parameters), similarly to Wu *et al.* [7]. This approach combines both methodologies and aims at extracting the most important information from the (possibly redundant) parameters.

D. Statistical Analysis

Our first aim was to investigate the suitability of different parameterization techniques for subsequent gait classification. For this purpose we analyzed the variance and discriminative power of each DP and TDP across the different classes by descriptive statistics in a first step. We calculated the median, interquartile-range (IQR) and range of each parameter within each class and visualized them as boxplots. This enabled us to visually inspect variances and distributions in and across the classes, thereby allowing a first estimation of the discriminative power of each parameter.

In a second step, we investigated the discriminative power of the parameters and the global PCA-based representations by linear discriminant analysis (LDA). A natural measure to describe the separation of two distributions (classes) is the Fisher criterion, which represents the core of LDA [29]. We applied (multi-class) LDA to assess the discriminative power of individual parameters for two (or more) classes. The advantage of this approach is that the discriminative power of a parameter (even across multiple classes) can be expressed by one scalar value that directly reflects the statistical properties of the input data. Hence, there is no need to apply additional modelling and data transformations (which may influence results) prior to LDA, which would be necessary for other methods such as SVM. Furthermore, this approach can easily be extended to estimate the discriminative power of a combination of several parameters by multi-dimensional LDA (e.g. in case of PCA-based representations). We computed the accuracy of LDA and reported the divergence from a random baseline [30] to quantify to which degree an input parameter or input representation is able to separate the underlying classes. The random baseline was estimated by the zero rule (always choosing the most frequent class in the dataset). Thus, in the case of five classes where the largest class contains 30% of the data the random baseline equals 30%.

E. Classification

We applied two classification tasks to the dataset by using SVMs as classifiers: (1) (binary) classification between normal gait and all gait disorders (N/GD) and (2) (multi-class) classification between N and each of the four GD classes ($N/C/A/K/H$). In the first task, the class priors are imbalanced, i.e. there are many more observations in the combined GD class than in the normal class (see Table I). The second task separates each type of disorder from each other and from the normal class.

For the classification experiments the dataset was split into a training (65%) and a test set (35%), thereby mutually disjoining the groups of patients in both sets. The training dataset in combination with a k-fold cross-validation approach served to train the classifiers and to optimize their parameters (model selection), whereas the test dataset was used to evaluate the generalization ability of the trained models (and was not considered during model selection and hyper-parameter optimization). The calculated DPs and TDPs as well as the PCA-based representations served as input to classification. The parameters (DPs and TDPs) were normalized (each independently) in a twofold way, by min-max normalization and z-standardization, in order to determine the more suitable approach. The PCA representations were z-standardized. We employed SVMs for the classification with linear and radial basis function (RBF) kernels, provided by the LIBSVM library [31]. For hyper-parameter selection we applied a grid search over the regularization parameter $C \in [2^{-5}, 2^{15}]$ for the linear SVM and over $C \in [2^{-1}, 2^{15}]$ and the kernel hyper-parameter $\gamma \in [2^{-15}, 2^{5}]$ for the RBF SVM. During the grid search, a 5-fold cross-validation was performed on the training dataset. Finally, an SVM with the best parameters estimated during model selection was trained on the entire training set and evaluated on the test set. Additionally a k-nearest neighbor (k-NN) classifier and a multi-layer perceptron (MLP) were employed as a reference to compare their results to the performance of the SVM. Grid search was performed over various values of k for the k-NN. For the MLPs different numbers and sizes of hidden layers were employed. As a performance measure we use the classification accuracy, which is the percentage of correct classifications among all classes and input samples. Since in different experiments the random baseline varies, the absolute values of accuracy are of limited expressiveness. To enable a fair comparison, we employ the *divergence from a random baseline* approach [30] and thus provide for each experiment the difference between the random baseline and the absolute classification accuracy.

III. Results and Discussion

This section presents and discusses the results of the statistical analysis and the classification experiments.

A. Statistical Analysis

The statistical analysis aimed at assessing the suitability of the individual GRF parameters (DPs and TDPs) for distinguishing different classes of gait disorders. In order to be considered a ”good” parameter, intra-class variation should be low (e.g. small IQR inside a given class), while the inter-class variation should be high (e.g. significantly different means or medians between the samples of different classes) [15].

The visual inspection of the boxplots for each parameter enables a first assessment of the intra- and inter-class variation and thereby gives an impression of the parameters' potential to differentiate between different classes. Fig. 2 shows boxplots for selected parameters. A presentation of boxplots for all 52 investigated parameters for all classes is provided in Fig. S3 (supplementary material). Parameters such as F_{V3} (see Fig. 2(a)) show a clear difference in the median and the IQR between the healthy controls and all four GD classes. This indicates a high potential to discriminate between normal gait and arbitrary gait disorders. However, the overlap of the distributions within the GD classes indicates a low potential to discriminate between them.

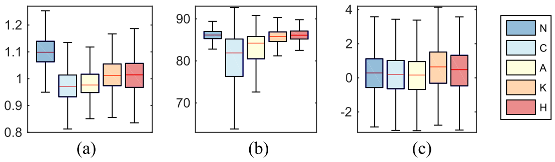


Fig. 2. Example of boxplots for three parameters. Each boxplot shows the median and the IQR (box) for each class (outliers were removed for better visualization). Box-whiskers correspond to 1.5 of the box-length, thus show approximately ± 2.7 standard deviations. The overlap of distributions between the classes gives an impression of the parameters' discriminative power. (a) F_{V3} [BW]. (b) T_{AP3} [%ST]. (c) IF_{AP} [%BW·s].

Other parameters such as T_{AP3} (see Fig. 2(b)) vary strongly in the IQR and the median across the classes. While the IQR is high for calcaneus and ankle, the classes hip, knee, and the normal controls exhibit a very similar distribution. Thus, such a parameter has solely limited potential to separate normal gait from general gait disorders. There may be, however, a certain potential to separate individual classes (in this case calcaneus) from other classes. Other parameters may lack in discriminative power. An example is IF_{AP} (see Fig. 2(c)), which shows a similar median and overlapping distributions with a similar IQR across all classes. Several parameters are discriminative for particular classes or a group of classes. However, none of the observed parameters discriminates well between all classes. Therefore, the combination of several parameters for the distinction between classes seems advisable. These assumptions are further corroborated by the LDA results.

LDA was applied to the individual parameters and their combination, as well as to the higher-dimensional PCA-based representations. This analysis aimed at quantifying the discriminative power of the investigated representations and thereby evaluating their suitability for automated classification. Fig. 3 illustrates discriminativity scores obtained by LDA in terms of deviation from the random baseline (*zero rule*). In detail, results for different combinations of classes (rows) are illustrated: rows 1-4 provide results for the discrimination of normal gait vs. ankle, calcaneus, hip or knee (each class separately). Row 5 shows how well all 5 classes can be differentiated from each other. Row 6 illustrates how well normal gait can be differentiated from all types of gait disorders. Rows 7–12 show how all possible pairs of gait disorder classes can be differentiated from each other. Positive discriminativity scores are represented by a color scale from blue (corresponding to low values) to yellow (representing high values), whereas negative values are colored in gray. Positive values mean that the random baseline is exceeded and that the respective input parameter or input representation exhibits a certain discriminative power (the higher the value the better). Negative values indicate the absence of discriminative power, i.e. the random baseline is not reached. It has to be noted that, since the different class partitions represented by the rows of Fig. 3 have different random baselines, the values across rows cannot be compared directly. Comparisons are solely valid along the rows. In general, however, columns including a larger number of high values indicate parameters or representations with a higher discriminative power. Similarly, rows with higher values represent tasks that are easier to solve than others.

The leftmost part of Fig. 3 illustrates the discriminativity scores for the individual parameters. Several parameters achieve high scores for individual classes or combinations of classes, e.g. F_{AP3}, F_{V3}, F_{VAVG}, F_{V1}, T_{V3}, T_{AP3}, GV, $STEPV$, DS, $STRLEN$, F_{V2}, $STEPWD$, CAD, BF, and $STRIDET$. No parameter, however, performs well across all tasks. This indicates that individual parameters are quite limited in expressiveness. The second part (ALL_PARAMS) of Fig. 3 illustrates the results from the combination of all parameters. The combination yields much better discrimination across all rows of Fig. 3. This demonstrates that the individual parameters contain complementary information and attain synergies when they are combined. The third part of Fig. 3 visualizes the results of the PCA-based representations of the five input signals F_V, F_{AP}, F_{ML}, COP_{AP}, and COP_{ML}. The three GRF components achieved higher scores compared to the COP signals. The rightmost part of Fig. 3 shows the discriminativity scores for combined PCA representations, i.e all three GRF components combined (PCA_F_{ALL}), both COP signals combined ($PCA\_COP_{AP,ML}$), and all five components combined (PCA_ALL). In general, the combination of components improved the results, which indicates that the individual GRF components are complementary to each other. The addition of the COP further improved the discriminative power. Thus, adding COP to a classification may contribute positively to the results. The representations (PCA_ALL and ALL_PARAMS) are combined able to contribute to all evaluated tasks (rows) of Fig. 3.

The evaluated representations are more suitable for differentiating between the healthy control group and a functional gait disorder (rows 1-4) than between two functional gait disorders (rows 7-12). Regarding the task N/GD, solely a few parameters are able to exceed the random baseline. This is due to the fact that the combined set of all gait disorders contains much more samples than the class of healthy controls (i.e. 279 vs. 161 samples). This yields a random baseline around 87.1% which is more difficult to exceed than random baselines in other tasks.

B. Classification

The results of the classification experiments, which were performed on data from the test set, are summarized in Table III. The test set was not presented to the classifier during the training phase and the selection of its parameters. Results are provided for the two classification tasks ($N/C/A/K/H$ and N/GD) and for three different parameterizations. The results of the additional experiments with other classifiers such as the multi-layer perceptron (MLP) and the k-nearest neighbors algorithm (k-NN) were all outperformed by the SVM results, which confirms also the results of Janssen, Schöllhorn *et al.* [32]. Therefore, and due to the limited space available, these results will not be discussed in detail.

The first evaluated parameterization comprises of 52 GRF parameters (DPs and TDPs) that are extracted from all five GRF input signals. Due to the strong variation in the parameters' value

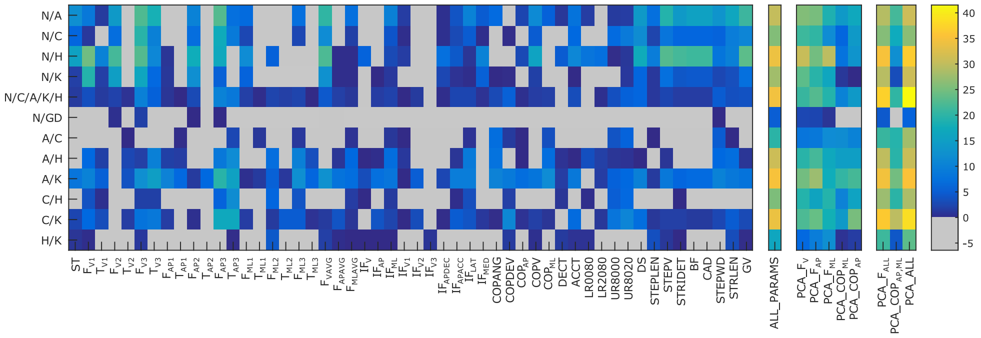


Fig. 3. Discriminativity scores obtained by LDA for different selections of classes (rows). The figure is divided into four blocks. Each column represents a different input parameter or higher-dimensional input representation. Best viewed in color.

TABLE III
CLASSIFICATION RESULTS (%) OF TWO TASKS - $N/C/A/K/H$ AND N/GD - AND THREE DIFFERENT PARAMETERIZATION APPROACHES

Parameterization	Norm.	Dim.	$N/C/A/K/H$ (RB: 31.8%)		N/GD (RB: 87.1%)	
			linear SVM	RBF SVM	linear SVM	RBF SVM
GRF Parameters (DPs and TDPs)	z-score	52	15.0 (46.8)	8.8 (40.6)	2.4 (89.5)	−0.8 (86.3)
GRF Parameters (DPs and TDPs)	min-max	52	14.3 (46.1)	9.5 (41.3)	1.6 (88.7)	−3.8 (83.3)
PCA on F_V, F_{AP}, F_{ML}	z-score	30	19.8 (51.6)	15.4 (47.2)	2.4 (89.5)	**2.0 (89.1)**
PCA on F_V, F_{AP}, F_{ML}, COP_{AP}, COP_{ML}	z-score	39	**22.5 (54.3)**	**19.4 (51.2)**	**3.7 (90.8)**	1.9 (89.0)
PCA on z-standardized GRF parameters	z-score	28	13.8 (45.6)	8.8 (40.6)	2.6 (89.7)	−0.6 (86.5)
PCA on min-max normalized GRF parameters	z-score	28	13.5 (45.3)	7.9 (39.7)	2.8 (89.9)	0.1 (87.2)

Note that the random baseline (RB) is stated next to the task name and that the values in the table represent the deviation from the random baseline (RB) and the corresponding absolute accuracy in brackets.

ranges, a suitable normalization of the data is essential. We evaluated min-max normalization as well as z-standardization. The use of z-standardization resulted in a slightly higher deviation from the RB for both tasks (except for the RBF SVM in task $N/C/A/K/H$) compared to min-max normalization. Furthermore, the RBF SVM failed to exceed the random baseline for both methods in the task N/GD.

The second parameterization was obtained by PCA of the raw GRF waveforms. PCAs obtained solely from the three force components clearly outperform the GRF parameters (DPs and TDPs). By adding the COP measurements the results were further improved for both tasks. Normalization of the PCA-based representations is crucial as performance otherwise drops significantly.

The third parameterization applied PCA on the z-standardized and min-max normalized DPs and TDPs. The dimensionality reduction resulted in a 28-dimensional vector which was also z-standardized prior to classification. In this case, results for both normalizations (last two rows of Table III) were improved for the task N/GD compared to the representation with the original GRF parameters (first two rows of Table III). However, this is not the case for task $N/C/A/K/H$, where the deviation from the RB slightly decreased.

In summary, the best performance (marked in bold in Table III) was achieved by applying PCA to all five GRF signals. The linear SVM achieved the highest deviation from the RB (22.5%) for task $N/C/A/K/H$ as well as for task N/GD (3.7%). Alternative classifiers which were also evaluated yielded a lower deviation from the RB: MLP 21.0% and k-NN 13.4% for task $N/C/A/K/H$ and MLP 2.6% and k-NN 2.2% for task N/GD. In terms of accuracy and deviation from the RB, the linear SVM performed better in all experiments. The RBF SVM has an advantage solely in terms of runtime.

The main reason for the great difference in the performance between the two tasks is the strong class imbalance in task N/GD, which makes this task particularly difficult to solve. One way of dealing with unbalanced datasets in SVMs is the use of different weights for different classes, thereby emphasizing the importance of the under-represented classes. Therefore, additional class-weighted experiments were performed. Results with different cost functions showed that no further performance increase can be achieved. The uniform cost function seems to work best on the data.

C. Discussion and Further Aspects

We presented a study on the classification of different functional gait disorders, stemming from a wide range of possible impairments, into categories that represent the main affected body region. The motivation for selecting these broad categories is that identifying the region of impairment is essential for clinical practice and may allow to pinpoint impairments already at an early stage. In addition, it could indicate secondary impairments which may easily be overlooked by the physician during clinical examination. The present study represents a first

TABLE IV
Results (%) of Analyses Assessing the Influence of Several Factors on the Results of the Two Tasks - $N/C/A/K/H$ and N/GD

Partitions of the dataset	$N/C/A/K/H$	N/GD
Session are balanced	23.7 (60.2)	20.6 (84.1)
Persons are balanced	28.3 (59.5)	5.3 (84.7)
Persons & sessions are balanced	39.2 (59.2)	35.4 (85.4)
Male population	20.9 (51.3)	0.6 (91.4)

The experiments are performed with a PCA on all five signals in combination with a linear SVM. Note that the values represent the deviation from the random baseline (RB) and the corresponding absolute accuracy in brackets.

performance baseline for the classification of gait disorders. Results are particularly promising for task N/GD. However, an absolute classification accuracy of 91% still lies below an acceptable threshold for clinical practice. For the classification of individual disorder categories, the results indicate that further improvements are necessary. To date, the proposed approach could, however, already serve as a support for clinicians indicating the presence of (additional) arthropathies or diseases. In order to reduce the classification complexity, while still providing support for clinicians, similar classes could be merged, i.e. the hip and knee classes into a thigh class and the ankle and calcaneus classes into a shank class. The results of this additional experiment showed a deviation from the RB of 26.8% (using a linear SVM, RB: 51%, absolute accuracy: 77.8%). Compared to the distinction of all five classes ($N/C/A/K/H$), this is a clear increase in accuracy and deviation from the RB.

Different influencing factors, i.e. the imbalance of the class priors, the variability in the number of sessions per person and gender-specific aspects may introduce a bias into the aforementioned analyses. To investigate the effect of these factors on classification performance, we performed additional experiments. To this end, we used the best configuration found so far as a baseline, i.e. PCA on all five signals with a linear SVM (4th parametrization in Table III) and applied it to different balanced subsets of our dataset. The results are presented in Table IV and are discussed in the following.

For the experiments in Section III-B we decided to use all available sessions of patients recorded in the course of their rehabilitation to account for different progression stages of impairments. This, however, may introduce a bias in the experiments as more trials exist for some patients than for others. To evaluate to which extent the varying number of recorded sessions per patient influences the overall result, we balanced the dataset by selecting only one random session per person. Interestingly, the deviation from the RB improved for task $N/C/A/K/H$ to 23.7% (+1.2%) and for task N/GD to 20.6% (+16.9%), as presented in the first row of Table IV. These results show that intra-patient variability needs to be taken into account and requires additional modeling in a classification approach.

Another factor causing an imbalance in the data are the different class cardinalities, i.e. different numbers of persons per class. In order to investigate the influence of this imbalance we performed an experiment for both tasks with a dataset containing the same number of participants per class (but keeping all sessions in the dataset). For task $N/C/A/K/H$ the balanced dataset is composed of data from 62 persons from each class (overall 310 persons, 7616 trials). For task N/GD the balanced dataset contained data from 160 healthy controls and 160 persons with a deficit (40 from each GD class, overall 320 persons and 6096 trials). The deviation from the RB improved for task $N/C/A/K/H$ to 28.3% (+5.8%) and for task N/GD to 5.3% (+1.6%), as shown in the second row of Table IV. Although the results show that balancing the number of patients among classes is beneficial, the results of task N/GD reveal the still existing imbalance in the dataset (due to the fact that healthy controls have only one session and patients up to several sessions).

The next question deals with the effect of balancing the number of patients and the number of sessions at the same time. We performed experiments with a completely balanced version of our dataset for each task, containing only one session per person and equal numbers of persons per class. For task $N/C/A/K/H$ the balanced dataset is composed of data from 62 persons from each class (overall 310 persons, 2480 trials). For task N/GD the balanced dataset contained data from 160 healthy controls and 160 persons with a deficit (40 from each GD class, overall 320 persons and 2560 trials). The results of our experiments showed clear performance improvements of +16.7% in the deviation from the RB compared to the baseline for task $N/C/A/K/H$ and +31% compared to the baseline for task N/GD (see the third row in Table IV).

Other biases in the data may be introduced by variations in gender, walking velocity, leg length and other parameters [33] leading to a variability of GRF parameterizations in the individual disorder classes. Additional normalization of the input data may be necessary to reduce intra-class variation and improve classification accuracy. Several studies have shown that in particular gender causes strong variability in gait signals [34], [35]. To assess the influence of gender on our results, an experiment was performed on a reduced dataset containing only data from male participants (note that the number of female participants in our dataset is not sufficient to perform separate experiments). Surprisingly, the results did not improve (see the last row in Table IV). This indicates that for our data, gender has rather little influence on the results, which, however, does not imply that the influence of gender can be neglected a priori. A detailed study on the influence of gender is subject to future investigation.

IV. Conclusions

The present study aimed at classifying patients with different orthopedic gait impairments at the hip, knee, ankle, and calcaneus from healthy controls using GRF measurements. For this purpose a dataset of 9,496 gait measurements from clinical practice was utilized. In a first step we investigated the suitability of state-of-the-art GRF parameterizations and analyzed their statistical properties and discriminative power among the classes. Based on these results, the use of entire GRF waveform parameterizations as input (such as PCA), rather than relying on GRF parameters (DPs and TDPs) seems advisable. Furthermore, the use of GRF force components paired with the respective COP measurements yielded the best results. Our experiments further showed that balancing the dataset significantly improves results (e.g. increasing the deviation from the random baseline by +16.7% for the classification into healthy controls and all four GD classes and by +31% for distinguishing between healthy controls and patients).

The presented study shows that results heavily depend on the employed GRF representation. Future work will investigate and evaluate adaptively learned signal representations [36], [37] to obtain more discriminative and expressive parameterizations of GRF measurements. Furthermore, we will focus on establishing a large, open-source, and balanced data set to foster further developments in this area. Our results thereby provide a first performance baseline for the classification of functional gait disorders and can serve as a reference for future improvements.

Acknowledgment

The authors want to thank Marianne Worisch and Szava Zoltán for their great assistance in data preparation and their great support in clinical and technical questions.

References

[1] R. Baker, *Measuring Walking: A Handbook of Clinical Gait Analysis*. London, U.K.: Mac Keith Press, 2013.

[2] 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.

[3] T. Chau, "A review of analytical techniques for gait data. Part 2: Neural network and wavelet methods," *Gait Posture*, vol. 13, no. 2, pp. 102–120, 2001.

[4] C. A. Lozano-Ortiz, A. M. S. Muniz, and J. Nadal, "Human gait classification after lower limb fracture using artificial neural networks and principal component analysis," in *Proc. IEEE 2010 Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2010, pp. 1413–1416.

[5] W. Zeng, F. Liu, Q. Wang, Y. Wang, L. Ma, and Y. Zhang, "Parkinson's disease classification using gait analysis via deterministic learning," *Neurosci. Lett.*, vol. 633, pp. 268–278, 2016.

[6] A. Vieira *et al.*, "Software for human gait analysis and classification," in *Proc. 2015 IEEE 4th Portuguese Meeting Bioeng.*, Porto, 2015, pp. 1–1.

[7] J. Wu, J. Wang, and L. Liu, "Feature extraction via KPCA for classification of gait patterns," *Hum. Movement Sci.*, vol. 26, no. 3, pp. 393–411, 2007.

[8] J. Wu and J. Wang, "PCA-based SVM for automatic recognition of gait patterns," *J. Appl. Biomech.*, vol. 24, no. 1, pp. 83–87, 2008.

[9] P. Levinger, D. T. H. Lai, R. K. Begg, K. E. Webster, and J. A. Feller, "The application of support vector machines for detecting recovery from knee replacement surgery using spatio-temporal gait parameters," *Gait Posture*, vol. 29, no. 1, pp. 91–96, 2009.

[10] N. Mezghani *et al.*, "Automatic classification of asymptomatic and osteoarthritis knee gait patterns using kinematic data features and the nearest neighbor classifier," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 3, pp. 1230–1232, Mar. 2008.

[11] M. Alaqtash, T. Sarkodie-Gyan, H. Yu, O. Fuentes, R. Brower, and A. Abdelgawad, "Automatic classification of pathological gait patterns using ground reaction forces and machine learning algorithms," in *Proc. IEEE 2011 Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2011, pp. 453–457.

[12] M. Ferrarin *et al.*, "Gait pattern classification in children with Charcot–Marie–Tooth disease type 1A," *Gait Posture*, vol. 35, no. 1, pp. 131–137, 2012.

[13] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[14] G. Giakas and V. Baltzopoulos, "Time and frequency domain analysis of ground reaction forces during walking: An investigation of variability and symmetry," *Gait Posture*, vol. 5, no. 3, pp. 189–197, 1997.

[15] R. Lafuente, J. M. Belda, J. Sánchez-Lacuesta, C. Soler, and J. Prat, "Design and test of neural networks and statistical classifiers in computer-aided movement analysis: A case study on gait analysis," *Clin. Biomech.*, vol. 13, no. 3, pp. 216–229, 1998.

[16] D. P. Soares, M. P. de Castro, E. A. Mendes, and L. Machado, "Principal component analysis in ground reaction forces and center of pressure gait waveforms of people with transfemoral amputation," *Prosthetics Orthotics Int.*, vol. 40, no. 6, pp. 729–738, 2016.

[17] A. M. S. Muniz and J. Nadal, "Application of principal component analysis in vertical ground reaction force to discriminate normal and abnormal gait," *Gait Posture*, vol. 29, no. 1, pp. 31–35, 2009.

[18] Z. Peng, C. Cao, Q. Liu, and W. Pan, "Human walking pattern recognition based on KPCA and SVM with ground reflex pressure signal," *Math. Probl. Eng.*, vol. 2013, 2013, Art. no. 143435.

[19] Y. Xu, D. Zhang, Z. Jin, M. Li, and J. Y. Yang, "A fast kernel-based nonlinear discriminant analysis for multi-class problems," *Pattern Recognit.*, vol. 39, no. 6, pp. 1026–1033, 2006.

[20] R. LeMoyne, W. Kerr, T. Mastroianni, and A. Hessel, "Implementation of machine learning for classifying hemiplegic gait disparity through use of a force plate," in *Proc. IEEE 2014 13th Int. Conf. Mach. Learn. Appl.*, 2014, pp. 379–382.

[21] G. Williams, D. Lai, A. Schache, and M. E. Morris, "Classification of gait disorders following traumatic brain injury," *J. Head Trauma Rehabil.*, vol. 30, no. 2, pp. E13–E23, 2015.

[22] W. I. Schöllhorn, B. M. Nigg, D. J. Stefanyshyn, and W. Liu, "Identification of individual walking patterns using time discrete and time continuous data sets," *Gait Posture*, vol. 15, no. 2, pp. 180–186, 2002.

[23] K. L. Goh, K. H. Lim, A. A. Gopalai, and Y. Z. Chong, "Multilayer perceptron neural network classification for human vertical ground reaction forces," in *Proc. 2014 IEEE Conf. Biomed. Eng. Sci.*, 2014, pp. 536–540.

[24] M. Köhle and D. Merkl, "Analyzing human gait patterns for malfunction detection," in *Proc. 2000 ACM Symp. Appl. Comput.*, 2000, vol. 1, pp. 41–45.

[25] M. Köhle and D. Merkl, "Identification of gait patterns with self-organizing maps based on ground reaction force," in *Proc. Eur. Symp. Artif. Neural Netw.*, 1996, vol. 96, pp. 24–26.

[26] K. C. Moisio, D. R. Sumner, S. Shott, and D. E. Hurwitz, "Normalization of joint moments during gait: A comparison of two techniques," *J. Biomech.*, vol. 36, no. 4, pp. 599–603, 2003.

[27] I. Jolliffe, *Principal Component Analysis*, 2nd ed. New York, NY, USA: Springer-Verlag, 2002.

[28] D. Slijepcevic *et al.*, "Ground reaction force measurements for gait classification tasks: Effects of different PCA-based representations," *Gait Posture*, vol. 57, pp. 4–5, 2017.

[29] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York, NY, USA: Wiley, 2012.

[30] C. M. De Vries, S. Geva, and A. Trotman, "Document clustering evaluation: Divergence from a random baseline," in Workshop Information Retrieval 2012 (IR-2012), Dortmund, Germany, 2012.

[31] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, pp. 27:1–27:27, 2011.

[32] D. Janssen, W. I. Schöllhorn, K. M. Newell, J. M. Jäger, F. Rost, and K. Vehof, "Diagnosing fatigue in gait patterns by support vector machines and self-organizing maps," *Hum. Movement Sci.*, vol. 30, no. 5, pp. 966–975, 2011.

[33] M. R. Pierrynowski and V. Galea, "Enhancing the ability of gait analyses to differentiate between groups: Scaling gait data to body size," *Gait Posture*, vol. 13, no. 3, pp. 193–201, 2001.

[34] B. M. Eskofier, M. Kraus, J. T. Worobets, D. J. Stefanyshyn, and B. M. Nigg, "Pattern classification of kinematic and kinetic running data to distinguish gender, shod/barefoot and injury groups with feature ranking," *Comput. Methods Biomech. Biomed. Eng.*, vol. 15, no. 5, pp. 467–474, 2012.

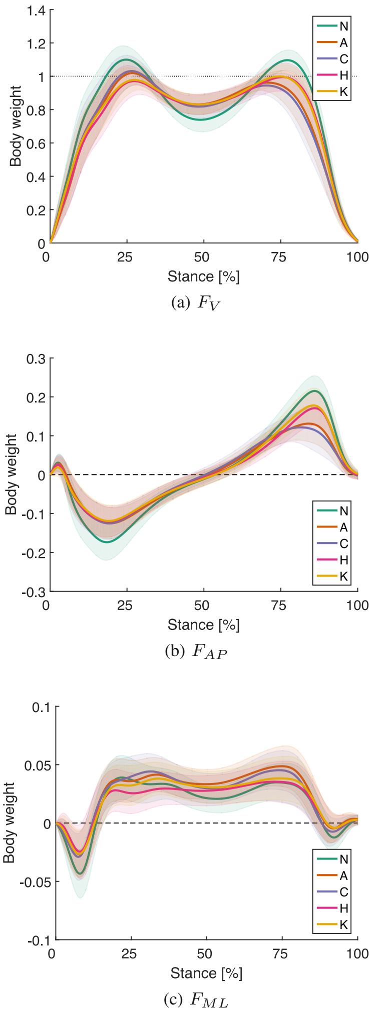
[35] M. C. Chiu and M. J. Wang, "The effect of gait speed and gender on perceived exertion, muscle activity, joint motion of lower extremity, ground reaction force and heart rate during normal walking," *Gait Posture*, vol. 25, no. 3, pp. 385–392, 2007.

[36] Y. Zhang, P. O. Ogunbona, W. Li, B. Munro, and G. G. Wallace, "Pathological gait detection of Parkinson's disease using sparse representation," in *Proc. IEEE 2013 Int. Conf. Digit. Image Comput., Technn. Appl.*, 2013, pp. 1–8.

[37] J. Hannink, T. Kautz, C. F. Pasluosta, K. G. Gaßmann, J. Klucken, and B. M. Eskofier, "Sensor-based gait parameter extraction with deep convolutional neural networks," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 1, pp. 85–93, Jan. 2017.

Supplementary Material for: Automatic Classification of Functional Gait Disorders

Djordje Slijepcevic, Matthias Zeppelzauer, Anna-Maria Gorgas, Caterine Schwab, Michael Schüller, Arnold Baca, Christian Breiteneder, Brian Horsak

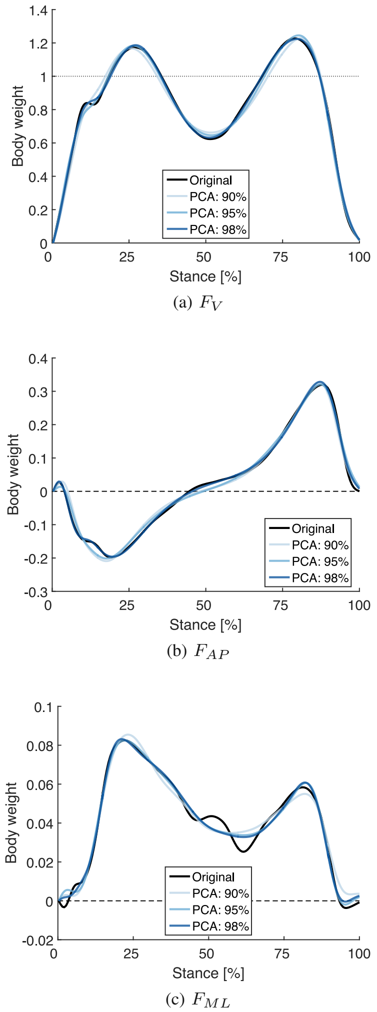


(a) F_V

(b) F_{AP}

(c) F_{ML}

Fig. S1. Mean pattern of the three ground reaction forces (GRF) enveloped by ± 1 standard deviations for each class. Data normalized by body weight and 100% stance.



(a) F_V

(b) F_{AP}

(c) F_{ML}

Fig. S2. Comparison of different PCA representations. The final dimensionality of the obtained representations is specified by the amount of variance preserved in a particular projection, i.e. 98%, 95%, and 90%. Data normalized by body weight and 100% stance.

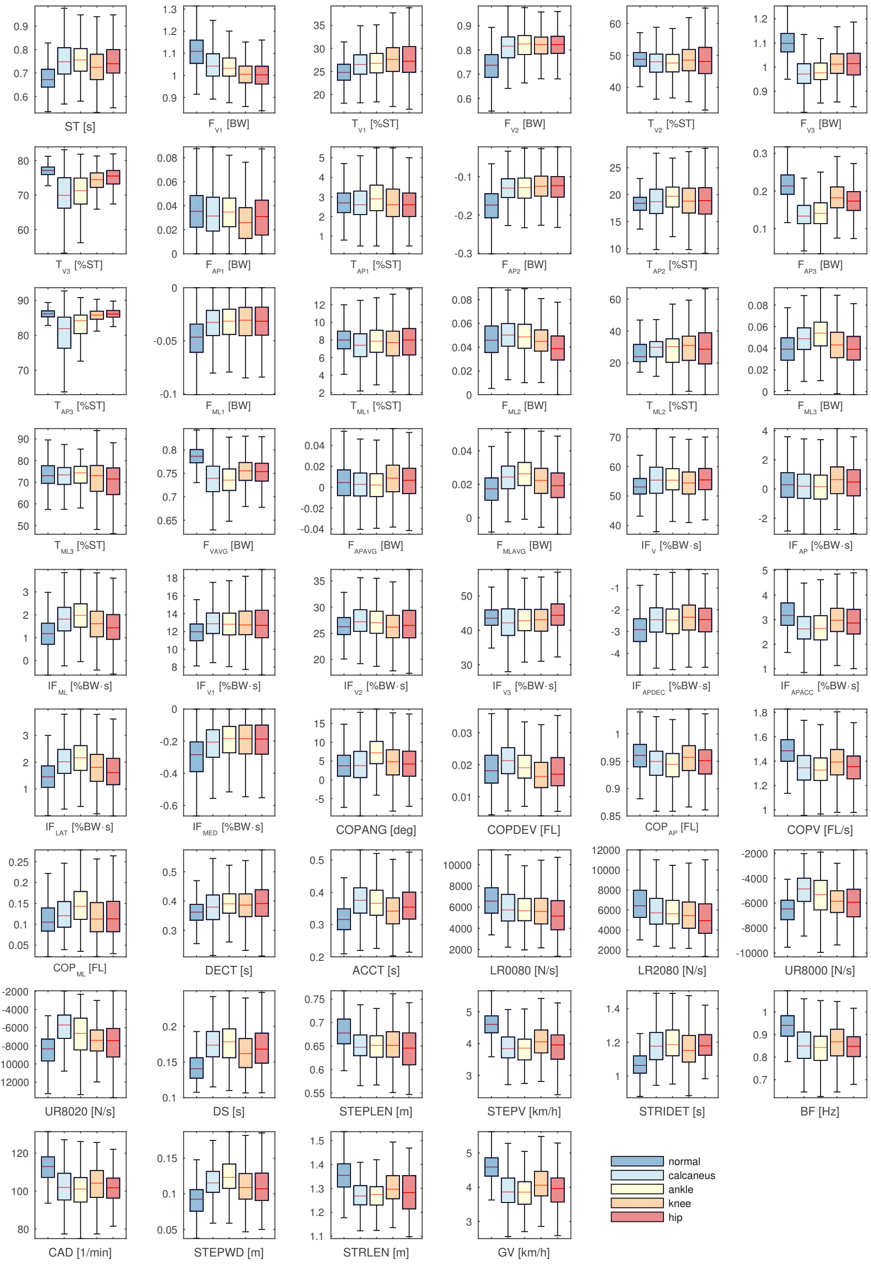


Fig. S3. Boxplots for all 52 GRF parameters. Each boxplot shows the median and the IQR (box) for each class (outliers were removed for better visualization). Box-whiskers correspond to 1.5 of the box-length, thus show approximately ± 2.7 standard deviations. The overlap of distributions between the classes gives an impression of the parameters' discriminative power (inter-class variation). Data normalized by body weight and 100% stance, prior to the calculation of the parameters.

2.3 Input Representations and Classification Strategies for Automated Human Gait Analysis

Djordje Slijepcevic, Matthias Zeppelzauer, Caterine Schwab, Anna-Maria Raberger, Christian Breiteneder, and Brian Horsak. **Input Representations and Classification Strategies for Automated Human Gait Analysis**. *Gait & Posture*, 76:198–203, 2020. DOI: 10.1016/j.gaitpost.2019.10.021

The final version of this publication is available at: `https://doi.org/10.1016/j.gaitpost.2019.10.021`. Permission for reprint granted, © 2019 Slijepcevic

Gait & Posture 76 (2020) 198–203

Contents lists available at ScienceDirect

Gait & Posture

journal homepage: www.elsevier.com/locate/gaitpost

Input representations and classification strategies for automated human gait analysis

Djordje Slijepcevic[a,*], Matthias Zeppelzauer[a], Caterine Schwab[b], Anna-Maria Raberger[b], Christian Breiteneder[c], Brian Horsak[b]

[a] St. Pölten University of Applied Sciences, Institute for Creative Media Technologies, St. Pölten, Austria
[b] St. Pölten University of Applied Sciences, Institute of Health Sciences, St. Pölten, Austria
[c] TU Wien, Institute of Visual Computing and Human-Centered Technology, Vienna, Austria

ARTICLE INFO

Keywords:
Ground reaction force
Gait classification
Machine learning
Gait disorders
Support vector machine

ABSTRACT

Background: Quantitative gait analysis produces a vast amount of data, which can be difficult to analyze. Automated gait classification based on machine learning techniques bear the potential to support clinicians in comprehending these complex data. Even though these techniques are already frequently used in the scientific community, there is no clear consensus on how the data need to be preprocessed and arranged to assure optimal classification accuracy outcomes.

Research question: Is there an optimal data aggregation and preprocessing workflow to optimize classification accuracy outcomes?

Methods: Based on our previous work on automated classification of ground reaction force (GRF) data, a sequential setup was followed: firstly, several aggregation methods – early fusion and late fusion – were compared, and secondly, based on the best aggregation method identified, the expressiveness of different combinations of signal representations was investigated. The employed dataset included data from 910 subjects, with four gait disorder classes and one healthy control group. The machine learning pipeline comprised principle component analysis (PCA), *z*-standardization and a support vector machine (SVM).

Results: The late fusion aggregation, i.e., utilizing majority voting on the classifier's predictions, performed best. In addition, the use of derived signal representations (relative changes and signal differences) seems to be advantageous as well.

Significance: Our results indicate that great caution is needed when data preprocessing and aggregation methods are selected, as these can have an impact on classification accuracies. These results shall serve future studies as a guideline for the choice of data aggregation and preprocessing techniques to be employed.

1. Introduction

Gait disorders can affect anyone, regardless of age, and often impede an individual's ability to participate in daily living activities such as walking and might even reduce movement efficiency in terms of energy consumption [1,2]. Gait analysis based on ground reaction force (GRF) assessment is a well-established method to diagnose the mechanisms that underlie gait disorders. The quantitative analysis of such data can provide relevant information for clinicians in diagnosing gait impairments, planning therapies and surgeries, supporting rehabilitation processes, or evaluating treatment outcomes [3]. However, quantitative gait analysis produces a vast amount of data, which are difficult to comprehend and analyze due to their high-dimensionality, temporal dependencies, strong variability, non-linear relationships, and inter-correlations [4]. Therefore, there is growing interest in employing machine learning techniques that allow for a cost-effective, fast and objective analysis of large amounts of gait measurements. Recently, automated gait classification has been successfully used for various patient groups [5] affected by stroke [6], Parkinson's disease [7], cerebral palsy [8], multiple sclerosis [9], osteoarthritis [10], or by age-related impairments [11].

Automated classification of gait is, however, a complex task consisting of many different processing steps which have to be carried out in a methodically correct way and for which various approaches exist. According to Figueiredo et al. [5] gait pattern recognition comprises the following main steps: (1) feature extraction, (2) feature normalization,

[*] Corresponding author at: St. Pölten University of Applied Sciences, Matthias Corvinus-Straße 15, 3100 St. Pölten, Austria.
E-mail address: djordje.slijepcevic@fhstp.ac.at (D. Slijepcevic).

https://doi.org/10.1016/j.gaitpost.2019.10.021
Received 22 December 2018; Received in revised form 8 October 2019; Accepted 14 October 2019

(3) feature selection, (4) forming a training and a testing dataset, (5) training a classification model, and (6) evaluating the performance. To date, there is no clear consensus on how to proceed in each of these steps. For tasks (2) to (6), the systematic review by Figueiredo et al. [5] might serve as a first guideline. For the first step of feature extraction, a variety of options can be found in the current literature, but so far no clear recommendation can be derived. However, different approaches in feature extraction might significantly effect classification accuracies.

Firstly, in the literature on gait classification, several recorded trials of a subject are usually either averaged to a single waveform [12–14] or all available trials are provided to a classifier [15,16,10]. To date, it is unclear which of these data aggregation strategies serves best for gait classification. Recently, a statistical method based on the notion of depth was suggested which identifies the most representative trial [17]. This approach, however, has not been employed in the gait classification community yet.

Secondly, there is no clear consensus on how the raw signals should be preprocessed and transformed to form an appropriate input feature vector for the machine learning algorithm. Regarding the available input data (ground reaction force (GRF) and center of pressure (COP) components), it is still unclear which form of representation (i.e. raw data, relative changes, or signal differences) is best suited for machine learning. Based on our earlier work [18] the two primary aims of this article are: to (i) evaluate the effects of different data aggregation methods on gait classification performance and to (ii) investigate which input representations and combinations of representations perform best for automated gait classification. To facilitate the comparability of machine learning approaches and to optimize performance, it is critical to identify best practice procedures for the individual steps of gait classification. The results of this article shall serve future studies as a guideline on machine learning for gait analysis.

2. Methods

2.1. Patients and dataset

The anonymized data used in this study are part of an existing clinical gait database maintained by a rehabilitation center of the Austrian Workers' Compensation Board (AUVA). The AUVA is the social insurance for occupational risks for more than 3.3 million employees and 1.4 million pupils and students in Austria. This retrospective study was approved by the local Ethics Committee of Lower Austria (GS1-EK-4/299-2014).

The dataset utilized comprises GRF measurements from 728 patients with gait disorders (GD) and data from 182 healthy controls, both of various physical composition and gender (see Table 1).The dataset is balanced regarding the number of persons per class, the number of recorded sessions per person and the number of trials per person. The dataset includes gait disorders associated with the calcaneus ($n = 182$), ankle ($n = 182$), knee ($n = 182$), and hip ($n = 182$). A well-experienced physical therapist (with more than a decade of clinical experience) has manually labeled the dataset based on the available medical diagnosis of each patient. The individual GD classes include patients after joint replacement surgery, fractures, ligament ruptures, and related disorders associated with the above-mentioned anatomical areas. The most common injuries present in the hip class are fractures of the pelvis and thigh as well as luxation of the hip joint, coxarthrosis, and total hip replacement. The knee class comprises patients after patella, femur or tibia fractures, ruptures of the cruciate or collateral ligaments or the meniscus, and total knee replacements. The ankle class includes patients after fractures of the malleoli, talus, tibia or lower leg, and ruptures of ligaments or the Achilles tendon. The calcaneus class comprises patients after calcaneus fractures or ankle fusion surgery. All of the injuries mentioned above may occur individually or in combinations within each class.

Table 1
Details on the dataset employed, the demography of the participants and the pre-defined classes.

		Age (yrs.)	Body Mass (kg)	Sex	
Class	n	Mean (SD)	Mean (SD)	(m/f)	Num. trials
Healthy controls	182	34.3 (14.0)	74.6 (15.8)	94/88	1,456
Calcaneus	182	44.3 (10.5)	86.3 (16.4)	167/15	1,456
Ankle	182	40.6 (10.9)	88.3 (18.2)	151/31	1,456
Knee	182	40.4 (12.3)	86.2 (20.3)	133/49	1,456
Hip	182	40.6 (12.8)	81.5 (15.0)	153/29	1,456
Total	**910**	**40.0 (12.1)**	**83.4 (17.1)**	**698/212**	**7,280**

2.2. Data recording and preprocessing

Gait analysis was performed on a 10 m walkway with two centrally embedded force plates (Kistler, Type 9281B12). The force plates were placed in consecutive order, allowing a person to walk across by placing one foot on each plate. Both plates were flush with the ground and covered with the same walkway surface material, so that targeting was not an issue. During a session, participants walked unassisted and without walking aid at self-selected walking speed until a minimum of eight valid recordings were available.

All processing steps and subsequent analyses were performed in Matlab 2017b (The MathWorks Inc., Natick, MA, USA). The three analog GRF signals, as well as the two COP signals, were converted to digital signals using a sampling rate of 2000 Hz and a 12-bit analog-digital converter (DT3010, Data Translation Incorporation, Marlboro, MA, USA) with a signal input range of ± 10 V. A threshold of 10 N was used for step detection and 30 N for COP calculation. Raw signals were filtered using a 2nd order low-pass Butterworth filter with a cut-off frequency of 20 Hz. All gait measurements were time-normalized to 1000 points (100% stance). Amplitude values of the three force components, i.e., vertical (V), medio-lateral (ML), and anterior–posterior (AP), were expressed as a multiple of body weight by dividing the force by the product of body mass times acceleration due to gravity.

2.3. Gait classification

The present paper builds upon the general gait classification pipeline established by Slijepcevic et al. [18] and uses it as a baseline for the performed experiments. A schematic illustration of the pipeline is shown in Fig. 1. In a first step, Principal Component Analysis (PCA) is applied to the raw input data, i.e. to each input representation separately (feature extraction).[1] Next, the resulting features, i.e., the principal components that retain 98% of the overall variance in the input data, are concatenated and z-standardized (feature normalization). The features are provided to a classifier which is trained and evaluated in a cross-validation manner. For the best parameters found during cross-validation the model is trained on the entire training set. To account for generalizability of this model we evaluated it on a completely independent and unseen dataset (see Figure S1 in the supplementary material). As demonstrated in [18], Support Vector Machines (SVM) are a suitable classifier for gait data outperforming several competitors, e.g., multi-layer perceptrons and the k-nearest neighbors algorithm. The SVM is trained in a multi-class fashion using a one-vs-one strategy.

2.3.1. Data aggregation methods

Usually, several trials per person are recorded during gait analysis. Thus, the question arises whether and how the information from these different trials can be aggregated. Such an aggregation step could be

[1] For each original input signal and derived representation a PCA is performed on a matrix of size 334 (samples) × t (trials), where t depends on the considered dataset.

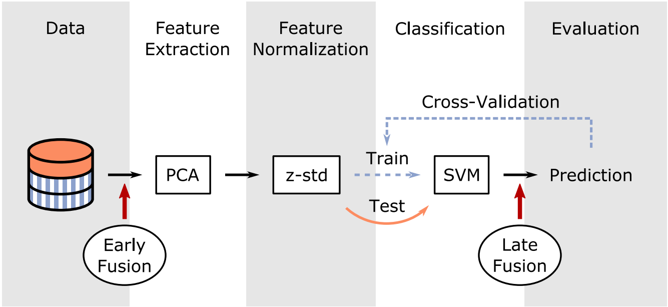


Fig. 1. Illustration of the employed gait classification framework. The dataset consisted of a training set (blue, dashed) and an independent test set (orange, solid). The latter was used to evaluate the generalizability of our classification.

implemented in an *early fusion* or a *late fusion* manner (see Fig. 1). The former directly affects the input data and thus precedes the feature extraction step, whereas the latter is directly applied to the classifier's predictions and affects mostly the classification scheme. Popular early fusion approaches include: (i) mean waveform, (ii) median waveform, and (iii) the most representative trial.

The *mean waveform* approach consists of averaging each measurement from a session (in this case, eight trials) pointwise. The resulting waveform should result in a more robust representation than the original signals by removing inter-trial variations and retaining the overall characteristic shape. The *median waveform* approach is similar but utilizes the point-wise median instead. It is more robust to outliers but may generate less smooth waveforms than the mean waveform approach. Both approaches could diminish informative waveform characteristics, or even cause artifacts that provide a distorted representation [19]. To overcome this problem, Sangeux et al. [17] proposed a statistical method to determine the *most representative* trial. Thereby, this approach assures that original measurement data is used. For machine learning, however, performance might be affected by the fact that not all available and potentially essential information is considered. A schematic illustration of the early fusion approaches is given in Fig. 2.

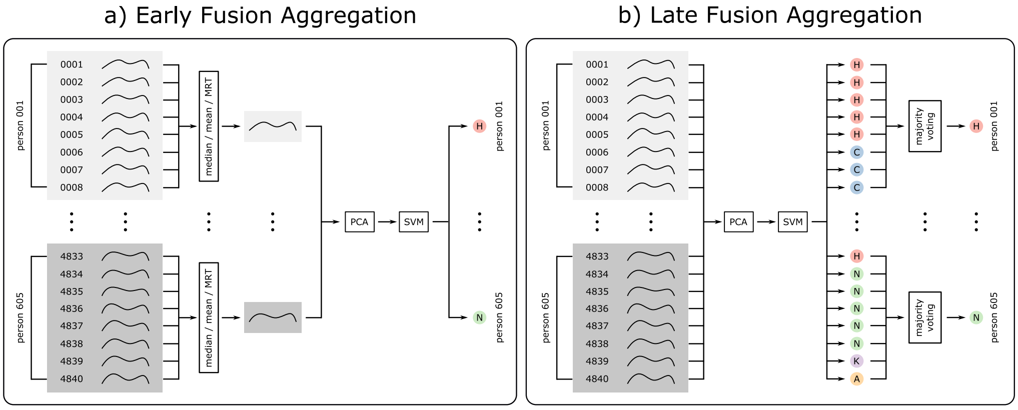
The late fusion approach utilizes all available original trials for the training of the model. As a result, the classifier returns one prediction per trial. These predictions are considered weak because they are based on individual measurements. The late fusion approach combines these weak predictions into a strong prediction. A robust approach for the combination of several predictions is majority voting. The majority vote is calculated based on the statistical mode, which returns the element (class label) that occurs most often in a set of predictions. For majority voting, only predictions with a likelihood of more than 40% for one of the five classes are used. Thereby, the negative influence of ambiguous trials is reduced. A schematic illustration of the late fusion approach is presented in Fig. 2.

To provide a baseline without aggregation of the available data, we employ all eight trials per person individually during the training and testing. Thus, each trial was predicted separately, and the information about the membership of the trial to a specific person was not utilized.

2.3.2. Input representations

We further investigate the expressiveness and suitability of different input representations for gait classification. Two different types of input representations are distinguished here: original *input signals* and *derived signals*.

Original input signals comprise the time and body weight


a) Early Fusion Aggregation

b) Late Fusion Aggregation

Fig. 2. (a) Schematic for the early fusion aggregation, i.e., mean, median, and most representative trial (MRT) approaches. Prior to training, the eight signals of one subject are aggregated by calculating a mean or median waveform, respectively or one trial is selected by MRT. (b) Schematic for the late fusion aggregation, which employs majority voting. For the training of an SVM, all recorded trials of the subjects are used. For the actual prediction of the test set, majority voting is applied to obtain a decision at subject level.

Table 2
Classification results (%) of the experiment investigating different aggregation methods over several trials (RB: 20%). Highest achieved results are highlighted bold.

Trial selection	Five-fold cross-validation on training set				Independent test set			
	Acc	P	R	F1	Acc	P	R	F1
Baseline without aggregation	52.0 (2.0)	51.8 (1.8)	52.7 (2.6)	51.4 (1.7)	56.5 (2.3)	56.2 (2.7)	56.6 (2.3)	56.0 (2.6)
Mean waveform approach	53.9 (5.2)	53.5 (6.5)	54.4 (5.9)	52.7 (5.9)	58.9 (2.8)	59.5 (1.0)	58.6 (2.2)	58.5 (1.8)
Median waveform approach	51.4 (3.3)	51.1 (3.7)	52.0 (3.7)	50.3 (3.8)	56.7 (3.5)	57.9 (2.7)	56.9 (2.9)	56.4 (2.6)
Most representative trial (MRT)	50.1 (2.0)	50.4 (2.1)	50.7 (2.4)	49.0 (2.0)	56.9 (5.3)	57.7 (4.6)	57.0 (4.3)	56.5 (4.8)
Majority voting	**55.5 (2.4)**	**54.4 (2.5)**	**56.6 (2.3)**	**54.3 (2.5)**	**61.0 (2.4)**	**60.9 (2.9)**	**61.1 (2.4)**	**60.1 (2.7)**

normalized waveforms, i.e., F_V, F_{AP}, F_{ML}, COP_{AP}, and COP_{ML} components of the affected (A) and unaffected (U) lower extremity. The affected and unaffected body side were defined by the physical therapist during data annotation. In case of healthy controls or bi-laterally affected patients the affected side was chosen randomly to avoid a bias.

The derived signal representations are calculated based on the original input signals. Two types of derived signals are investigated: the approximate first derivative (D_A,D_U) of each original input signal and the absolute difference between the input signals of the affected and unaffected lower extremity (Δ).

Furthermore, the expressive power of different combinations of the individual signal representations is examined, i.e., the combination of the original input signals and the derived representations of the affected and unaffected sides.

2.4. Experimental setup

Prior to the experiments, the dataset was randomly divided into a training set (65%) and an independent test set (35%), see Fig. 1. This split remained unchanged for all experiments. The classification experiments utilized a probabilistic SVM with a linear kernel (provided by the LIBSVM library [20]). For hyper-parameter selection, a grid search over the regularization parameter $C \in [2^{-5};2^{10}]$ was employed. During the grid search, a five-fold cross-validation was performed on the training set. After hyper-parameter selection an SVM with the best parameters was trained on the entire training set. To assess the generalizability of the methods, the test set was divided into three equally large and balanced test splits, on which we evaluated the SVM. By using multiple splits, it was possible to estimate not only the generalization ability but also the expected variation in performance for different subsets of test samples. The evaluation was conducted by calculating four performance measures, i.e. classification accuracy (Acc), precision (P), recall (R), and F1-score (F1), defined in terms of number of true positives (*TP*), true negatives (*TN*), false positives (*FP*), and false negatives (*FN*) as follows:

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

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad R = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad F1 = 2 \times \frac{P \times R}{P + R}$$

Furthermore, a sequential setup was followed: first, different aggregation methods were examined, and second, based on the best aggregation method the expressiveness of different (combinations of) signal representations was investigated. All results are reported as mean (SD), unless otherwise stated.

3. Results

The results of the first experiment investigating different aggregation methods over several trials are summarized in Table 2. The performances of the five-fold cross-validation and the evaluation on the independent test set showed similar trends. This demonstrates the generalization ability of our method. In the following, we discuss the results of the independent test set, which are more objective than the results on the training set. The baseline approach, where all available trials are employed (without aggregation), yielded an accuracy of 56.5% (2.3) (RB[2] : 20%). The use of the median waveform and MRT did not outperform the baseline performance. Within the group of early fusion approaches the mean waveform approach showed the greatest improvement with an accuracy of 58.9% (2.8). The late fusion approach, i.e., majority voting, achieved the highest absolute scores in all performance measures (although not statistically significant in this experiment).

The results of the second experiment in which we investigated the expressiveness of different signal representations on the independent test set are presented in Table 3 (further performance measures can be found in the supplementary material). The results obtained during the five-fold cross-validation follow a trend similar to that of the independent test set and are presented in the supplementary material. The first column in Table 3 indicates which components were used in each experiment: (1) each input signal separately (first five rows), (2) the combination of all three GRF components (row 6), (3) the combination of both COP components (row 7), and (4) the combination of all signals (GRF + COP, last row). For each of these selections, the columns show which (combinations of) derived representations were employed for both affected (A) and unaffected (U) sides. Most notably, column {A, Δ, D_A} shows the highest performance for most input configurations (in six of the eight rows), including also the overall best result with a classification accuracy of 62% (GRF + COP). For the F_{ML} component (row three), combination {A, D_A} provides the best result. The combination {A, D_A, U, D_U} provides the best results for the COP_{ML} component (row four) and the combination of both COP components (row seven). The comparison between the individual GRF and COP components (first five rows) and the three combinations, GRF, COP, and GRF + COP (last three rows) indicates that the combination of all components (last row) performed best. Furthermore, incorporating the information from both legs (via Δ) as well as using the first derivative (in particular of the affected leg) shows to be beneficial.

4. Discussion

From the first experiment (see Table 2) we observe that achieved performances of all approaches are higher for the test set than for the training set. The reason for this is that for experiments on the test set the SVM was trained on the entire training data with the optimal parameters determined during cross-validation and grid search. The improved results on the test set show that additional training data are beneficial for the classifier. Furthermore, the first experiment indicates that the inclusion of membership information can be beneficial. Two aggregation methods, i.e. the mean waveform approach and majority

[2] RB refers to the analytical “random baseline” and represents the theoretical accuracy obtained when assigning class labels randomly, i.e. the case where nothing is learned from the data. For a balanced dataset the analytical RB is the reciprocal of the number of classes, i.e. 20% in our case. The empirically estimated RB according to [21] which further takes the sample size into account is approximately 26% in our case. Every increase over the RB means that the underlying model has learned something from the data.

Table 3
Classification accuracies (%) for different combinations of input signals and derived representations (RB: 20%). Highest achieved results are highlighted bold.

Signals	**{A}**	**{U}**	**{A,U}**	**{A,Δ}**	**{U,Δ}**	**{A,D_A}**	**{U,D_U}**	**{A,D_A,U,D_U}**	**{A,Δ,D_A}**	**{U,Δ,D_U}**
F_V	42.6	38.7	47.2	44.9	47.5	47.5	37.1	46.6	**48.9**	44.3
F_{AP}	44.3	40.7	45.6	42.0	42.3	42.6	40.3	44.3	**46.6**	45.3
F_{ML}	44.3	32.5	44.6	43.3	34.8	**45.6**	37.7	44.6	44.3	38.4
COP_{ML}	28.2	26.9	31.2	26.6	25.3	43.6	34.1	**44.9**	44.6	35.4
COP_{AP}	36.4	26.9	35.1	40.0	33.1	45.3	30.8	45.3	**46.2**	35.1
GRF	56.7	45.6	54.4	55.7	46.9	55.1	45.6	55.4	**60.0**	48.2
COP	37.1	30.8	43.0	41.6	32.1	48.2	34.1	**52.8**	51.8	36.1
GRF + COP	*61.0*	*47.2*	*58.7*	*60.3*	*49.8*	*59.3*	*49.8*	*61.3*	***62.0***	*51.2*

voting, achieved an improvement compared to the baseline where no aggregation was performed. Specifically, the late fusion approach, i.e., majority voting, achieved better results in absolute scores than the early fusion approaches.

To evaluate the robustness of our approach in more detail, we repeated the experiment with 10 different (randomly selected) train-test splits. The results are presented in the supplementary material in Table S1 due to space limitations. For all 10 repetitions, the previously determined optimal SVM parameters (obtained from grid search on the original train-test split) remained unchanged to avoid overfitting. Additional statistical comparisons on the F1-scores from Table S1 in the supplementary material revealed that majority voting and the mean waveform approach significantly outperformed all other methods (see supplementary material for details). In total numbers and on average, majority voting showed the best performance results. We assume that this is because in early fusion large parts of the available input information are removed at an early stage and are not available during the training process. For late fusion, this is not the case. Furthermore, a comparison of the baseline method and the late fusion approach revealed that the aggregation of weak predictions by majority voting allows for a more accurate prediction at subject level. Majority voting adds a layer of abstraction to the outputs of the classifier, which seems to increase robustness. The performance level from the results in Table S1 (supplementary material) are equivalent to that of Table 2. This shows that the employed training-test split does not bias the test result in Table 2.

The conclusion from the first experiment is that as much information as possible should be retained during the classification process and thus late fusion is recommended. Aggregation of information at later stages of the process seems to be superior to aggregation at an early stage, as relevant information of the individual trials is lost. The second experiment suggests that using only the original input signals might not always be the best choice. In most of our experiments, a combined representation of input signals and derived representations is advantageous, especially the combinations $\{A, \Delta, D_A\}$ and $\{A, D_A, U, D_U\}$ in Table 3. Considerably lower accuracy was achieved when only the individual signals (first five rows in Table 3) were used. The use of a single COP signal (rows 4 and 5) lead to degeneration of the classifier in some cases, i.e., one class could not be modeled at all by the classifier. The combination of the three GRF components is considerably more expressive than the combination of the COP components. The best choice seems to be a combination of all signals (GRF + COP). This also supports our previous findings [16,18,22].

We further observed that the signals of the affected side are more expressive than those of the unaffected side ($\{A\}$ vs. $\{U\}$ in Table 3). This observation contradicts the findings of Williams et al. [23]. The combination of affected and unaffected input signals improved the results in five out of eight cases.

The Δ-waveform represents the difference between the affected and the unaffected side and thus explicitly captures the symmetry between both sides. When combined with the signals of the affected side, a moderate increase in accuracy was present in three of eight cases ($\{A\}$ vs. $\{A, \Delta\}$). This result suggests that the classifier is able to derive symmetry-related information also from the raw input signals and does not necessarily need it to be explicitly provided. For the unaffected side, the Δ-waveform provides an improvement in seven of eight cases ($\{U\}$ vs. $\{U, \Delta\}$). Therefore, the Δ-waveform seems to carry important information. Adding the first derivative as an additional input representation to the signals of the affected or unaffected side showed improvements in 30 out of 40 cases (evident by comparing the first five columns with the last five columns in Table 3).

To obtain additional indicators for the usefulness of the representations, we conducted further experiments with the overall best input representation (GRF + COP, last row in Table 3). We have calculated all ($2^5 - 1 = 31$) possible combinations of A, D_A,U,D_U and Δ for the case GRF + COP and examined how often each representation occurs within the best 10 results. The most useful representations seem to be D_A (contained in 8 of the 10 best results) as well as Δ and A (each contained in 6 of the 10 best results). D_U (5/10 results) and U (4/10 results) seem less important.

The overall recommendation that can be derived from these experiments is that the combination of *more* input signals and input representations (even when they contain redundant information) can lead to better results. This is especially true for combining GRF and COP components but also for using the derivatives of the affected and unaffected sides. Even though the derivatives represent redundant information to the original signals, they might still help the classifier to better grasp class differences. Furthermore, the combination of the affected and unaffected side (either explicitly or implicitly trough Δ) seems to be beneficial as well. The results of our study provide a first indication of which signals to use and how to fuse them. Further investigations with alternative datasets are required to corroborate these findings.

5. Conclusions

The presented work aims at clarifying which aggregation method and which signal representations are best suited for the classification of data obtained from gait analysis (based on GRF assessment).The results show that the aggregation of several trials of one subject is beneficial especially when late fusion or mean waveform is used. Furthermore, the results indicate that the combination of the original signals with derived representations increases the expressive power of the data during feature extraction and classification. The combination of GRF and COP components with derived representations, even though they may be partially redundant, improved classification performance on our data.

Future research will investigate adaptively-learned feature representations as well as the modeling of relationships within a gait cycle to derive more expressive representations.

Acknowledgments

This work was partly funded by the NFB – Lower Austrian Research and Education Company (NFB) and the Provincial Government of Lower Austria, Department of Science and Research (LSC14-005 and FTI17-014) and the Austrian Research Promotion Agency (FFG) and the BMDW within the COIN-program (866855). We want to thank Marianne Worisch and Szava Zoltán for their great assistance in data preparation and their great support in clinical and technical questions.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.gaitpost.2019.10.021.

References

[1] W. Pirker, R. Katzenschlager, Gait disorders in adults and the elderly, Wiener Klinische Wochenschrift 129 (3–4) (2017) 81–95.

[2] P. Mahlknecht, S. Kiechl, B.R. Bloem, J. Willeit, C. Scherfler, A. Gasperi, G. Rungger, W. Poewe, K. Seppi, Prevalence and burden of gait disorders in elderly men and women aged 60–97 years: a population-based study, PLoS ONE 8 (7) (2013) e69627.

[3] R. Baker, Measuring Walking: A Handbook of Clinical Gait Analysis, Mac Keith Press, London, 2013.

[4] T. Chau, A review of analytical techniques for gait data. Part 1: Fuzzy, statistical and fractal methods, Gait Posture 13 (1) (2001) 49–66, https://doi.org/10.1016/S0966-6362(00)00094-1.

[5] J. Figueiredo, C.P. Santos, J.C. Moreno, Automatic recognition of gait patterns in human motor disorders using machine learning: a review, Med. Eng. Phys. (2018), https://doi.org/10.1016/j.medengphy.2017.12.006.

[6] H. Lau, K. Tong, H. Zhu, Support vector machine for classification of walking conditions of persons after stroke with dropped foot, Hum. Mov. Sci. 28 (4) (2009) 504–514, https://doi.org/10.1016/j.humov.2008.12.003.

[7] F. Wahid, R.K. Begg, C.J. Hass, S. Halgamuge, D.C. Ackland, Classification of Parkinson's disease gait using spatial-temporal gait features, IEEE J. Biomed. Health Inform. 19 (6) (2015) 1794–1802, https://doi.org/10.1109/JBHI.2015.2450232.

[8] L. Van Gestel, T. De Laet, E. Di Lello, H. Bruyninckx, G. Molenaers, A. Van Campenhout, E. Aertbelin, M. Schwartz, H. Wambacq, P. De Cock, K. Desloovere, Probabilistic gait classification in children with cerebral palsy: a Bayesian approach, Res. Dev. Disabilit. 32 (6) (2011) 2542–2552, https://doi.org/10.1016/j.ridd.2011.07.004.

[9] M. Alaqtash, T. Sarkodie-Gyan, H. Yu, O. Fuentes, R. Brower, A. Abdelgawad, Automatic classification of pathological gait patterns using ground reaction forces and machine learning algorithms, 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2011, pp. 453–457.

[10] C. Nesch, V. Valderrabano, C. Huber, V. von Tscharner, G. Pagenstert, Gait patterns of asymmetric ankle osteoarthritis patients, Clin. Biomech. 27 (6) (2012) 613–618, https://doi.org/10.1016/j.clinbiomech.2011.12.016.

[11] J. Wu, J. Wang, PCA-based SVM for automatic recognition of gait patterns, J. Appl. Biomech. 24 (1) (2008) 83–87.

[12] D. Soares, M. de Castro, E. Mendes, L. Machado, Principal component analysis in ground reaction forces and center of pressure gait waveforms of people with transfemoral amputation, Prosthet. Orthot. Int. 40 (6) (2016) 729–738.

[13] J. Christian, J. Krll, G. Strutzenberger, N. Alexander, M. Ofner, H. Schwameder, Computer aided analysis of gait patterns in patients with acute anterior cruciate ligament injury, Clin. Biomech. 33 (2016) 55–60, https://doi.org/10.1016/j.clinbiomech.2016.02.008.

[14] B.M. Eskofier, P. Federolf, P.F. Kugler, B.M. Nigg, Marker-based classification of young-elderly gait pattern differences via direct PCA feature extraction and SVMs, Comput. Methods Biomech. Biomed. Eng. 16 (4) (2011) 435–442.

[15] P. Levinger, D. Lai, R. Begg, K. Webster, J. Feller, The application of support vector machines for detecting recovery from knee replacement surgery using spatio-temporal gait parameters, Gait Posture 29 (1) (2009) 91–96.

[16] D. Slijepcevic, B. Horsak, C. Schwab, A. Raberger, M. Schüller, A. Baca, C. Breiteneder, M. Zeppelzauer, Ground reaction force measurements for gait classification tasks: effects of different PCA-based representations, Gait Posture 57 (2017) 4–5.

[17] M. Sangeux, J. Polak, A simple method to choose the most representative stride and detect outliers, ResearchGate 41 (2) (2014), https://doi.org/10.1016/j.gaitpost.2014.12.004.

[18] D. Slijepcevic, M. Zeppelzauer, A.-M. Gorgas, C. Schwab, M. Schüller, A. Baca, C. Breiteneder, B. Horsak, Automatic classification of functional gait disorders, IEEE J. Biomed. Health Inform. 22 (5) (2018) 1653–1661.

[19] T. Chau, S. Young, S. Redekop, Managing variability in the summary and comparison of gait data, J. NeuroEng. Rehabil. 2 (1) (2005) 22, https://doi.org/10.1186/1743-0003-2-22.

[20] C. Chang, C. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 27:1-27:27.

[21] E. Combrisson, K. Jerbi, Exceeding chance level by chance: the caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy, J. Neurosci. Methods 250 (2015) 126–136.

[22] D. Slijepcevic, M. Zeppelzauer, C. Schwab, A. Raberger, B. Dumphart, A. Baca, C. Breiteneder, B. Horsak, P 011-towards an optimal combination of input signals and derived representations for gait classification based on ground reaction force measurements, Gait Posture 65 (2018) 249.

[23] G. Williams, D. Lai, A. Schache, M. Morris, Classification of gait disorders following traumatic brain injury, J. Head Trauma Rehabil. 30 (2) (2015) E13–E23.

Supplementary Material for: Input representations and classification strategies for automated human gait analysis

Djordje Slijepcevic[a,*], Matthias Zeppelzauer[a], Caterine Schwab[b], Anna-Maria Raberger[b],
Christian Breiteneder[c], Brian Horsak[b]

[a] *St. Pölten University of Applied Sciences, Institute for Creative Media Technologies, St. Pölten, Austria*
[b] *St. Pölten University of Applied Sciences, Institute of Health Sciences, St. Pölten, Austria*
[c] *TU Wien, Institute of Visual Computing and Human-Centered Technology, Vienna, Austria*

Abstract

Background: Quantitative gait analysis produces a vast amount of data, which can be difficult to analyze. Automated gait classification based on machine learning techniques bear the potential to support clinicians in comprehending these complex data. Even though these techniques are already frequently used in the scientific community, there is no clear consensus on how the data need to be preprocessed and arranged to assure optimal classification accuracy outcomes.
Research question: Is there an optimal data aggregation and preprocessing workflow to optimize classification accuracy outcomes?
Methods: Based on our previous work on automated classification of ground reaction force (GRF) data, a sequential setup was followed: firstly, several aggregation methods - early fusion and late fusion - were compared, and secondly, based on the best aggregation method identified, the expressiveness of different combinations of signal representations was investigated. The employed dataset included data from 910 subjects, with four gait disorder classes and one healthy control group. The machine learning pipeline comprised principle component analysis (PCA), z-standardization and a support vector machine (SVM).
Results: The late fusion aggregation, i.e., utilizing majority voting on the classifier's predictions, performed best. In addition, the use of derived signal representations (relative changes and signal differences) seems to be advantageous as well.
Significance: Our results indicate that great caution is needed when data preprocessing and aggregation methods are selected, as these can have an impact on classification accuracies. Our results shall serve future studies as a guideline for the choice of data aggregation and preprocessing techniques to be employed.

1. Methods

Figure S1 shows how the dataset is split for evaluation purposes during the conducted experiments. We have randomly divided the dataset into a training (65%) and a test set (35%).

[*] Corresponding author
Email address: djordje.slijepcevic@fhstp.ac.at (Djordje Slijepcevic)

5-fold cross validation is performed on the training set to determine the best parameters for the employed classifier, i.e. a probabilistic linear support vector machine (SVM), and to evaluate its dependency on the training data. Once the best configuration for the SVM was determined, it was trained on the entire training data. The test set was divided into three equal and balanced splits, which are used to evaluate the generalization ability of the best obtained model on previously unseen data.

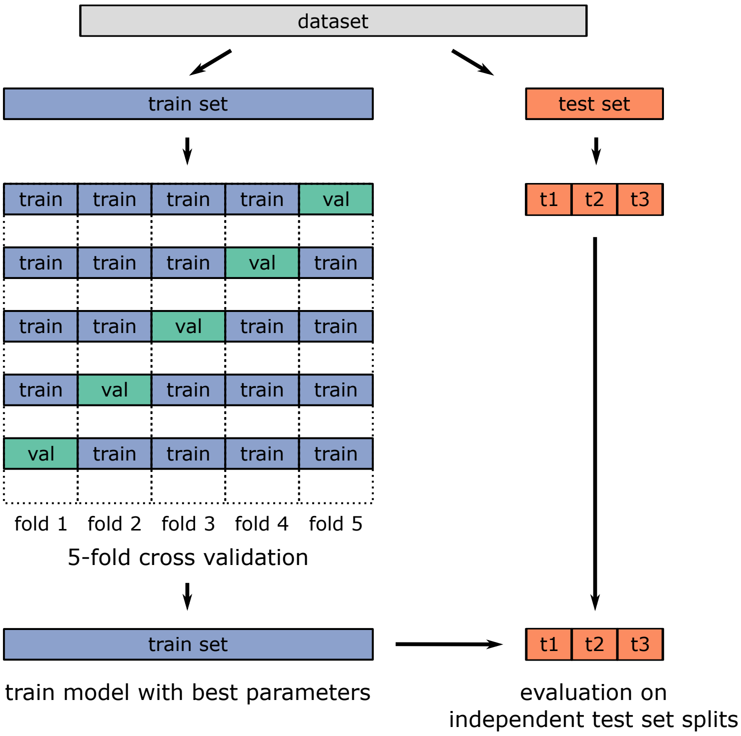

dataset
train set
test set
train train train train val
train train train val train
train train val train train
train val train train train
val train train train train
t1 t2 t3
fold 1 fold 2 fold 3 fold 4 fold 5
5-fold cross validation
train set
t1 t2 t3
train model with best parameters
evaluation on
independent test set splits

Figure S1: The dataset was randomly divided into a training set and an independent test set. For hyper-parameter selection, a 5-fold cross-validation grid search was employed on the training set. The SVM with the best parameters was trained on the entire training set and evaluated on the three test set splits.

2. Results

To further assess the generalizability of the first experiment, i.e. investigating different aggregation methods over several trials, we repeatedly evaluated the experiment with 10 different train-test splits. Results of these experiments are presented in Table S1. A repeated measures ANOVA with a Greenhouse-Geisser correction determined that the F1-score differed statistically significantly between the five methods ($F_{(2.157,19.411)} = 15.969$, $p < 0.001$). A Shapiro-Wilk test confirmed the normal distribution of all variables. Bonferroni-Holm corrected post hoc tests revealed that majority voting was superior to all other methods ($p < 0.01$), except for the mean waveform approach. Here the difference between majority voting and the mean waveform approach slightly missed significance ($p = 0.102$). The mean waveform approach was superior to the baseline and median waveform approach, but slightly missed significance for the most representative trial approach ($p = 0.053$).

Table S1: Classification results (%) of the experiment investigating different aggregation methods over several trials evaluated on 10 different train-test splits (RB: 20%).

Trial selection	Measure	Independent test set (10 different train-test splits)										Mean (SD)
		1	2	3	4	5	6	7	8	9	10	
Baseline without aggregation	Acc	56.5	53.2	53.2	53.0	56.1	53.7	52.9	55.4	57.2	59.1	55.0 (2.2)
	P	55.9	51.7	53.1	52.7	55.8	53.7	52.8	54.3	57.0	58.3	54.5 (2.1)
	R	56.5	53.2	53.2	53.1	56.1	53.7	53.0	55.5	57.2	59.1	55.1 (2.1)
	F1	56.0	52.1	53.1	52.8	55.8	53.7	52.9	54.6	57.1	58.5	54.7 (2.1)
Mean waveform approach	Acc	59.0	55.9	54.6	56.2	56.4	55.4	52.6	55.6	58.4	61.6	56.6 (2.5)
	P	58.8	55.0	54.9	56.2	56.3	55.6	52.0	54.9	58.6	60.9	56.3 (2.5)
	R	59.0	56.0	54.6	56.3	56.4	55.4	52.7	55.6	58.4	61.6	56.6 (2.5)
	F1	58.7	55.2	54.5	55.9	56.2	55.4	52.0	55.1	58.4	61.1	56.3 (2.6)
Median waveform approach	Acc	56.7	53.6	52.9	52.9	57.1	53.4	51.0	54.3	59.3	57.4	54.9 (2.6)
	P	56.5	52.7	52.9	52.3	57.1	53.7	50.6	53.2	59.0	56.6	54.5 (2.7)
	R	56.7	53.7	53.0	53.0	57.1	53.4	51.0	54.3	59.3	57.4	54.9 (2.6)
	F1	56.2	52.8	52.7	52.3	56.7	53.5	50.6	53.5	59.1	56.6	54.4 (2.6)
Most representative trial (MRT)	Acc	57.1	48.0	50.7	51.3	55.4	56.1	54.3	52.0	53.4	52.1	53.0 (2.8)
	P	57.1	47.1	50.7	51.1	55.3	55.3	53.6	51.0	52.7	51.4	52.5 (2.9)
	R	57.1	48.1	50.7	51.4	55.4	56.1	54.3	52.1	53.4	52.1	53.1 (2.7)
	F1	56.7	47.3	50.7	51.2	55.3	55.6	53.8	51.4	53.0	51.4	52.6 (2.8)
Majority voting	Acc	62.6	54.9	55.9	56.2	60.7	56.7	59.5	59.5	62.3	63.6	59.2 (3.1)
	P	62.2	52.8	55.5	55.7	60.1	56.5	58.9	58.6	62.0	62.9	58.5 (3.4)
	R	62.6	55.0	55.9	56.3	60.7	56.7	59.6	59.6	62.3	63.6	59.2 (3.1)
	F1	62.0	52.9	55.5	55.3	60.1	56.5	59.1	58.5	61.9	62.9	58.5 (3.3)

The following tables contain additional results from our second experiment, where effects of different combinations of input signals and derived representations were examined. Table S2, Table S3, Table S4, and Table S5 show the classification accuracy, precision, recall and F1-score results obtained during 5-fold cross-validation. Table S6, Table S7, and Table S8 show precision, recall and F1-score results obtained during the evaluation on the independent test set.

Table S2: Mean (SD) of **classification accuracy** (%) obtained within 5-fold cross-validation for different combinations of input signals and derived representations (RB: 20%).

Signals	**{A}**	**{U}**	**{A,U}**	**{A,Δ}**	**{U,Δ}**	**{A,D_A}**	**{U,D_U}**	**{A,D_A,U,D_U}**	**{A,Δ,D_A}**	**{U,Δ,D_U}**
F_V	44.8 (2.7)	35.9 (1.5)	45.3 (1.8)	47.6 (2.2)	47.1 (2.2)	46.1 (2.7)	35.7 (3.5)	47.6 (1.9)	**49.8 (3.1)**	47.3 (2.2)
F_{AP}	44.5 (3.8)	38.0 (4.3)	**48.3 (1.1)**	46.9 (2.1)	44.0 (4.5)	42.5 (2.7)	37.2 (2.2)	45.5 (2.1)	45.6 (2.1)	44.0 (3.2)
F_{ML}	40.8 (4.1)	32.2 (4.6)	43.3 (5.4)	**45.1 (4.5)**	34.9 (2.2)	41.0 (2.1)	33.2 (2.8)	40.7 (4.9)	44.1 (4.4)	36.0 (3.5)
COP_{ML}	27.1 (4.2)	25.0 (2.9)	30.2 (5.6)	28.1 (4.4)	25.3 (3.2)	37.9 (4.8)	33.6 (4.1)	**40.5 (4.3)**	38.0 (6.1)	34.2 (4.0)
COP_{AP}	35.9 (6.2)	26.6 (6.6)	35.7 (6.8)	38.0 (4.1)	31.2 (4.3)	39.2 (2.8)	32.9 (2.2)	**43.0 (3.2)**	41.0 (2.4)	36.9 (3.8)
GRF	55.2 (2.3)	46.0 (2.1)	55.0 (4.8)	**55.9 (3.3)**	49.6 (3.3)	54.9 (2.5)	45.8 (1.2)	53.1 (2.5)	55.7 (1.9)	51.2 (4.3)
COP	38.0 (4.6)	29.4 (5.3)	39.8 (4.0)	38.8 (2.9)	31.2 (4.5)	42.1 (3.9)	34.4 (3.4)	**46.4 (2.3)**	43.1 (3.3)	36.2 (4.0)
GRF + COP	*55.4 (3.0)*	*49.1 (3.6)*	*55.9 (3.2)*	**57.9 (3.1)**	*52.9 (1.9)*	*55.9 (4.4)*	*48.6 (1.7)*	*54.9 (4.1)*	**57.9 (3.3)**	*52.7 (2.1)*

Table S3: Mean (SD) of **precision** (%) obtained within 5-fold cross-validation for different combinations of input signals and derived representations (RB: 20%). If the performance measure is specified as 0, one of the five classes is not modeled at all by the classifier.

Signals	**{A}**	**{U}**	**{A,U}**	**{A,Δ}**	**{U,Δ}**	**{A,D_A}**	**{U,D_U}**	**{A,D_A,U,D_U}**	**{A,Δ,D_A}**	**{U,Δ,D_U}**
F_V	41.1 (3.5)	34.0 (1.7)	43.7 (1.8)	46.1 (2.0)	45.6 (2.3)	44.7 (3.8)	33.1 (4.0)	46.8 (2.2)	**48.4 (4.1)**	45.5 (1.9)
F_{AP}	41.5 (3.5)	37.5 (3.8)	**46.6 (2.4)**	44.9 (2.5)	41.6 (5.5)	40.0 (3.2)	34.7 (1.9)	43.5 (2.7)	44.5 (1.3)	40.9 (3.4)
F_{ML}	36.8 (3.5)	29.3 (7.7)	41.8 (5.5)	**42.3 (5.5)**	30.5 (3.0)	38.3 (2.6)	32.6 (4.4)	39.2 (6.0)	42.0 (4.0)	34.5 (3.1)
COP_{ML}	5.1 (10.2)	0 (0)	22.1 (12.4)	0 (0)	0 (0)	37.3 (5.1)	32.5 (9.2)	**39.3 (4.1)**	37.1 (6.7)	33.3 (8.4)
COP_{AP}	34.3 (4.0)	0 (0)	31.0 (7.0)	29.0 (14.7)	7.7 (9.6)	39.2 (2.2)	33.6 (5.6)	**41.9 (3.8)**	40.3 (1.4)	36.8 (6.7)
GRF	54.2 (2.7)	46.9 (2.7)	54.7 (5.6)	**55.4 (3.5)**	48.7 (4.2)	54.3 (2.4)	45.7 (1.1)	53.1 (2.6)	54.7 (2.0)	51.1 (5.2)
COP	38.0 (6.6)	0 (0)	38.9 (3.1)	36.3 (2.7)	12.3 (10.3)	41.9 (3.4)	32.8 (2.1)	**46.5 (4.0)**	43.0 (2.8)	34.0 (3.8)
GRF + COP	*54.3 (2.8)*	*49.0 (2.5)*	*55.9 (3.3)*	***57.2 (3.5)***	*53.0 (3.1)*	*55.7 (4.5)*	*48.7 (2.0)*	*54.7 (4.1)*	**57.2 (3.1)**	*52.4 (1.8)*

Table S4: Mean (SD) of **recall** (%) obtained within 5-fold cross-validation for different combinations of input signals and derived representations (RB: 20%).

Signals	**{A}**	**{U}**	**{A,U}**	**{A,Δ}**	**{U,Δ}**	**{A,D_A}**	**{U,D_U}**	**{A,D_A,U,D_U}**	**{A,Δ,D_A}**	**{U,Δ,D_U}**
F_V	45.2 (1.6)	37.0 (3.2)	45.7 (1.8)	48.3 (0.8)	48.2 (3.9)	46.3 (3.1)	36.7 (3.3)	48.0 (1.7)	**50.0 (2.6)**	48.9 (1.0)
F_{AP}	44.9 (1.9)	39.2 (2.7)	**48.9 (2.4)**	47.4 (0.5)	45.1 (4.7)	43.0 (2.2)	37.3 (2.0)	45.8 (2.5)	46.3 (2.9)	44.6 (4.1)
F_{ML}	41.3 (2.7)	34.1 (5.9)	43.8 (4.5)	**45.6 (2.9)**	36.0 (2.7)	41.9 (1.5)	35.0 (5.0)	41.5 (4.0)	44.8 (3.0)	37.6 (2.4)
COP_{ML}	28.3 (3.0)	25.6 (1.9)	31.0 (2.8)	28.8 (3.1)	26.1 (1.3)	38.6 (4.0)	34.3 (2.1)	**41.5 (3.6)**	38.6 (4.7)	35.4 (2.0)
COP_{AP}	36.4 (3.7)	26.8 (2.2)	36.0 (4.1)	38.6 (2.6)	31.7 (1.4)	39.5 (2.0)	33.5 (2.1)	**43.4 (3.7)**	41.5 (2.1)	37.5 (2.0)
GRF	56.1 (3.4)	47.5 (3.1)	56.0 (6.1)	**56.6 (2.2)**	51.0 (3.8)	55.6 (2.1)	46.4 (2.3)	53.6 (2.9)	56.3 (1.3)	52.8 (3.2)
COP	39.1 (3.9)	29.8 (1.9)	40.6 (3.0)	39.6 (3.0)	31.5 (1.4)	42.3 (4.0)	35.0 (2.0)	**47.3 (4.3)**	43.6 (3.5)	36.8 (3.1)
GRF + COP	*56.3 (3.0)*	*50.0 (2.2)*	*56.5 (4.3)*	*58.2 (2.1)*	*53.9 (2.9)*	*56.4 (4.8)*	*49.2 (2.7)*	*55.6 (5.3)*	**58.4 (3.0)**	*53.7 (2.1)*

Table S5: Mean (SD) of **F1-scores** (%) obtained within 5-fold cross-validation for different combinations of input signals and derived representations (RB: 20%). If the performance measure is specified as 0, one of the five classes is not modeled at all by the classifier.

Signals	**{A}**	**{U}**	**{A,U}**	**{A,Δ}**	**{U,Δ}**	**{A,D_A}**	**{U,D_U}**	**{A,D_A,U,D_U}**	**{A,Δ,D_A}**	**{U,Δ,D_U}**
F_V	33.9 (17.0)	32.7 (1.9)	43.0 (1.9)	45.4 (2.0)	44.8 (2.6)	44.4 (3.5)	32.6 (4.1)	46.3 (2.1)	**48.1 (3.6)**	45.4 (2.1)
F_{AP}	41.2 (3.5)	34.8 (4.9)	**45.8 (1.3)**	44.5 (2.4)	41.0 (4.5)	39.7 (2.9)	27.4 (13.8)	43.4 (2.4)	43.7 (1.9)	32.9 (16.7)
F_{ML}	36.5 (4.2)	5.8 (11.6)	40.7 (6.2)	**41.8 (5.3)**	18.7 (15.4)	37.5 (2.8)	30.8 (2.9)	38.6 (5.9)	41.1 (4.8)	33.3 (3.7)
COP_{ML}	0 (0)	0 (0)	11.0 (13.7)	0 (0)	0 (0)	35.8 (5.3)	17.8 (14.8)	**37.8 (4.2)**	35.8 (6.6)	12.0 (14.9)
COP_{AP}	23.8 (12.9)	0 (0)	14.0 (17.1)	20.6 (16.9)	0 (0)	36.9 (2.3)	28.1 (2.2)	**40.6 (3.1)**	38.3 (2.1)	26.5 (13.6)
GRF	54.1 (2.6)	44.8 (2.5)	54.3 (5.3)	**54.9 (3.2)**	48.4 (3.6)	53.8 (2.2)	44.7 (1.4)	52.6 (2.8)	54.5 (1.8)	50.4 (4.8)
COP	27.8 (14.5)	0 (0)	36.7 (2.8)	34.3 (3.0)	0 (0)	40.4 (3.1)	30.6 (3.5)	**45.0 (2.9)**	41.6 (2.8)	32.8 (4.5)
GRF + COP	*54.1 (3.0)*	*48.0 (3.2)*	*55.3 (3.2)*	*56.7 (3.1)*	*51.8 (2.2)*	*55.0 (4.3)*	*47.8 (1.7)*	*54.3 (4.3)*	**56.9 (2.9)**	*51.8 (1.8)*

Table S6: **Precision** (%) for different combinations of input signals and derived representations obtained on the independent test set. If the performance measure is specified as 0, one of the five classes is not modeled at all by the classifier.

Signals	**{A}**	**{U}**	**{A,U}**	**{A,Δ}**	**{U,Δ}**	**{A,D_A}**	**{U,D_U}**	**{A,D_A,U,D_U}**	**{A,Δ,D_A}**	**{U,Δ,D_U}**
F_V	40.3	35.9	45.4	42.8	46.0	46.1	35.4	46.7	**48.9**	44.3
F_{AP}	40.6	40.3	43.8	37.6	42.2	41.6	42.1	43.6	45.2	**45.5**
F_{ML}	40.2	29.5	42.0	38.9	32.8	**43.5**	37.2	42.5	41.6	35.7
COP_{ML}	0	0	23.4	0	0	42.9	34.2	42.6	**43.7**	34.3
COP_{AP}	33.7	0	30.8	35.4	27.6	44.1	25.9	43.8	**44.3**	31.3
GRF	55.8	44.8	54.0	54.6	45.8	54.5	44.8	55.8	**59.6**	47.6
COP	32.0	49.7	41.5	40.6	34.4	47.7	32.8	**52.7**	50.9	30.8
GRF + COP	*60.9*	*45.8*	*58.2*	*59.5*	*49.2*	*59.0*	*48.9*	**61.5**	**61.5**	*50.8*

Table S7: **Recall** (%) for different combinations of input signals and derived representations obtained on the independent test set (RB: 20%).

Signals	**{A}**	**{U}**	**{A,U}**	**{A,Δ}**	**{U,Δ}**	**{A,D_A}**	**{U,D_U}**	**{A,D_A,U,D_U}**	**{A,Δ,D_A}**	**{U,Δ,D_U}**
F_V	42.6	38.7	47.2	44.9	47.5	47.5	37.1	46.6	**48.9**	44.3
F_{AP}	44.3	40.7	45.6	42.0	42.3	42.6	40.3	44.3	**46.6**	45.3
F_{ML}	44.3	32.5	44.6	43.3	34.8	**45.6**	37.7	44.6	44.3	38.4
COP_{ML}	28.2	26.9	31.2	26.6	25.3	43.6	34.1	**44.9**	44.6	35.4
COP_{AP}	36.4	26.9	35.1	40.0	33.1	45.3	30.8	45.3	**46.2**	35.1
GRF	56.7	45.6	54.4	55.7	46.9	55.1	45.6	55.4	**60.0**	48.2
COP	37.1	30.8	43.0	41.6	32.1	48.2	34.1	**52.8**	51.8	36.1
GRF + COP	*61.0*	*47.2*	*58.7*	*60.3*	*49.8*	*59.3*	*49.8*	*61.3*	**62.0**	*51.2*

Table S8: **F1-score** (%) for different combinations of input signals and derived representations obtained on the independent test set. If the performance measure is specified as 0, one of the five classes is not modeled at all by the classifier.

Signals	**{A}**	**{U}**	**{A,U}**	**{A,Δ}**	**{U,Δ}**	**{A,D_A}**	**{U,D_U}**	**{A,D_A,U,D_U}**	**{A,Δ,D_A}**	**{U,Δ,D_U}**
F_V	39.3	35.6	45.8	43.5	46.0	46.4	35.5	46.3	**47.9**	43.0
F_{AP}	41.2	38.6	43.5	38.8	40.4	41.5	38.8	43.7	**45.2**	44.2
F_{ML}	38.6	27.6	41.6	38.8	30.2	**43.3**	35.2	42.3	41.5	35.7
COP_{ML}	0	0	0	0	0	41.4	27.8	**42.5**	42.3	29.5
COP_{AP}	28.8	0	29.7	33.5	22.8	43.2	25.2	43.4	**43.8**	29.7
GRF	55.4	45.0	54.1	55.0	46.1	54.6	44.9	55.4	**59.6**	47.7
COP	30.9	21.4	39.0	38.0	23.4	46.5	31.7	**51.3**	50.5	31.4
GRF + COP	*60.7*	*46.1*	*58.3*	*59.7*	*49.3*	*59.0*	*49.1*	*61.2*	**61.7**	*50.7*

2.4 Explaining Machine Learning Models for Clinical Gait Analysis

Djordje Slijepcevic, Fabian Horst, Sebastian Lapuschkin, Brian Horsak, Anna-Maria Raberger, Andreas Kranzl, Wojciech Samek, Christian Breiteneder, Wolfgang Immanuel Schöllhorn, and Matthias Zeppelzauer. **Explaining Machine Learning Models for Clinical Gait Analysis**. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(2):1–27, 2021. DOI: 10.1145/3474121

The final version of this publication is available at: `https://doi.org/10.1145/3474121`. Permission for reprint granted, © 2021 Slijepcevic

Explaining Machine Learning Models for Clinical Gait Analysis

DJORDJE SLIJEPCEVIC, Institute of Creative Media Technologies, Department of Media & Digital Technologies, St. Pölten University of Applied Sciences, Austria
FABIAN HORST, Department of Training and Movement Science, Institute of Sport Science, Johannes Gutenberg-University Mainz, Germany
SEBASTIAN LAPUSCHKIN, Department of Artificial Intelligence, Fraunhofer Heinrich Hertz Institute, Germany
BRIAN HORSAK, Institute of Health Sciences, Department of Health Sciences, St. Pölten University of Applied Sciences, Austria and Center for Digital Health and Social Innovation, St. Pölten University of Applied Sciences, Austria
ANNA-MARIA RABERGER, Institute of Health Sciences, Department of Health Sciences, St. Pölten University of Applied Sciences, Austria
ANDREAS KRANZL, Laboratory for Gait and Movement Analysis, Orthopaedic Hospital Vienna-Speising, Austria
WOJCIECH SAMEK, Department of Artificial Intelligence, Fraunhofer Heinrich Hertz Institute, Germany
CHRISTIAN BREITENEDER, Institute of Visual Computing and Human-Centered Technology, TU Wien, Austria
WOLFGANG IMMANUEL SCHÖLLHORN, Department of Training and Movement Science, Institute of Sport Science, Johannes Gutenberg-University Mainz, Germany
MATTHIAS ZEPPELZAUER, Institute of Creative Media Technologies, Department of Media & Digital Technologies, St. Pölten University of Applied Sciences, Austria

Djordje Slijepcevic and Fabian Horst contributed equally to this research.

This work was partly funded by the Austrian Research Promotion Agency (FFG) and the Austrian Federal Ministry for Digital and Economic Affairs (BMDW) within the COIN-program (ReMoCapLab – #866855 and BigDataAnalytics – #866880), the Lower Austrian Research and Education Company (NFB), the Provincial Government of Lower Austria (IntelliGait3D – #FTI17-014). Further support was received from the German Ministry for Education and Research as BIFOLD (#01IS18025A and #01IS18037A) and TraMeExCo (#01IS18056A), as well as the European Union's Horizon 2020 research and innovation programme through the iToBoS project (#965221).

Authors' addresses: D. Slijepcevic and M. Zeppelzauer, Institute of Creative Media Technologies, Department of Media & Digital Technologies, St. Pölten University of Applied Sciences, St. Pölten, Austria; email: Djordje.Slijepcevic@fhstp.ac.at; F. Horst and W. I. Schöllhorn, Department of Training and Movement Science, Institute of Sport Science, Johannes Gutenberg-University Mainz, Mainz, Germany; email: horst@uni-mainz.de; S. Lapuschkin and W. Samek, Department of Artificial Intelligence, Fraunhofer Heinrich Hertz Institute, Berlin, Germany; B. Horsak and A.-M. Raberger, Institute of Health Sciences, Department of Health Sciences, St. Pölten University of Applied Sciences, St. Pölten, Austria; A. Kranzl, Laboratory for Gait and Movement Analysis, Orthopaedic Hospital Vienna-Speising, Vienna, Austria; C. Breiteneder, Institute of Visual Computing and Human-Centered Technology, TU Wien, Vienna, Austria.

2637-8051/2021/12-ART14
https://doi.org/10.1145/3474121

Machine Learning (ML) is increasingly used to support decision-making in the healthcare sector. While ML approaches provide promising results with regard to their classification performance, most share a central limitation, their black-box character. This article investigates the usefulness of *Explainable Artificial Intelligence* (XAI) methods to increase transparency in automated *clinical gait classification* based on time series. For this purpose, predictions of state-of-the-art classification methods are explained with a XAI method called Layer-wise Relevance Propagation (LRP). Our main contribution is an approach that explains class-specific characteristics learned by ML models that are trained for gait classification. We investigate several gait classification tasks and employ different classification methods, i.e., Convolutional Neural Network, Support Vector Machine, and Multi-layer Perceptron. We propose to evaluate the obtained explanations with two complementary approaches: a statistical analysis of the underlying data using Statistical Parametric Mapping and a qualitative evaluation by two clinical experts. A gait dataset comprising ground reaction force measurements from 132 patients with different lower-body gait disorders and 62 healthy controls is utilized. Our experiments show that explanations obtained by LRP exhibit promising statistical properties concerning inter-class discriminativity and are also in line with clinically relevant biomechanical gait characteristics.

CCS Concepts: • **Computing methodologies → Neural networks**; • **Applied computing → Health care information systems;**

Additional Key Words and Phrases: Explainable artificial intelligence, clinical gait analysis, human gait classification, layer-wise relevance propagation, statistical parametric mapping, ground reaction forces, convolutional neural networks

ACM Reference format:
Djordje Slijepcevic, Fabian Horst, Sebastian Lapuschkin, Brian Horsak, Anna-Maria Raberger, Andreas Kranzl, Wojciech Samek, Christian Breiteneder, Wolfgang Immanuel Schöllhorn, and Matthias Zeppelzauer. 2021. Explaining Machine Learning Models for Clinical Gait Analysis. *ACM Trans. Comput. Healthcare* 3, 2, Article 14 (December 2021), 27 pages.
https://doi.org/10.1145/3474121

1 INTRODUCTION

Artificial Intelligence (AI) and **Machine Learning (ML)** techniques have become almost ubiquitous in our daily lives by supporting or guiding our decisions and providing recommendations. Impressively, there are certain medical tasks, such as the detection of skin or breast cancer, that ML approaches have already been able to solve more efficiently and effectively than humans [16, 21, 42]. Therefore, it is not surprising that ML approaches are currently becoming popular in the healthcare sector [73]. This trend has also been recognized in the field of **clinical gait analysis (CGA)** [18, 61]. CGA focuses on the quantitative description and analysis of human gait from a kinematic (i.e., joint angles), kinetic (i.e., ground reaction forces and joint moments), and muscular (i.e., electromyographic activity) point of view [9, 79]. Thereby, CGA produces a vast amount of data [22, 54], which are difficult to comprehend due to their multi-dimensional and multi-correlated nature [13, 80]. In recent years, ML methods have been successfully employed in CGA for the classification of patient groups [18, 61], such as stroke [36], Parkinson's disease [76], cerebral palsy [74], multiple sclerosis [3], osteoarthritis [50], and patients suffering from different functional gait disorders [66]. While ML approaches yield promising results regarding classification performance, most share a central limitation, which is their black-box character [1]. This means that even if the underlying mathematical principles of these methods are understood, it is often unclear why a particular prediction has been made and if meaningfully grounded patterns have led to this prediction. Additionally, the black-box character hinders ML approaches to provide justifications of their predictions. This is, however, necessary for compliance with legislation such as the **General Data Protection Regulation (GDPR**, EU 2016/679**)** [1, 17, 23]. These factors currently limit the application of ML-based decision-support systems in medical practice [26, 59].

Due to the aforementioned reasons, the field of ***Explainable Artificial Intelligence*** **(XAI)** gained increasing attention in recent years. Different approaches have been proposed (see Section 2: Related work). In general, XAI methods intend to illustrate how complex and non-linear ML models operate and how they produced their predictions. However, explanation is understood in the sense of providing more differentiated insights into the behaviour of ML models to fathom the dependence of the results on input variables (without claiming to give causation). Even though research in XAI is still in an early stage, the application of such approaches in medicine has already raised attention [26, 72]. The motivation is to increase the traceability of ML models and trust in them among medical professionals [27]. However, the application of XAI methods to the field of CGA remains to be investigated. A first step in this direction has recently been taken by Horst et al. [29] for explaining predictions in gait-based person recognition.

The primary aim of this article is to investigate and explain which class-specific characteristics ML models learn from CGA data, i.e., time series. For this purpose, we train several classification models for different gait classification tasks and extract prediction explanations from the trained models via **Layer-wise Relevance Propagation (LRP)**. Subsequently, the explanations of the individual predictions are aggregated to obtain class-specific model explanations. The assessment of the resulting explanations is, however, a challenge, since no ground truth exists for automatically generated explanations in CGA. In contrast to images, which are more frequently subject to explainability studies [2, 19, 57, 58], the evaluation of explanations becomes particularly challenging when the input signals are more abstract and thus not straightforward to interpret, as often is the case with biomedical signals. Recently, it has been shown that XAI approaches do not necessarily refer to the actual prediction of the classification model and sometimes even build upon unrelated information [2]. Thus, a more comprehensive investigation of explanations obtained by XAI methods is necessary to verify whether they are meaningful and justified. To account for the above-mentioned challenges, we suggest a two-step approach for the evaluation of the obtained explanations. First, we analyze the discriminatory power of the obtained explanations from a statistical perspective. For this purpose, we leverage **Statistical Parametric Mapping (SPM)** [51], a method building upon random field theory, to derive statistical measures along with the input signals and thereby investigate how statistically justified the obtained explanations are. Second, two experienced clinical experts interpret the explainability results from a clinical perspective to evaluate whether obtained explanations match characteristics from clinical practice.

Our investigation focuses on two leading research questions:

(1) Which input features or signal regions are most relevant for automatic gait classification?
(2) To what extent are input features or signal regions identified as being relevant for a given gait classification task statistically justified and in line with clinical assessment?

In addition to these two leading questions, we investigate several further aspects that may influence classification performance as well as explainability in more detail, including the influence of different classification methods, the impact of data normalization, and the role of different input signal components (i.e., the horizontal forces, measurements of the affected leg, and measurements of the unaffected leg). We perform our investigation on the GaitRec dataset [28], which contains **ground reaction force (GRF)** measurements from clinical practice. We design prediction models for different gait classification tasks and derive possible explanations from the resulting models that are based on relevance scores. These relevance scores are directly related to specific regions in the input signal. Subsequently, we analyze the explanations from a statistical as well as a clinical perspective. The results show that explanations share promising statistical properties concerning class discriminativity and thus indicate that predictions are grounded on statistically justified information for the task. Further, we show that input features considered as relevant can also be interpreted as meaningful and clinically relevant biomechanical gait characteristics. Overall, our investigation demonstrates the usefulness of XAI in the domain of gait classification, exemplifies how to apply XAI methods to gait measurement data, and suggests approaches to evaluate their quality. The performed study suggests that XAI methods can be useful to better understand and interpret automatic predictions in clinical gait analysis and thus has the potential to yield an added value for clinical practice in the future.

2 RELATED WORK

Methods from XAI can be grouped according to the type of explanation they provide. We distinguish between XAI approaches for (i) **data exploration**, (ii) **prediction explanation**, and (iii) **model explanation** based on an adaptation of the taxonomy introduced by Arya et al. [6]. In the following, we briefly introduce the three different types of approaches and their capabilities.

Data exploration includes methods from the fields of visual analytics, statistics and unsupervised machine learning. As such, the methods are not capable of explaining a model but rather the data on which the model is trained. These methods focus on projecting the data into a space where it is possible to find meaningful structures or clusters and thus understand the data in more detail. A popular approach for data exploration introduced by Maaten and Hinton [39] is **t-distributed Stochastic Neighbor Embedding (t-SNE)**, which projects high-dimensional data into a lower-dimensional and visualizable space. The projection is performed in a way that the cluster structure in the original data space is optimally exposed. Thereby, an understanding of the data and the identification of typical patterns and clusters in the data is facilitated. Other approaches in this category are visual analytics approaches that employ advanced techniques for the interactive visualization of data to support data exploration, i.e., finding characteristic patterns or dependencies within data [75, 77].

Prediction explanation aims at explaining the local behavior of a model, i.e., the prediction for a given input instance. For a classification task, these methods can provide, for example, explanations about which part of the input influenced the classifier's prediction the most. For classification of gait data, the explanation should highlight all relevant signal regions and characteristic signal shapes in the input data, which are associated with a particular gait disorder. Two main categories can be distinguished for explaining the local behavior of a machine learning model: (i) self-explaining models and (ii) post-hoc methods.

Self-explaining models integrate components that learn relationships between input data and predictions during training. Simultaneously, they learn how these relationships relate to terms from a predefined dictionary and consequently generate explanations from them. A self-explaining approach that does not visually highlight relevant regions in input data but generates textual explanations was proposed by Hendricks et al. [24]. This self-explaining model combines a **Convolutional Neural Network (CNN)** and a **Recurrent Neural Network (RNN)**. The CNN learns discriminative features to perform a classification task, while the RNN generates textual explanations of the prediction. This approach cannot be applied to a previously trained model in a post-hoc manner, which limits its practical applicability.

Post-hoc methods provide much greater applicability, as they can be applied to already-trained models. These methods can be further categorized into (i) propagation-based, (ii) perturbation-based, and (iii) Shapley-value-based methods. *Propagation-based methods* determine the contributions of each input feature by (back-)propagating some quantity of interest from the model's output layer to the input layer. Sensitivity Analysis [82, 83] has been introduced to **Support Vector Machines (SVMs)** [8] and CNNs [65] in the form of saliency maps. **Layer-wise Relevance Propagation (LRP)** [7, 44] and **Deep Learning Important FeaTures (DeepLIFT)** [63] are methods that propagate importance scores from the output layer back to the input, thereby enabling the identification of positive and negative evidences for a specific prediction. Sensitivity Analysis and the therewith obtained explanations, in general, suffer from the effects of shattered gradients [10], especially so in more complex (deeper) networks. Modern approaches to CNN explainability, such as LRP or DeepLift, do not have this problem and work well for a wider range of network architectures and models in general [32, 46]. *Perturbation-based methods*, such as those introduced by Fong and Vedaldi [19] or Zintgraf et al. [81], treat the model as a black box and estimate the importance of input features by (partially) occluding the input and measuring the effect on the model output. While some methods produce explanations directly from a perturbation process, others employ a learning component, e.g., the **Interpretable Model-agnostic Explanations (LIME)** method [55], to estimate locally interpretable surrogate models mimicking the prediction process of the black-box model. Perturbation-based methods can be considered to be model-agnostic, as they do not require access to internal model parameters or structures to operate. However, this model-agnosticism is bought at a considerable computational cost, compared to propagation-based approaches. *Shapley-value-based methods* are rooted in game theory [84] and attempt to approximate the Shapley values of a given prediction. For this purpose, the effect of omitting an input feature is examined, taking into account all possible combinations of other input features, which can be included or excluded [71]. Lundberg and Lee [38] proposed the **SHapley Additive exPlanations (SHAP)** method, which is a unified approach building upon the theory of Shapley values and existing propagation-based and perturbation-based methods, e.g., LIME, DeepLIFT, and LRP.

Model explanation provides an interpretation of what a trained model has learned, i.e., the most characteristic representations or prototypes for an entire class are visualized (e.g., a class of gait disorders in CGA). These methods can indicate which classes overlap and point out ambiguous input features. In addition to saliency maps, Simonyan et al. [65] proposed a method for generating a representative visualization for a specific class that was learned by a CNN. For this purpose, they applied activation maximization, i.e., starting with a blank image, each pixel is changed by utilizing back-propagation so the activity of a neuron is increased. The resulting visualizations give a first impression about the patterns learned but are highly abstract and can only be interpreted to a limited extent. To generate visualizations that are easier to interpret, Nguyen et al. [48] proposed a method to constrain the optimization process by image priors that were learned automatically. Lapuschkin et al. [35] proposed the **Spectral Relevance Analysis (SpRAy)**, which summarizes a model's learned strategies by analyzing similarities and dissimilarities over large quantities of input relevance maps computed with respect to a category of interest.

For gait classification, prediction explanation is desirable to provide clinical experts with detailed information about which patterns in the input signals are important for a specific prediction. Additionally, based on aggregations of these explanations, differences between patient groups can be assessed, i.e., in terms of class-specific model explanations. In this context, post-hoc methods are preferable, because they provide a classifier-agnostic approach (can be applied to any classification model) and do not require retraining or additional labels. We, therefore, choose an established post-hoc explainability method, i.e., LRP, in our experiments.

3 APPROACH AND METHODOLOGY

The general approach we followed in this study was to design and train classification models for automated gait classification tasks (see Figure 1(B)) based on three-dimensional **ground reaction forces (GRFs)** of both legs (see Figure 1(A)), to explain the predictions of these models based on relevance scores that are related to the input signal space by using LRP (see Figure 1(C)) and to evaluate these results from a statistical (see Figure 1(D)) and a clinical perspective (see Figure 1(E)). The experimental setup, including a detailed description of the data (pre-) processing and classification pipeline, can be found in Section 4.

3.1 Gait Classification

The main task in automated gait classification is to determine whether a person has a healthy or pathological gait pattern based on gait measurements. We employed three-dimensional GRFs of the affected and unaffected sides as input signals and investigated the classification performance of several state-of-the-art classification methods. Furthermore, the input signals were fed directly into the classification models. This ensures that the results of the employed explainability method (LRP) can be directly mapped to the original signals. For easier interpretation of the XAI results, we refrained from using data reduction techniques such as, e.g., **Principal Component Analysis (PCA)**, which is a common practice in automated gait classification [12, 22, 68].

3.2 Prediction Explanation

We employed **Layer-wise Relevance Propagation (LRP)** for prediction explanation [7] as a propagation-based post-hoc method that provides explanations in the input space, which is the space where the signals are usually interpreted by experts in clinical practice. LRP reversely iterates over the layered structure of an ML model to produce an explanation. Consider a neural network:

$$f(x) = f_L \circ \cdots \circ f_1(x). \tag{1}$$

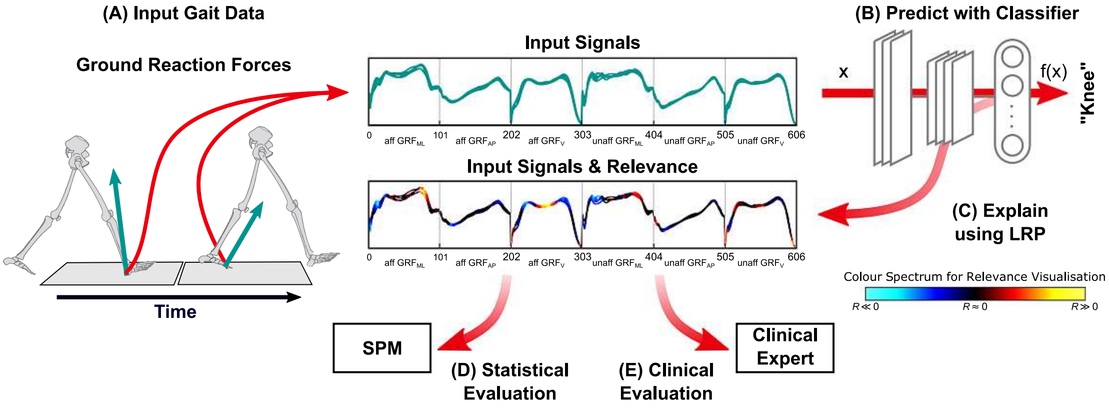

(A) Input Gait Data
Ground Reaction Forces
Time
Input Signals
0 aff GRF$_{ML}$ 101 aff GRF$_{AP}$ 202 aff GRF$_{V}$ 303 unaff GRF$_{ML}$ 404 unaff GRF$_{AP}$ 505 unaff GRF$_{V}$ 606
Input Signals & Relevance
0 aff GRF$_{ML}$ 101 aff GRF$_{AP}$ 202 aff GRF$_{V}$ 303 unaff GRF$_{ML}$ 404 unaff GRF$_{AP}$ 505 unaff GRF$_{V}$ 606
(B) Predict with Classifier
x
f(x)
"Knee"
(C) Explain using LRP
Colour Spectrum for Relevance Visualisation
R≪0
R≈0
R≫0
SPM
(D) Statistical Evaluation
(E) Clinical Evaluation
Clinical Expert

Fig. 1. Overview of our workflow for data acquisition, prediction, and prediction explanation in automated gait classification, showing the data of one participant belonging to the knee disorder class. (A) The clinical gait analysis consists of five recordings of each participant walking barefoot (unassisted) a distance of 10 m at a self-selected walking speed. Two centrally embedded force plates capture the three-dimensional ground reaction forces (GRFs) during the stance phase of the right and left foot. (B) The GRF comprising the medio-lateral (GRF_{ML}), anterior-posterior (GRF_{AP}), and vertical (GRF_V) force components of the affected and unaffected side are used as time-normalized and concatenated input vector x (1×606-dimensional) for the prediction of the knee disorder class using a classifier (e.g., CNN). (C) Decomposition of input relevance scores is achieved using LRP. The color spectrum for the visualization of input relevance scores of the model predictions is shown in the bottom right corner. Black line segments are irrelevant to the model's prediction. Warm hues identify input segments causing a prediction corresponding to the class label, while cool hues are features contradicting the class label. (D) Statistical and (E) Clinical evaluation of class-specific (averaged) relevance scores.

An SVM model can be regarded as a single-layer neural network and thus a special case of Equation (1). In a forward pass, activations are computed at each layer f_l of the neural network, depending on the learned parameters of the model and the previous layers' activations. The activation score in the output layer f_L forms the prediction $f(x)$, which is then, for a specific class and neuron of interest, back-propagated and redistributed layer by layer until the input is reached. The method yields time- and signal-resolved input relevance scores R_i for each individual value of the input vector x_i. The redistribution process follows a conservation principle analogous to Kirchhoff's laws in electrical circuits, i.e., all relevance assigned to any neuron during the back-propagation process is redistributed without loss to its inputs in the underlying layer. The relevance back-propagation flow is illustrated in Figure 2.

Various purposeful propagation rules have been proposed in the literature [7, 32, 44]. For example, the LRP$_\varepsilon$ rule [7] is defined as:

$$R_{j \leftarrow k} = \frac{z_{jk}}{z_k + \varepsilon \cdot \text{sign}(z_k)} R_k, \tag{2}$$

where $z_{jk} = a_j w_{jk}$ is the quantity propagated from the jth input neuron to the kth output neuron within a given layer, depending on the input activation a_j and the learned weight parameters w_{jk}. The $z_k = \sum_j z_{jk}$ is the pre-activation of the kth output neuron, aggregating all forward-propagated z_{jk}, which includes any potential bias terms. The variable $\varepsilon \geq 0$ is a free parameter to tune the decomposition rule with the intent to suppress noisy forward activations z_{jk} and divisions by zero.[1] Equation (2) redistributes R_k proportionally based on the relative contribution of z_{jk} to z_k towards all input components j. After the step of relevance decomposition, lower layer

[1] Note that for this purpose the sign function is defined as: $\text{sign}(x) = 1$ iff. $x \geq 0$; else -1; [7].

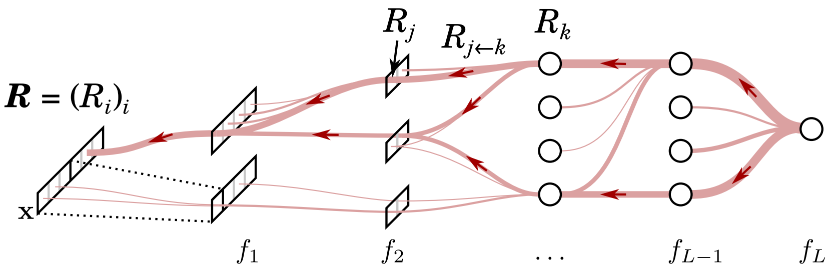


Fig. 2. Illustration of the LRP back-propagation procedure applied to a neural network function $f(x) = f_L \circ \cdots \circ f_1(x)$. The prediction at the output is propagated backward in the network, until the input features are reached and relevance scores are obtained for all input features and hidden units as R_i, R_j, and R_k, respectively. The propagation flow is shown in red color.

neuron relevance is aggregated from incoming relevance messages as $R_j = \sum_k R_{j\leftarrow k}$. Other propagation rules, such as LRP$_\gamma$ [44], LRP$_{\alpha\beta}$, LRP$_{z^\mathcal{B}}$, or LRP$_\flat$, are suitable for other application scenarios, layer types, or particularly deeper neural networks [32, 44, 58] and have been shown to work well in practice [57].

LRP enables to explain the prediction of an ML model as partial contributions of an individual input value. LRP indicates which information a model uses to predict in favor or against an output class. Thereby, it enables the interpretation of input relevance scores and their dynamics as representation for a certain class (i.e., healthy controls or functional disorders in ankle, knee, or hip).

For the explanation of predictions, we decomposed the input relevance scores of each gait trial with LRP. To analyze patterns learned for a specific class, we used LRP to decompose the ground truth label (and not necessarily the predicted value) of the trial. For the visualization of the explanations, we averaged the underlying GRF signals and the resulting input relevance scores over all trials of a class.

Given that the models investigated in this study are comparatively shallow and are largely unaffected by detrimental effects such as gradient shattering [10, 44, 45], we performed relevance decomposition according to LRP$_\varepsilon$ with $\varepsilon = 10^{-5}$ in all layers across the different models (except for the CNN for which we employed the LRP$_\flat$ rule at the input layer, which uniformly distributes a neuron's relevance score R_k across its receptive field, disregarding any applied transformations w_{jk} or input activations a_j) [32].

3.3 Statistical Evaluation

To evaluate the derived relevance scores of LRP, we employed **Statistical Parametric Mapping (SPM)** [51, 52], which recently received increased attention in the gait analysis community [11, 49]. While standard inference statistical approaches tend to reduce time-continuous signals to single time-discrete values for statistical testing, SPM allows to use the entire time-continuous signals to make probabilistic conclusions. It follows the same notion and logic as classical inference statistics. The main advantages of SPM are that the statistical results are presented in the original sampling space and that there is no need for a (potentially biasing) parameterization technique [51, 52]. Since the LRP explanations and the results of SPM reside in the same space (the input signal space), we can leverage SPM to demonstrate the meaningfulness of LRP explanations from a statistical point of view.

LRP and SPM can both be considered explainability approaches, however, they target different goals. SPM fits linear models (e.g., general linear models) to the data and tries to explain differences in the data (i.e., differences between groups or classes). SPM can thus be considered a data-centric explainability method. LRP tries to explain the inner working of complex (non-linear) models and can thus be considered a model-centric explainability method. Both methods are thus complementary to each other. Another difference is that LRP can explain

Table 1. Demographic Details of the Employed Dataset for Each Pre-defined Class

Classes	N	Age (yrs.) Mean (SD)	Body Mass (kg) Mean (SD)	Gender (m/f)	Walking Speed (m/s)	Num. Trials
Healthy Control	62	36.0 (10.8)	72.3 (15.0)	28/34	4.1 (0.3)	310
Hip	37	44.2 (12.5)	81.4 (14.1)	31/6	3.7 (0.3)	185
Knee	52	43.5 (13.8)	85.6 (16.4)	37/15	3.5 (0.4)	260
Ankle	43	42.6 (10.9)	91.6 (20.4)	36/7	3.4 (0.4)	215
Total	**194**	**41.1 (12.4)**	**81.9 (18.0)**	**132/62**	**3.7 (0.5)**	**970**

individual model predictions (even without using ground-truth information), while SPM explains data characteristics by taking the ground truth information (group or class information) into account. As part of Section 6.3, we will discuss the results obtained with both approaches to address the additional value of LRP in CGA.

For the statistical evaluation, we computed independent t-tests using the SPM1D[2] package provided by Pataky [52] for Matlab and investigate differences between each GRF signal between two classes (for visualization purposes, we concatenated the results obtained on each GRF component). To take into account the dependence of SPM results on the choice of a distinct alpha level, we performed experiments with three different alpha levels: 0.01, 0.05, and 0.1. The output of SPM provides t-values for each point of the investigated time series and the threshold corresponding to the chosen alpha level. The t-values exceeding this threshold indicate statistically significant differences in the corresponding sections of the time series. For a better visibility, we depicted these significant sections as gray-shaded areas in Figure 5 and Figure 6. We used three different shades of gray for the three different alpha levels, i.e., dark gray for 0.01, gray for 0.05, and light gray for 0.1. Additionally, we computed the *effect size* by transforming the resulting t-values to Pearson's correlation coefficient r using the definition by Rosenthal [56]. The effect size provides an indicator for the discriminativeness of a given signal region independent of the alpha level.

3.4 Clinical Evaluation

To evaluate the derived relevance scores of LRP from a clinical perspective, two clinical experts with more than 10 and more than 25 years' experience in human gait analysis analyzed the explainability results. The experts evaluated the extent to which regions with the highest input relevance scores correspond to GRF characteristics from clinical practice and assessed the usefulness of explainability approaches for CGA.

4 EXPERIMENTAL SETUP

4.1 Data Recording and Dataset

For the gait classification task, we utilized a subset of the large-scale GaitRec dataset [28]. This dataset is part of an existing clinical gait database maintained by a local Austrian rehabilitation center. Before conducting our experiments approval was obtained from the local Ethics Committee (#GS1-EK-4/299-2014). The employed dataset contains bilateral three-dimensional GRF recordings of patients and healthy controls walking unassisted at self-selected walking speed on an approximately 10 m walkway with two centrally embedded force plates (Kistler, Type 9281B12, Winterthur, CH). Data were recorded at 2,000 Hz, filtered with a zero-lag Butterworth filter of 2nd order with a cut-off frequency of 20 Hz, time-normalized to 101 points (100% stance phase), and amplitude-normalized to 100% body weight. During one session, participants walked barefoot or in socks until a minimum of five valid recordings were available. Recordings were defined as valid by an experienced assessor.

[2] SPM1D v.0.4, http://www.spm1d.org/.

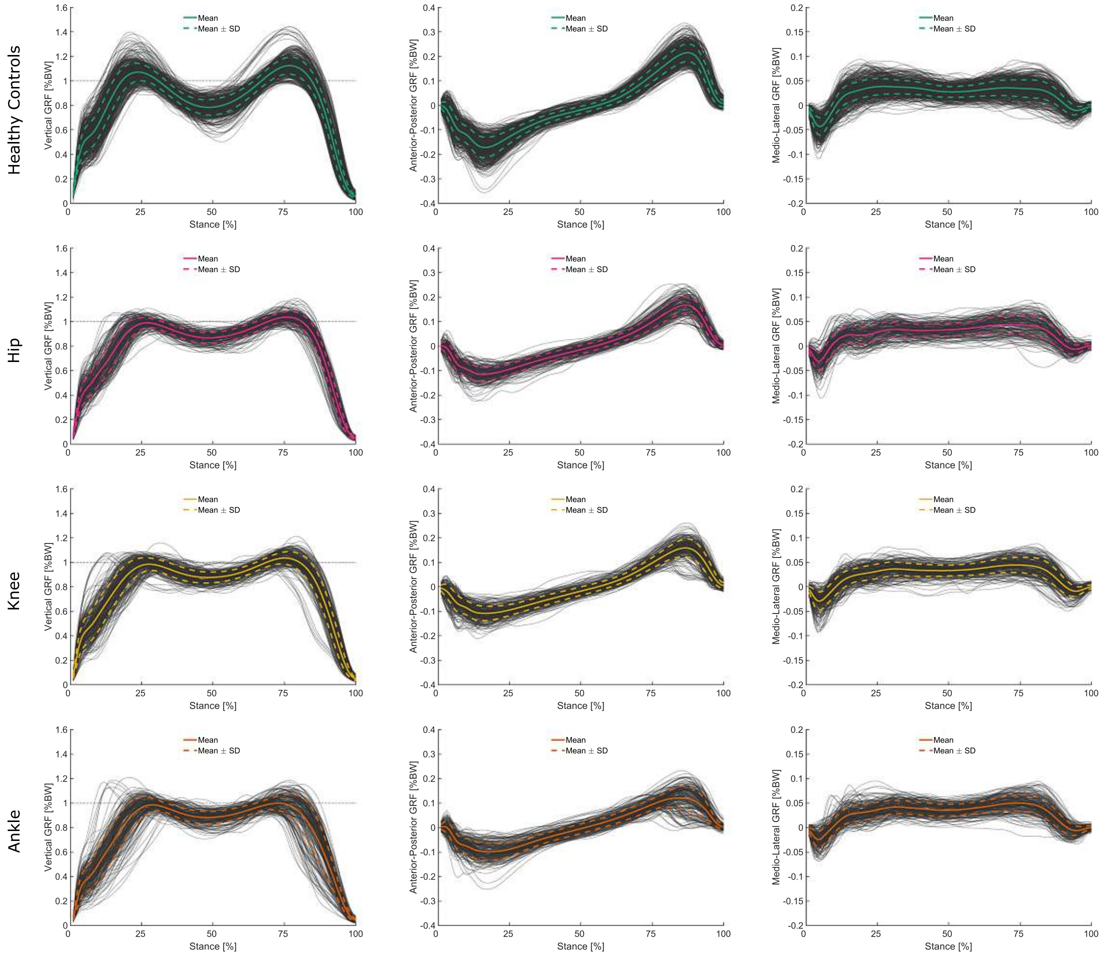


Fig. 3. Visualization of vertical (left panel), anterior-posterior (central panel), and medio-lateral (right panel) force components of the body weight-normalized GRF measurements of the affected side available per participant and class. For healthy controls, all available measurements are visualized. Mean and standard deviation signals (calculated per class) are highlighted as solid and dashed colored lines.

In total, the dataset comprises GRF measurements from 132 patients with lower-body **gait disorders (*GD*)** and data from 62 **healthy controls (*HC*)**, both of various physical composition and gender. The dataset includes three classes of orthopaedic gait disorders associated with the **hip** (***H***, N = 37), **knee** (***K***, N = 52), and **ankle** (***A***, N = 43). For class-specific demographic details of the data, refer to Table 1. The dataset is balanced regarding the number of recorded sessions per person and the number of trials per person. Figure 3 shows an overview of all GRF measurements of the affected side (except for healthy controls where each step is visualized) per class and the associated mean and standard deviation. The *GD* classes (*A*, *H*, and *K*) include patients after joint replacement surgery, fractures, ligament ruptures, and related disorders associated with the above-mentioned anatomical areas. A well-experienced physical therapist with more than a decade of clinical experience manually labeled the dataset based on the available medical diagnosis of each patient.

4.2 Input Data Preparation

The input data for each classification task is a concatenated version of the three-dimensional GRF signals from both force plates (see Figure 1). The concatenation of all six GRF signals (three force components per force plate) results in a 1 × 606-dimensional input vector for each gait trial. The three-dimensional GRF signals are the medio-lateral horizontal force (GRF_{ML}), anterior-posterior horizontal force (GRF_{AP}), and vertical force (GRF_V). The dataset includes only unilateral gait disorders, i.e., disorders where the main physical limitation can be attributed to one leg (the *affected leg/side* in the following). The signal components of the affected leg (input features: 1 to 303) are concatenated first and are followed by the signal components of the unaffected leg (input features: 304 to 606) in the input vector. For the healthy controls there is no affected and unaffected side (both sides are unaffected). Thus, the order of the signals was randomly assigned, while ensuring an equal distribution, to avoid any bias regarding the side.

4.3 Data Normalization

Normalization of input vectors is applied to ensure an equal contribution of all six GRF signals to the classification models and thus avoids that signals with larger numeric ranges dominate those with smaller numeric ranges [14, 31]. We applied min-max normalization to the input signals and thereby scaled each signal to the range $[0, 1]$. The global minimum and maximum values were determined separately for each of the six GRF signals over all trials.

4.4 Classification Tasks

We investigate different classification tasks on the dataset introduced above to provide a more comprehensive picture of the investigated problem and to enable the differentiation between task-specific and general observations. Classification tasks include:

- binary classification between healthy controls and all gait disorders (HC/GD),
- binary classification between healthy controls and each gait disorder separately (i.e., HC/H, HC/K, and HC/A),
- multi-class classification between healthy controls and all gait disorders ($HC/H/K/A$),
- and multi-class classification between all gait disorders ($H/K/A$).

4.5 Classification Methods

In our experiments, three representative machine learning approaches, i.e., (linear) SVM, MLP, and CNN were compared in terms of prediction accuracy and learned input relevance patterns. The SVM models were trained using a standard quadratic optimization algorithm, with an error penalty parameter $C = 0.1$ and ℓ_2-constrained regularization of the learned weight vector w. The MLP models comprised three consecutive fully connected layers with ReLU non-linearities activating the hidden neurons and a final SoftMax activation in the output layer. The size of both hidden layers is 768, whereas the size of the output layer is c, where c is the number of target classes. The CNN models process the given data via three consecutive convolutional layers, with a <filter size>-<stride>-<output channel> configuration of 8-2-24, 8-2-24, and 6-3-48, and ReLUs for non-linear neuron activation. The resulting 48 × 48 feature mapping is then unrolled into a 2,304-dimensional vector and fed into a fully connected layer, which directly maps to the model output. This fully connected layer is topped with a SoftMax output activation, which is acting as a multi-class predictor output towards the c target classes. Both, the MLP and CNN models, have been trained via standard error back-propagation using stochastic gradient descent [37] and a mean absolute (ℓ_1) loss function. The training procedure was executed for $3 \cdot 10^4$ iterations of mini batches of five randomly selected training samples and an initial learning rate of $5 \cdot 10^{-3}$. The learning rate was gradually decreased after every 10^4-th training iteration to 10^{-3} by a factor of 0.2 and then to $5 \cdot 10^{-4}$ by a factor of 0.5. Model weights were initialized with random values drawn from a normal distribution with $\mu = 0$

and $\sigma = m^{-\frac{1}{2}}$, where m is the number of inputs to each output neuron of the layer [37]. Since the CNN receives as input a 1 × 606-dimensional input vector, its convolution operations can be understood as 1D convolutions, moving over the time axis only. We used 1D convolutions to maintain comparability with the two other classification methods (MLP and SVM). Preliminary experiments demonstrated negligible differences between 1D and 2D CNNs.

4.6 Performance Evaluation

The prediction accuracies were reported over a stratified 10-fold cross-validation configuration, where eight partitions of the data are used for training, one partition is used as validation set, and the remaining partition is reserved for testing. The samples from each class were distributed evenly while ensuring that all gait trials from an individual participant were placed in the same partition of the data to rule out person-related information influencing the measured model performance during testing. All results are reported as mean with **standard deviation (SD)**, unless otherwise stated. Additionally, we calculated the **Zero-Rule baseline (ZRB)** for each classification task. The ZRB refers to the theoretical accuracy obtained by assigning class labels according to the prior probabilities of the classes, i.e., the target labels are always set to the class with the greatest cardinality in the training dataset.

4.7 Implementation

The implementation of the three ML methods and the LRP method was conducted within the software framework Python 3.7 (Python Software Foundation, USA). Data preprocessing, SPM, and the visualization of the results were performed in Matlab 2017b (MathWorks, USA). Our source code and the utilized dataset are publicly available at: https://github.com/sebastian-lapuschkin/explaining-deep-clinical-gait-classification.

5 RESULTS

We first present the results obtained in our classification experiments as well as from the explainability analysis and then discuss them in detail in Section 6. We start with a presentation of the classification accuracies achieved for the different classification methods, tasks, and normalization methods (Section 5.1) and continue with a presentation of the explainability results obtained by LRP (Section 5.2).

5.1 Classification Results

The mean prediction accuracy showed a clear superiority over the ZRB for all three classification methods (CNN, SVM, and MLP) and all classification tasks (see Figure 4 and supplementary Table S1). A 2 × 2 repeated measures **analysis of variance (ANOVA)** (classification method: CNN, SVM, and MLP; normalization: min-max and non-normalized) conducted for each classification task only indicated a significant difference in classification accuracy between the three classifiers for task HC/GD ($F_{2,18}$ = 4.038, p = 0.036, η_p^2 = 0.310). However, differences were in general not relevant (<2%) and additional pairwise Bonferroni-corrected post-hoc tests failed to identify any differences as significant. No other significant differences were found for the classifiers' performances. Regarding normalization, ANOVA revealed two simple main effects of normalization for task $H/K/A$ ($F_{1,9}$ = 7.269, p = 0.025, η_p^2 = 0.447) and task $HC/H/K/A$ ($F_{1,9}$ = 9.054, p = 0.015, η_p^2 = 0.502). Estimated marginal means for normalization during Bonferroni-corrected post-hoc tests showed a performance increase of 6% and 3% for $H/K/A$ and $HC/H/K/A$, respectively. No further significant effects and differences were found.

5.2 Explainability Results

In the following, we present in detail the results for classification task HC/GD together with respective result visualizations. Figure 5 shows an exemplary result for prediction explanation by LRP, i.e., the averaged signals together with the color-coded averaged relevance values for each of the 606 input values for task HC/GD with

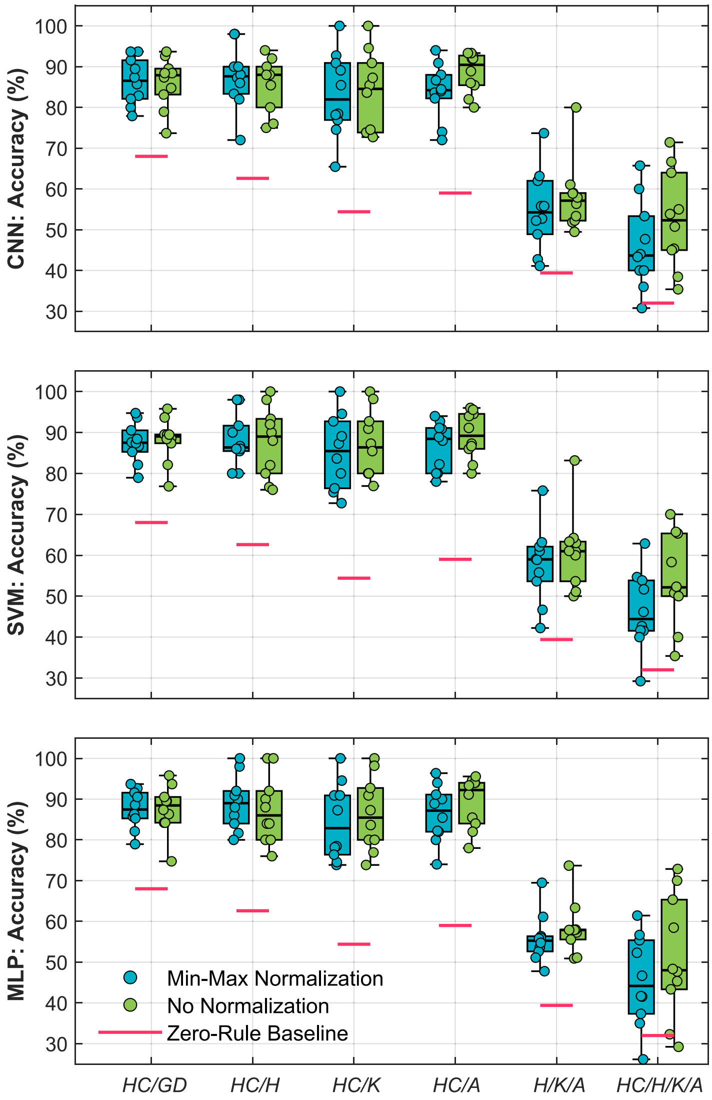


Fig. 4. Overview of the prediction accuracy obtained for the three employed classification methods (CNN, SVM, and MLP) and all classification tasks with min-max normalized and non-normalized input signals, reported as boxplots enhanced with the classification accuracies obtained over 10-fold cross-validation (represented as individual dots).

min-max normalized GRF signals. The input relevance values point out which GRF characteristics were most relevant for (or contradictory to) the classification of a certain class (*HC* or *GD*). For visualization, input values neutral to the prediction ($R_i \approx 0$) are shown in black color, while warm hues indicate input values supporting the prediction ($R_i \gg 0$) of the analyzed class and cool hues identify contradictory input values ($R_i \ll 0$). For binary classification tasks (*HC/GD*, *HC/H*, *HC/K*, and *HC/A*), note that a high input relevance value for one class results in a contradictory input relevance value for the other class. Therefore, the total relevance, which is the absolute sum of the relevance scores of both classes, is a good indicator for the overall relevance of an

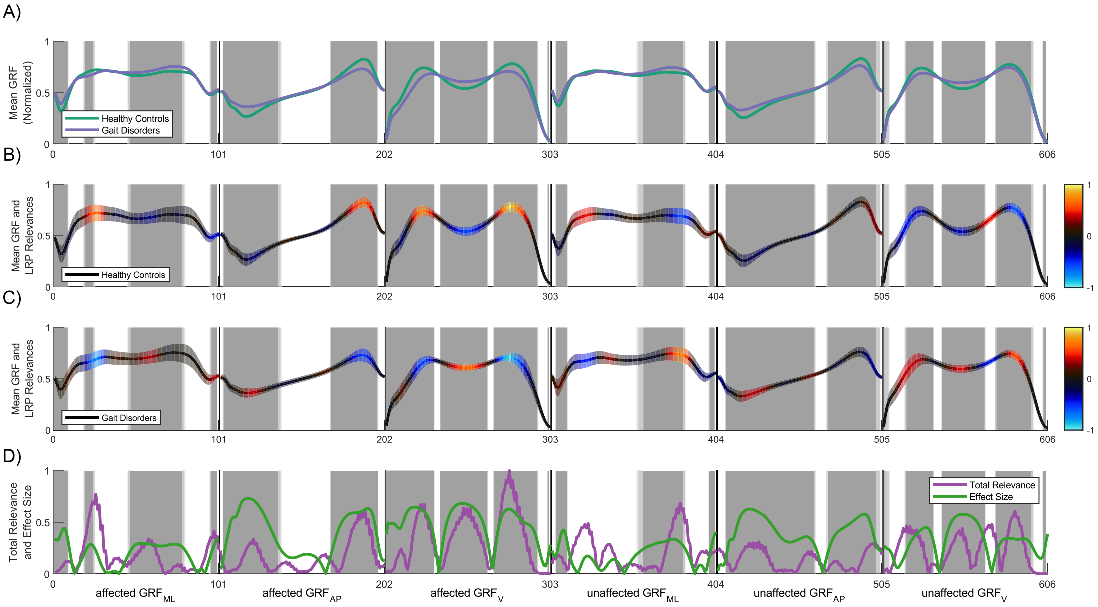


Fig. 5. Results overview for the classification of healthy controls (*HC*) and the aggregated class of all three gait disorders (*GD*) based on min-max normalized GRF signals using a CNN as classifier. (A) Averaged GRF signals for *HC* and *GD*. The first three signals represent the three GRF components of the affected side and are followed by the three GRF components of the unaffected side. Note that the data for both sides are composed of three GRF components (e.g., input features of the affected side: 1 to 101 (GRF_{ML}), 102 to 202 (GRF_{AP}), and 203 to 303 (GRF_V)). This means, for example, that input features 21 (GRF_{ML}), 122 (GRF_{AP}), and 233 (GRF_V) all correspond to the relative time of 20% of the same stance phase. The areas that are depicted in three different shades of gray for the three different alpha levels, i.e., dark gray for 0.01, gray for 0.05, and light gray for 0.1, highlight regions in the input signals where SPM indicates statistically significant differences between both classes (i.e., *HC* and *GD*). (B) Averaged GRF signals of all test trials as a line plot for the healthy controls class, with a band of one standard deviation, color-coded via input relevance values for the class (*HC*) obtained by LRP. (C) Averaged GRF signals of all test trials are shown as a line plot for the class of all the gait disorders (*GD*), in the same format as in (B). (D) Line plots showing the effect size computed as Pearson's correlation coefficient and total relevance based on the absolute sum of the LRP relevance values of both classes (*HC* and *GD*). The total relevance correlates with the local discriminativity of the input signal for the classification task.

input value for a respective classification task. The higher the total relevance at a certain signal region, the more discriminative is this region for the two underlying classes.

Figure 5 illustrates the signal regions of high input relevance for the classification between the *HC* and *GD* class. These regions are prevalent within all GRF signal components. The most relevant regions for distinguishing between the two classes have been found to include the local minima and maxima in the affected GRF_V signal. A similar pattern, though less pronounced, appears in the unaffected GRF_V. For GRF_{AP}, LRP identified relevant regions in the affected and unaffected signals, with the maximum peak in the affected signal being the most pronounced. For GRF_{ML}, relevant information appears to be predominantly located around the first lateral peak of the affected side and the second lateral peak of the unaffected side. The identified regions of high total relevance according to LRP agree to a large extent with the signal regions assessed as significantly different by SPM (gray-shaded areas in Figure 5).

Figure 6 shows the effect size obtained via SPM and the total relevance according to LRP for the task *HC*/*GD* (with min-max normalized GRF signals as in Figure 5) and all three employed classification methods (CNN, SVM,

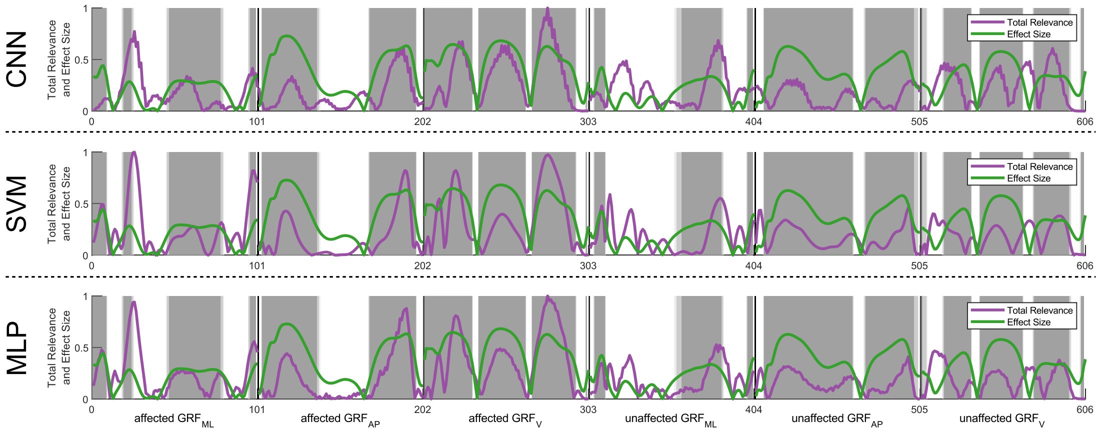


Fig. 6. Comparison of different classification methods (CNN, SVM, and MLP) for the classification of healthy controls and the class of all three gait disorders (*HC*/*GD*) based on min-max normalized GRF signals. The comparison is based on the total relevance of the LRP results as well as statistically significant differences (gray-shaded areas) and effect size computed as Pearson's correlation coefficient. Note that the gray-shaded areas and the effect size (green curve) are the same, while the total relevance varies between the three classification methods.

and MLP). The relevance scores agree strongly between the three classification methods. In fact, only some signal regions are prioritized differently, e.g., the affected and unaffected GRF_{ML} at the beginning and the end of the signal. These results show that the investigated classification methods rely on the same regions in the input data with only small exceptions.

For the sake of brevity, only the results for the classification task *HC*/*GD* were presented. For results of the other classification tasks, we refer the reader to the supplementary Figures S4, S7, S10 (CNN), Figures S6, S9, S12 (SVM), and Figures S5, S8, S11 (MLP). In the following, the discussion will incorporate all binary classification tasks.

6 DISCUSSION

The primary aim of this article is to investigate whether XAI methods can enhance explainability of ML predictions in clinical gait classification. In this section, the classification results are analyzed, compared, and interpreted in terms of classification accuracy and relevance-based explanations. These explanations are, furthermore, evaluated from a statistical and clinical viewpoint. Additionally, we discuss dependencies, influences, and interesting observations with respect to different classification methods, tasks, normalization methods, and signal components (horizontal forces and affected/unaffected leg signals).

6.1 Classification Results

The results expressed in terms of classification accuracy (presented in Figure 4 and supplementary Table S1) demonstrate a comparable level of performance between the three different machine learning methods (CNN, SVM, and MLP). The achieved performance level is not only interesting by itself but also important information for further explainability experiments. The reason is that an objective analysis of explainability by a post hoc method like LRP is only meaningful if the classification model can robustly differentiate between the target classes, i.e., a certain model quality is necessary to draw meaningful conclusions from explainability results. An analysis of unreliable classification models bears the potential risk that unstable patterns, noise, and spurious correlations bias the explainability results. For this reason, we excluded the classification tasks *HC*/*H*/*K*/*A* and

$H/K/A$ from our further investigation, as the tasks could not be solved with sufficient accuracy (average classification accuracy above 80%). For the binary classification tasks this risk is much lower, because the higher classification accuracies (and deviations from ZRB) obtained suggest that robust features can be found in the input data.

Another aspect we assessed is the influence of normalization on the input data (see Figure 4 and supplementary Table S1). The normalization of the input data is important for machine learning, since highly differing value ranges can have a negative influence on the classification model, i.e., input variables with a higher value range have a stronger influence on the predictions [14, 31]. The same appears to be the case for gait data, where the normalization of the input data strongly influences the classification models, as can be observed from the relevance scores of the horizontal forces in Figure 5 and supplementary Figure S13. Surprisingly, however, min-max normalization does not significantly improve the classification results (see Figure 4 and supplementary Table S1) for the investigated classification tasks. This raises the question of whether the use of GRF_V alone would already be sufficient to solve the classification tasks. We discuss this seemingly contradictory behavior in the following section.

6.2 Explainability Results

In the following, we discuss different related aspects with regard to our first leading research question: ***Which input features or signal regions are most relevant for automatic gait classification?*** The visualizations for all classification tasks and classification methods can be found in the supplementary Figures S1–S12.

Which input features are relevant for the classification of functional gait disorders? LRP identified several regions of high relevance in the GRF signals for all classification tasks. The ML models often used regions (and not single time-discrete values) encompassing peaks and valleys in the GRF signals to distinguish between the different classes, e.g., for task HC/GD using the CNN (see Figure 5) in the affected and unaffected GRF_V (all three local maxima and minima), affected GRF_{AP} (both peaks), unaffected GRF_{AP} (first peak), affected GRF_{ML} (first lateral peak), and unaffected GRF_{ML} (both lateral peaks). The highest total relevance scores are present in the signals of the affected side and most commonly in GRF_V for all investigated classification tasks. This is in line with earlier studies, e.g., where the peaks and valley (as time-discrete parameters) of the affected GRF_V showed the highest discriminatory power [66].

Are signal regions of the unaffected side important for the classification of functional gait disorders? Across all classification tasks, relevant regions are also pronounced in the GRF signals of the unaffected side, but less than in those of the affected side. In earlier studies [67, 68], we showed that the omission of the unaffected side during classification negatively affected classification accuracy. The explainability results confirm this observation. The unaffected side seems to capture complementary information relevant to the classification task under consideration. In particular, the identified relevant regions in the GRF signals occur at similar relative (e.g., in both peaks of GRF_V) or absolute (e.g., the second peak of the affected GRF_{AP} and the first peak of the unaffected GRF_{AP}) time points of the stance phases of the unaffected and affected side.

Are the anterior-posterior and medio-lateral forces relevant for the task? While the highest total relevance scores can be observed in GRF_V in most cases, relevant regions are always also observed in the horizontal GRF signals (GRF_{AP} and GRF_{ML}). However, the locations and degree of relevance within the horizontal signals vary for different classification tasks, e.g., for task HC/A, the highest relevance scores occur in the affected GRF_{AP} (and GRF_V) and hardly any relevant regions exist in GRF_{ML} (see supplementary Figure S10), while the highest relevance score for the task HC/H appears at the beginning of the affected GRF_{ML} (see supplementary Figure S4).

What is the impact of normalization on explainability results? Normalization of input data is a standard procedure prior to classification with ML models to ensure equal numerical ranges of different signals [14, 31]. XAI methods such as LRP allow to visualize the effects of normalization on the predictions of ML models directly at the level of the input signals. To gain a deeper understanding of these effects and the underlying data, we also conducted experiments without normalization of input data (see supplementary Figures S13–S24). For the classification of non-normalized GRF signals, the most relevant input values are located in GRF_V, i.e., especially the two peaks and the valley in between are relevant for the tasks. A minimal degree of relevance can be observed in the peaks of the affected and unaffected GRF_{AP} signals. The reason for the absence of relevant regions in the horizontal forces could be their small value range. The rather small range compared to the GRF_V component may lead to a smaller influence on the training of the classification models. Explainability results for min-max normalized input data show that highly relevant regions are identified in the horizontal forces of the affected and unaffected side (e.g., Figure 5). Thus, normalization amplifies the relevance of values in the horizontal forces and thereby makes them similarly important as GRF_V. Based on the LRP relevance scores, we conclude that normalization is important to obtain unbiased predictions of ML models (bias introduced by different signal amplitudes).

Are all identified relevant regions necessary for the task? For all classification tasks and classification methods, with min-max normalized input data, many regions of the GRF signals are identified to be relevant for classification according to LRP. The classification performance with and without normalization does, however, not vary significantly for the binary classification tasks (see classification results in Section 5.1). This raises the question of whether all regions identified as relevant are necessary to achieve peak performance in classification or whether some of them are redundant (i.e., not yielding an increase in classification performance when combined). Note that the assumption of redundancy is supported by the fact that the three GRF components represent individual dimensions of the same three-dimensional physical process. Thus, a strong correlation is *a priori* given in the data.

To answer the question, we conducted additional experiments with occluded parts of the input vector and evaluated the changes in classification performance. Occlusion is realized by replacing the horizontal forces (GRF_{AP} and GRF_{ML}) of both sides (affected and unaffected) with zero values. Table 2 shows the classification results for the experiments with occluded input signals as deviation from the mean classification accuracy of the experiments with non-occluded input signals. The results decrease on average when the horizontal forces are occluded (except for tasks HC/GD and HC/A using the CNN). Thus, relevant regions in the horizontal forces cannot be completely redundant to those in GRF_V and, therefore, represent also complementary information. This is in line with previous quantitative performance evaluations [67, 68]. However, the classification results of the binary classification tasks are not influenced by the occlusion of horizontal forces in a statistically significant way. This was confirmed by several dependent t-tests ($p > 0.05$) with Bonferroni-Holm [25] correction. Our results indicate that the relevant regions identified by LRP may represent an over-complete set, which exhibits a certain degree of redundancy, as removing relevant sections does not necessarily lead to reduced classification performance. However, redundancy is not necessarily a negative property, as it may help to achieve higher robustness to noise and possibly also to outliers and missing data [29].

Do different ML methods rely on different patterns? A comparison of the three employed classification methods is depicted in Figure 6. Across all binary classification tasks, relevant signal regions for all three classification methods are largely consistent, especially with respect to their location. Minor differences exist in the amplitude of the relevance scores, e.g., at the beginning of the affected GRF_V or the second peak in the affected GRF_{AP} (see Figure 6). The similarities between MLP and SVM are more pronounced. The remaining binary classification tasks, i.e., HC/H (see supplementary Figures S4, S5, and S6), HC/K (see supplementary Figures S7, S8, and S9), and HC/A (see supplementary Figures S10, S11, and S12) confirm these findings. Although LRP clearly shows where the prediction is grounded, it cannot explain *why* these patterns are important. However, it allows to identify and compare the learning strategies of different classification methods.

Can we derive additional properties of the models from the explanations, e.g., different learning strategies? Explanations provided by local XAI methods, such as LRP, inform about a model's reasoning on individual samples. A more general understanding about the model's learned patterns can be obtained via the evaluation of

Table 2. Classification Results for the Experiment with Occluded Horizontal Forces (GRF_{AP}, GRF_{ML}), in Percent

Task	Normalization	CNN	SVM	MLP
HC/GD	min-max	0.2	−1.4	−1.4
HC/H	min-max	−4.5	−6.5	−4.9
HC/K	min-max	−2.1	−3.7	−4.2
HC/A	min-max	1.5	−0.9	−1.3

The results are reported as mean deviation from the prediction accuracy of the original input signals presented in Figure 4 and supplementary Table S1, i.e., negative values signify a decrease and positive values an improvement in classification performance.

larger sets of sample-specific explanations [34]. In the previous sections, we achieved this by averaging relevance patterns across all samples of a given class. To perform a more detailed analysis that is able to identify different learning strategies of the ML models, we propose the use of SpRAy [35] as described in [5] for clinical gait data. The basic idea of this approach is to cluster the relevance patterns obtained for different samples and classes and to analyze the resulting clusters and subclusters.

SpRAy is a statistical analysis method for the explorative discovery of a model's characteristic prediction strategies from XAI-based relevance patterns. With its core in Spectral Clustering [43, 47], the method discovers structure within the set of given relevance patterns and yields, among its outputs, a spectral embedding Φ together with suggested groupings within the embedding in form of k cluster labels. Here, the embedding Φ directly corresponds to the individual relevance patterns, under consideration of their local, global, and potentially non-linear affinity structure. Sets of samples with similar relevance patterns are tightly grouped together in the spectral embedding space, while samples with dissimilar patterns are located far apart. Together with the suggested cluster labels, the analytically derived solution in Φ can then be visualized in $\mathbb{R}^2$, e.g., via a t-SNE projection [5, 39]. We implemented and evaluated SpRAy using the CoRelAy[3] framework [4] for Python.

Figure 7 shows exemplary SpRAy results for task HC/GD (with min-max normalized GRF signals) using the CNN as classification method. Based on the clustering provided in Figure 7(C) and 7(F), we see that the relevance patterns are grouped into clusters. This indicates that the ML model learned different classification strategies. Considering the ground truth class labels (see Figure 7(D)), we see that the model's explanations for the overall gait disorder (GD) class are grouped into distinct clusters that contain samples from the individual gait disorder classes (H, K, and A), even though the model was never explicitly trained to do so in this classification task. This means that the model learned different strategies for different pathological subclasses in GD. Considering the participant labels (see Figure 7(B) and Figure 7(E)), we can see that the relevance patterns of the five trials of a participant are often clustered together (Figure 7(B) and 7(E)). This means that the model learns similar strategies for the samples belonging to one participant. From a biomechanical perspective, this is plausible because each individual person has unique gait patterns that differ from the gait patterns of other individuals [30]. For clinical experts, it is important to see that the model is able to reflect such patterns.

In conclusion, SpRAy demonstrates the ability of ML models to learn patterns and dependencies in the data without explicit label information. For the clinical domain, this ability is of great value, since pathologies have various manifestations (that are sometimes even beyond the expertise of a clinical expert).

6.3 Statistical Evaluation

In the following, we investigate the statistical properties of the signal regions found to be relevant by LRP to answer the second leading research question: ***To what extent are input features or signal regions identified***

[3] https://github.com/virelay/corelay.

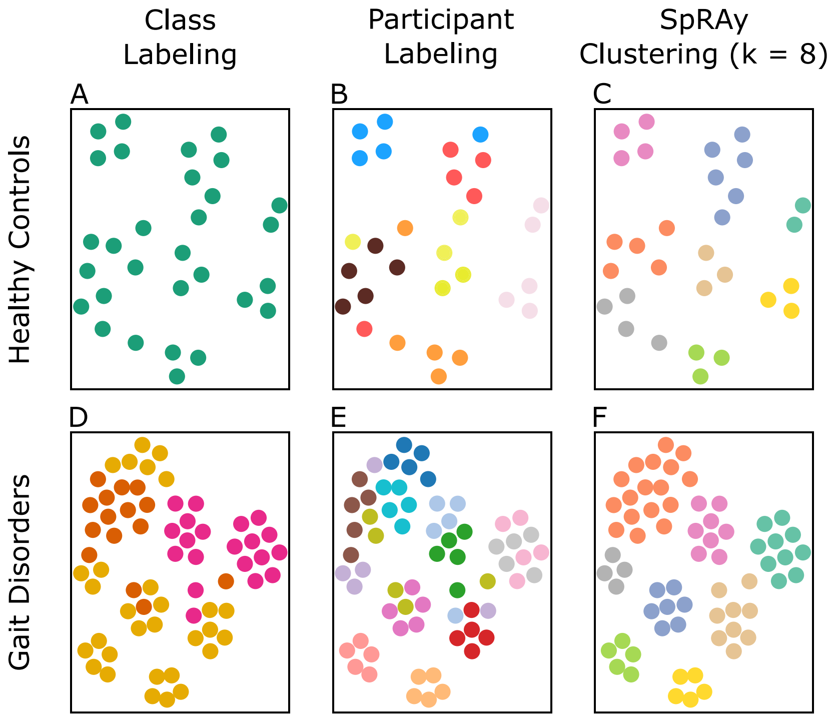


Fig. 7. The spectral embedding Φ derived via SpRAy from LRP explanations for the CNN model on test data, visualized via t-SNE for samples labeled as healthy controls (HC; N = 30; subfigures A-C) and the aggregated class of all three gait disorders ($GD = \{H, K, A\}$; N = 65; subfigures D-F). Each column of panels marks the embedded sample explanations with respect to different sets of labels as indicated by color: (subfigures A/D) ground truth class labels (HC, H, K, A), (subfigures B/E) ground truth participant labels, and (subfigures C/F) cluster labels inferred via SpRAy for $k = 8$ clusters on Φ before projecting the spectral embedding into $\mathbb{R}^2$ via t-SNE. The figure shows that the relevance patterns are grouped into clusters, indicating that the ML model learned different classification strategies.

as being relevant for a given gait classification task statistically justified? To answer this question, we leverage SPM, which provides statistical inference estimates for each value of the input vector. We compare the LRP regions with those considered as significantly different by SPM. Results show that in the vast majority of cases, the SPM analysis shows statistically significant differences in regions that are also highly relevant for classification according to LRP. Thus, for binary classification tasks, it seems that ML models base their predictions primarily on features that are also significantly different between the two classes. This can be observed across all classification tasks (e.g., see Figure 5(D) for task HC/GD). As the total relevance increases, the effect size usually also increases. We performed a cross-correlation to determine the relationship between the effect size and the total relevance. Both curves show highly correlated behavior for the min-max normalized input data for all classification tasks: HC/GD (r = 0.76), HC/H (r = 0.66), HC/K (r = 0.76), and HC/A (r = 0.78). However, minimal differences between the results of LRP and SPM can be detected, e.g., the location of the first relevant signal region in the unaffected GRF_V. For all classification tasks, we observed that LRP already considers the slope to the first GRF_V peak of the unaffected leg as relevant for the classification, whereas SPM, slightly shifted, emphasizes the region encompassing the peak itself with a high effect size. Future research is needed to address this observation and examine differences between LRP and SPM in more detail.

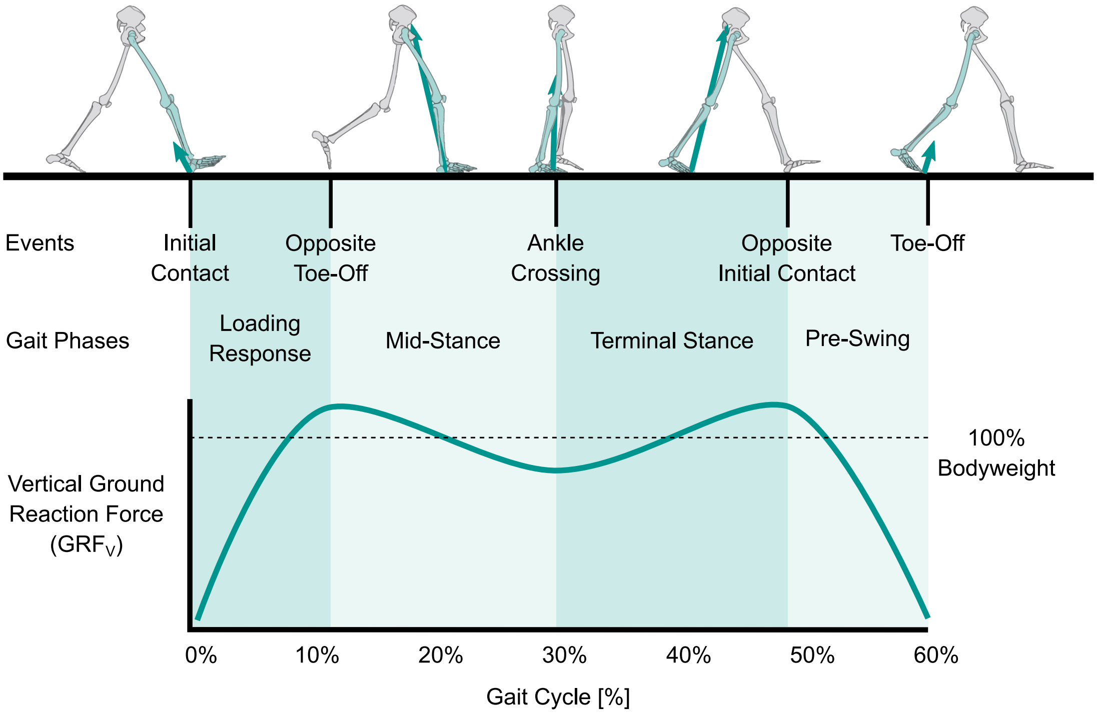


Fig. 8. Overview of the most relevant gait events during the stance phase. In clinical gait analysis, a gait cycle (100%) is defined from the initial contact of one foot to the subsequent initial contact of the same foot. During the first approximately 60% of the gait cycle, referenced as the stance phase (relevant time range for the present work), the foot has contact to the ground. The beginning of the stance phase is defined as initial contact with the ground (typically by the heel), then body weight is shifted to the supporting leg (loading response and mid-stance), followed by terminal stance (forward propulsion), pre-swing (preparation of the swing phase), and toe-off. Adapted from References [9, 62].

Concerning our second research question, we conclude that the relevance estimates according to LRP are to the greatest extent statistically justified. The second part of the research question regarding the validity of the explanations with respect to clinical assessment is investigated in the following section.

6.4 Clinical Evaluation

To what extent are input features or signal regions identified as being relevant for a given gait classification task in line with clinical assessment? This question is answered in the following by two clinical experts in human gait analysis. To assist the reader in following the discussion and to facilitate the interpretation of the input signals, the domain-specific terms and gait cycle definitions are described in Figure 8. For further details on the principles of human gait and its clinical implications, the interested reader is referred to literature such as Perry and Burnfield [53] or Winter [79].

The explainability results for classification of healthy controls (*HC*) and the aggregated class of all three gait disorders (*GD*) based on min-max normalized GRF signals illustrate clinically meaningful patterns (see Figure 5). High LRP relevance scores occurred during loading response, terminal stance, and pre-swing in GRF_{AP} and GRF_{ML} as well as in loading response, mid-stance, terminal stance, and pre-swing in GRF_V. These phases are especially sensitive toward gait anomalies as loading response requires the absorption of body weight and terminal stance plays an essential role for forward propulsion [33]. Both aspects are affected in case of gait impairments due to a diminished walking speed (requiring less absorption or push-off) as well as factors that go along with an injury, such as the presence of pain, a decreased range of motion, and/or lessened muscle strength [64, 78]. When analyzing the explainability results in more detail, one can identify specific gait dynamics that can be traced back to an impairment at a certain joint level.

For classification task HC/A (see supplementary Figure S10), we can observe pronounced peaks in the total relevance curves of GRF_{AP} and GRF_V caused by alterations in the terminal stance and pre-swing phase of the affected side. This is in agreement with the observations of Son et al. [69], who found a significantly increased propulsive force (GRF_{AP} in terminal stance) for patients with chronic ankle instability. They also identified an increased GRF_V during late terminal stance (push-off) compared to healthy controls, which is also in line with the relevance scores obtained in our study. Both our explainability results and the study of Son et al. [69] did not indicate any relevance or difference to healthy controls in the GRF_{ML}.

For classification task HC/K, the highest LRP relevance scores are present in GRF_V, GRF_{AP}, and GRF_{ML} (see supplementary Figure S7). Changes in GRF_V may result from lessened knee flexibility that hinders typical knee dynamics over the entire course of the stance phase. More precisely, healthy walking requires a slightly flexed knee joint during initial contact followed by a knee flexion thereafter, by definition called loading response. During the mid-stance phase the walker's center of gravity is shifted forward and thus demands further knee extension. This is in line with the study of Cook et al. [15], who analyzed the effects of restricted knee flexion and walking speed on the GRF_V. According to their results, the loading rate (slope during loading response), unloading rate (slope during pre-swing), and peak GRF_V of the restricted leg showed significant speed-knee flexion restriction interactions.

Highest LRP relevance values for the classification task HC/H are obtained during loading response and terminal stance in GRF_V of the affected side (see supplementary Figure S4). McCrory et al. [41] and Martinez-Ramirez et al. [40] identified the GRF_V as an objective measure of gait for patients following hip arthroplasty. McCrory et al. [41] found significant differences between patients and healthy controls in several variables of the GRF_V such as the first and second local peaks, impulse, and stance time. They also identified that the unaffected side holds relevant information, as significant differences were found in the GRF_V either compared to the control group or the affected side. This is also seen in our obtained LRP relevance scores for the classification task HC/H where two distinct relevance peaks are present for GRF_V for the first and second GRF_V peak of the affected side. These results are also in agreement with Martinez-Ramirez et al. [40], who demonstrated that patients after successful hip arthroplasty still show significantly altered GRF_V for both the affected and unaffected leg including a continuing GRF_V asymmetry between both sides.

With regard to our second research question, we conclude that signal regions with high relevance according to LRP can be largely associated with clinical gait analysis literature and are plausible from a clinical point of view according to two domain experts.

6.5 On the Usefulness of XAI Methods for Clinical Gait Analysis

XAI methods increase transparency and can make the decision process of ML models more comprehensible for clinical experts. Transparency of state-of-the-art ML models is crucial to promote the acceptance of such systems in clinical practice, allowing clinicians to benefit from high, and in some cases already better than human [16, 21, 42], classification accuracy that ML models achieve.

In the previous subsections (i.e., Sections 6.3 and 6.4), we showed that explainability results are consistent from a statistical and domain experts' point of view. In particular, regions of high relevance according to LRP are highly discriminatory according to SPM, and the clinical experts associated these regions with clinical explanations. Having evaluated the explainability results, we now want to address the question: ***What is the added value that XAI methods can provide to clinical practice?***

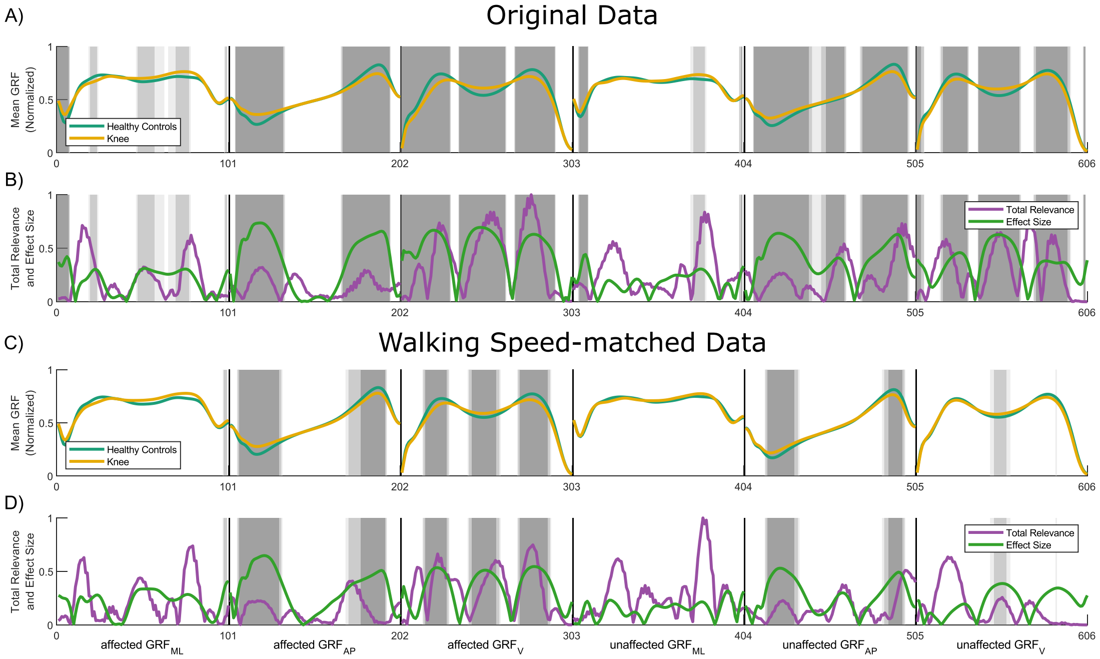


Fig. 9. Comparison of explainability results of the original (top) and walking speed-matched (bottom) data for the classification task HC/K based on the min-max normalized GRF signals using CNN.

The two experts reported that they mainly focus on regions in the GRF_V signals during the evaluation process of patients in clinical practice. In particular, the evaluation of the unaffected GRF_V is very important for the clinicians. The main motivation for this is that many compensatory patterns manifest in this signal, i.e., as patients try to put as little weight on the affected leg as possible, they take shorter steps with the unaffected leg. This is reflected in a reduced slope in the unaffected GRF_V during loading response.

Our explainability results show that in addition to regions in GRF_V, regions in GRF_{ML} and GRF_{AP} are also highly relevant for the classification tasks. These signals are less considered in clinical practice. However, the relevant regions in GRF_{ML} and GRF_{AP} indicate additional information about the classification of pathological gait patterns.

Explainability approaches can lead to novel insights and a deeper understanding of the models and the underlying data as illustrated in the following example. In the clinical evaluation of the explainability results, the experts identified also relevant regions for the ML models that are not directly related to the specific functional gait disorders, according to their personal expertise and the literature. The experts assumed that, e.g., the relevant regions in the affected and unaffected GRF_V, in particular during mid-stance, terminal stance, and pre-swing, are strongly influenced by differences in walking speed between healthy controls and patients. From this observation the clinical experts derived the hypothesis that the trained ML models might be biased by the walking speed.

Using the HC/K classification task as an example, we examined whether there is a significant difference in walking speed between HC and K. An independent samples t-test revealed a statistically significant difference in walking speed between HC and K ($p < 0.001$). The differences in walking speed affect the shape of the signals (although the signals were time-normalized) and the ML models could have learned these dissimilarities. To assess the influence of walking speed on the ML models, we repeated the experiment for the task HC/K on a subsample of the original data. This subsample does not exhibit statistically significant differences with respect to walking speed (independent samples t-test; p = 0.068). A comparison of the explainability results obtained for task HC/K (with min-max normalized GRF signals) using CNNs that were trained on the original and walking speed-matched data are presented in Figure 9. The results for the walking speed-matched data clearly show that most of the relevant regions according to LRP agree with the regions obtained for the original data (with only small changes in amplitude). However, relevant regions in the unaffected GRF_V after loading response are less relevant for the model trained on walking speed-matched data. Thus, in contrast to the model trained on the original data, this model barely takes these regions into account. The conclusion that can be drawn is that these regions are related to differences in walking speed.

Using our XAI approach, we have been able to show that some degree of walking speed-related bias was learned in the original models, but that this influence was not as strong as assumed by the clinical experts. Another interesting aspect of the experiment concerns the SPM results. While the trend of effect size and the total relevance remain similar, the statistically significant regions are clearly reduced (compare gray-shaded areas for both settings in Figure 9), showing the sensitivity of SPM to the alpha level.

Overall, we showed that our proposed XAI approach exhibits substantial usefulness for the clinical setting, as we were able to demonstrate that: (i) regions in the signals that are less focused on in the literature and clinical evaluation, i.e., GRF_{AP} and GRF_{ML}, also contain informative and relevant regions that can be associated to the underlying pathology, (ii) ML models learn different strategies for different samples and patient groups (experiment with SpRAy; see Section 6.2), and (iii) XAI methods allow the identification of biases in ML models, e.g., with respect to normalization or walking speed-related differences between classes.

The increased transparency provides additional insights into the working mechanisms of the trained ML models, enabling clinicians to better understand them and increase their level of trust [70].

6.6 Limitations and Future Work

A fundamental problem in evaluating the explainability results is the absence of a ground truth. A challenge in interpreting the explainability results is that alterations of the input signals can be caused not only by the influence of a pathology, but also by other independent parameters, e.g., a lower walking speed or an increased body mass. To minimize potential biases introduced by independent parameters on prediction explanations, future research should attempt to develop normalization procedures for input signals that compensate such influencing factors or develop classification models that inherently learn the relationship between influencing factors and input signals.

Another limiting factor is that we solely used GRF signals for classification. This does not perfectly reflect best practice in clinical gait analysis where clinicians usually base medical decisions on a combination of GRF and 3D kinematic data [9]. The additional use of kinematic data is expected to improve the classification accuracy to an appropriate level for clinical application, in particular for multi-class classification tasks. However, 3D kinematic data are prone to several difficulties such as inconsistencies due to inter-assessor and inter-laboratory differences [20, 60]. This makes it more difficult to create a homogeneous, large-scale, and real-world dataset compared to using simple data, such as GRF signals. Thus, the utilized GaitRec data [28] provide a large-scale dataset with an easy to comprehend clinical example, which allows to showcase how XAI methods can support transparency of ML models and their predictions.

Besides visual explanations as presented in this article, a translation into human-understandable textual explanations would be desired for clinical application. An interesting direction for future research is the generation of textual explanations based on biomechanical parameters estimated from the input signals. This would enable approaches that exceed pure explainability and provide deeper interpretations for clinical experts in the form of, e.g., “there is a high probability of a pathology in the knee due to a limited knee extension during the mid-stance phase.”

We will conduct further research to compare different explanation methods and rule-based approaches [32] for different classification tasks and datasets. In addition, we want to point out that quantitative and objective methods are necessary to assess the quality of prediction explanations [57] including datasets with respective ground truth explanations.

7 CONCLUSION

The present findings highlight that the investigated ML models base their predictions on meaningful features of GRF signals in various clinical gait classification tasks. These features are in accordance with a statistical and clinical evaluation. Hence, XAI methods that provide explainability for predictions provided by ML models, such as LRP, can be promising to increase justification of automatic classification predictions in CGA and can help to make the prediction processes comprehensible to clinical experts. Thereby, XAI may facilitate the application of ML-based decision-support systems in clinical practice. Within the scope of our analysis, we were able to show that:

- Highly relevant regions were identified in the signals of the affected and unaffected sides. Thus, the unaffected side captures additional information that are relevant for automated gait classifications.
- For time-series data such as GRF signals, SPM has shown to be a suitable statistical reference. Highly relevant regions in the input data (according to LRP) are in most cases also significantly different (according to SPM) and in line with clinical evaluation.
- In addition to GRF_V, the horizontal forces contain regions of high relevance, which is consistent with clinical gait analysis literature.
- ML models seem to learn an over-complete set of features that may contain redundant information. This might explain why the occlusion of horizontal forces and input normalization in our experiments had negligible influence on the classification accuracies.
- ML models for gait classification are able to learn different strategies for individual persons and patient groups.
- Explainability approaches can help to detect bias in ML models and help to assess their correct working, which is important for clinicians to enable building trust in the predictions of these models.

This article represents a first step towards establishing explainability of ML approaches for time-series classification. Thereby, we want to promote the application of ML in clinical gait analysis to support medical decision-making in the future.

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

DS, BH, A-MR prepared the dataset. DS, FH, SL, BH, WS, WIS conceived the presented idea. BH, WS, WIS, MZ raised the funding. DS, FH, SL, BH, A-MR, AK, WS, CB, WIS, MZ participated in the data analysis. DS, FH, SL, BH, A-MR, MZ wrote the manuscript. DS, FH, SL, BH, A-MR, AK, WS, MZ designed the figures. DS, FH, SL, BH, A-MR, AK, WS, CB, WIS, MZ reviewed and approved the final manuscript.

DATA AVAILABILITY STATEMENT

For our analyses, we used a subset of the GaitRec dataset [28]. Our source code and the utilized dataset are publicly available at: https://github.com/sebastian-lapuschkin/explaining-deep-clinical-gait-classification.

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160. DOI: https://doi.org/10.1109/ACCESS.2018.2870052

[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18)*. Curran Associates Inc., Red Hook, NY, 9525–9536.

[3] Murad Alaqtash, Thompson Sarkodie-Gyan, Huiying Yu, Olac Fuentes, Richard Brower, and Amr Abdelgawad. 2011. Automatic classification of pathological gait patterns using ground reaction forces and machine learning algorithms. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*. IEEE, 453–457. DOI: https://doi.org/10.1109/IEMBS.2011.6090063

[4] Christopher J. Anders, David Neumann, Wojciech Samek, Klaus-Robert Müller, and Sebastian Lapuschkin. 2021. Software for Dataset-wide XAI: From local explanations to global insights with Zennit, CoRelAy, and ViRelAy. *CoRR* abs/2106.13200 (2021).

[5] Christopher J. Anders, Leander Weber, David Neumann, Wojciech Samek, Klaus-Robert Müller, and Sebastian Lapuschkin. 2022. Finding and removing clever Hans: Using explanation methods to debug and improve deep models. *Information Fusion* 77 (2022), 261–295. DOI:https://doi.org/10.1016/j.inffus.2021.07.015

[6] Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilovic, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John T. Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. 2019. One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques. *CoRR* abs/1909.03012 (2019).

[7] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS One* 10, 7 (2015), e0130140. DOI: https://doi.org/10.1371/journal.pone.0130140

[8] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. 2010. How to explain individual classification decisions. *J. Mach. Learn. Res.* 11 (2010), 1803–1831. Retrieved from http://portal.acm.org/citation.cfm?id=1859912.

[9] Richard Baker. 2013. *Measuring Walking: A Handbook of Clinical Gait Analysis*. Mac Keith Press, London.

[10] David Balduzzi, Marcus Frean, Lennox Leary, J. P. Lewis, Kurt Wan-Duo Ma, and Brian McWilliams. 2017. The shattered gradients problem: If resnets are the answer, then what is the question? In *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, 342–350.

[11] Brian G. Booth, Noël L. W. Keijsers, Jan Sijbers, and Toon Huysmans. 2018. STAPP: Spatiotemporal analysis of plantar pressure measurements using statistical parametric mapping. *Gait Post.* 63 (2018), 268–275.

[12] Johannes Burdack, Fabian Horst, Sven Giesselbach, Ibrahim Hassan, Sabrina Daffner, and Wolfgang I. Schöllhorn. 2020. Systematic comparison of the influence of different data preprocessing methods on the performance of gait classifications using machine learning. *Front. Bioeng. Biotechnol.* 8 (2020), 260. DOI: https://doi.org/10.3389/fbioe.2020.00260

[13] Tom Chau. 2001. A review of analytical techniques for gait data. Part 1: Fuzzy, statistical and fractal methods. *Gait Post.* 13, 1 (Feb. 2001), 49–66. DOI: https://doi.org/10.1016/S0966-6362(00)00094-1

[14] François Chollet. 2017. *Deep Learning with Python*. Manning Publications Company, Shelter Island, NY.

[15] Thomas M. Cook, Kevin P. Farrell, Iva A. Carey, Joan M. Gibbs, and Gregory E. Wiger. 1997. Effects of restricted knee flexion and walking speed on the vertical ground reaction force during gait. *J. Orthop. Sports Phys. Therap.* 25, 4 (1997), 236–244. DOI: https://doi.org/10.2519/jospt.1997.25.4.236

[16] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 7639 (2017), 115–118. DOI: https://doi.org/10.1038/nature21056

[17] European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Offic. J. Eur. Union* L 119 (2016), 1–88. Retrieved from https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[18] Joana Figueiredo, Cristina P. Santos, and Juan C. Moreno. 2018. Automatic recognition of gait patterns in human motor disorders using machine learning: A review. *Med. Eng. Phys.* 53 (2018), 1–12. DOI: https://doi.org/10.1016/j.medengphy.2017.12.006

[19] Ruth C. Fong and Andrea Vedaldi. 2017. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 3429–3437. DOI: https://doi.org/10.1109/ICCV.2017.371

[20] George E. Gorton, David A. Hebert, and Mary E. Gannotti. 2009. Assessment of the kinematic variability among 12 motion analysis laboratories. *Gait Post.* 29, 3 (2009), 398–402. DOI: https://doi.org/10.1016/j.gaitpost.2008.10.060

[21] Holger A. Haenssle, Christine Fink, R. Schneiderbauer, Ferdinand Toberer, Tobias Buhl, A. Blum, A. Kalloo, A. Ben Hadj Hassen, Luc Thomas, A. Enk, et al. 2018. Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann. Oncol.* 29, 8 (2018), 1836–1842.

[22] Eni Halilaj, Apoorva Rajagopal, Madalina Fiterau, Jennifer L. Hicks, Trevor J. Hastie, and Scott L. Delp. 2018. Machine learning in human movement biomechanics: Best practices, common pitfalls, and new opportunities. *J. Biomech.* 81 (2018), 1–11.

[23] Jianxing He, Sally L. Baxter, Jie Xu, Jiming Xu, Xingtao Zhou, and Kang Zhang. 2019. The practical implementation of artificial intelligence technologies in medicine. *Nat. Med.* 25, 1 (2019), 30–36. DOI: https://doi.org/10.1038/s41591-018-0307-0

[24] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating visual explanations. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 3–19. DOI: https://doi.org/10.1007/978-3-319-46493-0_1

[25] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* 6, 2 (1979), 65–70.

[26] Andreas Holzinger, Chris Biemann, Constantinos S. Pattichis, and Douglas B. Kell. 2017. What do we need to build explainable AI systems for the medical domain? *CoRR* abs/1712.09923 (2017).

[27] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. 2019. Causability and explainability of artificial intelligence in medicine. *Data Mining Knowl. Discov.* 9, 4 (July 2019), e1312. DOI: https://doi.org/10.1002/widm.1312

[28] Brian Horsak, Djordje Slijepcevic, Anna-Maria Raberger, Caterine Schwab, Marianne Worisch, and Matthias Zeppelzauer. 2020. GaitRec, a large-scale ground reaction force dataset of healthy and impaired gait. *Sci. Data* 7, 1 (May 2020), 1–8. DOI: https://doi.org/10.1038/s41597-020-0481-z

[29] Fabian Horst, Sebastian Lapuschkin, Wojciech Samek, Klaus-Robert Müller, and Wolfgang I. Schöllhorn. 2019. Explaining the unique nature of individual gait patterns with deep learning. *Sci. Rep.* 9, 1 (2019), 2391. DOI: https://doi.org/10.1038/s41598-019-38748-8

[30] Fabian Horst, Markus Mildner, and Wolfgang I. Schöllhorn. 2017. One-year persistence of individual gait patterns identified in a follow-up study—A call for individualised diagnose and therapy. *Gait Post.* 58 (2017), 476–480. DOI: https://doi.org/10.1016/j.gaitpost.2017.09.003

[31] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. 2016. *A Practical Guide to Support Vector Classification.* Technical Report. National Taiwan University. Retrieved from https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

[32] Maximilian Kohlbrenner, Alexander Bauer, Shinichi Nakajima, Alexander Binder, Wojciech Samek, and Sebastian Lapuschkin. 2020. Towards best practice in explaining neural network decisions with LRP. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN).* IEEE, 1–7.

[33] Arthur D. Kuo and J. Maxwell Donelan. 2010. Dynamic principles of gait and their clinical implications. *Phys. Ther.* 90, 2 (2010), 157–174. DOI: https://doi.org/10.2522/ptj.20090125

[34] Sebastian Lapuschkin, Alexander Binder, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2016. Analyzing classifiers: Fisher vectors and deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (CVPR).* IEEE Computer Society, 2912–2920.

[35] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2019. Unmasking clever Hans predictors and assessing what machines really learn. *Nat. Commun.* 10 (2019), 1096. DOI: https://doi.org/10.1038/s41467-019-08987-4

[36] Hong-yin Lau, Kai-yu Tong, and Hailong Zhu. 2009. Support vector machine for classification of walking conditions of persons after stroke with dropped foot. *Hum. Movem. Sci.* 28, 4 (Aug. 2009), 504–514. DOI:https://doi.org/10.1016/j.humov.2008.12.003

[37] Yann LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. 2012. Efficient BackProp. In *Neural Networks: Tricks of the Trade - Second Edition.* Springer, 9–48. DOI: https://doi.org/10.1007/978-3-642-35289-8_3

[38] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems (NIPS).* Curran Associates, Inc., 4765–4774. Retrieved from http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[39] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, Nov. (2008), 2579–2605.

[40] Alicia Martínez-Ramírez, Dirk Weenk, Pablo Lecumberri, Nico Verdonschot, Dean Pakvis, and Peter H. Veltink. 2014. Assessment of asymmetric leg loading before and after total hip arthroplasty using instrumented shoes. *J. NeuroEng. Rehabil.* 11, 1 (2014), 20. DOI: https://doi.org/10.1186/1743-0003-11-20

[41] Jean L. McCrory, Scott C. White, and Robert M. Lifeso. 2001. Vertical ground reaction forces: Objective measures of gait following hip arthroplasty. *Gait Post.* 14, 2 (2001), 104–109. DOI: https://doi.org/10.1016/S0966-6362(01)00140-0

[42] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg C. Corrado, Ara Darzi, et al. 2020. International evaluation of an AI system for breast cancer screening. *Nature* 577, 7788 (2020), 89–94.

[43] Marina Meila and Jianbo Shi. 2001. A random walks view of spectral segmentation. In *Proceedings of the International Workshop on Artificial Intelligence and Statistics (AISTATS).*

[44] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. 2019. Layer-wise relevance propagation: An overview. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning.* Springer, 193–209. DOI: https://doi.org/10.1007/978-3-030-28954-6_10

[45] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. 2017. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recogn.* 65 (2017), 211–222.

[46] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2018. Methods for interpreting and understanding deep neural networks. *Dig. Sig. Process.* 73 (2018), 1–15. DOI: https://doi.org/10.1016/j.dsp.2017.10.011

[47] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. 2002. On spectral clustering: Analysis and an algorithm. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*. 849–856.

[48] Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. 2016. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*. Curran Associates, Inc., 3387–3395. Retrieved from http://papers.nips.cc/paper/6519-synthesizing-the-preferred-inputs-for-neurons-in-neural-networks-via-deep-generator-networks.pdf.

[49] Angela Nieuwenhuys, Eirini Papageorgiou, Kaat Desloovere, Guy Molenaers, and Tinne De Laet. 2017. Statistical parametric mapping to identify differences between consensus-based joint patterns during gait in children with cerebral palsy. *PLoS One* 12, 1 (2017).

[50] Corina Nüesch, Victor Valderrabano, Cora Huber, Vinzenz von Tscharner, and Geert Pagenstert. 2012. Gait patterns of asymmetric ankle osteoarthritis patients. *Clin. Biomech.* 27, 6 (July 2012), 613–618. DOI: https://doi.org/10.1016/j.clinbiomech.2011.12.016

[51] Todd C. Pataky. 2010. Generalized n-dimensional biomechanical field analysis using statistical parametric mapping. *J. Biomech.* 43, 10 (July 2010), 1976–1982. DOI: https://doi.org/10.1016/j.jbiomech.2010.03.008

[52] Todd C. Pataky. 2012. One-dimensional statistical parametric mapping in Python. *Comput. Meth. Biomech. Biomed. Eng.* 15, 3 (Mar. 2012), 295–301. DOI: https://doi.org/10.1080/10255842.2010.527837

[53] Jacquelin Perry and Judith M. Burnfield. 2010. *Gait Analysis: Normal and Pathological Function* (2nd ed.) Slack, Thorofare, NJ.

[54] Angkoon Phinyomark, Giovanni Petri, Esther Ibáñez-Marcelo, Sean T. Osis, and Reed Ferber. 2018. Analysis of big data in gait biomechanics: Current trends and future directions. *J. Med. Biol. Eng.* 38, 2 (2018), 244–260. DOI: https://doi.org/10.1007/s40846-017-0297-2

[55] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-agnostic interpretability of machine learning. *CoRR* abs/1606.05386 (2016).

[56] Robert Rosenthal. 1991. Meta-Analytic Procedures for Social Research. SAGE Publications Inc. DOI : 10.4135/9781412984997

[57] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. 2017. Evaluating the visualization of what a deep neural network has learned. *IEEE Trans. Neural Netw. Learn. Syst.* 28, 11 (Nov. 2017), 2660–2673. DOI: https://doi.org/10.1109/TNNLS.2016.2599820

[58] Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J. Anders, and Klaus-Robert Müller. 2021. Explaining deep neural networks and beyond: A review of methods and applications. *Proc. IEEE* 109, 3 (2021), 247–278. DOI: https://doi.org/10.1109/JPROC.2021.3060483

[59] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *ITU J.: ICT Discov.* 1, 1 (2017), 39–48.

[60] Emilia Scalona, Roberto Di Marco, Enrico Castelli, Kaat Desloovere, Marjolein Van Der Krogt, Paolo Cappa, and Stefano Rossi. 2019. Inter-laboratory and inter-operator reproducibility in gait analysis measurements in pediatric subjects. *Int. Biomech.* 6, 1 (2019), 19–33. DOI: https://doi.org/10.1080/23335432.2019.1621205

[61] Wolfgang I. Schöllhorn. 2004. Applications of artificial neural nets in clinical biomechanics. *Clin. Biomech.* 19, 9 (2004), 876–898. DOI: https://doi.org/10.1016/j.clinbiomech.2004.04.005

[62] Huijuan Shi, Hongshi Huang, Yuanyuan Yu, Zixuan Liang, Si Zhang, Bing Yu, Hui Liu, and Yingfang Ao. 2018. Effect of dual task on gait asymmetry in patients after anterior cruciate ligament reconstruction. *Sci. Rep.* 8, 1 (2018), 1–10.

[63] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, 3145–3153.

[64] Maureen J. Simmonds, C. Ellen Lee, Bruce R. Etnyre, and G. Stephen Morris. 2012. The influence of pain distribution on walking velocity and horizontal ground reaction forces in patients with low back pain. *Pain Res. Treatm.* (2012), 11. DOI: https://doi.org/10.1155/2012/214980

[65] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Proceedings of the International Conference on Learning Representations (ICLR)*. Retrieved from http://arxiv.org/abs/1312.6034.

[66] Djordje Slijepcevic, Matthias Zeppelzauer, Anna-Maria Gorgas, Caterine Schwab, Michael Schüller, Arnold Baca, Christian Breiteneder, and Brian Horsak. 2017. Automatic classification of functional gait disorders. *IEEE J. Biomed. Health Inf.* 22, 5 (2017), 1653–1661. DOI : https://doi.org/10.1109/JBHI.2017.2785682

[67] Djordje Slijepcevic, Matthias Zeppelzauer, Caterine Schwab, Anna-Maria Raberger, Christian Breiteneder, and Brian Horsak. 2020. Input representations and classification strategies for automated human gait analysis. *Gait & Posture* 76 (2020), 198–203. DOI : https://doi.org/10.1016/j.gaitpost.2019.10.021

[68] Djordje Slijepcevic, Matthias Zeppelzauer, Caterine Schwab, Anna-Maria Raberger, Bernhard Dumphart, Arnold Baca, Christian Breiteneder, and Brian Horsak. 2018. P 011-Towards an optimal combination of input signals and derived representations for gait classification based on ground reaction force measurements. *Gait Post.* 65 (2018), 249. DOI: https://doi.org/10.1016/j.gaitpost.2018.06.155

[69] S. Jun Son, Hyunsoo Kim, Matthew K. Seeley, and J. Ty Hopkins. 2019. Altered walking neuromechanics in patients with chronic ankle instability. *J. Athlet. Train.* 54, 6 (2019), 684–697. DOI: https://doi.org/10.4085/1062-6050-478-17

[70] Fabian Sperrle, Mennatallah El-Assady, Grace Guo, Rita Borgo, Duen Horng Chau, Alex Endert, and Daniel Keim. 2021. A survey of human-centered evaluations in human-centered machine learning. *Comput. Graph. Forum* 40, 3 (2021).

[71] Erik Štrumbelj and Igor Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.* 41, 3 (2014), 647–665. DOI: https://doi.org/10.1007/s10115-013-0679-x

[72] Erico Tjoa and Cuntai Guan. 2019. A survey on explainable artificial intelligence (XAI): Towards medical XAI. *CoRR* abs/1907.07374 (2019).

[73] Eric J. Topol. 2019. High-performance medicine: The convergence of human and artificial intelligence. *Nat. Med.* 25, 1 (2019), 44–56. DOI: https://doi.org/10.1038/s41591-018-0300-7

[74] Leen Van Gestel, Tinne De Laet, Enrico Di Lello, Herman Bruyninckx, Guy Molenaers, Anja Van Campenhout, Erwin Aertbeliën, Mike Schwartz, Hans Wambacq, Paul De Cock, and Kaat Desloovere. 2011. Probabilistic gait classification in children with cerebral palsy: A Bayesian approach. *Res. Devel. Disab.* 32, 6 (Nov. 2011), 2542–2552. DOI:https://doi.org/10.1016/j.ridd.2011.07.004

[75] Markus Wagner, Djordje Slijepcevic, Brian Horsak, Alexander Rind, Matthias Zeppelzauer, and Wolfgang Aigner. 2018. KAVAGait: Knowledge-assisted visual analytics for clinical gait analysis. *IEEE Trans. Visualiz. Comput. Graph.* 25, 3 (2018), 1528–1542.

[76] Ferdous Wahid, Rezaul K. Begg, Chris J. Hass, Saman Halgamuge, and David C. Ackland. 2015. Classification of Parkinson's disease gait using spatial-temporal gait features. *IEEE J. Biomed. Health Inf.* 19, 6 (2015), 1794–1802.

[77] Nils Wilhelm, Anna Vögele, Rebeka Zsoldos, Theresia Licka, Björn Krüger, and Jürgen Bernard. 2015. FuryExplorer: Visual-interactive exploration of horse motion capture data. In *Visualization and Data Analysis 2015*. International Society for Optics and Photonics, 93970F. DOI: https://doi.org/10.1117/12.2080001

[78] Carin Willén, Katarina Stibrant Sunnerhagen, Claes Ekman, and Gunnar Grimby. 2004. How is walking speed related to muscle strength? A study of healthy persons and persons with late effects of polio. *Arch. Phys. Med. Rehabil.* 85, 12 (2004), 1923–1928. DOI: https://doi.org/10.1016/j.apmr.2003.11.040

[79] David A. Winter. 2009. *Biomechanics and Motor Control of Human Movement* (4th ed.). Wiley, Hoboken, NJ.

[80] Sebastian Wolf, Tobias Loose, Matthias Schablowski, Leonhard Döderlein, Rüdiger Rupp, Hans Jürgen Gerner, Georg Bretthauer, and Ralf Mikut. 2006. Automated feature assessment in instrumented gait analysis. *Gait Post.* 23, 3 (2006), 331–338. DOI: https://doi.org/10.1016/j.gaitpost.2005.04.004

[81] Luisa M. Zintgraf, Taco S. Cohen, Tameem Adel, and Max Welling. 2017. Visualizing deep neural network decisions: Prediction difference analysis. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[82] Jacek M. Zurada, Aleksander Malinowski, and Ian Cloete. 1994. Sensitivity analysis for minimization of input data dimension for feedforward neural network. In *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 447–450. DOI: https://doi.org/10.1109/ISCAS.1994.409622

[83] Niels J. S. Morch, Ulrik Kjems, Lars Kai Hansen, Claus Svarer, Ian Law, Benny Lautrup, Steve Strother, and Kelly Rehm. 1995. Visualization of neural networks using saliency maps. In *Proceedings of ICNN'95-International Conference on Neural Networks*, vol 4. IEEE, 2085–2090. DOI: 10.1109/ICNN.1995.488997

[84] Erik Strumbelj and Igor Kononenko. 2010. An efficient explanation of individual classifications using game theory. *The Journal of Machine Learning Research* 11 (2010), 1–18. JMLR. org.

Received July 2020; revised June 2021; accepted July 2021

Supplementary Material for: Explaining Machine Learning Models for Clinical Gait Analysis

DJORDJE SLIJEPCEVIC, Institute of Creative Media Technologies, Department of Media and Digital Technologies, St. Pölten University of Applied Sciences
FABIAN HORST, Department of Training and Movement Science, Institute of Sport Science, Johannes Gutenberg-University Mainz
SEBASTIAN LAPUSCHKIN, Department of Artificial Intelligence, Fraunhofer Heinrich Hertz Institute
BRIAN HORSAK, Institute of Health Sciences, Department of Health Sciences, St. Pölten University of Applied Sciences and Center for Digital Health and Social Innovation, St. Pölten University of Applied Sciences, Austria
ANNA-MARIA RABERGER, Institute of Health Sciences, Department of Health Sciences, St. Pölten University of Applied Sciences
ANDREAS KRANZL, Laboratory for Gait and Movement Analysis, Orthopaedic Hospital Vienna-Speising
WOJCIECH SAMEK, Department of Artificial Intelligence, Fraunhofer Heinrich Hertz Institute
CHRISTIAN BREITENEDER, Institute of Visual Computing and Human-Centered Technology, TU Wien
WOLFGANG IMMANUEL SCHÖLLHORN, Department of Training and Movement Science, Institute of Sport Science, Johannes Gutenberg-University Mainz
MATTHIAS ZEPPELZAUER, Institute of Creative Media Technologies, Department of Media and Digital Technologies, St. Pölten University of Applied Sciences

The supplementary material presents additional results we generated for the article

"Explaining Machine Learning Models for Clinical Gait Analysis".

The primary aim of this article is to explain which class-specific characteristics **Machine Learning (ML)** models learn from **clinical gait analysis (CGA)** data. For this purpose, we investigate different gait classification tasks, employ a representative set of classification methods, i.e., (linear) **Support Vector Machine** (**SVM**), **Multi-layer Perceptron** (**MLP**), and **Convolutional Neural Network** (**CNN**), and an **Explainable Artificial Intelligence** (**XAI**) method, i.e., **Layer-wise Relevance Propagation** (**LRP**), to explain predictions at the signal (input) level. Subsequently, the explanations of the individual predictions are aggregated to obtain class-specific model explanations. Since there is no ground truth for automatically generated explanations in this context, we we suggest a two-step approach for the evaluation of the obtained explanations. First, we analyze the discriminatory power of the obtained explanations from a statistical perspective. For this purpose, we leverage **Statistical Parametric Mapping** (**SPM**) to derive statistical measures along with the input signals and thereby investigate how statistically justified the obtained explanations are. Second, two experienced clinical experts interpret the explainability results from a clinical perspective, to evaluate whether obtained explanations match characteristics from clinical practice.

2637-8051/2021/12-ART14 $15.00
https://doi.org/10.1145/3474121

The dataset employed, comprises **ground reaction force** (**GRF**) measurements from 132 patients with **gait disorders** (*GD*) and data from 62 **healthy controls** (*HC*). The *GD* class is furthermore differentiated into three classes of gait disorders associated with the **hip (*H*)**, **knee (*K*)**, and **ankle (*A*)**. The classification tasks, which represent the basis of the XAI investigation, due to high classification accuracies obtained, include a binary classification between healthy controls and all gait disorders (*HC*/*GD*), and a binary classification between healthy controls and each gait disorder separately, i.e., *HC*/*H*, *HC*/*K*, and *HC*/*A*. The classification results obtained for all classification tasks, are presented in supplementary Table S1.

The following figures visualize the relevance-based explanations obtained with LRP. The input vector for the classifiers comprises concatenated affected and unaffected GRF signals. These GRF signals are time-normalized to 101 points (100% stance phase), thus the input vector contains 606 values. For each value, LRP provides whether they are relevant or not for the classification. Sub-figure (A) shows mean GRF signals averaged over each class of the classification task. The shaded areas in all sub-figures highlight areas in the input signals where SPM resulted in a statistically significant difference between both classes. Sub-figure (B) shows mean GRF signals (including a band of one standard deviation) for the *HC* class. The input relevance indicates, which GRF characteristics were most relevant for (or contradictory to) the classification of a certain class. For visualization, input values neutral to the prediction ($R_i \approx 0$) are shown in black, while warm hues indicate input values supporting the prediction ($R_i \gg 0$) of the analyzed class and cool hues identify contradictory input values ($R_i \ll 0$). Sub-figure (C) depicts mean GRF signals averaged over a pathological class (*H*, *K*, or *A*) or all gait disorders (*GD*), in the same format as in sub-figure (B). Sub-figure (D) shows the effect size computed as Pearson's correlation coefficient and the total relevance, which is calculated as the sum of the absolute input relevance values of both classes. The total relevance indicates the common relevance of the input signal for the classification task.

CLASSIFICATION RESULTS

Table S1. Overview of the Prediction Accuracy Obtained for the Three Employed Classification Methods (CNN, SVM, and MLP) and All Classification Tasks with Min–Max Normalized and Non-Normalized Input Signals, Reported as Mean (Standard Deviation) Over the Ten-Fold Cross Validation in Percent

Task	Normalization	ZRB	CNN	SVM	MLP
HC/GD	no norm.	68.0	87.8 (4.5)	88.6 (4.9)	88.1 (4.8)
HC/GD	min-max	68.0	88.0 (5.0)	88.4 (5.3)	88.8 (5.0)
HC/H	no norm.	62.6	85.1 (8.2)	85.9 (8.4)	86.6 (7.9)
HC/H	min-max	62.6	85.5 (8.0)	87.1 (7.6)	86.7 (8.5)
HC/K	no norm.	54.4	84.8 (9.9)	85.7 (9.0)	86.1 (7.9)
HC/K	min-max	54.4	85.9 (9.3)	88.5 (7.2)	88.5 (7.6)
HC/A	no norm.	59.0	88.7 (5.5)	89.1 (5.9)	88.3 (6.3)
HC/A	min-max	59.0	86.7 (8.3)	87.6 (7.4)	86.5 (8.1)
H/K/A	no norm.	39.4	48.0 (10.1)	46.4 (9.5)	45.9 (11.0)
H/K/A	min-max	39.4	50.7 (9.8)	51.8 (9.6)	47.4 (10.9)
HC/H/K/A	no norm.	32.0	55.0 (8.7)	58.7 (7.5)	55.6 (7.6)
HC/H/K/A	min-max	32.0	57.5 (7.0)	59.5 (8.5)	59.2 (7.6)

Note that the **Zero-Rule Baseline** (**ZRB**) is task-specific.

EXPLAINABILITY RESULTS

Classification Task: *HC*/*GD* | Classification method: *CNN*

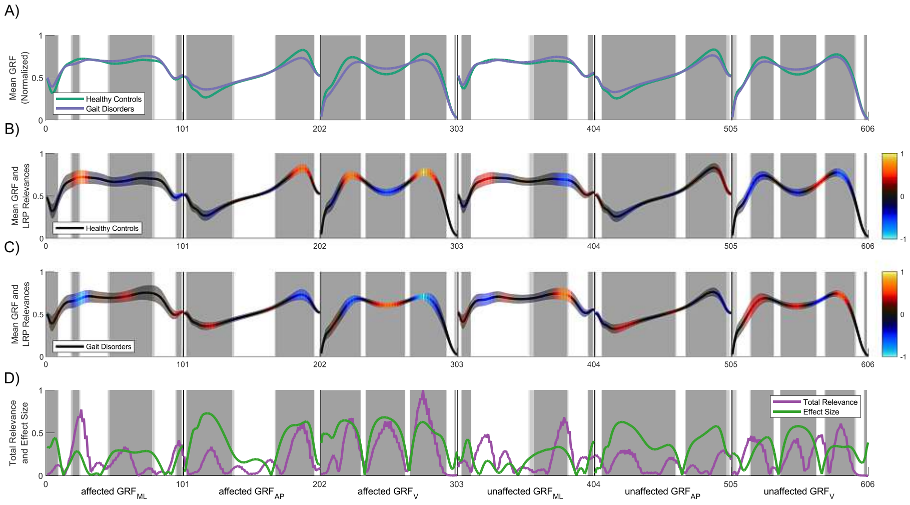


Fig. S1. Result overview for the classification of healthy controls and the aggregated class of all three gait disorders (*HC*/*GD*) based on min–max normalized GRF signals using a CNN as classifier.

Classification Task: *HC*/*GD* | Classification method: *MLP*

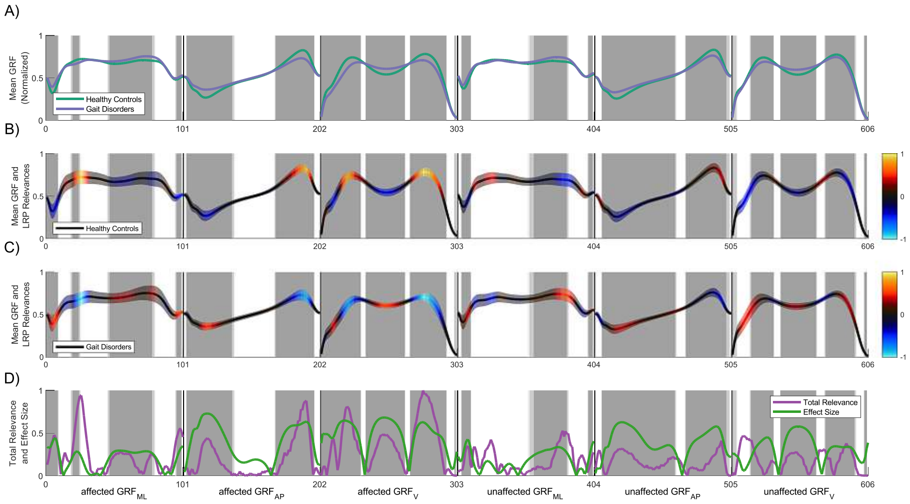


Fig. S2. Result overview for the classification of healthy controls and the aggregated class of all three gait disorders (*HC*/*GD*) based on min–max normalized GRF signals using an MLP as classifier.

Classification Task: *HC/GD* | Classification method: *SVM*

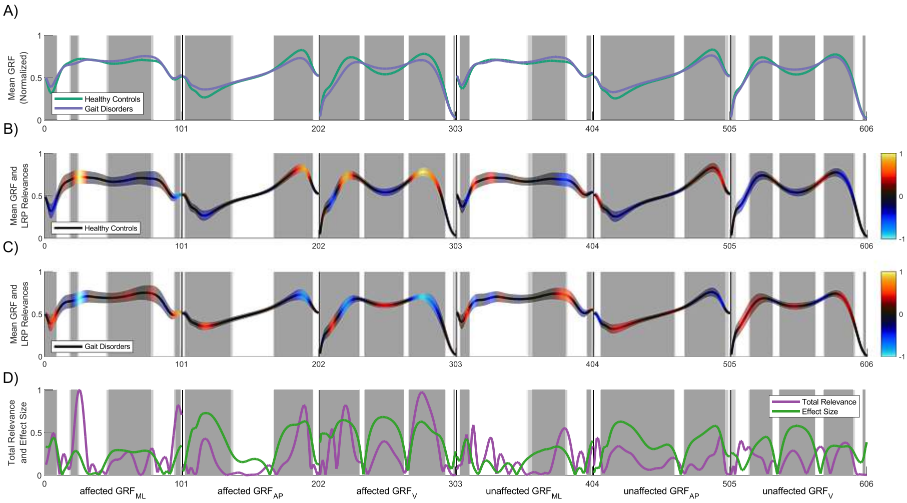


Fig. S3. Result overview for the classification of healthy controls and the aggregated class of all three gait disorders (*HC/GD*) based on min–max normalized GRF signals using an SVM as classifier.

Classification Task: *HC/H* | Classification method: *CNN*

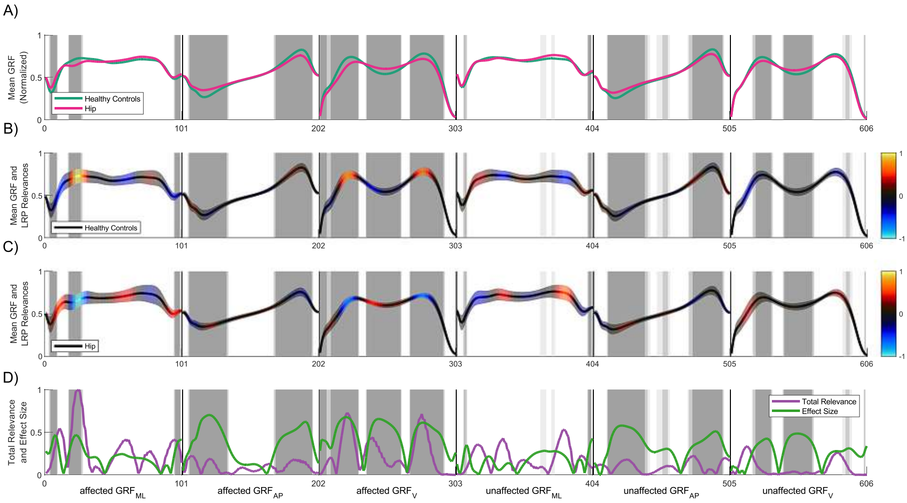


Fig. S4. Result overview for the classification of healthy controls (*HC*) and hip injury class (*H*) based on min–max normalized GRF signals using a CNN as classifier.

Classification Task: HC/H | Classification method: MLP

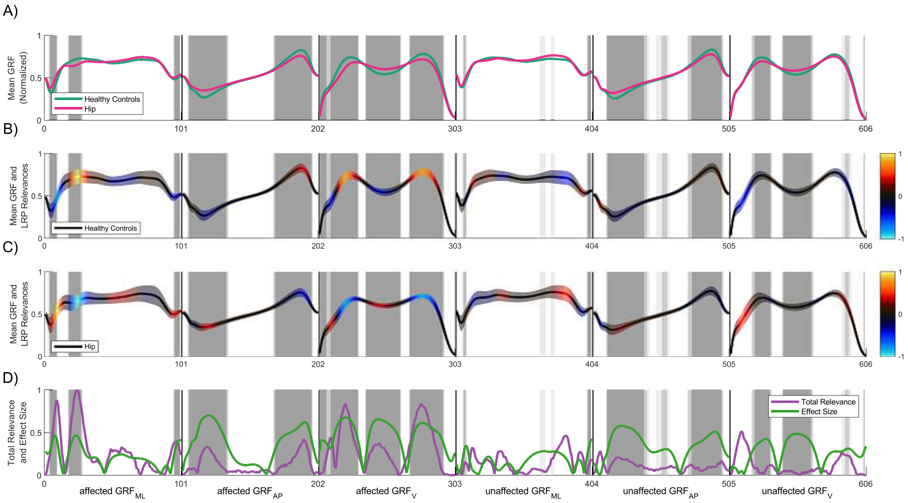


Fig. S5. Result overview for the classification of healthy controls (HC) and hip injury class (H) based on min–max normalized GRF signals using an MLP as classifier.

Classification Task: HC/H | Classification method: SVM

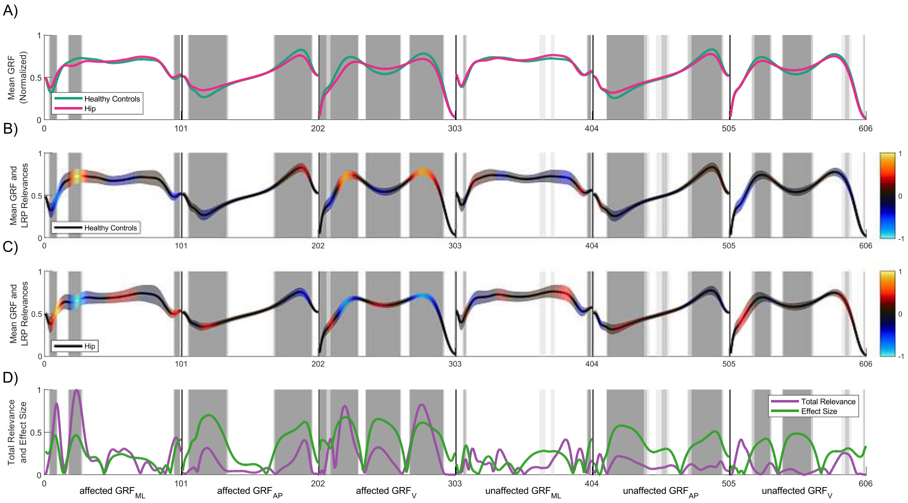


Fig. S6. Result overview for the classification of healthy controls (HC) and hip injury class (H) based on min–max normalized GRF signals using an SVM as classifier.

Classification Task: HC/K | Classification method: CNN

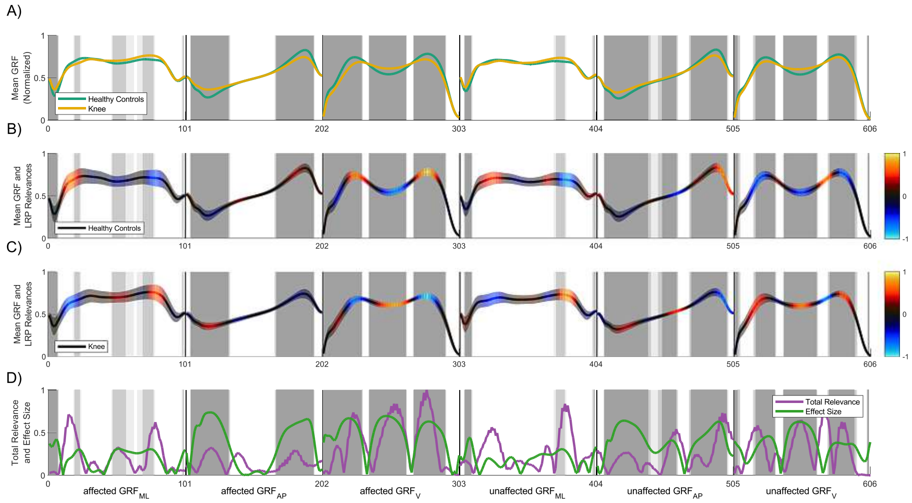


Fig. S7. Result overview for the classification of healthy controls (HC) and knee injury class (K) based on min–max normalized GRF signals using a CNN as classifier.

Classification Task: HC/K | Classification method: MLP

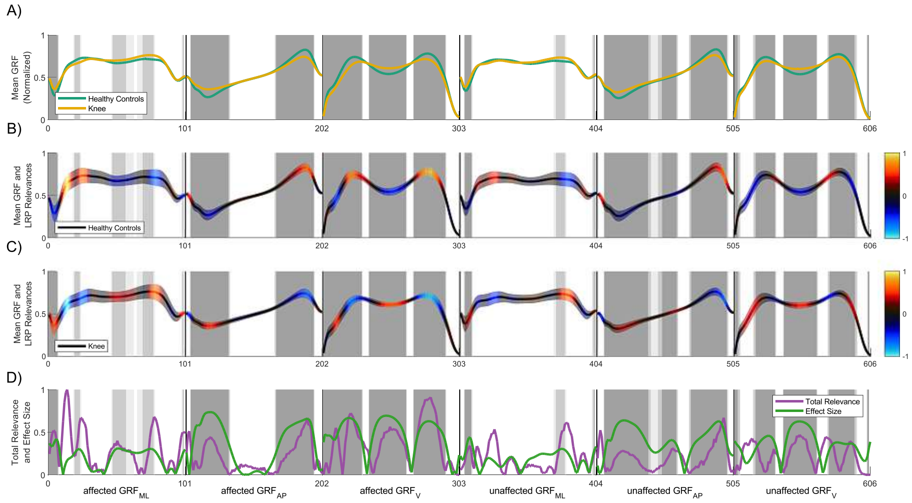


Fig. S8. Result overview for the classification of healthy controls (HC) and knee injury class (K) based on min–max normalized GRF signals using an MLP as classifier.

Classification Task: HC/K | Classification method: SVM

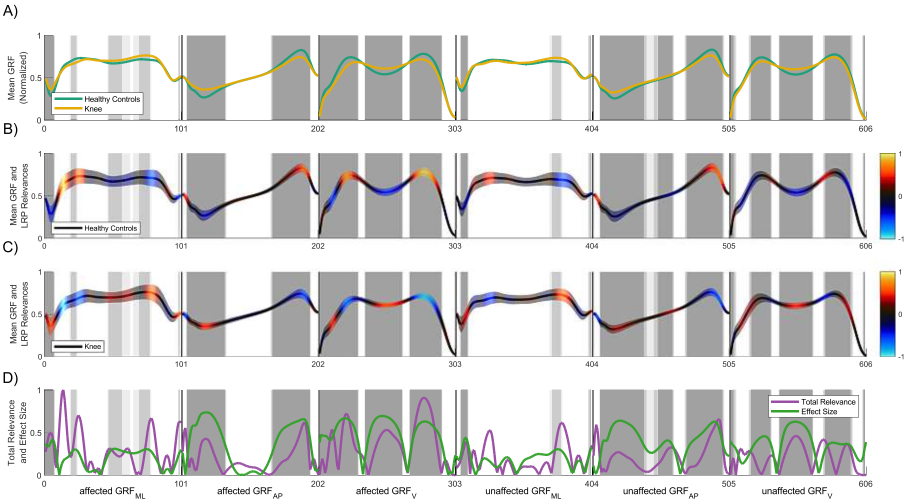


Fig. S9. Result overview for the classification of healthy controls (HC) and knee injury class (K) based on min–max normalized GRF signals using an SVM as classifier.

Classification Task: HC/A | Classification method: CNN

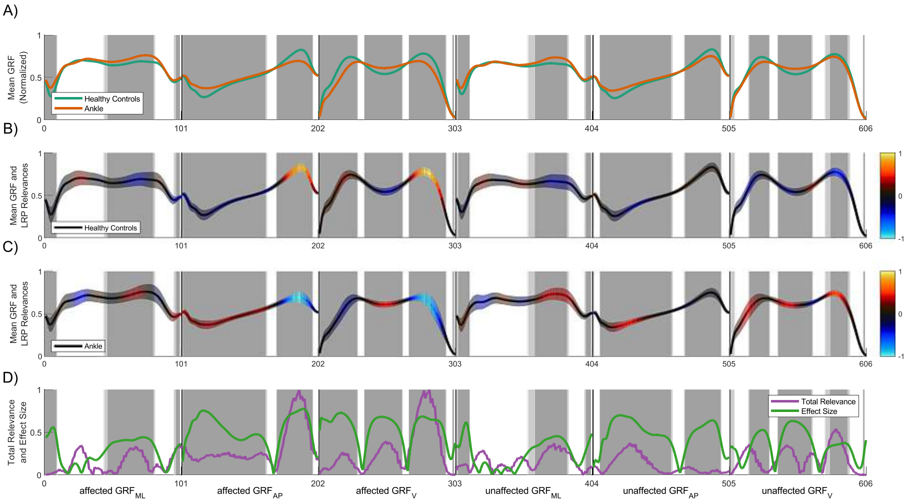


Fig. S10. Result overview for the classification of healthy controls (HC) and ankle injury class (A) based on min–max normalized GRF signals using a CNN as classifier.

Classification Task: *HC*/*A* | Classification method: *MLP*

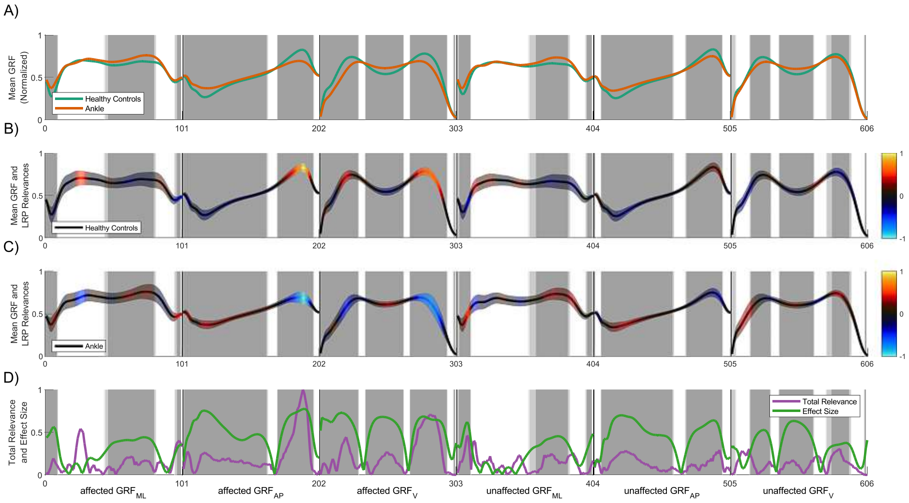


Fig. S11. Result overview for the classification of healthy controls (*HC*) and ankle injury class (*A*) based on min–max normalized GRF signals using an MLP as classifier.

Classification Task: *HC*/*A* | Classification method: *SVM*

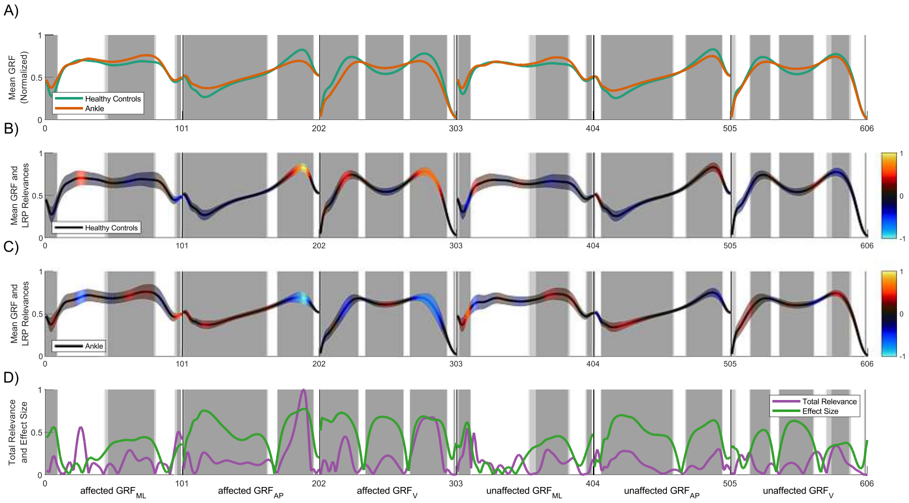


Fig. S12. Result overview for the classification of healthy controls (*HC*) and ankle injury class (*A*) based on min–max normalized GRF signals using an SVM as classifier.

EXPLAINABILITY RESULTS – NON-NORMALIZED DATA

Classification Task: *HC*/*GD* | Classification method: *CNN*

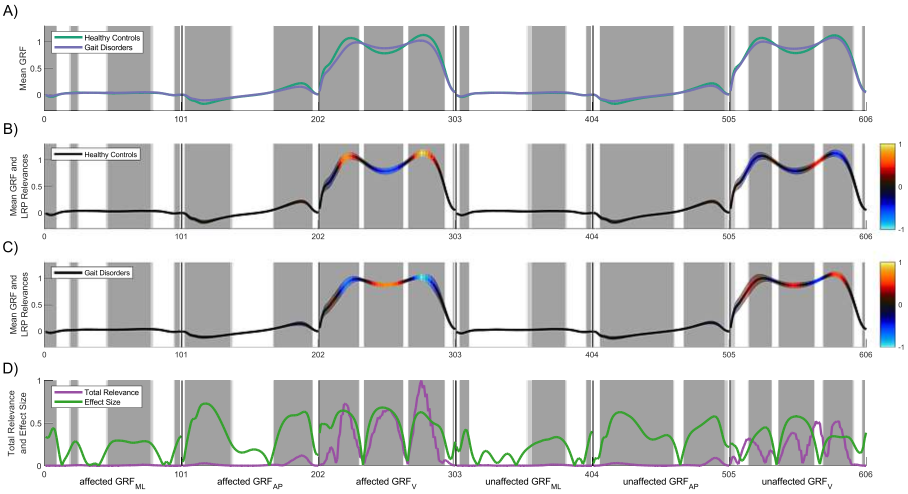


Fig. S13. Result overview for the classification of healthy controls and the aggregated class of all three gait disorders (*HC*/*GD*) based on non-normalized GRF signals using a CNN as classifier.

Classification Task: *HC*/*GD* | Classification method: *MLP*

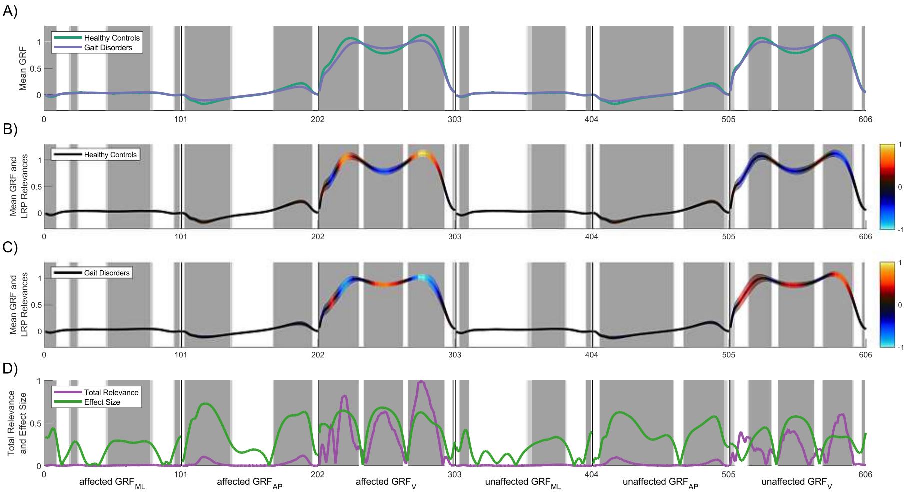


Fig. S14. Result overview for the classification of healthy controls and the aggregated class of all three gait disorders (*HC*/*GD*) based on non-normalized GRF signals using an MLP as classifier.

Classification Task: *HC*/*GD* | Classification method: *SVM*

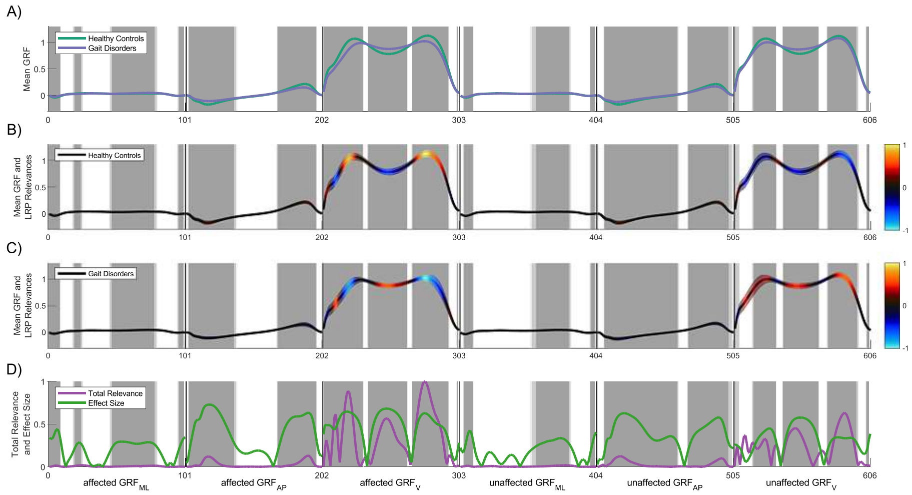


Fig. S15. Result overview for the classification of healthy controls and the aggregated class of all three gait disorders (*HC*/*GD*) based on non-normalized GRF signals using an SVM as classifier.

Classification Task: *HC*/*H* | Classification method: *CNN*

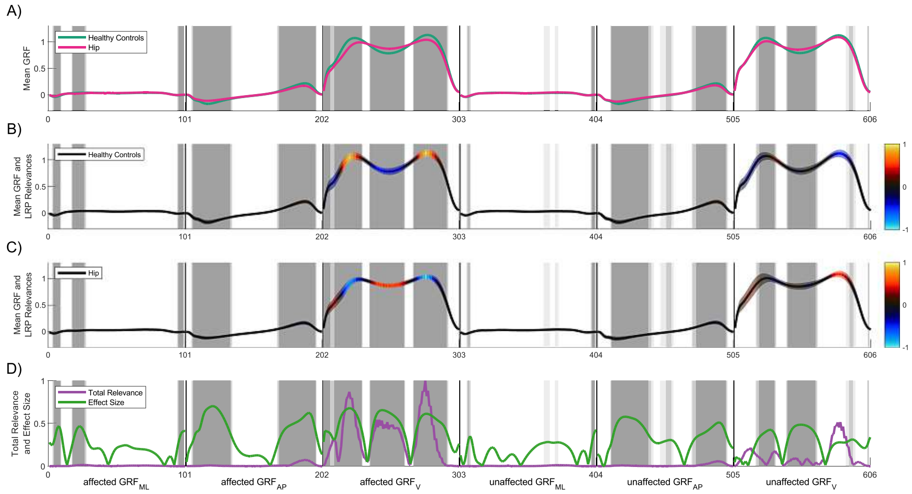


Fig. S16. Result overview for the classification of healthy controls (*HC*) and hip injury class (*H*) based on non-normalized GRF signals using a CNN as classifier.

Classification Task: *HC*/*H* | Classification method: *MLP*

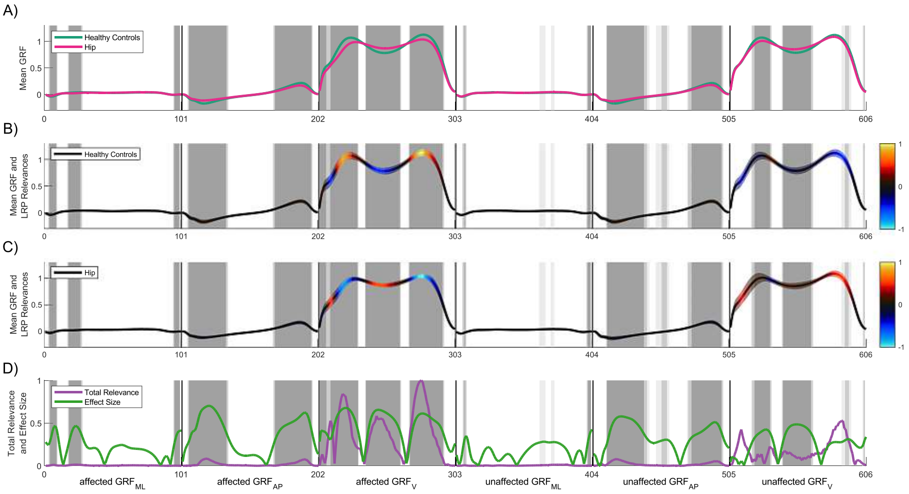


Fig. S17. Result overview for the classification of healthy controls (*HC*) and hip injury class (*H*) based on non-normalized GRF signals using an MLP as classifier.

Classification Task: *HC*/*H* | Classification method: *SVM*

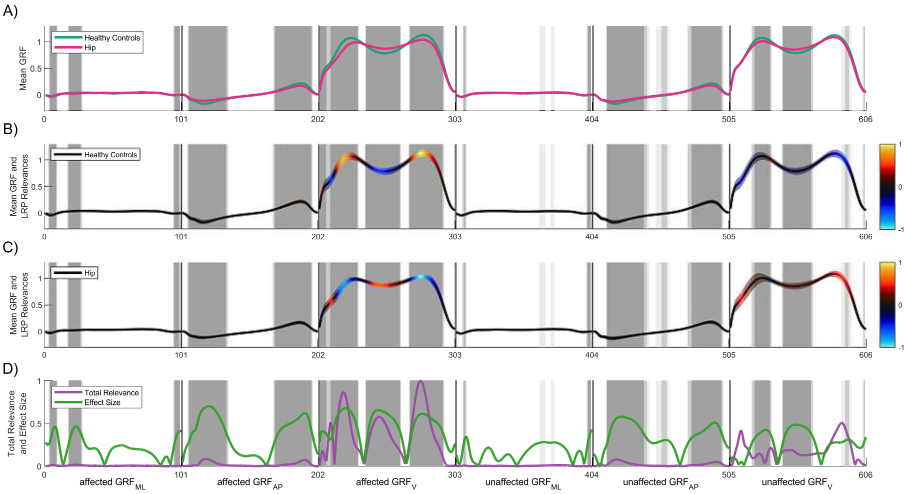


Fig. S18. Result overview for the classification of healthy controls (*HC*) and hip injury class (*H*) based on non-normalized GRF signals using an SVM as classifier.

Classification Task: *HC*/*K* | Classification method: *CNN*

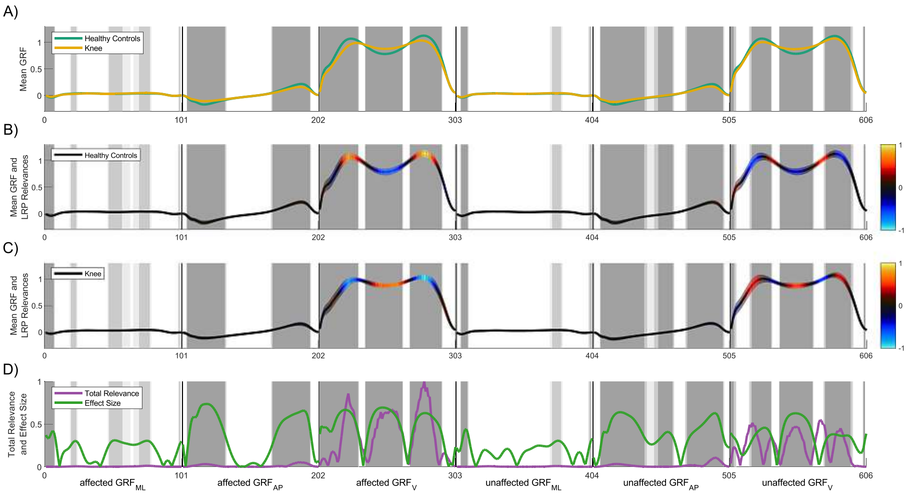


Fig. S19. Result overview for the classification of healthy controls (*HC*) and knee injury class (*K*) based on non-normalized GRF signals using a CNN as classifier.

Classification Task: *HC*/*K* | Classification method: *MLP*

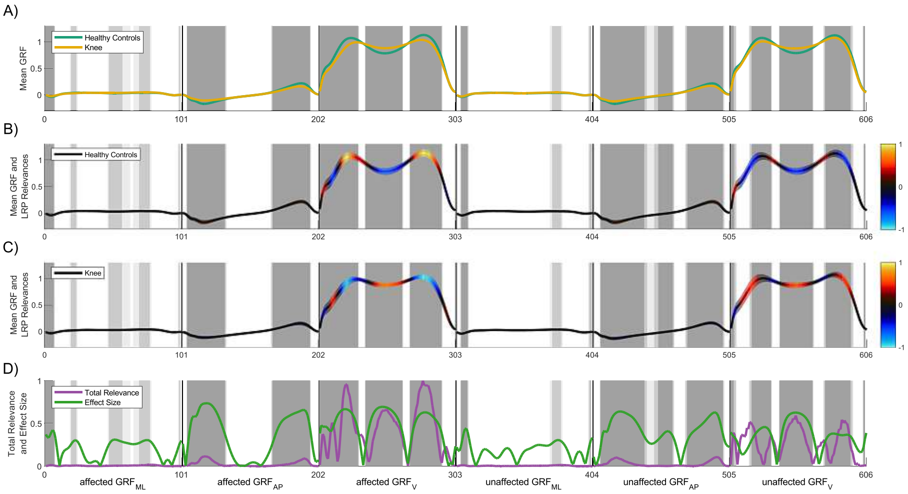


Fig. S20. Result overview for the classification of healthy controls (*HC*) and knee injury class (*K*) based on non-normalized GRF signals using an MLP as classifier.

Classification Task: *HC*/*K* | Classification method: *SVM*

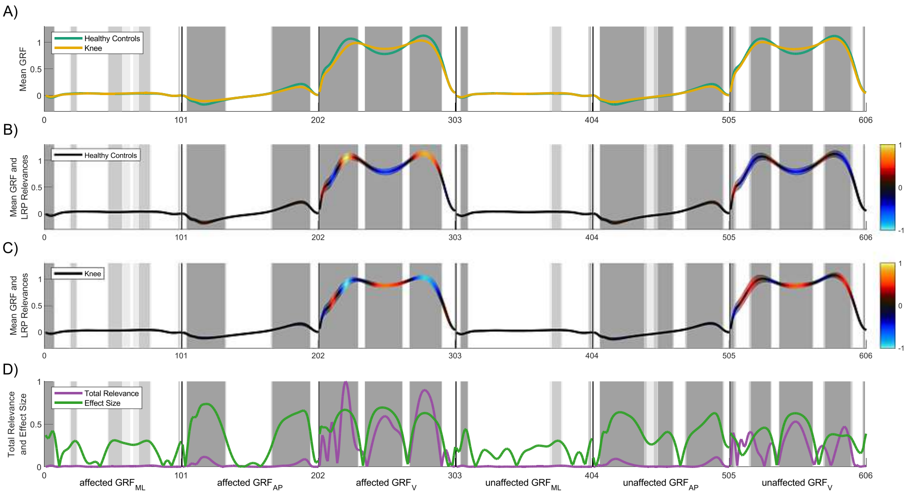


Fig. S21. Result overview for the classification of healthy controls (*HC*) and knee injury class (*K*) based on non-normalized GRF signals using an SVM as classifier.

Classification Task: *HC*/*A* | Classification method: *CNN*

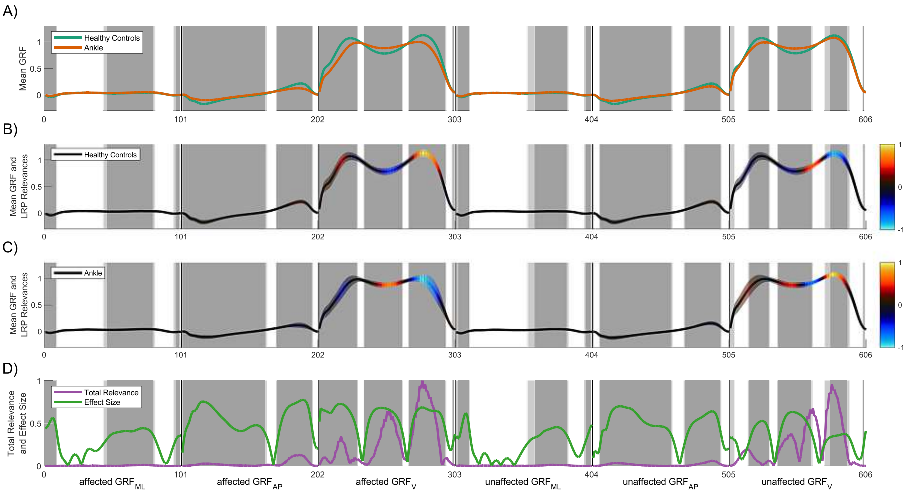


Fig. S22. Result overview for the classification of healthy controls (*HC*) and ankle injury class (*A*) based on non-normalized GRF signals using a CNN as classifier.

Classification Task: *HC*/*A* | Classification method: *MLP*

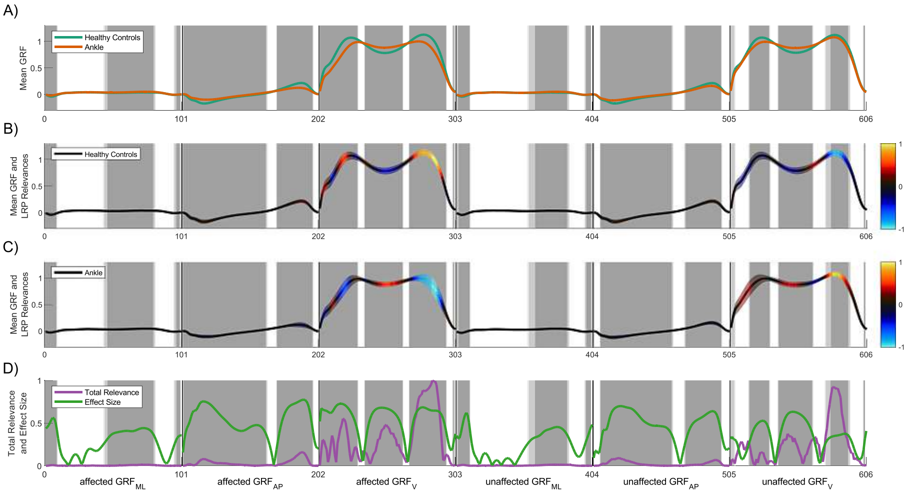


Fig. S23. Result overview for the classification of healthy controls (*HC*) and ankle injury class (*A*) based on non-normalized GRF signals using an MLP as classifier.

Classification Task: *HC*/*A* | Classification method: *SVM*

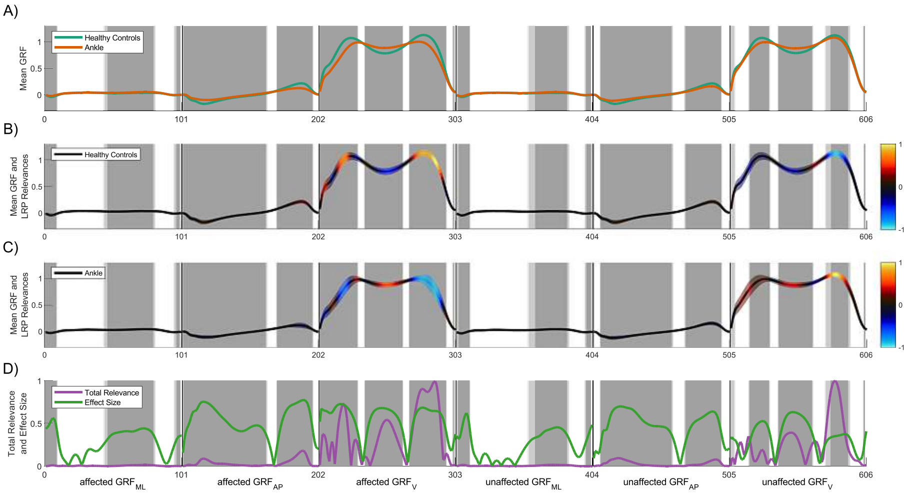


Fig. S24. Result overview for the classification of healthy controls (*HC*) and ankle injury class (*A*) based on non-normalized GRF signals using an SVM as classifier.

2.5 Explainable Machine Learning in Human Gait Analysis: A Study on Children With Cerebral Palsy

Djordje Slijepcevic, Matthias Zeppelzauer, Fabian Unglaube, Andreas Kranzl, Christian Breiteneder, and Brian Horsak. **Explainable Machine Learning in Human Gait Analysis: A Study on Children With Cerebral Palsy**. *IEEE Access*, 11:65906–65923, 2023. DOI: 10.1109/ACCESS.2023.3289986

The final version of this publication is available at: `https://doi.org/10.1109/ACCESS.2023.3289986`. Permission for reprint granted, © 2023 Slijepcevic

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

IEEE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY SECTION

Received 10 May 2023, accepted 11 June 2023, date of publication 27 June 2023, date of current version 6 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3289986

RESEARCH ARTICLE

Explainable Machine Learning in Human Gait Analysis: A Study on Children With Cerebral Palsy

DJORDJE SLIJEPCEVIC[1], MATTHIAS ZEPPELZAUER[1], FABIAN UNGLAUBE[2], ANDREAS KRANZL[2], CHRISTIAN BREITENEDER[3], AND BRIAN HORSAK[4,5]

[1]Institute of Creative Media Technologies, St. Pölten University of Applied Sciences, 3100 Sankt Pölten, Austria
[2]Laboratory for Gait and Movement Analysis, Orthopaedic Hospital Speising, 1130 Vienna, Austria
[3]Institute of Visual Computing and Human-Centered Technology, TU Wien, 1040 Vienna, Austria
[4]Institute of Health Sciences, St. Pölten University of Applied Sciences, 3100 Sankt Pölten, Austria
[5]Center for Digital Health and Social Innovation, St. Pölten University of Applied Sciences, 3100 Sankt Pölten, Austria

Corresponding author: Djordje Slijepcevic (djordje.slijepcevic@fhstp.ac.at)

This work was supported in part by the Research Promotion Agency of Lower Austria and the Provincial Government of Lower Austria within IntelliGait3D under Grant FTI17-014, and in part by the Endowed Professorship for Applied Biomechanics and Rehabilitation Research under Grant SP19-004.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of the City of Vienna under Application No. EK 19-083-VK.

ABSTRACT This work investigates the effectiveness of various machine learning (ML) methods in classifying human gait patterns associated with cerebral palsy (CP) and examines the clinical relevance of the learned features using explainability approaches. We trained different ML models, including convolutional neural networks, self-normalizing neural networks, random forests, and decision trees, and generated explanations for the trained models. For the deep neural networks, Grad-CAM explanations were aggregated on different levels to obtain explanations at the decision, class and model level. We investigate which subsets of 3D gait analysis data are particularly suitable for the classification of CP-related gait patterns. The results demonstrate the superiority of kinematic over ground reaction force data for this classification task and show that traditional ML approaches such as random forests and decision trees achieve better results and focus more on clinically relevant regions compared to deep neural networks. The best configuration, using sagittal knee and ankle angles with a random forest, achieved a classification accuracy of 93.4 % over all four CP classes (crouch gait, apparent equinus, jump gait, and true equinus). Deep neural networks utilized not only clinically relevant features but also additional ones for their predictions, which may provide novel insights into the data and raise new research questions. Overall, the article provides insights into the application of ML in clinical practice and highlights the importance of explainability to promote trust and understanding of ML models.

INDEX TERMS Explainable artificial intelligence, explainability, human gait analysis, biomechanical gait data, kinematics, ground reaction forces, convolutional neural network, self-normalizing neural network, random forest, decision tree.

I. INTRODUCTION

Walking impairments can severely affect a person's ability to participate in social activities and work life and negatively impact quality of life. Causes of walking impairments can range from traumatic events to various diseases such as stroke, Parkinson's disease, or cerebral palsy (CP). One of the most common causes of physical disability in children is CP, which occurs in approximately 2.5 out of every 1,000 births in developed countries [1]. This group of

The associate editor coordinating the review of this manuscript and approving it for publication was Kin Fong Lei.

neurological disorders can cause tremors, muscle weakness, stiffness, and spasticity, which can affect a child's motor functions and ability to walk [2]. One of the most frequent causes of CP are brain lesions that occur before, during, or shortly after birth, mostly leading to musculoskeletal impairments that can worsen throughout childhood and adolescence [3].

Accurate examination and quantification of underlying movement mechanisms are necessary to ensure the best possible treatment for children with CP. Such information is essential for clinicians to offer targeted treatment plans to their patients. The worldwide established gold standard for this purpose is clinical 3D gait analysis (3DGA). This method allows to objectively and quantitatively describe and analyze the human motor function of patients from a kinematic (joint angles) and kinetic (ground reaction forces, joint reaction forces, and joint moments) point of view [4]. The basis for 3DGA is motion capturing and assessment of ground reaction forces (GRF). Both data sources capture complementary information on walking behavior.

This work investigates the automated classification of gait patterns associated with CP using data from clinical 3DGA, i.e., kinematic data, GRF data, and a combination of both. Recordings obtained during 3DGA produce a vast amount of data. A typical report in clinical practice contains up to a few dozen discrete parameters along with more than 20 waveforms describing kinematic and kinetic gait variables across a gait cycle. A gait cycle refers to the interval between the initial contact of a foot and the subsequent initial contact of the same foot and is the standard "time frame" used in clinical practice to describe human gait. Due to the complexity, characterized by high dimensionality, temporal dependence, strong variability, non-linear relationship, and inter-correlation [5], these data are challenging to comprehend and analyze manually (see Figure 1 for an example of a data record obtained by 3DGA). Hence, data interpretation in clinical practice is highly challenging, and a lot of experience is required to draw valid medical conclusions.

The complexity of the data in 3DGA combined with the need for timely and precise decision-making has motivated research to utilize Machine Learning (ML) to aid decision-making [6]. ML approaches increasingly leverage non-linear classification models, such as multi-layer (deep) neural networks, which have shown to provide promising results concerning classification accuracy in the field of clinical gait analysis [7], [8]. However, such complex classification models share a major limitation: their black-box nature [9]. This means that it is hard to trace back and understand how a certain model has reached a specific decision, how it is grounded in the input data, and what kind of patterns and rules it actually learned from the data. Consequently, even well-performing ML models are rarely used in clinical practice [10].

Given an ML model trained on 3DGA data, it is a non-trivial task to trace back which patterns in the signal are responsible for its predictions. Furthermore, it is unclear whether predictions are based on clinically relevant patterns or rather on signals that relate to the targeted pathologies due to a spurious correlation or a bias in the data but are not causally related to them. The experts' skepticism regarding automatically generated predictions and diagnostic suggestions is, therefore, well justified. At the same time, the strong performance obtained by state-of-the-art ML models shows great potential to significantly support the diagnostic process and, thus, to save costs and time in everyday clinical practice. Therefore, their application in clinical practice would be of great value. This, however, requires ML approaches to become more transparent and traceable, e.g., via explainability mechanisms [8]. In addition, this would further help to fulfill legal requirements, such as the EU General Data Protection Regulation (GDPR) [11], that require the traceability of ML predictions.

The primary aim of this work is to investigate the effectiveness of various ML methods in automatically classifying gait patterns associated with CP and to employ explainability approaches to examine whether the features learned by these models are clinically relevant. To this end, we trained different ML models, including convolutional neural networks (CNNs), self-normalizing neural networks (SNNs) [12], random forests (RFs), and decision trees (DTs) and generated and compared explanations for the trained models. We utilized model-specific explanations for DTs and RFs in terms of Gini impurity-based feature importance. For the investigated deep neural networks (DNNs), we adapted the well-known Grad-CAM algorithm [13] to be applicable to one-dimensional time series input data. This explainability method has been shown to be robust [14] in explaining the internal workings of DNNs.

Our investigation focuses on the following leading research questions:

1) How advantageous is the use of kinematic data over GRF data in the automated classification of gait patterns associated with CP, and are the two inputs more effective in combination than used individually?
2) How do traditional ML models compare to state-of-the-art DNNs for the automated classification of clinical 3DGA data in terms of performance and explainability?
3) To what extent do the investigated ML models base their decisions on clinically meaningful features when classifying CP-related gait patterns?
4) To what extent are the explanations obtained from DNNs robust to variations in architecture?

We performed experiments on a dataset of 302 patients with CP (375 limbs) and four different gait patterns related to this condition. Our results show an unexpected outcome. DNNs have not met initial expectations and fall behind traditional methods in classification performance. Compared to traditional methods that provide concise explanations and identify and utilize clinically relevant regions in the input data for the classification task, DNNs are less informative in their explanations.

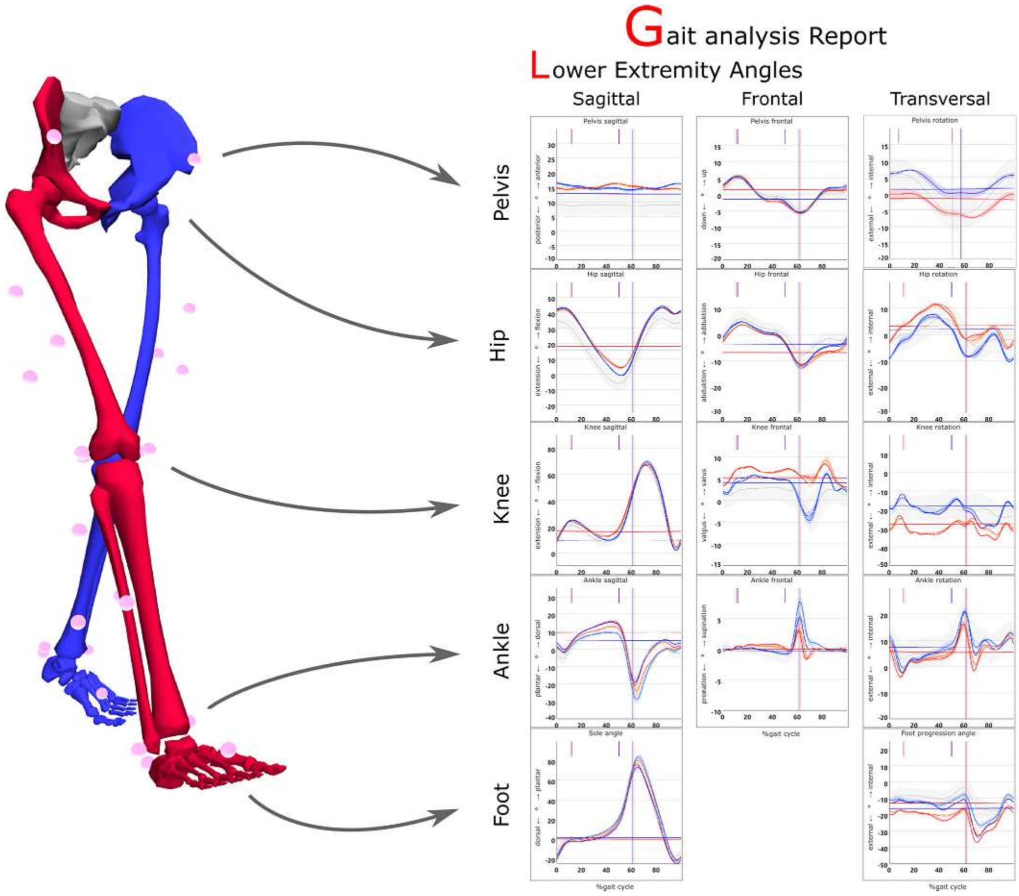


FIGURE 1. Visualization of a 3DGA data record as used in clinical practice. Several retro-reflective markers (pink spheres) are attached to specific anatomical landmarks of the human body and allow quantifying human locomotion using 3D motion-capturing techniques. The 3D trajectories of these markers combined with geometrical biomechanical models are used to calculate, e.g., joint angles. In clinical practice, this information is used to inform medical decision-making. The data from clinical 3DGA are typically reported in simple line plots. Blue and red colors encode the right and left body sides, respectively. Deriving a diagnosis from these abstract line plots is a challenging task that requires trained medical personnel. Thus, machine learning models are highly desirable to assist decision-making.

II. RELATED WORK

The research in this paper combines methodology from multiple disciplines, namely automated classification in clinical gait analysis and explainable machine learning, a branch of explainable artificial intelligence (XAI). For this reason, the related work is structured in two subsections, one for each field.

A. AUTOMATED CLASSIFICATION OF PATIENTS WITH CP

There is a growing interest in using ML in the field of clinical gait analysis due to its ability to analyze large amounts of gait data in a cost-effective, fast, and objective manner [6], [15], [16]. ML methods have been successfully applied to analyze gait patterns of patients with different conditions, such as stroke [17], Parkinson's disease [18], multiple sclerosis [19], osteoarthritis [20], and various functional gait disorders [21], [22]. One area that has received particular attention in the literature is the use of ML for automated classification of gait patterns associated with CP [23].

Several studies have compared the performance of different ML approaches for this task. Ferrari et al. [24] compared the use of multi-layer perceptrons (MLPs), support vector machines (SVMs), and long short-term memory networks (LSTMs) for the classification of four CP-related gait patterns defined by Ferrari et al. [25]. According to their results, the LSTMs achieved the highest classification accuracy of 67.4 %. The authors utilized kinematic data and frequency information (obtained via fast Fourier transform) from 174 patients. Zhang and Ma [26] compared seven ML methods, including MLP, SVM, DT, and RF, for the classification of CP-related gait patterns as defined by Rodda and Graham [27], i.e., crouch gait, apparent equinus, jump gait, and true equinus. The MLPs performed best with a classification accuracy of 93.5 %. The DT, RF, and SVM had considerably lower accuracy rates of 84.3 %, 83.6 %, and 85.0 %, respectively. The dataset comprised discrete parameters from kinematic waveforms of 200 children. Darbandi et al. [28] also used discrete parameters of kinematic data of 66 children in a stochastic approach to translate expert knowledge into rules and to perform fuzzy clustering. For the classification of gait patterns as defined by Rodda and Graham [27], their approach achieved a classification accuracy of 94.0 %. Chia et al. [29] developed a decision support system that used discrete parameters from kinematic waveforms, physical examinations, and anthropometric data to identify 14 different CP-related impairments (e.g., hamstring spasticity, gastrocnemius spasticity, and gluteal weakness) and provide surgical recommendations. The dataset comprised 689 3DGA recordings of 423 children. The authors evaluated the performance of a stratified and standard RF, with the latter achieving better results, i.e., a misclassification rate of 0.13 (corresponding to a classification accuracy of 87.0 %). Furthermore, feature importance served as decision explanation and partial dependence plots as model explanation.

B. EXPLAINABLE MACHINE LEARNING

The inherent non-transparency of modern ML models, in particular DNNs, has greatly advanced research on explainability methods in the field of XAI in recent years. These methods are designed to provide explanations for automated predictions and to help clinical experts understand how and why a particular prediction was made. XAI methods can be categorized according to the type of explanation they provide. Following the taxonomy of Arya et al. [30], we distinguish between XAI approaches for (i) data exploration, (ii) decision explanation, and (iii) model explanation.

Data exploration methods cannot explain an ML model, but rather the data on which the model was trained. These methods include techniques from the field of visual analytics [31], statistics (e.g., statistical parametric mapping) [8], and unsupervised machine learning [32], [33]. The goal is to visualize and adequately transform the data, thereby enabling domain experts to find meaningful structures and patterns that will allow them to better understand the data, their distribution, and cluster structures. This process should result in novel insights from the data. Data exploration is generally recommended before an ML model is trained.

Decision explanation methods explain the *local* behavior of an ML model, i.e., providing an explanation for the prediction of an individual data sample. For a classification task, such an explanation can, for example, indicate which parts of the input are responsible for the prediction. In the case of gait classification, such methods can identify characteristic sections in the input time series related to a specific gait disorder [8]. The majority of decision explanation methods are *post-hoc* methods, which offer great flexibility as they can be directly applied to previously trained classification models [30]. Typical results of post-hoc methods are saliency maps that highlight which input features are most relevant to a particular prediction [13]. Post-hoc methods can be divided into propagation-based and perturbation-based approaches. Propagation-based methods determine the effect of input features on the model's prediction by (partially) back-propagating an entity of interest (e.g., gradients) from the output to the input of the model. Popular examples for such approaches are SmoothGrad [34], Grad-CAM [13], and Layer-wise Relevance Propagation (LRP) [35]. Perturbation-based methods, e.g., Local Interpretable Model-Agnostic Explanations (LIME) [36] and SHapley Additive exPlanations (SHAP) [37], estimate the importance of input features by partially masking the input and measuring the effect on the model output. Perturbation-based methods are model-agnostic since no access to the internal architecture of the models is necessary. Compared to propagation-based methods, however, they require a significantly higher computational effort. Propagation-based methods are computationally more efficient and allow the explanation of classifier-specific characteristics, thus enabling a more profound analysis.

Besides decision explanations, there are model explanation methods that aim to explain what a trained model has learned at a *global* level, e.g., by providing class-specific prototypes [38] or synthesized samples reflecting the characteristic patterns learned for a certain class [39]. Consequently, ambiguous features that the model learned can be identified and overlaps between classes can be detected. Model explanation allows to check whether a model has been trained correctly and whether the predicted classes are based on meaningful patterns. Decision and model explanation, thus, complement each other.

In clinical gait analysis, only a few studies have used XAI to shed light on the underlying black-box models and promote their use within the clinical setting. We have recently proposed several approaches for model and decision explanation based on LRP to explain the functioning of different ML models (i.e., linear SVMs, MLPs, and CNNs) for the classification of GRF data into different functional gait disorders [8]. The investigated ML models utilized GRF waveforms as input. Consequently, to obtain class-specific explanations, the averaged relevance scores were superimposed over the averaged GRF waveforms. Furthermore, we proposed the use of a model explanation based on SpRAy [40], which is a method for identifying clusters within the explanations. These approaches have been further investigated to explain sex- and age-dependent gait patterns learned by ML models [41], [42]. Dindorf et al. [43] used LIME to explain a linear SVM that was trained to distinguish between healthy controls and patients after total hip arthroplasty. In their study, two input scenarios were examined, one with kinematic and kinetic waveforms, while the other employed discrete parameters derived from these waveforms. The explainability results showed that the SVMs were highly sensitive to the input representation employed and that each of the models often focused on different biomechanical features. Kokkotis et al. [44] leveraged SHAP to explain which discrete kinematic and kinetic parameters contributed most to the decisions of an SVM for the classification of patients with anterior cruciate ligament injury (with and without reconstruction surgery) and healthy controls. The authors noted that SHAP highlighted several discrete parameters that were consistent with biomechanical findings reported in the literature. However, a discrepancy was observed between the explainability results and the results of conventional statistical analysis. Recently, we proposed *gaitXplorer* [45], a visual analytics approach for the classification of CP-related gait patterns that employed Grad-CAM [13] to explain predictions of CNNs. This work employs the same dataset and explainability method as the present paper, but focuses solely on decision explanations. Figure 2 shows the interactive visual interface of the *gaitXplorer*.

The present work investigates for the first time the suitability of different DNN architectures and their explainability in terms of decision and model explanations for the classification of CP-related gait patterns as defined by Rodda et al. [47]. Moreover, our research addresses the ability of DNNs and conventional ML approaches to capture clinically significant features from kinematic and kinetic waveforms.

III. METHODS

A. CLINICAL USE CASE AND TARGET CLASSES

The clinical use case of this paper is the identification of clinically well-defined gait patterns in children with CP and neuro-muscular disorders. This patient group is associated with varying symptoms such as muscle weakness and stiffness, tremors, and limited joint range of motion, among other impairments [2]. All of these can strongly affect motor function and the ability to walk. The present study uses 3DGA data from patients who can walk independently and have well-recognizable gait deviations such as toe-walking, flexed-stiff knees, flexed hips, and an anteriorly tilted pelvis [47]. The correct classification and identification of these underlying impairments are essential as clinicians base their decisions about optimal treatment interventions on this information.

All patients in our study were categorized into four pathological gait patterns by a clinically established procedure, the so-called *ankle plantarflexor-knee extension couple (PFKE) index* [48]. This categorization served as the ground truth during the training and evaluation of the ML model. The method compares the sagittal knee and ankle angles of patients with those of a speed-matched healthy control cohort and automatically determines the four classes using a set of rules. These rules provide a well-suited reference for the evaluation of the appropriateness of the explainability results. In our experiment we expect that a trustworthy classification model for CP would base its decisions on the same signal regions as the PFKE method.

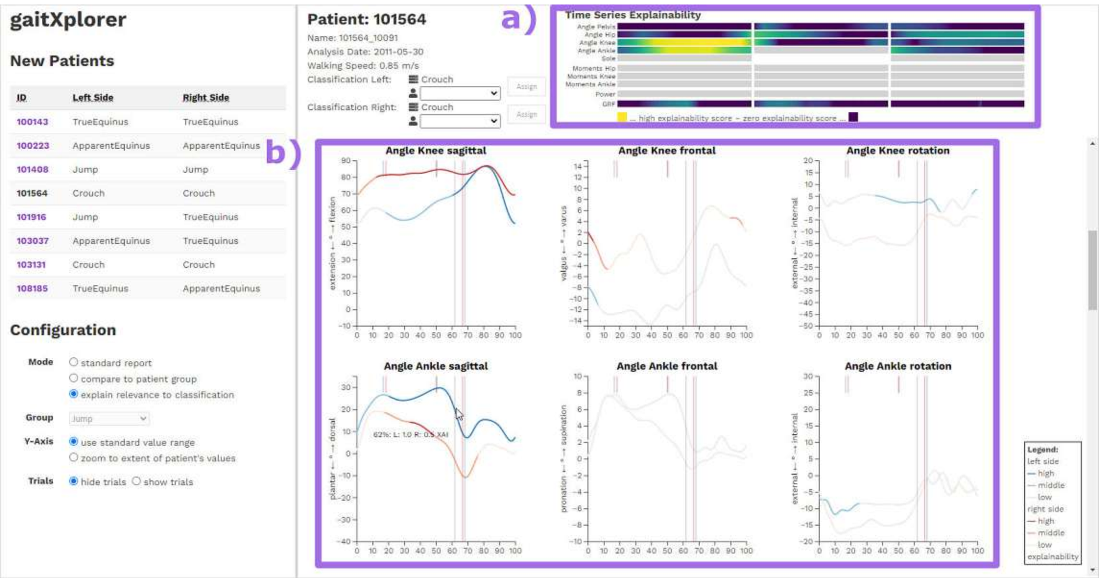
The gait patterns associated with CP are illustrated in Figure 3 and briefly described in the following [47]:

- True equinus: The ankle is in plantar flexion throughout the stance phase (''toe-walking'').
- Jump gait: Equinus at the ankle (partly in late stance), flexion at knee and hip (especially in early stance), anterior pelvis tilt, and increased lumbar lordosis.
- Apparent equinus: The ankle has a normal range, but the knee and hip are excessively flexed throughout the stance, and the heel is off the ground during walking.
- Crouch gait: The ankle is excessively dorsiflexed throughout the stance, and the knee and hip are excessively flexed.

B. DATASET

The data used for this study are retrospective gait analysis data from an existing clinical database maintained by the Laboratory for Gait and Movement Analysis at the Orthopaedic Hospital Vienna-Speising. Gait analysis data are, briefly described, obtained by motion capturing techniques where spherical retroreflective markers with a diameter of approximately 1 cm are placed directly on the patient's skin above anatomical landmarks. Then the patient is asked to walk freely up and down a walkway of roughly ten meters in a gait laboratory. A motion capture system comprising several infrared-based cameras then records the 2D trajectories of each reflective marker for each camera. These redundant 2D coordinates are then triangulated to derive the 3D coordinates in space for each marker [49] at any instant of time. The obtained marker positions are then used to fit a multibody biomechanical model into these 3D trajectories by a least square algorithm. The model then allows describing kinematic and kinetic variables of human locomotion in detail [4].

The local ethics committee approved this retrospective study (EK 19-083-VK). The dataset comprises anonymized data from 302 patients with CP (375 affected legs) and includes the aforementioned four gait patterns: true equinus ($N = 129$), jump gait ($N = 72$), apparent equinus ($N = 92$), and crouch gait ($N = 82$). Table 1 presents class-specific demographic details. The 3D clinical gait analysis was performed on a 12 m walkway using a motion capture system (150 Hz, Vicon, Oxford, United Kingdom) comprising at least 14 infrared cameras and three force plates (1500 Hz, Advanced Mechanical Technology Inc., MA, USA). The force plates were embedded in the ground flush with the walkway and covered with the same surface material as the floor. Patients walked unassisted (without a walking aid) and at self-selected walking speed until at least five valid recordings had been obtained. A record was



gaitXplorer

New Patients

ID	Left Side	Right Side
100143	TrueEquinus	TrueEquinus
100223	ApparentEquinus	ApparentEquinus
101408	Jump	Jump
101564	Crouch	Crouch
101916	Jump	TrueEquinus
103037	ApparentEquinus	TrueEquinus
103131	Crouch	Crouch
108185	TrueEquinus	ApparentEquinus

Configuration

Mode ○ standard report ○ compare to patient group ● explain relevance to classification

Group Jump

Y-Axis ● use standard value range ○ zoom to extent of patient's values

Trials ● hide trials ○ show trials

Patient: 101564

Name: 101564_10091

Analysis Date: 2011-05-30

Walking Speed: 0.85 m/s

Classification Left: Crouch

Classification Right: Crouch

FIGURE 2. Visual interface of the gaitXplorer [45] showing the classification prediction and corresponding explanations for both legs of a patient. The top right corner (a) features a compact overview of the Grad-CAM-based explanations. The main panel (b) illustrates the patient's 3D gait analysis data as line plots, with color intensity indicating the relevance for the predictions (i.e., blue for the left and red for the right leg). Figure adapted from [46].

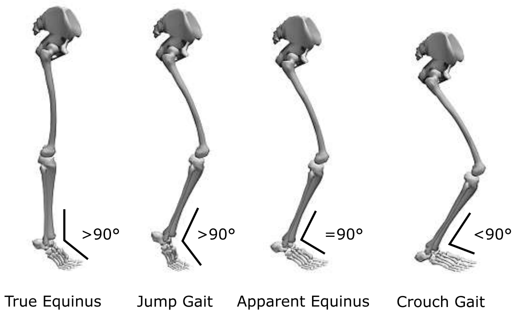


FIGURE 3. Four motion patterns in patients with cerebral palsy, i.e., true equinus (toe-walking), jump gait, apparent equinus, and crouch gait [47]. One indicator for these four gait patterns is the sagittal ankle angle (dorsi-plantarflexion). The typical value range of this angle is displayed for each gait pattern.

considered valid if the patient walked naturally and had a clean foot strike on one of the force plates. The raw data were preprocessed with Vicon Nexus (Vicon, Oxford, United Kingdom) and custom-made Matlab routines (The MathWorks, Inc., Matrick, MA, USA). Marker trajectories were filtered with a Woltring filter (mean square error of 15 mm^2) and GRF data with a third-order Savitzky-Golay filter. Joint angles and moments were calculated according to the modified Cleveland clinical marker set. Based on distinct gait events, i.e., initial contact and foot off, data of all valid gait cycles were linearly time normalized to 100 % of the respective gait cycle. Subsequently, the average curve was computed for each joint angle, joint moment, and GRF component by aggregating data from all gait cycles within one recording session.

The dataset includes information about the joint angles (kinematics) of the pelvis, hip, knee, and ankle and GRFs in all three planes of motion. For the gait kinematics, the *sagittal*, *frontal*, and *transverse* plane of motion correspond to the flexion/extension, abduction/adduction, and internal/external rotation of a joint, respectively. Consistent with standard practice in this domain, data were time-normalized to one gait cycle. As a result, each signal has 101 data samples after time-normalization, i.e., corresponding to 0–100 % of the gait cycle for joint angles and stand phase for GRFs. The data are multi-dimensional and consist of 13 signals in total, i.e., vertical (GRF_V), anterior-posterior (GRF_{AP}), and medio-lateral (GRF_{ML}) GRFs as well as sagittal ($Pelvis_S$), frontal ($Pelvis_F$), transversal ($Pelvis_T$) pelvis angles, sagittal (Hip_S), frontal (Hip_F), transversal (Hip_T) hip angles, sagittal ($Knee_S$), frontal ($Knee_F$), transversal ($Knee_T$) knee angles, and sagittal ankle angle ($Ankle_S$). Each signal represents either one of the GRF components or the kinematic profile at a particular joint and anatomical plane during one gait cycle. Several gait cycles were available for each patient. The signals from these gait cycles were averaged to one waveform per body side to account for intra-subject gait variability. The classification was conducted at the level of individual legs, i.e., only the affected legs were classified.

TABLE 1. Demographic information for each class within the employed dataset.

Class	Num. Patients	Num. Limbs	Sex (m/f)	Age (yrs.)	Body Mass (kg)	Height (m)	Speed (m/s)
Crouch Gait	61	82	29/32	14.7 (7.3)	43.4 (17.2)	1.5 (0.3)	1.0 (0.2)
Apparent Equinus	77	92	46/31	14.4 (8.2)	45.7 (19.1)	1.5 (0.2)	1.0 (0.2)
Jump Gait	61	72	46/15	13.6 (6.9)	43.6 (19.7)	1.5 (0.2)	1.1 (0.2)
True Equinus	103	129	61/42	12.0 (9.2)	36.5 (17.6)	1.4 (0.3)	1.1 (0.2)
TOTAL	302	375	182/120	13.7 (7.9)	42.3 (18.4)	1.5 (0.3)	1.1 (0.2)

For our experiments we employed different subsets of the captured data to investigate the influence of different signals on the classification performance and on the obtained explanations. For each subset we concatenated the respective signals into a one-dimensional vector. We defined the following four subsets for our experiments:

- the lower body kinematic and GRF data (i.e., a 1×1313-dimensional input vector for each patient, consisting of 13 signals, with 101 samples each);
- only lower body kinematic data (a 1×1010-dimensional input vector for each patient);
- only GRF data (a 1×303-dimensional input vector for each patient);
- the sagittal knee and ankle joint angles which are the signals that are actually used to determine the ground truth (a 1×202-dimensional input vector for each patient).

Since the signal amplitudes differ in their dynamic ranges, we normalized the input features component-wise to the range [0,1]. Hence, it is ensured that each signal can contribute equally to the decision process and signals with a smaller amplitude range are not disadvantaged.

C. CLASSIFICATION METHODS

For the classification task, we examined various ML models, including CNNs, SNNs, RFs, and DTs, and generated and compared explanations for the trained models. We selected CNNs because they have not been previously employed in the literature for this purpose, despite their success in other gait analysis tasks [8]. SNNs utilize scaled exponential linear units (SELUs) as activation function, which exhibit self-normalizing properties, causing the output of the layers to converge to zero mean and unit variance [12]. Klambauer et al. [12] demonstrated that SNNs exhibit great robustness due to these properties, since vanishing and exploding gradients are eliminated by construction. As traditional ML methods, RFs and DTs performed well for classifying CP-related gait patterns, and thus we utilized them as baseline approaches. We evaluated the ML models in a stratified five-fold cross-validation approach. Hence, three folds served as training data, one fold served as a validation set on which the optimal architecture and hyperparameters were determined, and the remaining fold served as a test set.

1) NEURAL NETWORKS

CNNs and SNNs learn abstract feature representations for the provided data via several consecutive 1D convolutional layers. The filter size and stride[1] remained fixed for all convolutional layers. We investigated whether compressing information across the convolutional stack provides an advantage in both performance and explainability. To this end, we examined a stride of one (baseline without compression) and two (compression by half). For both model types, CNNs and SNNs, the filter size was set to three. Simonyan and Zisserman [50] showed the advantage of using a stack of 3 × 3 convolutional layers over filters with larger receptive fields in image classifiers. This approach employs multiple non-linearities, resulting in a more discriminative decision function as well as a reduction in the number of parameters [50]. Non-linear neuron activations in terms of ReLUs for CNNs and SeLUs for SNNs were applied in each convolutional layer. The feature maps in the last convolutional layer were flattened and linked to a fully-connected (dense) layer stack. This stack consists of one dense layer (with ReLU for CNNs and SeLU for SNNs as an activation function) and an output layer situated on top that has four output neurons. To promote generalizability and counteract potential overfitting during training, a dropout was applied to the last two dense layers (including the output layer). The output layer has a softmax activation function attached to scale the outputs to class likelihoods. The fully-connected layers (including the output layer) can be considered a non-linear multi-class predictor, which operates on top of a hierarchically learned stack of 1D filters. The convolutional layer stack is strongly non-linear, which makes this part of the architecture very flexible in modeling but at the same time non-transparent.

During the training process, the weights were updated via back-propagation using the Adam optimizer (1000 training epochs with early stopping using the validation loss as the monitored metric and a patience of 100 epochs) and a categorical cross-entropy loss function. For each input setting, we determined the optimal hyperparameters via a grid search, i.e., stride $\{1, 2\}$, number of convolutional layers and number of filters $\{\{32, 32\}, \{32, 32, 32, 32\}, \{32, 32, 32, 32, 32, 32\}, \{32, 64\}, \{32, 32, 64, 64\}\}$, size of the dense layer $\{64, 128\}$, dropout rate $\{0.1, 0.25, 0.5\}$, batch size $\{32, 64\}$, and learning rate $\{10^{-4}, 10^{-3}\}$. The optimal hyperparameters for the CNNs are presented in Table 2 and for the SNNs in Table 3.

[1] The number of input features that the convolution filter moves across the input of the convolutional layer.

TABLE 2. Optimal hyperparameters, i.e., stride (S), number of layers and filters in the convolution stack, size of dense layer, dropout rate (DR), batch size (BS), and learning rate (LR), for the CNN architectures.

Input Signals	S	#Filters	Dense	DR	BS	LR
All	1	32, 64	128	0.1	64	10^{-4}
All	2	2x32, 2x64	64	0.1	64	10^{-4}
Kinematics	1	2x32	128	0.25	32	10^{-4}
Kinematics	2	2x32	64	0.25	32	10^{-4}
GRF	1	2x32	128	0.1	32	10^{-3}
GRF	2	2x32	64	0.25	32	10^{-3}
Ankles & Knees	1	2x32	64	0.1	32	10^{-4}
Ankles & Knees	2	4x32	128	0.25	64	10^{-3}

TABLE 3. Optimal hyperparameters, i.e., stride (S), number of layers and filters in the convolution stack, size of dense layer, dropout rate (DR), batch size (BS), and learning rate (LR), for the SNN architectures.

Input Signals	S	#Filters	Dense	DR	BS	LR
All	1	2x32	128	0.1	64	10^{-3}
All	2	2x32, 2x64	128	0.1	64	10^{-3}
Kinematics	1	6x32	128	0.1	64	10^{-3}
Kinematics	2	6x32	128	0.1	32	10^{-4}
GRF	1	32, 64	64	0.1	64	10^{-3}
GRF	2	32, 64	128	0.1	32	10^{-3}
Ankles & Knees	1	2x32, 2x64	64	0.1	64	10^{-3}
Ankles & Knees	2	2x32, 2x64	64	0.1	64	10^{-3}

2) TREE-BASED MODELS

In addition to DNNs, we explored traditional ML methods such as DTs and RFs. The non-parametric, non-linear, and intrinsically interpretable nature of DTs makes them popular for gait classification [6]. DTs have a tree structure containing decision nodes and leaf nodes. A decision node uses a suitable input feature to try to split the data into two homogeneous subsets. To determine which feature is suitable for a decision node, different metrics can be used. The most commonly used metrics are *Gini impurity* and *information gain* based on entropy. These metrics can also be used to calculate the importance of each input feature for the entire model, which can be used as a model explanation. Since a feature can be used at different levels of the tree, the importance of a feature is determined as the total contribution in reducing the impurity.

Different algorithms exist for the construction of a DT, e.g., ID3 [51], C4.5 [52], and CART [53]. For our experiments, we utilized the DT implementation of Scikit-learn [54], which is based on the CART algorithm. To construct the DT, we used Gini impurity and the feature importance based on this metric serves as explanation method.

Since individual DTs can be sensitive to even small changes in the input data, we also investigated RFs. RFs are also supervised non-parametric ML methods built on a set of simple DTs. To generate an RF, a predetermined number of DTs are first trained on different subsets of the training data, and then the predictions of these simpler models are combined. As RFs are sensitive to the number of individual DTs, we performed a grid search over this hyperparameter $N_{DT} \in \{100, 200, 300\}$. The number of individual DTs that performed best in all settings was 100. For our experiments, we utilized the RF implementation of Scikit-learn [54] with Gini impurity as metric. Similar to an individual DT, the feature importance can be calculated for the entire RF using Gini impurity. This can directly serve as an explanation for the trained model.

D. EXPLAINABILITY METHODS FOR NEURAL NETWORKS

Once the networks were trained, a central question was how to explain these models to examine their internal functioning and plausibility. With regard to DNNs, the ever-growing ecosystem of XAI methods offers many choices. However, as demonstrated by Adebayo et al. [14], not all of the proposed XAI methods are robust and the validity of obtained explanations should be questioned. Unfortunately, most of the popular gradient-based methods, which are particularly well-suited for DNNs, are heavily exposed to artifacts caused by the problem of gradient shattering [55]. Thus, the explanations in the input space are not continuous, and single input features show considerably different or even opposite importance values (regarding a given prediction) compared to input features in their immediate neighborhood [34]. Furthermore, it is questionable whether the information obtained for individual input features represents an adequate abstraction level to explain a decision, especially when the input signal is a continuous time series. Justifications based on local and consecutive signal features in the time series may lead to more comprehensible and intuitive explanations. For this reason, we decided to perform the explanation at a higher semantic level. A suitable level is the last layer of the convolutional filter stack. This layer represents higher-level signal filters with a larger receptive field and, thus, potentially captures more meaningful signal features for human observers.

A method that provides explanations at this level is Grad-CAM [13]. This method does not propagate the gradients back to the input space, but the final prediction of the network is directly explained in terms of the abstract features learned in the last convolutional layer. Grad-CAM weighs the activation map of the last convolutional layer with the gradients (which flow into this layer) with respect to the target class to be explained. The weighted activation map is averaged over all channels of the layer. This results in an activation pattern that reflects higher-level signal patterns and captures contextual information. For easier interpretation of the results, the activation pattern can be upscaled (via interpolation) and mapped (e.g., via color coding) to the input signal. The upscaled activation pattern highlights continuous but local sections in the input signal that have a strong relation to the target class under investigation. An overview of how Grad-CAM functions for 1D gait analysis data is provided in Figure 4.

In our experimental setup, we employed a five-fold cross-validation, which results in five distinct models. We decided to explain the model that performed closest to the median

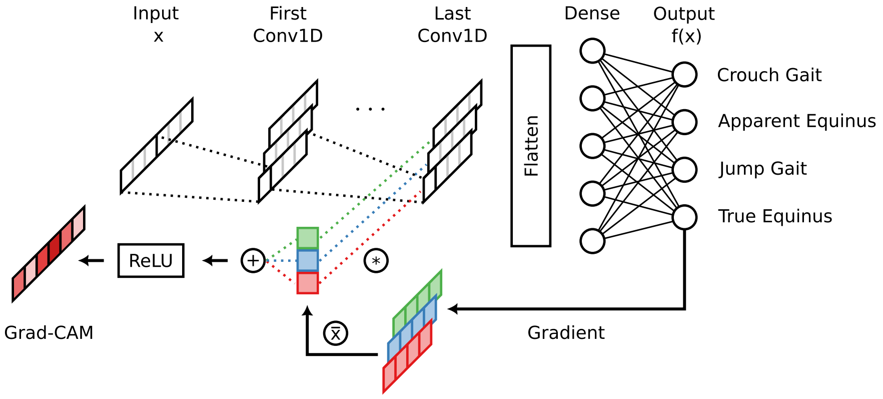


FIGURE 4. A schematic representation of the Grad-CAM method adapted for deep neural networks trained on 1D gait data. This example illustrates the process of generating a decision explanation for the true equinus class. To generate a Grad-CAM explanation, the gradients of the feature maps of the last convolutional layer are averaged channel-wise ($\bar{x}$) and used as weights to calculate a weighted sum of the activations of this layer. Then, a ReLU function is applied to obtain only positive values, because these contribute to the prediction of the specific target class. In a final step, the Grad-CAM explanation is scaled up to match the size of the input.

among the five models because we intended to emulate real-life scenarios where usually a single model is used in practice rather than multiple models. For this model, we generated explanations at different levels, i.e., at the decision level and subsequently by aggregating these explanations at the class and model level. The motivation for this approach is to gain a comprehensive understanding of the model by explaining not only a single decision, but also which features are important for each class and the overall model.

First, we computed *decision explanations* for each record in the dataset (training, validation, and test samples) using the ground truth label as target for the explanation. Next, we calculated the median over the Grad-CAM activations of all records assigned to a particular class to obtain *explanations on the class level*. The obtained activation pattern highlights class-specific patterns for an entire target class. The calculation of the median, however, can cause interpretation difficulties by obscuring the existence of different decision strategies that the model may have learned for different patient subgroups within a class. As shown in our previous research [8], CNNs have the ability to learn different strategies for distinct patient subgroups. Consequently, if a model learns complementary strategies to distinguish different patient subgroups of a particular class, the median plot of that class may not provide sufficient information to understand the model's functioning. Therefore, similar to individual conditional expectation (ICE) [56] plots, we propose an additional subplot that also visualizes the individual Grad-CAM activations.

Furthermore, we propose an *explanation at the model level* that goes beyond the interpretation of individual classes. For this type of model explanation, we calculated the total relevance as the sum of Grad-CAM activations over all samples for all four classes. This model explanation should serve as an informative indicator for the overall relevance of an input feature for the underlying classification task.

The implementation of all classification and explainability methods was conducted within the software framework Python 3.7.10 (Python Software Foundation, USA), TensorFlow 2.3.0 (Google Brain Team, Google LLC, USA), and Scikit-learn 1.0.2 [54].

IV. RESULTS

Subsection IV-A presents the quantitative results in terms of classification accuracy for all investigated ML models which were used to classify the four CP-related patterns presented in Subsection III-A. The explainability results for the examined ML models, which aim to explain the functioning of the models on the class and model level, are presented in Subsection IV-B.

A. CLASSIFICATION RESULTS

Classification results are provided for the four classification methods from Section III-C, i.e., CNN (with stride of one and two), SNN (with stride of one and two), RF, and DT. Each model has been trained and evaluated on the four different input configurations (signal sub-selections) defined in Section III-B, i.e., i) all 3DGA signals, ii) only kinematic signals, iii) only sagittal knee and ankle angles, and iv) only GRF signals. The zero rule baseline (ZRB), which refers to the theoretical accuracy obtained by assigning always the class label with the highest prior probability, is 34.4 % for this classification task. Since we evaluated the ML models via stratified five-fold cross-validation, we report the averaged classification accuracy over all five folds for the training, validation and test set. Table 4 summarizes all quantitative results.

In addition to the presented methods, we also conducted experiments with linear support vector machines (SVMs) and gradient boosting classifiers to evaluate their potential. For the SVMs, we performed a grid search on the hyperparameter C = $\{10^{-4}, 10^{-3}, \ldots, 10^{3}, 10^{4}\}$. The results of the SVMs (All: 75.5%, Kinematics: 72.3%, and *Ankles$_S$* & *Knees$_S$*: 78.8%) showed suboptimal performance for the employed dataset and classification task and thus the classifier was not further investigated. We also evaluated the performance of a gradient boosting classifier using a grid search on the hyperparameters: learning rate $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ and number of trees {100, 200, 300}. Since RF outperformed gradient boosting (All: 91.4%, Kinematics: 91.2%, and *Ankles$_S$* & *Knees$_S$*: 92.0%), we focus on the RF results in the following.

The results in Table 4 show that all models overfitted on the training data, as evidenced by their significantly higher performance on the training set compared to the test set, even though the optimal parameters were selected using a validation split. RF outperformed all other models, with CNNs and SNNs showing the lowest test accuracies. The performance of DTs lies between that of the DNNs and RFs. RFs consistently achieved peak performance for all three input configurations where kinematic data were present. RFs are outperformed by CNNs and SNNs only when exclusively GRF data are used.

CNNs performed slightly better than SNNs across all input configurations. For CNNs and SNNs, there was little difference in performance between using a stride of one and two. There is no consistent trend in performance with respect to stride. Only in the first input configuration where all 3DGA signals were used, a stride of one performed slightly better.

For CNNs and SNNs, reducing the data to signals that are most relevant for the classification task, i.e., sagittal ankle (*Ankles$_S$*) and knee (*Knees$_S$*) angles, showed a significant advantage. The pre-selection of input signals seems to help the networks to find the most relevant information for solving the task. This effect can also be observed with DTs, but it is less pronounced. Remarkably, RFs demonstrate a high degree of invariance towards the pre-selection of input signals. For all input configurations where sagittal ankle and knee angles are included, RFs achieved a similarly high performance level independent of input dimensionality. This shows that RFs are very good at identifying the most relevant information and are hardly distracted by information in unrelated signals.

Furthermore, the results in Table 4 allow to compare the classification performance achievable with GRF and kinematic data. This is an important question for clinical practice, as the acquisition of 3D data is much more demanding than capturing GRF data via force plates. Our results show that 3DGA data (i.e., kinematic data) are essential for automated gait classification. We observed a significant drop in performance when classification is restricted to the use of GRF data. We further discuss the results as well as their relevance in the context of the driving research questions of our study in Section V.

TABLE 4. Classification accuracy (averaged over all five folds, in %) for the four classification methods (CNN, SNN, RF, and DT) and two variations of CNN and SNN (each with different stride S) and different input signal selections.

Signals	Model	Train Acc.	Val. Acc.	Test Acc.
All	CNN (S=1)	96.9	83.1	79.8
All	CNN (S=2)	94.8	83.1	76.3
All	SNN (S=1)	91.5	83.2	78.9
All	SNN (S=2)	91.8	82.4	76.0
All	RF	100.0	88.7	*92.9*
All	DT	100.0	87.4	84.8
Kinematics	CNN (S=1)	97.9	82.5	80.0
Kinematics	CNN (S=2)	97.7	82.5	81.1
Kinematics	SNN (S=1)	92.5	84.6	78.3
Kinematics	SNN (S=2)	95.1	80.4	76.9
Kinematics	RF	100.0	90.2	**93.4**
Kinematics	DT	100.0	85.3	86.6
Ankles$_S$ & Knees$_S$	CNN (S=1)	95.6	88.1	86.6
Ankles$_S$ & Knees$_S$	CNN (S=2)	95.5	90.8	86.6
Ankles$_S$ & Knees$_S$	SNN (S=1)	94.3	93.7	85.7
Ankles$_S$ & Knees$_S$	SNN (S=2)	94.3	93.7	85.7
Ankles$_S$ & Knees$_S$	RF	100.0	89.5	**93.4**
Ankles$_S$ & Knees$_S$	DT	100.0	86.7	89.7
GRF	CNN (S=1)	64.0	52.0	*47.2*
GRF	CNN (S=2)	66.0	54.9	46.6
GRF	SNN (S=1)	62.8	52.8	44.7
GRF	SNN (S=2)	58.7	50.1	44.3
GRF	RF	100.0	45.0	42.0
GRF	DT	100.0	34.6	37.1

B. EXPLAINABILITY RESULTS

In the following, we present the explainability results of the investigated classification models. We start with CNNs and show their explanations at the class level (Figure 5) and model level (Figure 6). Subsequently, we compare them with the explanations of DTs and RFs at the model level.

1) EXPLANATIONS OF NEURAL NETWORKS

Here we focus on the CNN with a stride of two, since the explanations generated by this model were found to be the most satisfactory compared to the other DNNs. In addition, we focus on the scenario in which the kinematic signals were utilized as input data, as these are the signals most commonly considered for CP. Figure 5 shows the respective explainability results on the class level. The results are presented in four panels, each showing the results of a particular gait class. The top of each panel shows the averaged input signals (in this case 10 concatenated kinematic signals) per class. These are colored with a sequential red palette based on the median of the Grad-CAM relevances of all samples assigned to the particular class. The higher the degree of red coloration of a region in the averaged input signal, the greater its relevance is to the corresponding class. The bottom part of each panel shows the Grad-CAM explanations for individual input samples, with the median visualized in blue. The lower part of each panel provides a more comprehensive overview by showing the distribution of individual decision explanations, from which we can derive additional information. For instance, considering the $Hips_S$ signal in the jump gait class in Figure 5C, the lower visualization reveals at least two strategies the model learned for classifying this class, i.e., one focuses on the central regions of $Hips_S$, while the other focuses on the regions at the beginning and the end of $Hips_S$.

In general, the explanations on the class level show different relevance patterns for the different classes. Very similar regions were considered relevant for crouch gait and apparent equinus. The relevant regions for these two classes primarily reside during the stance phase (approximately the first 60 % of the signal) of $Ankles_S$ (markers b and d in Figure 5), the swing phase (approximately the last 40 % of the signal) of $Knees_S$ (markers a and c in Figure 5), and the beginning and the end of $Hips_S$. In addition to these regions, for the apparent equinus class, the other signals also exhibit moderate relevance (e.g., most prominent in Hip_F, $Pelvis_F$, and $Knee_F$). The jump gait class shares some of the relevant regions (i.e., the swing phase of $Knees_S$ (marker e in Figure 5), as well as the start and end of $Hips_S$) with the two classes described previously. However, there are certain differences, especially to crouch gait, such as the moderate relevance in $Pelvis_S$, $Hips_S$, Hip_T, $Knee_F$, and $Knee_T$ as well as for some samples the stand phase in $Hips_S$ is highly relevant, whereas for others it is not. The true equinus class is clearly dominated by relevant regions during the stance phase in $Hips_S$ and $Knees_S$ (markers f and g in Figure 5). This relevance pattern is clearly different from all other classes.

The highlighted regions in Figure 5 represent class-specific activations and do not have to be discriminative for the classification task per se. To identify which regions are most relevant for the overall classification task we calculate the total relevance as the sum of Grad-CAM activations of all samples over all four classes. The higher this overall activation, the more relevant is a given signal portion for the classification task. Figure 6 shows the min-max normalized overall activation as blue lines. Figure 6A shows the mean and standard deviation of the raw input signals (calculated per class) as solid and dashed lines, respectively. Figure 6B-E shows the overall activation for models trained using the four different input configurations (Subsection IV-A). Figure 6B-E shows the Gini impurity-based feature importance for the RFs and DTs in orange and red, respectively.

As with the class level explanations, we observed very high relevances for the CNN in the signals $Hips_S$, $Knees_S$, and $Ankles_S$ in the corresponding model explanation in Figure 6C. This is independent of whether GRF signals are used (Figure 6B) or not (Figure 6C). In case GRF is used in addition, the propulsion peak in GRF_{AP} (marker d in Figure 6) is very relevant for the classification task.

When using only the signals that are most relevant to clinicians for classification (Figure 6D), we can see that i) for $Knees_S$ the relevance shifts more to the stance phase (marker h in Figure 6D) while decreasing for the swing phase (marker i in Figure 6D) and ii) for $Ankles_S$ the relevance shifts to the swing phase (marker j in Figure 6D) while decreasing overall.

Similar regions exhibit high relevance when only GRF data are utilized (Figure 6E) as in the case when all 3DGA signals are used (Figure 6B).

2) EXPLANATIONS OF TREE-BASED MODELS

Figure 6 shows relevance scores (Gini impurity-based feature relevances) for both DTs (red) and RFs (orange) at the model level. RFs and DTs both exhibit locally similar regions that are considered highly relevant. The DT places a high emphasis on a single input feature, while the RF distributes the relevance to a more widespread area, which is strongly related to the relevant features of DT. For the first three input configurations (Figure 6B-D) there is a strong correspondence between these regions, showing that both DT and RF find the most relevant information for the classification task in a highly targeted manner. The identified regions further correlate with those of the CNN (except for $Ankles_S$ in Figure 6D where the CNN activation is shifted towards the swing phase). Interestingly, RF and DT provide explanations which are much more focused on the clinically relevant signals, compared to the CNNs, for which the activations are distributed across a broad range of input signals.

Finally, we want to point out that for the case where only GRF data are used (Figure 6E), a very noisy pattern is observed in the relevances of RF and DT, focusing mainly on the beginnings of the signals, which is in contrast to the regions relevant to the CNN. We assume that the low expressiveness of the GRF signals for the CP-related gait patterns is the reason for the unfocused activation patterns.

V. DISCUSSION

In the following, we analyze and interpret the classification and explainability results from a technical and clinical perspective. Additionally, we discuss the influence of different input configurations on the performance of the investigated classification methods. We structurally organize this section according to the research questions and provide answers to each of them.

1) HOW ADVANTAGEOUS IS THE USE OF KINEMATIC DATA OVER GRF DATA IN THE AUTOMATED CLASSIFICATION OF GAIT PATTERNS ASSOCIATED WITH CP, AND ARE THE TWO INPUTS MORE EFFECTIVE IN COMBINATION THAN USED INDIVIDUALLY?

The classification results in Table 4 demonstrate a significant difference in classification performance between the use of kinematic data and GRF data as input in all experiments. The absolute differences range from 32.6 % for the SNN (stride of two) to 51.4 % for the RF. The results on GRF data also show that DNNs are more effective than traditional ML approaches in modeling this type of data (the absolute difference between CNN and RF is 5.2 %). The model level explanation provides a possible rationale for this observation: The CNNs use very similar regions in the GRF signals for both input conditions, which is not the case for traditional

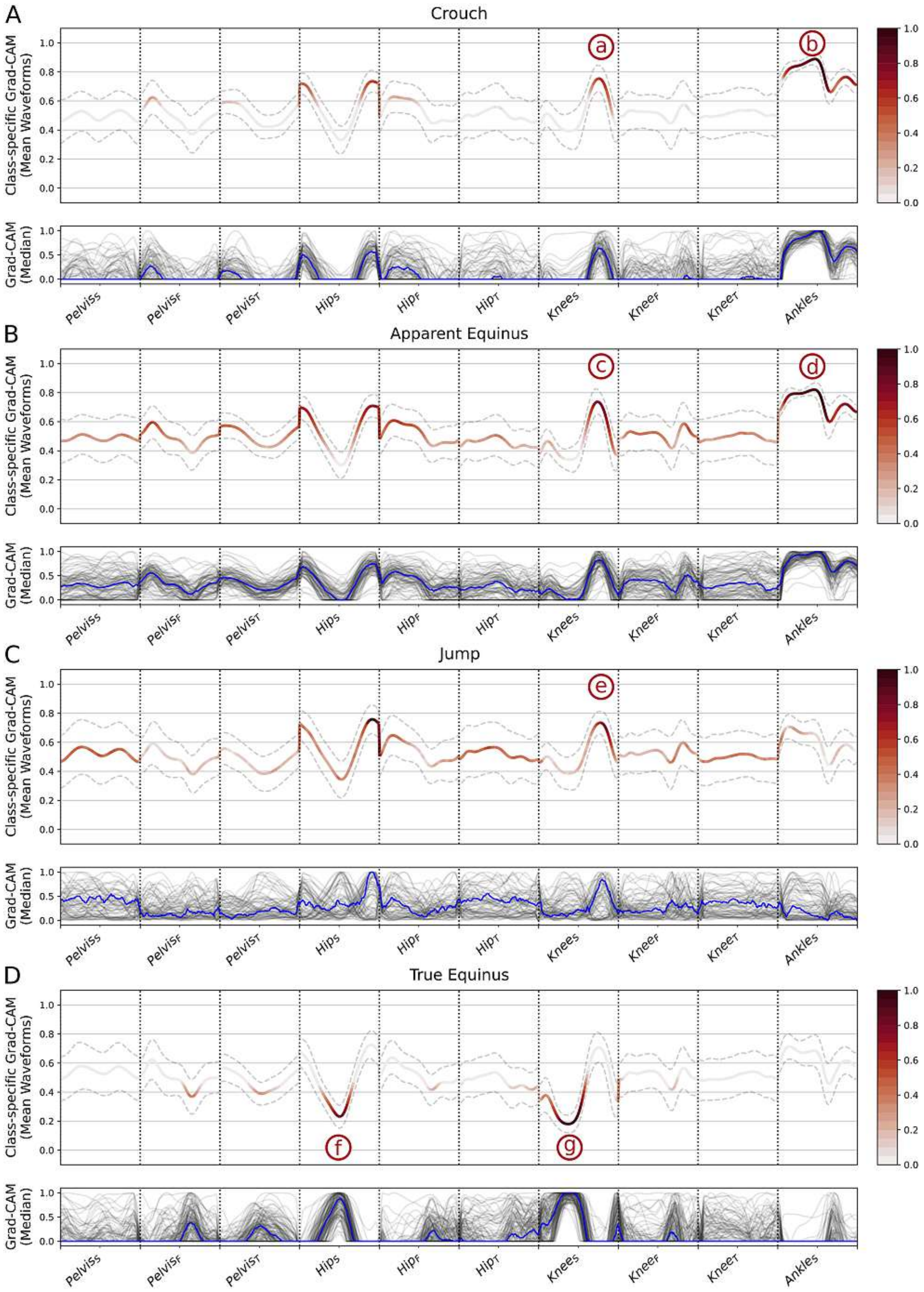


FIGURE 5. Explainability results on the class level for the classification of gait patterns associated with cerebral palsy based on min-max normalized kinematic data using a CNN (with stride of two). The results are shown in four panels, each showing the results for a pathological gait pattern. The top of each panel displays the class-averaged kinematic signals that are colored with a sequential red palette, based on the median Grad-CAM relevance for that class (i.e., the redder, the more relevant). The bottom of each panel shows the Grad-CAM relevances for individual samples, with the median plotted in blue.

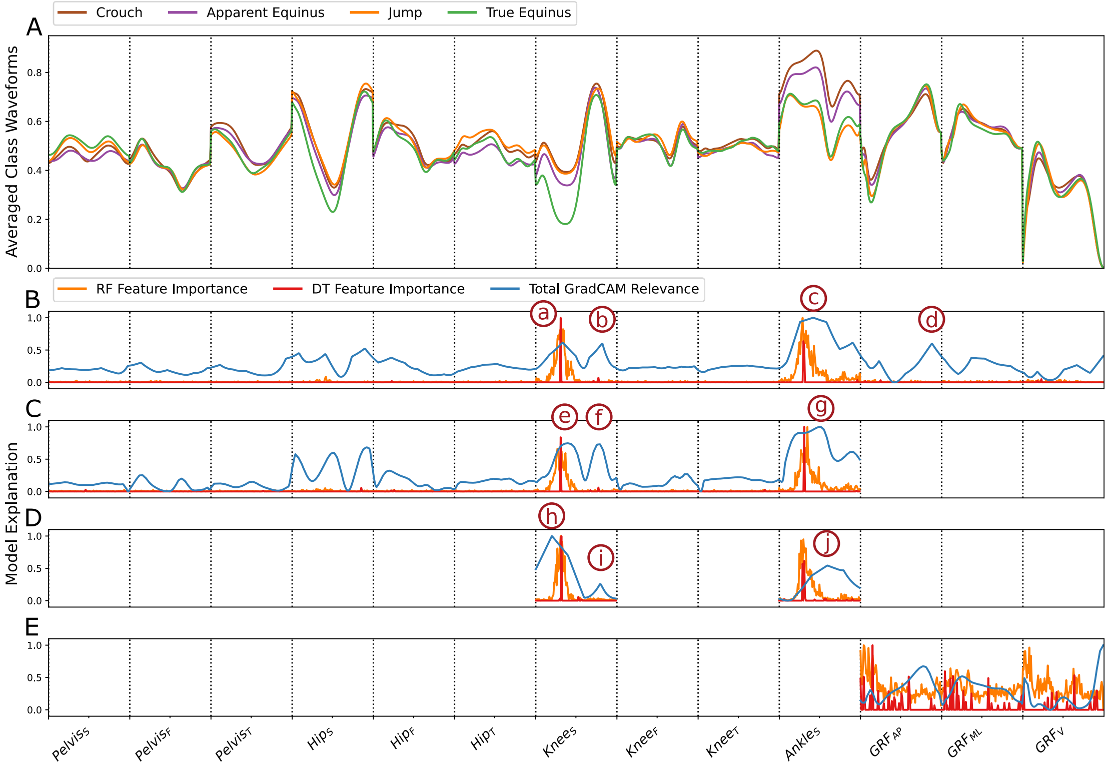


FIGURE 6. Explainability results on the model level for the classification of gait patterns associated with cerebral palsy based on min-max normalized kinematic data using a CNN (stride of two), a DT, and an RF. The results are presented in five subfigures (A-E). Subfigure A) displays the mean and standard deviation of the raw input signals as solid and dashed lines, respectively. The following four subfigures show the model explanations in terms of relevance for models trained on four different input configurations: B) all signals from 3DGA, C) only kinematic signals, D) only sagittal knee and ankle angles, and F) only ground reaction forces. The model explanations (relevances) for the CNN (blue) are calculated by adding the Grad-CAM relevances for all input samples and then applying min-max normalization. For DT and RF we provide the Gini impurity-based feature relevances in orange (RF) and red (DT).

ML approaches. For illustration, refer to Figure 6B and E, where the relevance of the CNNs (blue curves) are similar in both subfigures (apart from the high relevance at the end of GRF_V, which is not particularly meaningful from a clinical perspective). The distribution of feature importance for RF and DT shown in Figure 6E is highly scattered and noisy, and there is also a lack of agreement between the two models. This suggests that RF and DT exhibit difficulties in learning the most important input features from GRF data.

When considering kinematic data, using only $Ankle_S$ and $Knee_S$ for the classification task results in a significant improvement in performance. Given that these signals are considered by the clinicians to be most relevant to the classification task, these improvements in performance are not surprising. Our experiments show that these two signals are also the most important and useful ones for the ML models. Using all kinematic signals as input slightly decreases performance (except for RF), which indicates that (i) the models are distracted to a certain degree by the additional input and (ii) the other signals do not contribute additional information to the task.

Combining kinematic and GRF data does not provide any advantage and leads to a slight degradation in performance in the majority of cases. This suggests that there is no complementary information in the GRF signals compared to the kinematic data for the evaluated task.

In the literature, GRF data have been utilized in multiple studies examining pathological gait patterns associated with Parkinson's disease [18], [57], [58], cerebral palsy [19], multiple sclerosis [19], osteoarthritis [59], transfemoral amputation [60], and lower limb fracture [61]. However, notable success in classification has been achieved primarily for the relatively simple classification tasks of distinguishing between one or two pathological gait patterns and healthy controls. Furthermore, the majority of previous research employed relatively small datasets. For the few studies that addressed more complex research questions, such as the classification of various functional gait disorders [8], [21], [22], the exclusive use of GRF data has yielded less promising results. This tendency is also evident in our results. The lower body motion information aggregated in GRF signals is (i) insufficient when used alone and (ii) does not contribute complementary information for the classification of gait patterns associated with CP.

2) HOW DO TRADITIONAL ML MODELS COMPARE TO STATE-OF-THE-ART DNNS FOR THE AUTOMATED CLASSIFICATION OF CLINICAL 3DGA DATA IN TERMS OF PERFORMANCE AND EXPLAINABILITY?

For the given task and data, RFs performed significantly better compared to the other ML methods including DNNs. When analyzing all input scenarios, RFs have consistently shown the best performance ranging from 92.9 % to 93.4 %, except for the scenarios involving only GRF signals. DTs demonstrated the second-best performance in almost all input scenarios, with the exception of scenarios involving only GRF signals. One explanation for the superior performance of the traditional ML methods can be attributed to the good generalization ability of these methods to a small number of training samples. This is not the case for CNNs and SNNs, which have a strong tendency to overfit on smaller datasets.

In comparison with related work focusing on the classification of the four CP-related gait patterns as defined by Rodda et al. [27], we achieved similar classification performance with the traditional ML models. Zhang and Ma [26] reported that MLPs achieved the highest classification accuracy of 93.5%, while DTs and RFs achieved lower accuracy rates of 84.3% and 83.6%, respectively, using a dataset of 200 children and the four classes. Similarly, Darbandi et al. [28] achieved a classification accuracy of 94.0% with their stochastic approach, using a dataset of 66 children and the four classes. In our study, utilizing a significantly larger dataset of 302 children, RFs demonstrated the highest performance.

The explainability results show that DNNs attempt to learn features from a broad range of signals for classification (e.g., Figure 6B shows high relevances for $Pelvis_S$, $Knee_S$, $Ankle_S$, and GRF_{AP}, while other signals are also considered relevant to a certain degree). A cause for this behaviour can be the limited dataset size. In contrast, RFs and DTs focus much more on the regions in $Knee_S$ and $Ankle_S$ that are actually relevant. We assume that their lower complexity in terms of numbers of parameters compared to DNNs is beneficial for the task and dataset. Interestingly, the feature importance for DT and RF is consistent for all input configurations (Figure 6B-D), which confirms that both models are not distracted by additional (obviously mostly unrelated) input signals. The feature relevance for the DNNs is more sensitive and varies stronger between the different input configurations.

3) TO WHAT EXTENT DO THE INVESTIGATED ML MODELS BASE THEIR DECISIONS ON CLINICALLY MEANINGFUL FEATURES WHEN CLASSIFYING CP-RELATED GAIT PATTERNS?

In clinical practice, the four investigated CP-related gait patterns mainly differ in the sagittal knee ($Knee_S$) and sagittal ankle ($Ankle_S$) angles during the stance phase. The model explanations for the three input configurations with kinematic data show the highest relevance in these signals (markers a/e/h and c/g/j in Figure 6). This matches expectations from clinical practice and is in agreement with other studies [48] which identified both signals as the most promising to distinguish crouch gait, apparent equinus, jump gait, and true equinus.

As previously discussed, DNNs tend to learn patterns from a broad range of input signals, in contrast to RFs and DTs, which focus only on the most clinically relevant signals, i.e., $Knee_S$ and $Ankle_S$. Considering the case where all kinematic data are used (Figure 6C), the CNN shows the highest activations in $Knee_S$ and $Ankle_S$ as well, but also in Hip_S, which is clinically reasonable. Interestingly, although clinically relevant, Hip_S does not contribute to the classification performance in our experiments.

From a clinical point of view, the main characteristic of the gait pattern true equinus is an increased plantarflexion during stance. One could expect a strong activation in $Ankle_S$ for true equinus. However, true equinus has a plantarflexed pattern in the ankle, which is similar to jump gait. Apparent equinus (neutral ankle angle) and crouch gait (dorsiflexed ankle angle) are more similar in $Ankle_S$ compared to jump gait and true equinus. However, they also differ from each other in $Ankle_S$. Therefore, the high activation in $Ankle_S$ for crouch gait and apparent equinus (markers b and d in Figure 5) and the non-activation in this signal in jump gait and true equinus seem plausible from a clinical point of view.

In $Knee_S$ we see less activation for crouch gait and apparent equinus during stance (Figure 5A and B). This seems plausible from a clinical perspective because both gait patterns are associated with increased knee flexion during stance compared to normative data. Both jump gait and especially true equinus are gait patterns associated with a decrease in knee flexion or even hyperextension during stance. The classification algorithm clearly picked up this pattern associated with knee hyperextension as we observed increased relevance scores during stance in $Knee_S$ for both classes, i.e., for some jump gait samples, but especially for true equinus (marker g in Figure 5).

Our experiments have further revealed that the explainability approaches highlight certain signal regions as highly relevant, which may not be considered important from a clinical perspective. We observed a high activation during early to mid-swing in $Knee_S$ (markers b, f, and i in Figure 6), which is not expected from a clinical point of view. The reason might be twofold, either this attribution is due to a bias in the data (e.g., a spurious correlation with the respective classes), or it indicates a potentially useful signal region that the ML model has discovered during learning, which is not considered in clinical practice (either because it shows too subtle differences that have not been considered as clinically relevant yet, or it has not been observed in practice yet). These results demonstrate that explainability approaches have the potential to assess not only the correctness of the trained models but also to gain new clinical insights about the data and the investigated task.

Overall, we conclude that the ML models successfully learn clinically relevant patterns for the distinction of different CP-related gait patterns. All three model explanations have a high activation in the clinically most relevant signal regions, i.e., $Knees_S$ (markers a, e, and h in Figure 6) and $Ankles_S$ (markers c and g in Figure 6) during stance. In addition, the CNN also considered regions that were not expected from a clinical perspective, e.g., $Knees_S$ during swing (markers b,f, and i in Figure 6) and $Ankles_S$ during swing (marker j in Figure 6D). Explainability approaches can reveal such unexpected patterns and are essential for clinicians to verify the correct working of the model, to gain trust in its decisions and to support gaining new insights into the data. We can conclude that all employed methods primarily focus on clinically relevant input signals, whereas this pattern is much more distinct for DT and RF than for the DNN models.

4) TO WHAT EXTENT ARE THE EXPLANATIONS OBTAINED FROM DNNS ROBUST TO VARIATIONS IN ARCHITECTURE?

CNNs and SNNs show significant differences in their explanations. A visualization of the explanations on the class level for all input configurations can be found in the supplementary material in Figure S1. A direct comparison shows that SNNs lack activations for entire classes in the scenario where all signals (kinematic and GRF data) are used. This means that in some situations the Grad-CAM explanation does not highlight any signal as important, which is counter-intuitive and not credible. However, when we reduce the number of signals, we obtain more reasonable explanations. This behavior may be attributed to the high dimensionality of the input data, which potentially leads to an over-parameterization of the model and which in turn impedes Grad-CAM to identify distinct features. For CNNs (with stride of one) we also observe this problem, but only for the apparent equinus class in the configuration where all signals are used. This indicates that the normalization introduced by SNNs is not the (only) reason for this behavior. It is more likely that the high input dimensionality is responsible for the partly meaningless explanations. Further experiments with stronger regularization are needed to investigate this problem in more detail.

For the other input configurations (except for SNN with a stride of one and the kinematic data as input), there are more similarities in the explanations of CNNs and SNNs, especially for the crouch gait, apparent equinus, and true equinus classes. In general, there is also more similarity between the two CNNs with different strides except for GRF signals, where the CNN with a stride of 1 learns patterns that are more similar to those of SNNs. The two SNNs with different strides show very high similarities for two input configurations, i.e., GRF and $Ankles_S$ & $Knees_S$. We conclude that small changes in the model architecture may lead to larger variations in the explanations, which should not be the case if the models are able to robustly model the given task. The input dimensionality seems to be one factor that impedes the robustness of explainability, but not the only one. Different network architectures (components, layers and connection schemes between layers) may further either impede or facilitate the explainability of the model.

VI. FUTURE WORK & LIMITATIONS

An important prerequisite for reliable and well-functioning ML models, particularly DNNs, is a sufficient amount of data, but this is often a limitation in practice. The more data are available, the more robust patterns may be learned by the classifiers leading to more intuitive explanations. Especially regarding the use of DNNs for the domain of human gait analysis we are optimistic for two reasons. First, depending on the model type and architecture it is possible to train models which provide meaningful explanations for clinical experts even with the currently available data. Second, it is very likely that further advances in performance and generalizability will be made in the future, as new data are constantly being recorded (just as in other domains where DNNs have proven their superiority after being trained on large datasets). The limitation of training data underlines the need to analyze and combine 3DGA data from multiple laboratories in the future. Merging 3DGA data from different laboratories would provide a much larger and heterogeneous dataset that could improve the generalizability of the models. Subsequently, this could lead to the development of more robust and diverse models that can be employed across multiple laboratories.

During our study and in previous research [45], we observed discrepancies between clinicians' expectations and the explanations of the trained models. Clinicians expected ML models to use all regions that are characteristic (from a clinical perspective) for a particular class (independent of which other classes are modeled). However, the models often used only a subset of these regions. This discrepancy is exemplified by the true equinus class in our experiments, in which the model used regions in $Hips_S$ and $Knees_S$, whereas the clinicians expected the model to use also regions in $Ankle_S$. From an ML perspective, it is logical that ML models mainly use features that exhibit large differences between classes. The high similarity of $Ankles_S$ between the true equinus and jump gait class (Figure 5A) seems to be the reason why its features are not used for the classification of true equinus. On the other hand, there are significant differences in $Hips_S$ and $Knees_S$ that are relevant to true equinus in contrast to the other classes. Therefore, the high relevance in these two signals (and missing relevance in $Ankle_S$) is reasonable from an ML perspective.

A further reason for this discrepancy could originate from the diagnostic approach of clinicians, which often compares a patient's walking pattern to the walking pattern of healthy controls. To this end, clinicians typically employ methods such as statistical parametric mapping (SPM) [62], which allows to identify statistically significant differences in the 3DGA data between a patient group and healthy controls. The experience with such methods may contribute to the expectation that an ML model's relevant regions should include all regions that are considered different between a pathological gait pattern and healthy controls. However, in our case, the ML model learns to differentiate the four different pathological gait patterns in a discriminative manner. The ML model does not use a reference to physiological gait and mainly focuses on discriminative patterns that effectively separate two or more pathological classes.

A future direction may be to develop novel approaches that mimic the diagnostic approach of clinicians, while still being explainable and trustworthy. One possible approach is to regularize the ML model with input from clinicians, which would force the network to use specific regions in the data during the training process. Still, there is a trade-off between sacrificing potential insights into the data and building trust in the ML model that should be explored in future work.

VII. CONCLUSION

Building trust in ML models is essential in the medical field to facilitate their use in clinical practice. Explainability approaches provide a useful tool to explain on which information a model bases its predictions. Building upon the post-hoc explainability method Grad-CAM – initially introduced for images and adapted by us to time series – we generated explanations for DNNs trained to differentiate CP-related gait patterns on several levels, i.e., on the decision, class and model level. Furthermore, we trained traditional models (DTs and RFs) for the given problem and explained them via feature importance.

We investigated which subsets of 3DGA data are particularly suitable for the classification of gait patterns associated with CP. Our results confirm the superiority of kinematic over GRF data for this complex classification task, with the former achieving a classification accuracy of up to 93.4 % compared to 47.2 % with GRFs. Our results further demonstrate that the employed ML models base their predictions on clinically relevant features. Traditional ML approaches such as RFs and DTs achieve not only better results in classifying CP-related gait patterns, but also focus more on the clinically relevant regions in the 3DGA data compared to DNNs. An interesting point from the clinical perspective is that DNNs use additional (initially unexpected) features for their predictions. This may facilitate providing novel insights into the data, and thereby raise novel questions in the field.

REFERENCES

[1] S. McIntyre, "The continually changing epidemiology of cerebral palsy," *Acta Paediatrica*, vol. 107, no. 3, pp. 374–375, Mar. 2018.

[2] N. Pérez and A. Rodríguez, "Cerebral palsy: Hope through research," NIH NINDS, Bethesda, MD, USA, 2013.

[3] H. K. Graham, P. Rosenbaum, N. Paneth, B. Dan, J.-P. Lin, L. D. Damiano, G. J. Becher, D. Gaebler-Spira, A. Colver, D. S. Reddihough, K. E. Crompton, and R. L. Lieber, "Cerebral palsy," *Nature Rev. Disease Primers*, vol. 2, no. 1, pp. 1–25, 2016.

[4] R. Baker, *Measuring Walking: A Handbook of Clinical Gait Analysis*. London, U.K.: Mac Keith Press, 2013.

[5] 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, Feb. 2001.

[6] J. Figueiredo, C. P. Santos, and J. C. Moreno, "Automatic recognition of gait patterns in human motor disorders using machine learning: A review," *Med. Eng. Phys.*, vol. 53, pp. 1–12, Mar. 2018.

[7] F. Horst, S. Lapuschkin, W. Samek, K.-R. Müller, and W. I. Schöllhorn, "Explaining the unique nature of individual gait patterns with deep learning," *Sci. Rep.*, vol. 9, no. 1, pp. 1–13, Feb. 2019.

[8] D. Slijepcevic, F. Horst, S. Lapuschkin, B. Horsak, A.-M. Raberger, A. Kranzl, W. Samek, C. Breiteneder, W. I. Schöllhorn, and M. Zeppelzauer, "Explaining machine learning models for clinical gait analysis," *ACM Trans. Comput. Healthcare*, vol. 3, no. 2, pp. 1–27, Apr. 2022.

[9] 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.

[10] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable AI systems for the medical domain?" 2017, *arXiv:1712.09923*.

[11] European Union, "Regulation (EU) 2016/679 of the European parliament and of the council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (general data protection regulation)," *Off. J. Eur. Union*, vol. 119, pp. 1–88, May 2016. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj

[12] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–10.

[13] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE ICCV*, Oct. 2017, pp. 618–626.

[14] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proc. Adv. NIPS*, 2018, pp. 9505–9515.

[15] 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," *J. Biomech.*, vol. 81, pp. 1–11, Nov. 2018.

[16] P. Khera and N. Kumar, "Role of machine learning in gait analysis: A review," *J. Med. Eng. Technol.*, vol. 44, no. 8, pp. 441–467, Nov. 2020.

[17] C. Cui, G. Bian, Z. Hou, J. Zhao, G. Su, H. Zhou, L. Peng, and W. Wang, "Simultaneous recognition and assessment of post-stroke hemiparetic gait by fusing kinematic, kinetic, and electrophysiological data," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 4, pp. 856–864, Apr. 2018.

[18] 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 J. Biomed. Health Informat.*, vol. 19, no. 6, pp. 1794–1802, Nov. 2015.

[19] M. Alaqtash, T. Sarkodie-Gyan, H. Yu, O. Fuentes, R. Brower, and A. Abdelgawad, "Automatic classification of pathological gait patterns using ground reaction forces and machine learning algorithms," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2011, pp. 453–457.

[20] M. J. Long, E. Papi, L. D. Duffell, and A. H. McGregor, "Predicting knee osteoarthritis risk in injured populations," *Clin. Biomech.*, vol. 47, pp. 87–95, Aug. 2017.

[21] D. Slijepcevic, M. Zeppelzauer, A. Gorgas, C. Schwab, M. Schüller, A. Baca, C. Breiteneder, and B. Horsak, "Automatic classification of functional gait disorders," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 5, pp. 1653–1661, Sep. 2018.

[22] D. Slijepcevic, M. Zeppelzauer, C. Schwab, A.-M. Raberger, C. Breiteneder, and B. Horsak, "Input representations and classification strategies for automated human gait analysis," *Gait Posture*, vol. 76, pp. 198–203, Feb. 2020.

[23] E. Papageorgiou, A. Nieuwenhuys, I. Vandekerckhove, A. Van Campenhout, E. Ortibus, and K. Desloovere, "Systematic review on gait classifications in children with cerebral palsy: An update," *Gait Posture*, vol. 69, pp. 209–223, Mar. 2019.

[24] A. Ferrari, L. Bergamini, G. Guerzoni, S. Calderara, N. Bicocchi, G. Vitetta, C. Borghi, R. Neviani, and A. Ferrari, "Gait-based diplegia classification using LSMT networks," *J. Healthcare Eng.*, vol. 2019, pp. 1–8, Jan. 2019.

[25] A. Ferrari, S. Alboresi, S. Muzzini, R. Pascale, S. Perazza, and G. Cioni, "The term diplegia should be enhanced. Part I: A new rehabilitation oriented classification of cerebral palsy," *Eur. J. Phys. Rehabil. Med.*, vol. 44, no. 2, p. 195, 2008.

[26] Y. Zhang and Y. Ma, "Application of supervised machine learning algorithms in the classification of sagittal gait patterns of cerebral palsy children with spastic diplegia," *Comput. Biol. Med.*, vol. 106, pp. 33–39, Mar. 2019.

[27] J. Rodda and H. K. Graham, "Classification of gait patterns in spastic hemiplegia and spastic diplegia: A basis for a management algorithm," *Eur. J. Neurol.*, vol. 8, no. 5, pp. 98–108, Nov. 2001.

[28] H. Darbandi, M. Baniasad, S. Baghdadi, A. Khandan, A. Vafaee, and F. Farahmand, "Automatic classification of gait patterns in children with cerebral palsy using fuzzy clustering method," *Clin. Biomech.*, vol. 73, pp. 189–194, Mar. 2020.

[29] K. Chia, I. Fischer, P. Thomason, H. K. Graham, and M. Sangeux, "A decision support system to facilitate identification of musculoskeletal impairments and propose recommendations using gait analysis in children with cerebral palsy," *Frontiers Bioeng. Biotechnol.*, vol. 8, Nov. 2020, Art. no. 529415.

[30] V. Arya, R. K. E. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. Vera Liao, R. Luss, A. Mojsilović, S. Mourad, P. Pedemonte, R. Raghavendra, J. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K. R. Varshney, D. Wei, and Y. Zhang, "One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques," 2019, *arXiv:1909.03012*.

[31] M. Wagner, D. Slijepcevic, B. Horsak, A. Rind, M. Zeppelzauer, and W. Aigner, "KAVAGait: Knowledge-assisted visual analytics for clinical gait analysis," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 3, pp. 1528–1542, Mar. 2019.

[32] A. Bois, B. Tervil, A. Moreau, A. Vienne-Jumeau, D. Ricard, and L. Oudre, "A topological data analysis-based method for gait signals with an application to the study of multiple sclerosis," *PLoS ONE*, vol. 17, no. 5, May 2022, Art. no. e0268475.

[33] N. Roche, D. Pradon, J. Cosson, J. Robertson, C. Marchiori, and R. Zory, "Categorization of gait patterns in adults with cerebral palsy: A clustering approach," *Gait Posture*, vol. 39, no. 1, pp. 235–240, Jan. 2014.

[34] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: Removing noise by adding noise," 2017, *arXiv:1706.03825*.

[35] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, Jul. 2015, Art. no. e0130140.

[36] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.

[37] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–10.

[38] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, "This looks like that: Deep learning for interpretable image recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.

[39] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013, *arXiv:1312.6034*.

[40] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking clever Hans predictors and assessing what machines really learn," *Nature Commun.*, vol. 10, no. 1, pp. 1–8, Mar. 2019.

[41] F. Horst, D. Slijepcevic, M. Zeppelzauer, A. M. Raberger, S. Lapuschkin, W. Samek, W. I. Schöllhorn, C. Breiteneder, and B. Horsak, "Explaining automated gender classification of human gait," *Gait Posture*, vol. 81, pp. 159–160, Sep. 2020.

[42] D. Slijepcevic, F. Horst, M. Simak, S. Lapuschkin, A. M. Raberger, W. Samek, C. Breiteneder, W. I. Schöllhorn, M. Zeppelzauer, and B. Horsak, "Explaining machine learning models for age classification in human gait analysis," *Gait Posture*, vol. 97, pp. S252–S253, Sep. 2022.

[43] 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, p. 4385, Aug. 2020.

[44] C. Kokkotis, S. Moustakidis, T. Tsatalas, C. Ntakolia, G. Chalatsis, S. Konstadakos, M. E. Hantes, G. Giakas, and D. Tsaopoulos, "Leveraging explainable machine learning to identify gait biomechanical parameters associated with anterior cruciate ligament injury," *Sci. Rep.*, vol. 12, no. 1, pp. 1–12, Apr. 2022.

[45] A. Rind, D. Slijepčević, M. Zeppelzauer, F. Unglaube, A. Kranzl, and B. Horsak, "Trustworthy visual analytics in clinical gait analysis: A case study for patients with cerebral palsy," in *Proc. IEEE Workshop Trust Expertise Vis. Anal. (TREX)*, Oct. 2022, pp. 8–15.

[46] A. Rind and D. Slijepcevic, "gaitXplorer screenshot," Zenodo, Pölten Univ. Appl. Sci., St. Pölten, Austria, Dec. 2022, doi: 10.5281/zenodo.7442945.

[47] J. M. Rodda, H. K. Graham, L. Carson, M. P. Galea, and R. Wolfe, "Sagittal gait patterns in spastic diplegia," *J. Bone Joint Surg.*, vol. 86-B, no. 2, pp. 251–258, Mar. 2004.

[48] M. Sangeux, J. Rodda, and H. K. Graham, "Sagittal gait patterns in cerebral palsy: The plantarflexor–knee extension couple index," *Gait Posture*, vol. 41, no. 2, pp. 586–591, Feb. 2015.

[49] D. A. Winter, *Biomechanics and Motor Control of Human Movement*, 3rd ed. Hoboken, NJ, USA: Wiley, 2005.

[50] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[51] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, Mar. 1986.

[52] J. R. Quinlan, *C4.5: Programs for Machine Learning* (Morgan Kaufmann Series in Machine Learning). San Mateo, CA, USA: Morgan Kaufmann, 1993.

[53] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. New York, NY, USA: Taylor & Francis, 1984.

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

[55] A. Mamalakis, E. A. Barnes, and I. Ebert-Uphoff, "Investigating the fidelity of explainable artificial intelligence methods for applications of convolutional neural networks in geoscience," *Artif. Intell. Earth Syst.*, vol. 1, no. 4, Oct. 2022, Art. no. e220012.

[56] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," *J. Comput. Graph. Statist.*, vol. 24, no. 1, pp. 44–65, Jan. 2015.

[57] I. El Maachi, G.-A. Bilodeau, and W. Bouachir, "Deep 1D-convnet for accurate Parkinson disease detection and severity prediction from gait," *Expert Syst. Appl.*, vol. 143, Apr. 2020, Art. no. 113075.

[58] W. Zeng, F. Liu, Q. Wang, Y. Wang, L. Ma, and Y. Zhang, "Parkinson's disease classification using gait analysis via deterministic learning," *Neurosci. Lett.*, vol. 633, pp. 268–278, Oct. 2016.

[59] C. Nüesch, V. Valderrabano, C. Huber, V. von Tscharner, and G. Pagenstert, "Gait patterns of asymmetric ankle osteoarthritis patients," *Clin. Biomech.*, vol. 27, no. 6, pp. 613–618, Jul. 2012.

[60] D. P. Soares, M. P. de Castro, E. A. Mendes, and L. Machado, "Principal component analysis in ground reaction forces and center of pressure gait waveforms of people with transfemoral amputation," *Prosthetics Orthotics Int.*, vol. 40, no. 6, pp. 729–738, 2016.

[61] A. M. S. Muniz and J. Nadal, "Application of principal component analysis in vertical ground reaction force to discriminate normal and abnormal gait," *Gait Posture*, vol. 29, no. 1, pp. 31–35, Jan. 2009.

[62] T. C. Pataky, "Generalized n-dimensional biomechanical field analysis using statistical parametric mapping," *J. Biomech.*, vol. 43, no. 10, pp. 1976–1982, Jul. 2010.

DJORDJE SLIJEPCEVIC received the M.Sc. degree in computer engineering from TU Wien, Austria, where he is currently pursuing the Ph.D. degree in technical sciences, with a focus on the development of machine learning methods in the field of clinical human gait analysis. He is a Researcher with the Institute of Creative Media Technologies (ICMT), St. Pölten University of Applied Sciences, Austria. He has extensive experience in research in the domain of automated human gait analysis, with particular focus on the topics gait recognition, gait pattern classification, gait event detection, and similarity retrieval of gait patterns. His research interests include machine learning, explainable artificial intelligence, computer vision, and time series analysis.

MATTHIAS ZEPPELZAUER received the Ph.D. and Habilitation degrees in computer science from the Vienna University of Technology. He is currently the Head of the Media Computing Research Group and a Coordinator of the Center for Artificial Intelligence, St. Pölten University of Applied Sciences, Austria. His research interests include computer vision, machine learning and multimedia analysis. He has a long track of research on automated human gait analysis focusing on machine learning architectures and features for the extraction of information in gait signals and the prediction of pathological gait patterns.

FABIAN UNGLAUBE received the master's degree in sports science. Since 2016, he has been a Research Assistant with the Laboratory for Gait and Movement Analysis, Orthopaedic Hospital Speising, Vienna, Austria. He has been participated in several national and international research projects in the field of clinical gait and movement analysis. On a daily basis, he works with orthopaedic and cerebral palsy related patients in the gait laboratory. He has been an active member in several professional societies, such as the European Society for Movement Analysis in Adults and Children (ESMAC).

ANDREAS KRANZL received the Ph.D. degree in sports science. Since 1996, he has been the Head of the Laboratory for Gait and Movement Analysis, Orthopaedic Hospital Speising, Vienna, Austria. He has more than 30 years experience in clinical gait and movement analysis with focus on orthopaedic and cerebral palsy related patients. He has been holding active membership in several professional societies, such as the European Society for Movement Analysis in Adults and Children (ESMAC). He has been active as a reviewer for several internationally renowned journals in the field of gait analysis.

CHRISTIAN BREITENEDER received the Diploma in Engineering degree in computer science from Johannes Kepler University Linz, in 1978, and the Ph.D. degree in computer science from TU Wien, in 1991. He studied history of art with the University of Vienna, from 1977 to 1981, and theatre directing with Max Reinhardt Seminar, Vienna, from 1981 to 1984. He was Postdoctoral Researcher with CUI, University of Geneva, Switzerland, from 1991 to 1993, and GMD (now Fraunhofer), Birlinghoven, Germany, from 1995 to 1996. He was an Associate Professor with the University of Vienna, from 1997 to 2000. He is currently a Retired Professor with the Institute of Visual Computing and Human-Centered Technology, TU Wien. His current research interests include interactive media systems, media processing systems, augmented and virtual reality, content-based multimodal information retrieval, and the analysis of high-dimensional data.

BRIAN HORSAK received the Ph.D. and Habilitation degrees in sport science from the University of Vienna. He is currently the Head of the Motor Rehabilitation Research Group and the Scientific Director of the Center for Digital Health and Social Innovation, St. Pölten University of Applied Sciences, Austria. He is an accomplished researcher whose vision is to combine technology and healthcare to provide advanced medical solutions in the field of gait analysis and rehabilitation. His research is geared toward enhancing clinical practice and facilitating medical decision-making through the utilization of cutting-edge technologies, including motion capturing, wearables, musculoskeletal simulations, machine learning, and augmented and virtual reality.

• • •