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
BRIAN HORSAK, Institute of Health Sciences, Department of Health Sciences, St. Pölten University of Applied Sciences, Austria
SEBASTIAN LAPUSCHKIN, Department of Artificial Intelligence, Fraunhofer Heinrich Hertz Institute, Germany
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

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

*Both authors contributed equally to this research.

Authors' addresses: Djordje Slijepcevic, Djordje.Slijepcevic@fhstp.ac.at, Institute of Creative Media Technologies, Department of Media & Digital Technologies, St. Pölten University of Applied Sciences, St. Pölten, Austria; Fabian Horst, horst@uni-mainz.de, Department of Training and Movement Science, Institute of Sport Science, Johannes Gutenberg-University Mainz, Mainz, Germany; Brian Horsak, Institute of Health Sciences, Department of Health Sciences, St. Pölten University of Applied Sciences, St. Pölten, Austria; Sebastian Lapuschkin, Department of Artificial Intelligence, Fraunhofer Heinrich Hertz Institute, Berlin, Germany; Anna-Maria Raberger, Institute of Health Sciences, Department of Health Sciences, St. Pölten University of Applied Sciences, St. Pölten, Austria; Andreas Kranzl, Laboratory for Gait and Movement Analysis, Orthopaedic Hospital Vienna-Speising, Vienna, Austria; Wojciech Samek, Department of Artificial Intelligence, Fraunhofer Heinrich Hertz Institute, Berlin, Germany; Christian Breiteneder, Institute of Visual Computing and Human-Centered Technology, TU Wien, Vienna, Austria; Wolfgang Immanuel Schöllhorn, Department of Training and Movement Science, Institute of Sport Science, Johannes Gutenberg-University Mainz, Mainz, Germany; Matthias Zeppelzauer, Institute of Creative Media Technologies, Department of Media & Digital Technologies, St. Pölten University of Applied Sciences, St. Pölten, Austria.

XXXX-XXXX/2020/7-ART $15.00
https://doi.org/10.1145/1122445.1122456

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: clinical gait analysis, human gait classification, explainable artificial intelligence, layer-wise relevance propagation, statistical parametric mapping, ground reaction forces, convolutional neural networks

ACM Reference Format:
Djordje Slijepcevic, Fabian Horst, Brian Horsak, Sebastian Lapuschkin, Anna-Maria Raberger, Andreas Kranzl, Wojciech Samek, Christian Breiteneder, Wolfgang Immanuel Schöllhorn, and Matthias Zeppelzauer. 2020. Explaining Machine Learning Models for Clinical Gait Analysis. 1, 1 (July 2020), 42 pages. https://doi.org/10.1145/1122445.1122456

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 [74]. This trend has also been recognized in the field of clinical gait analysis (CGA) [18, 62]. 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, 80]. Thereby, CGA produces a vast amount of data [22, 55], which are difficult to comprehend due to their multi-dimensional and multi-correlated nature [13, 81]. In the last years, ML methods have been successfully employed in CGA for the classification of patient groups [18, 62] such as stroke [36, 53], Parkinson's disease [77], cerebral palsy [75], multiple sclerosis [3], osteoarthritis [50], and patients suffering from different functional gait disorders [67]. 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 in 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 also 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, 60].

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 in order 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, 73]. The motivation is to increase the traceability and trust of medical professionals in ML approaches [27]. However, application of XAI methods to the field of CGA remains to be investigated. A first step in that 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, 58, 59], 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 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 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 [76, 78].

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 which 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 the practical applicability of such approaches.

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 [83] has been introduced to Support Vector Machines (SVM) [8] and CNNs [66] in the form of saliency maps. Layer-wise Relevance Propagation (LRP) [7, 44] and Deep Learning Important FeaTures (DeepLIFT) [64] 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. [82], 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 [56] – 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* 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, that can be included or excluded [72]. 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. [66] 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 that 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 a 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 1B) based on three-dimensional ground reaction forces (GRFs) of both legs (see Figure 1A), to explain the predictions of these models based on relevance scores that are related to the input signal space by using LRP (see Figure 1C), and to evaluate these results from a statistical (see Figure 1D) and a clinical perspective (see Figure 1E). 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 side 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 are a common practice in automated gait classification [12, 22, 69].

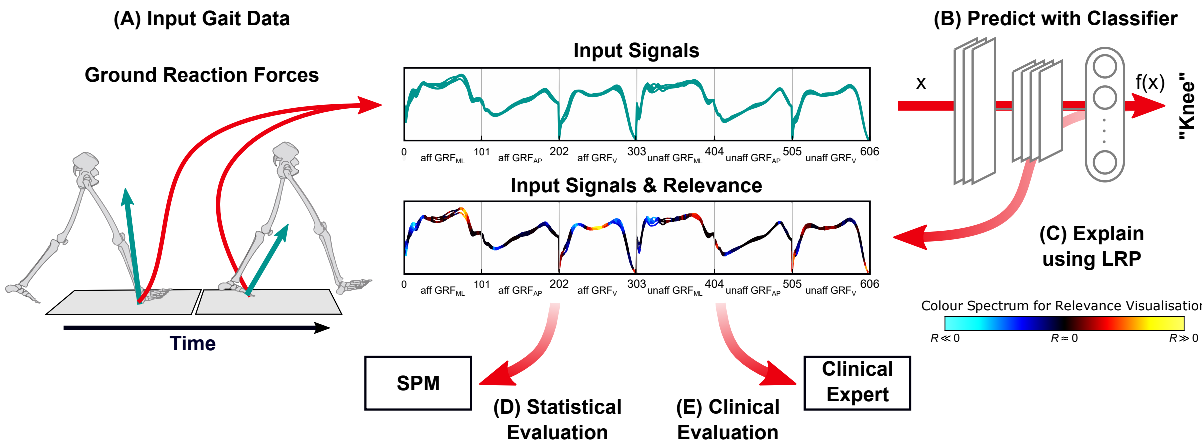

Fig. 1. Overview of our proposed workflow for data acquisition, prediction and prediction explanation in automated gait classification, showing the data of one participant belonging to the knee disorder class. (A) The clinical gait analysis consists of five recordings of each participant walking barefoot (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.

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:

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

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.

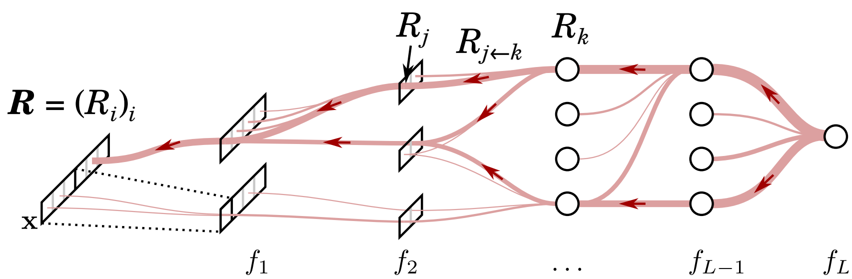

Fig. 2. Illustration of the LRP back-propagation procedure applied to a neural network function $f(x) = f_L \circ \cdots \circ f_1(x)$. The prediction at the output is propagated backward in the network, until the input features are reached and relevance scores are obtained for all input features and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

Various purposeful propagation rules have been proposed in the literature [7, 32, 44]. For example, the LRP_ε 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 j^{th} input neuron to the k^{th} 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 k^{th} 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 neuron relevance is aggregated from incoming relevance messages as $R_j = \sum_k R_{j\leftarrow k}$.

Other propagation rules such as LRP_γ [44], $\text{LRP}_{\alpha\beta}$, $\text{LRP}_{z^\mathcal{B}}$ or $\text{LRP}_\flat$, are suitable for other application scenarios, layer types, or particularly deeper neural networks [32, 44, 59] and have been shown to work well in practice [58].

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. In order 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_ε with $\varepsilon = 10^{-5}$ in all layers across the different models (except for the CNN for which we employed the $\text{LRP}_\flat$ 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].

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

3.3 Statistical Evaluation

To evaluate the derived relevance scores of LRP, we employ 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 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 compute 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 [57]. 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 ten and more than twenty-five 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 ground reaction force (GRF) recordings of patients and healthy

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

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 2000 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 number of 5 valid recordings were available. Recordings were defined as valid by an experienced assessor.

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**

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.

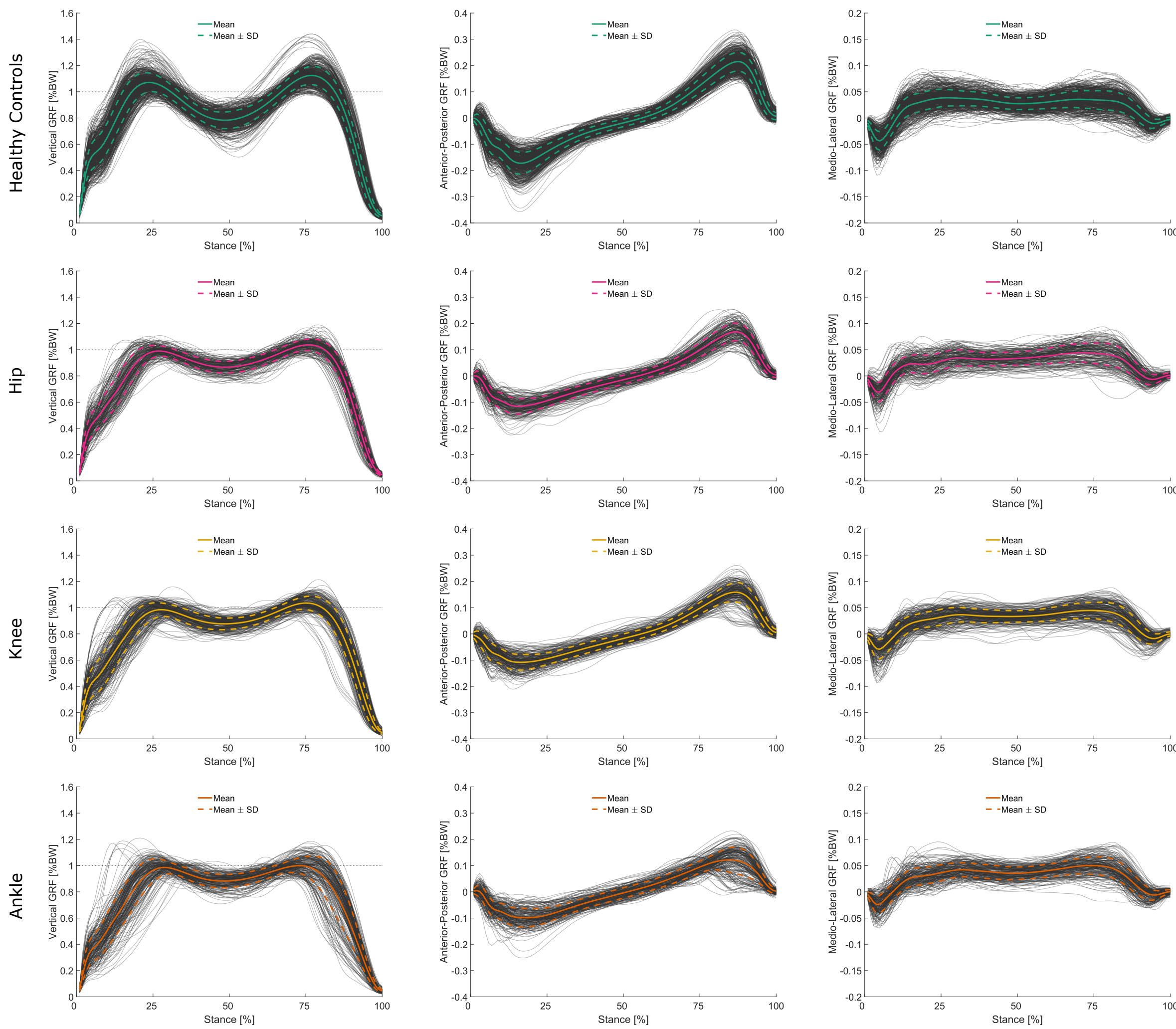

Fig. 3. Visualization of vertical (left panel), anterior-posterior (central panel), and medio-lateral (right panel) force components of the body weight-normalized GRF measurements of the affected side available per participant and class. For healthy controls all available measurements are visualized.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

4.4 Classification Tasks

We investigate different classification tasks on the dataset introduced above to provide a more comprehensive picture on 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) Support Vector Machine (SVM), Multi-layer Perceptron (MLP), and Convolutional Neural Network (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 of 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 2304-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 ten-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 are 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 utilised 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 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 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, 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 end of the signal.

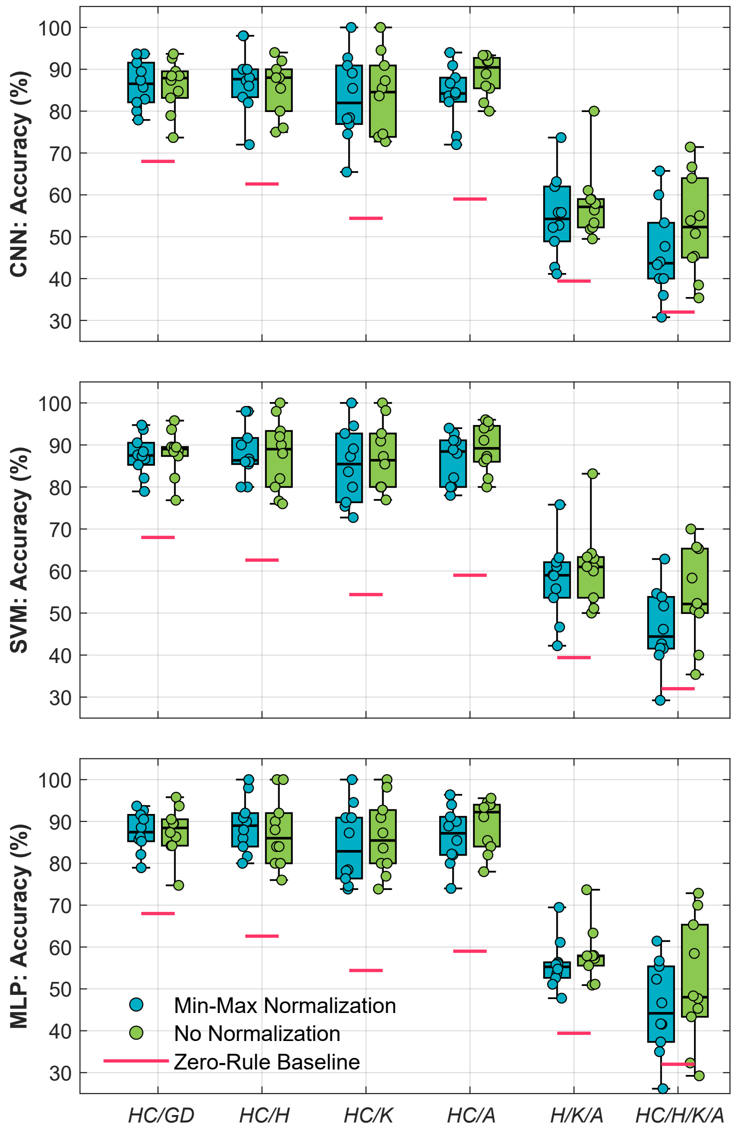

Fig. 4. Overview of the prediction accuracy obtained for the three employed classification methods (CNN, SVM and MLP) and all classification tasks with min-max normalized and non-normalized input signals, reported as boxplots enhanced with the classification accuracies obtained over ten-fold cross validation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

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). The discussion in the following will incorporate all classification tasks.

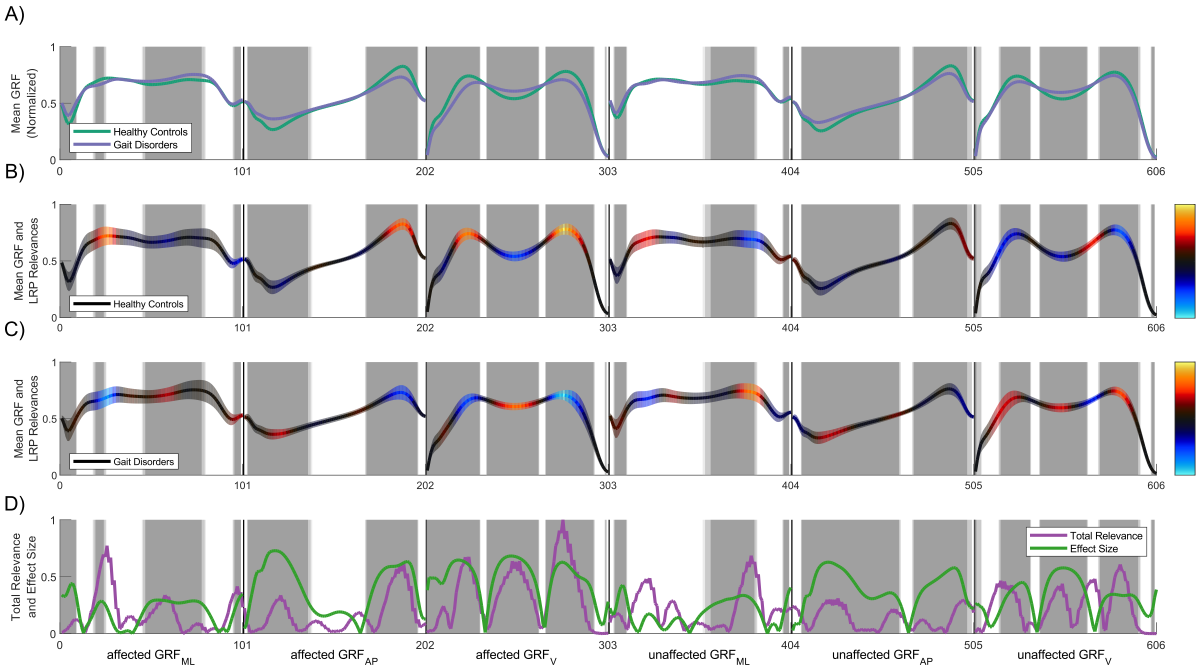

Fig. 5. Results overview for the classification of healthy controls (*HC*) and the aggregated class of all three gait disorders (*GD*) based on min-max normalized GRF signals using a CNN as classifier. (A) Averaged GRF signals for *HC* and *GD*. The first three signals represent the three GRF components 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, which are depicted in three different shades of grey for the three different alpha levels, i.e., dark grey for 0.01, grey for 0.05, and light grey 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.

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).

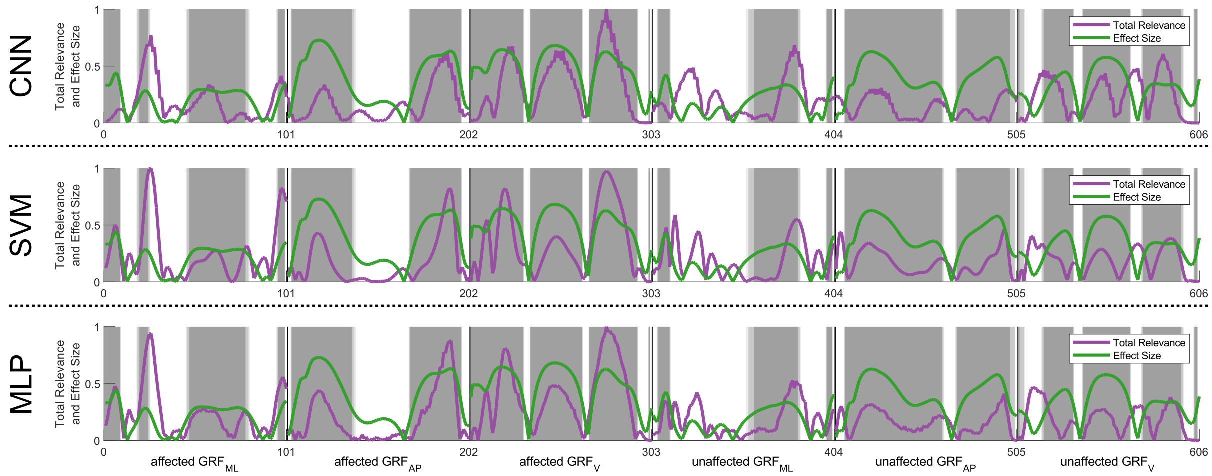

Fig. 6. Comparison of different classification methods (CNN, SVM, and MLP) for the classification of healthy controls and the class of all three gait disorders (HC/GD) based on min-max normalized GRF signals. The comparison is based on the total relevance of the LRP results as well as statistically significant 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.

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 found 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 [67].

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 [68, 69], 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 varies 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 region 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 [68, 69]. 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].

Table 2. Classification results for the experiment with occluded horizontal forces (GRF_{AP}, GRF_{ML}), in percent. 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.

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

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 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 Spectral Relevance Analysis (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 7C and 7F, 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 7D), 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 7B and Figure 7E), we can see that the relevance patterns of the five trials of a participant are often clustered together (Figure 7B and 7E). 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 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 which 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 5D 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,

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

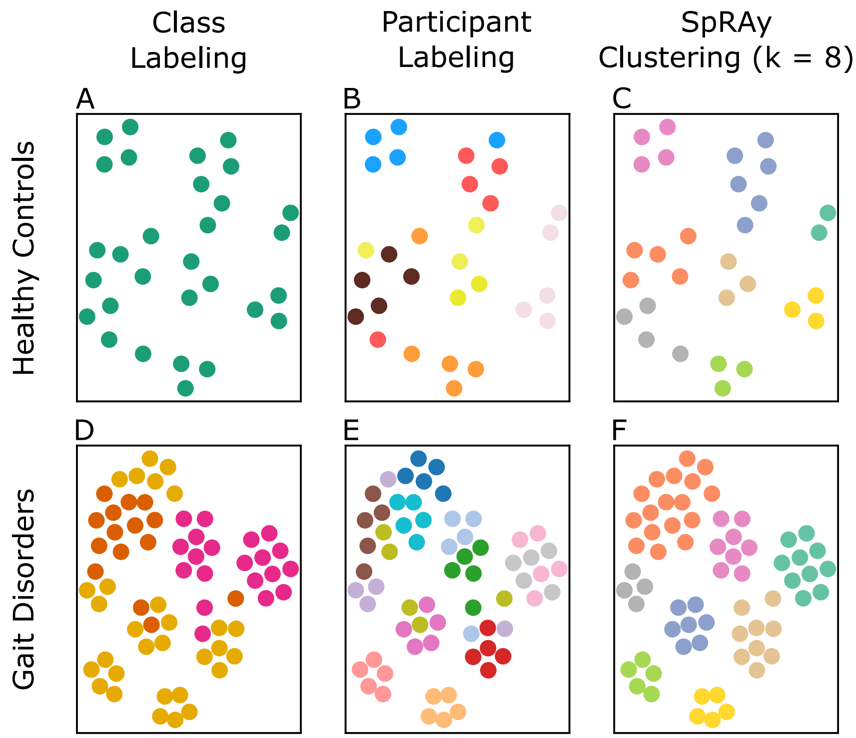

Fig. 7. The spectral embedding Φ derived via SpRAy from LRP explanations for the CNN model on test data, visualized via t-SNE for samples labeled as healthy controls (*HC*; N=30; subfigures A-C) and the aggregated class of all three gait disorders ($GD = \{H, K, A\}$; N=65; subfigures D-F). Each column 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.

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.

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 [54] or Winter [80].

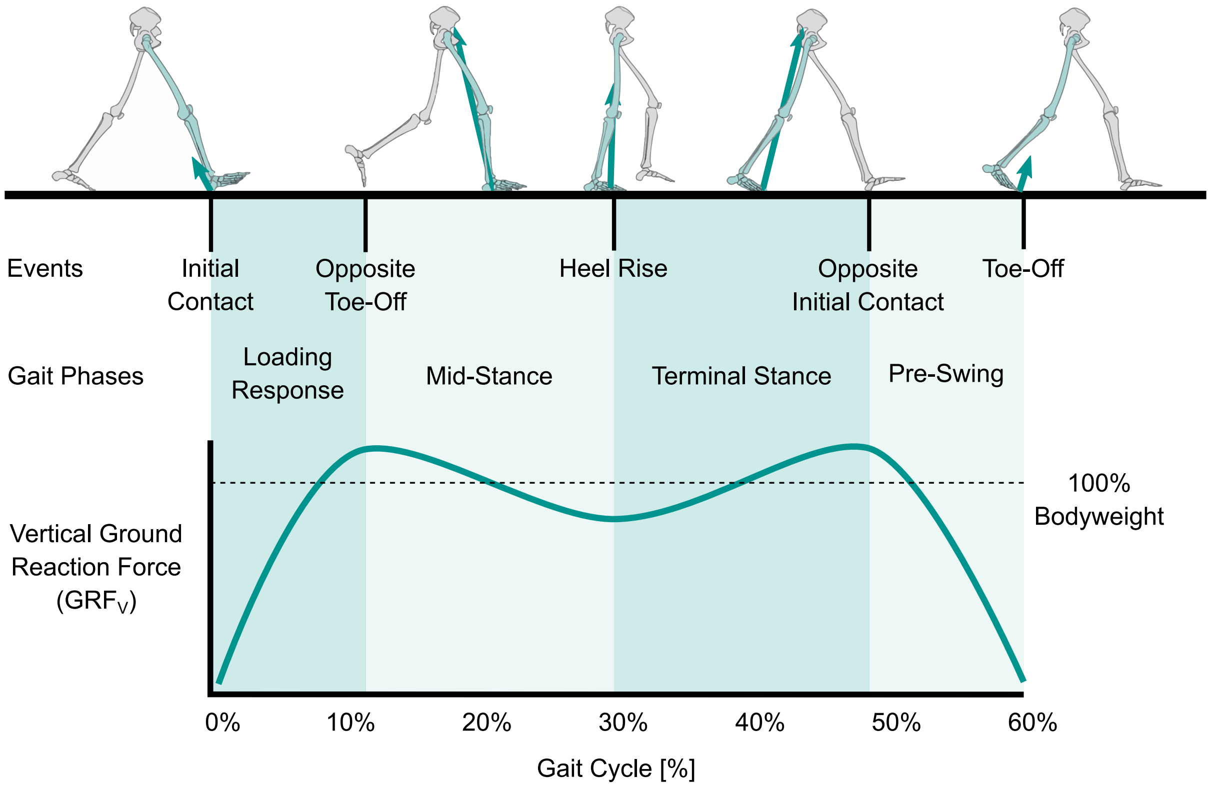

Fig. 8. Overview of the most relevant gait events during the stance phase. In clinical gait analysis, a gait cycle (100%) is defined from the initial contact of one foot to the subsequent initial contact of the same foot. During the first approximately 60% of the gait cycle, referenced as the stance 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 [9, 63].

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 [65, 79]. 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. [70] 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. [70] 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 could also associate 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?***

The two experts reported that they mainly focus on regions in the GRF_V signals during the evaluation process of patients in the 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 tasks as an example, we examine 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 clearly show that most of the relevant regions according to LRP for the walking speed-matched data 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 which are less focused 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 [71].

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.

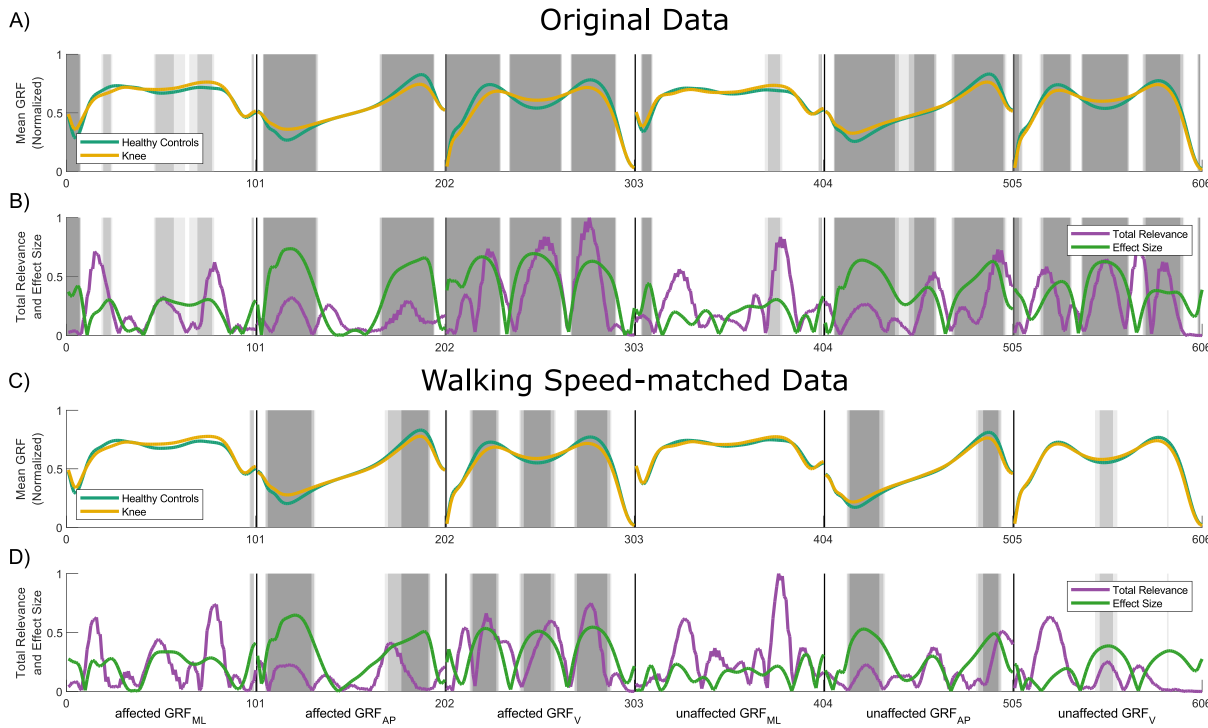

Fig. 9. Comparison of explainability results of the original (top) and walking speed-matched (bottom) data for the classification task HC/K based on the min-max normalized GRF signals using CNN.

Another limiting factor is that we solely used GRF signals for classification. This does not perfectly reflect the 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, 61]. This makes it more difficult to create a homogeneous, large-scale, and real-world data set 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 paper, 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 [58] including datasets with respective ground truth explanations.

7 CONCLUSION

The present findings highlight that machine learning models base their predictions on meaningful features of GRF signals in clinical gait classification tasks that are in accordance with a statistical and clinical evaluation. Hence, XAI methods which provide explainability for predictions made by machine learning models, such as LRP, can be promising solutions to increase justification of automatic classification predictions in CGA and can help to make the prediction processes comprehensible to clinical and legal 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 side. Thus, the unaffected side captures additional information which 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 the most cases also significantly different 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 paper 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, A-MR, MZ, BH prepared the dataset. FH, DS, SL, WS, WIS, BH conceived the presented idea. MZ, WS, WIS, BH raised the funding. FH, DS, SL, A-MR, MZ, WS, CB, WIS, BH participated in the data analysis. FH, DS, SL, A-MR, MZ, BH wrote the manuscript. FH, DS, SL, A-MR, MZ, WS, BH designed the figures. FH, DS, SL, A-MR, MZ, WS, CB, WIS, BH reviewed and approved the final manuscript.

FUNDING

This work was partly funded by the Austrian Research Promotion Agency (FFG) and the 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).

DATA AVAILABILITY STATEMENT

For our analyses, we used a subset of the GaitRec dataset [28]. Our source code and the utilised 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. 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 Advances in Neural Information Processing Systems 31, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.). Curran Associates, Inc., 9505–9515. http://papers.nips.cc/paper/8160-sanity-checks-for-saliency-maps.pdf

[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 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS). IEEE, 453–457. 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. 2020. Finding and Removing Clever Hans: Using Explanation Methods to Debug and Improve Deep Models. CoRR abs/1912.11425 (2020). http://arxiv.org/abs/1912.11425

[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). http://arxiv.org/abs/1909.03012

[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. 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. Journal of Machine Learning Research 11 (2010), 1803–1831. 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 Proc. of the International Conference on Machine Learning (ICML). PMLR, 342–350.

[11] Brian G Booth, Noël LW Keijsers, Jan Sijbers, and Toon Huysmans. 2018. STAPP: spatiotemporal analysis of plantar pressure measurements using statistical parametric mapping. Gait & Posture 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. Frontiers in Bioengineering and Biotechnology 8 (2020), 260. 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 & Posture 13, 1 (Feb. 2001), 49–66. 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. Journal of Orthopaedic & Sports Physical Therapy 25, 4 (1997), 236–244. 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. 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). Official Journal of the European Union L 119 (2016), 1–88. Available at: 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. Medical Engineering and Physics 53 (2018), 1–12. 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 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 3429–3437. 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 & Posture 29, 3 (2009), 398–402. https://doi.org/10.1016/j.gaitpost.2008.10.060

[21] Holger A Haenssle, Christine Fink, R Schneiderbauer, Ferdinand Toberer, Timo 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. Annals of Oncology 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. Journal of Biomechanics 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. Nature Medicine 25, 1 (2019), 30–36. 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 European Conference on Computer Vision (ECCV). Springer, 3–19. https://doi.org/10.1007/978-3-319-46493-0_1

[25] Sture Holm. 1979. A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics 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). http://arxiv.org/abs/1712.09923

[27] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. 2019. Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9, 4 (July 2019), e1312. 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. Scientific Data 7, 1 (May 2020), 1–8. 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. Scientific Reports 9, 1 (2019), 2391. https://doi.org/10.1038/s41598-019-38748-8

[30] Fabian Horst, M. Mildner, and W. I. Schöllhorn. 2017. One-year persistence of individual gait patterns identified in a follow-up study - A call for individualised diagnose and therapy. Gait Posture 58 (2017), 476–480. 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. Available at: 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 Proc. of International Joint Conference on Neural Networks (IJCNN) 2020. IEEE, 1–7.

[33] Arthur D. Kuo and J. Maxwell Donelan. 2010. Dynamic Principles of Gait and Their Clinical Implications. Physical Therapy 90, 2 (2010), 157–174. 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 Proc. 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. Nature Communications 10 (2019), 1096. 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. Human Movement Science 28, 4 (Aug. 2009), 504–514. 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. 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 Advances in Neural Information Processing Systems (NIPS). Curran Associates, Inc., 4765–4774. Available at: 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. Journal of Machine Learning Research 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. Journal of NeuroEngineering and Rehabilitation 11, 1 (2014), 20. 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 & Posture 14, 2 (2001), 104–109. 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 Proc. 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. 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 Recognition 65 (2017), 211–222.

[46] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2018. Methods for interpreting and understanding deep neural networks. Digital Signal Processing 73 (2018), 1–15. 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 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 Advances in Neural Information Processing Systems. Curran Associates, Inc., 3387–3395. Available at: 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. Clinical Biomechanics 27, 6 (July 2012), 613–618. 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. Journal of Biomechanics 43, 10 (July 2010), 1976–1982. https://doi.org/10.1016/j.jbiomech.2010.03.008

[52] Todd C. Pataky. 2012. One-dimensional statistical parametric mapping in Python. Computer Methods in Biomechanics and Biomedical Engineering 15, 3 (March 2012), 295–301. https://doi.org/10.1080/10255842.2010.527837

[53] P. Patil, K. S. Kumar, N. Gaud, and V. B. Semwal. 2019. Clinical Human Gait Classification: Extreme Learning Machine Approach. In Proc. of the International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT). 1–6. https://doi.org/10.1109/ICASERT.2019.8934463

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

[55] 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. Journal of Medical and Biological Engineering 38, 2 (2018), 244–260. https://doi.org/10.1007/s40846-017-0297-2

[56] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-Agnostic Interpretability of Machine Learning. CoRR abs/1606.05386 (2016). http://arxiv.org/abs/1606.05386

[57] Robert Rosenthal. 1986. Meta-Analytic Procedures for Social Science Research Sage Publications: Beverly Hills, 1984, 148 pp. Educational Researcher 15, 8 (Oct. 1986), 18–20. https://doi.org/10.3102/0013189X015008018

[58] 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 Transactions on Neural Networks and Learning Systems 28, 11 (Nov 2017), 2660–2673. https://doi.org/10.1109/TNNLS.2016.2599820

[59] 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. of the IEEE 109, 3 (2021), 247–278. https://doi.org/10.1109/JPROC.2021.3060483

[60] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. ITU Journal: ICT Discoveries 1, 1 (2017), 39–48.

[61] 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. International Biomechanics 6, 1 (2019), 19–33. https://doi.org/10.1080/23335432.2019.1621205

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

[63] 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. Scientific Reports 8, 1 (2018), 1–10.

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

[65] 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 Research and Treatment (2012), 11. https://doi.org/10.1155/2012/214980

[66] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In Proc. of the International Conference on Learning Representations (ICLR). http://arxiv.org/abs/1312.6034

[67] 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 Journal of Biomedical and Health Informatics 22, 5 (2017), 1653–1661. https://doi.org/10.1109/JBHI.2017.2785682

[68] Djordje Slijepcevic, Matthias Zeppelzauer, Caterine Schwab, Anna-Maria Raberger, Christian Breiteneder, and Brian Horsak. 2019. Input Representations and Classification Strategies for Automated Human Gait Analysis. Gait & Posture (2019). https://doi.org/10.1016/j.gaitpost.2019.10.021

[69] 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 & Posture 65 (2018), 249. https://doi.org/10.1016/j.gaitpost.2018.06.155

[70] S. Jun Son, Hyunsoo Kim, Matthew K. Seeley, and J. Ty Hopkins. 2019. Altered Walking Neuromechanics in Patients With Chronic Ankle Instability. Journal of Athletic Training 54, 6 (2019), 684–697. https://doi.org/10.4085/1062-6050-478-17

[71] F Sperrle, M El-Assady, G Guo, R Borgo, D Horng Chau, A Endert, and D Keim. 2021. A Survey of Human-Centered Evaluations in Human-Centered Machine Learning. STAR 40, 3 (2021).

[72] Erik Štrumbelj and Igor Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems 41, 3 (2014), 647–665. https://doi.org/10.1007/s10115-013-0679-x

[73] Erico Tjoa and Cuntai Guan. 2019. A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI. CoRR abs/1907.07374 (2019). http://arxiv.org/abs/1907.07374

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

[75] 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. Research in Developmental Disabilities 32, 6 (Nov. 2011), 2542–2552. https://doi.org/10.1016/j.ridd.2011.07.004

[76] Markus Wagner, Djordje Slijepcevic, Brian Horsak, Alexander Rind, Matthias Zeppelzauer, and Wolfgang Aigner. 2018. KAVAGait: Knowledge-assisted visual analytics for clinical gait analysis. IEEE Transactions on Visualization and Computer Graphics 25, 3 (2018), 1528–1542.

[77] 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 Journal of Biomedical and Health Informatics 19, 6 (2015), 1794–1802.

[78] 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. https://doi.org/10.1117/12.2080001

[79] 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. Archives of Physical Medicine and Rehabilitation 85, 12 (2004), 1923–1928. https://doi.org/10.1016/j.apmr.2003.11.040

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

[81] 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 & Posture 23, 3 (2006), 331–338. https://doi.org/10.1016/j.gaitpost.2005.04.004

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

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

SUPPLEMENTARY MATERIAL

The supplementary material presents additional results we generated for the paper

"Explaining Machine Learning Models for Clinical Gait Analysis".

The primary aim of this article is to explain which class-specific characteristics ML models learn from CGA data. For this purpose, we investigate different gait classification tasks, employ a representative set of classification methods – (linear) Support Vector Machine (SVM), Multi-layer Perceptron (MLP), and Convolutional Neural Network (CNN) –, and a Explainable Artificial Intelligence (XAI) method – 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.

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 in pairs of mean (standard deviation) over the ten-fold cross validation in percent. Note that the Zero-Rule Baseline (ZRB) is task-specific.

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)

EXPLAINABILITY RESULTS

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

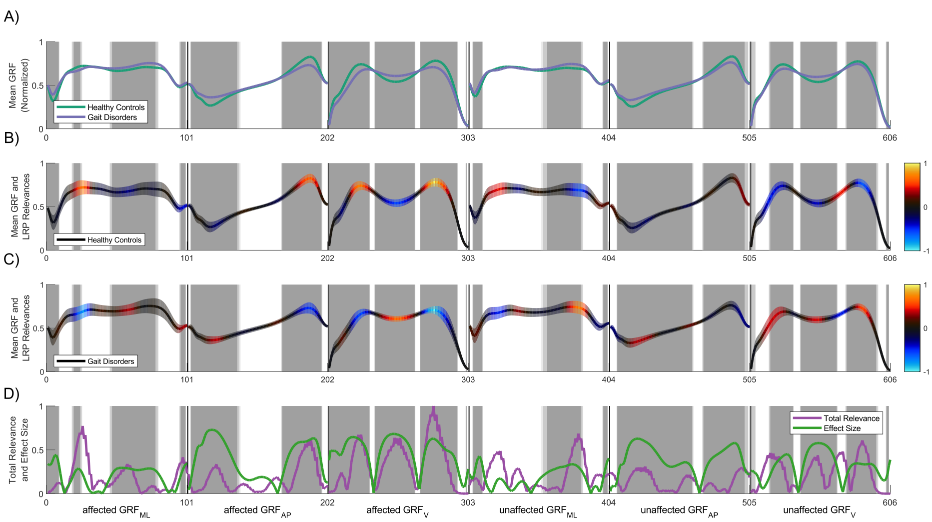

Fig. S1. Result overview for the classification of healthy controls and the aggregated class of all three gait disorders (*HC/GD*) based on min-max normalized GRF signals using a CNN as classifier.

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

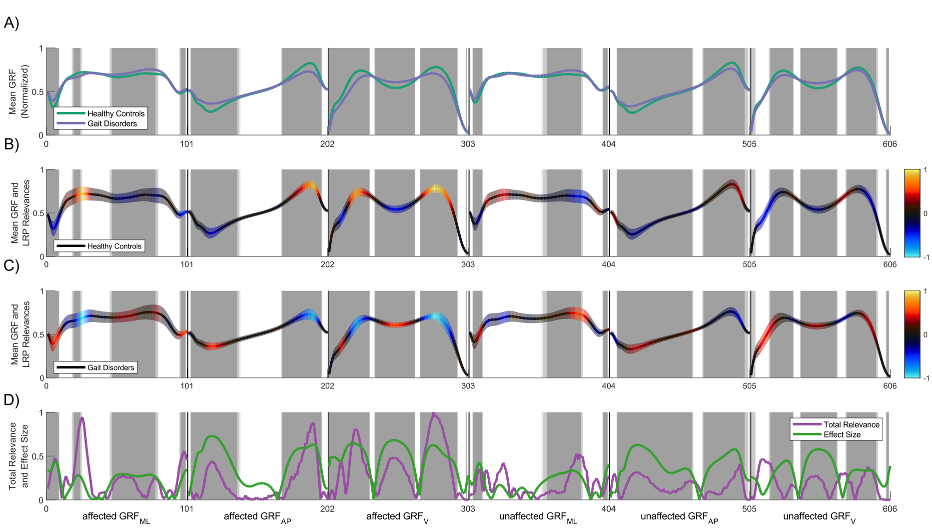

Fig. S2. Result overview for the classification of healthy controls and the aggregated class of all three gait disorders (*HC/GD*) based on min-max normalized GRF signals using an MLP as classifier.

Classification Task: HC/GD | Classification method: SVM

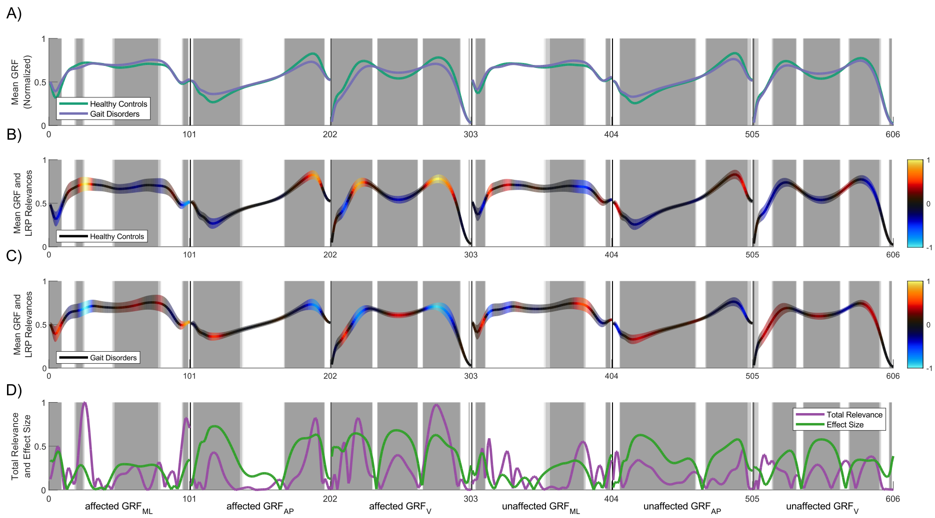

Fig. S3. Result overview for the classification of healthy controls and the aggregated class of all three gait disorders (HC/GD) based on min-max normalized GRF signals using an SVM as classifier.

Classification Task: HC/H | Classification method: CNN

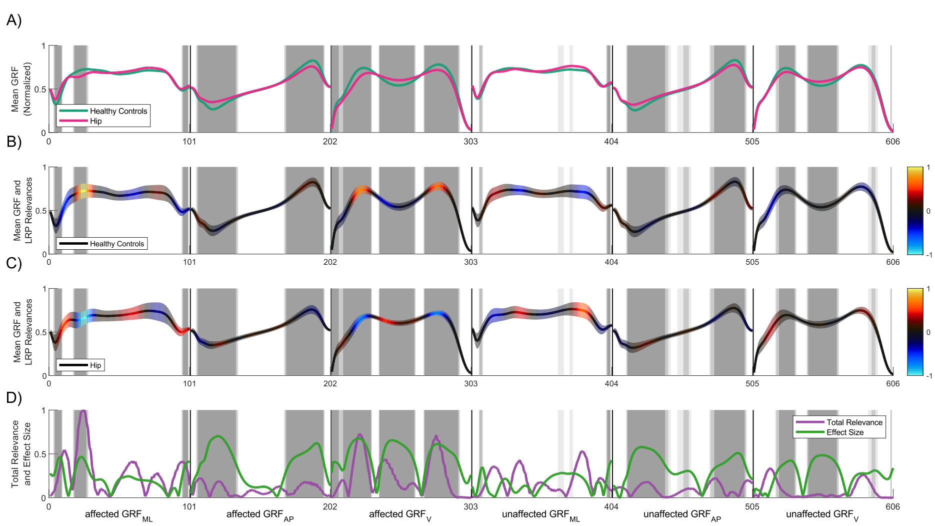

Fig. S4. Result overview for the classification of healthy controls (HC) and hip injury class (H) based on min-max normalized GRF signals using a CNN as classifier.

Classification Task: HC/H | Classification method: MLP

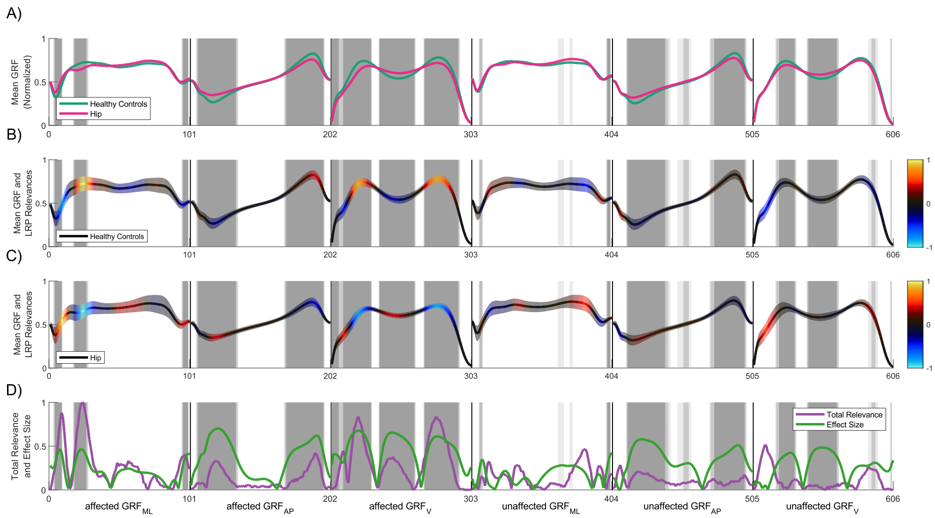

Fig. S5. Result overview for the classification of healthy controls (HC) and hip injury class (H) based on min-max normalized GRF signals using an MLP as classifier.

Classification Task: HC/H | Classification method: SVM

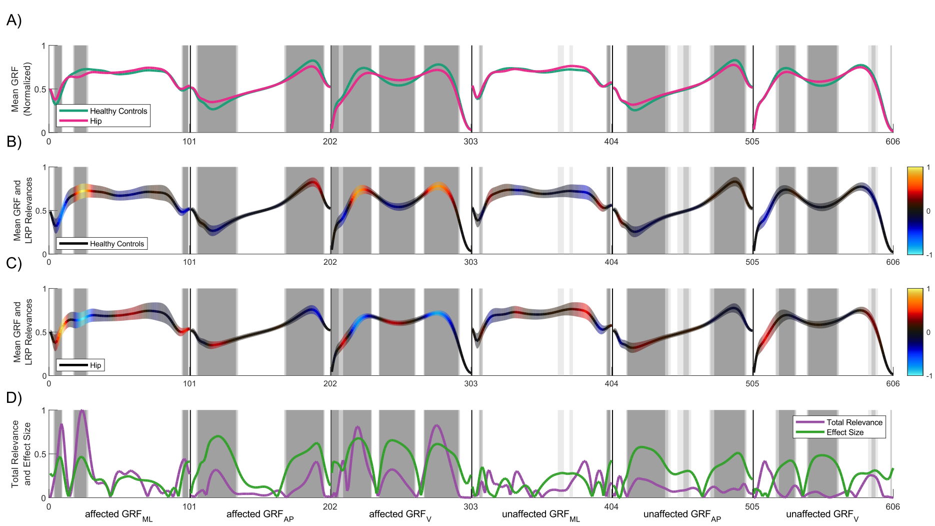

Fig. S6. Result overview for the classification of healthy controls (HC) and hip injury class (H) based on min-max normalized GRF signals using an SVM as classifier.

Classification Task: HC/K | Classification method: CNN

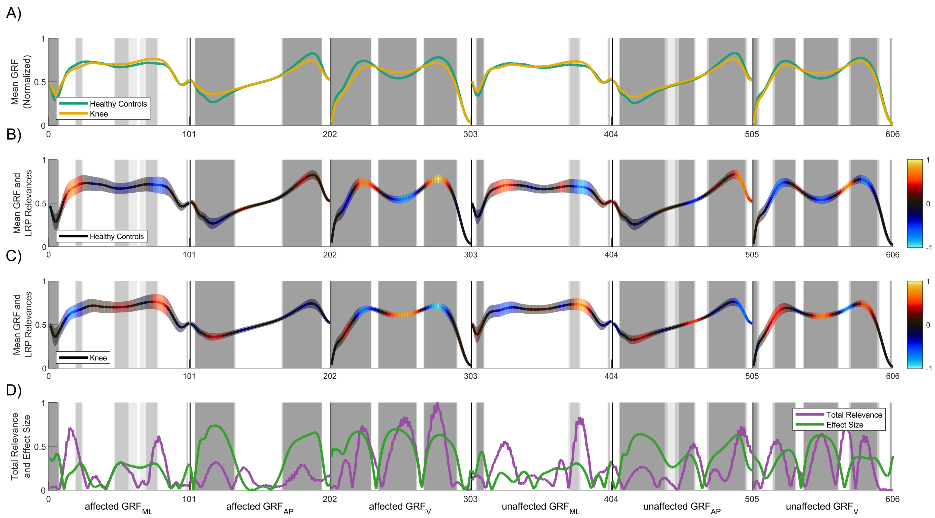

Fig. S7. Result overview for the classification of healthy controls (HC) and knee injury class (K) based on min-max normalized GRF signals using a CNN as classifier.

Classification Task: HC/K | Classification method: MLP

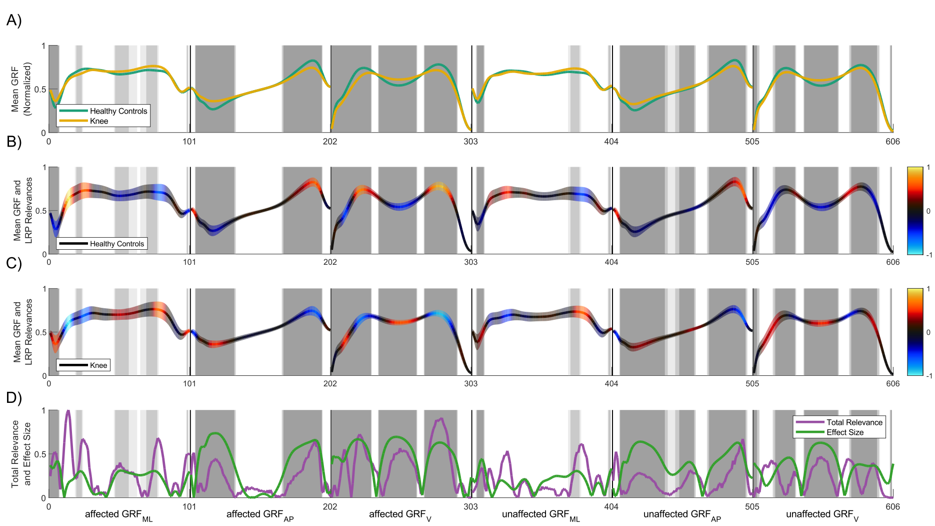

Fig. S8. Result overview for the classification of healthy controls (HC) and knee injury class (K) based on min-max normalized GRF signals using an MLP as classifier.

Classification Task: HC/K | Classification method: SVM

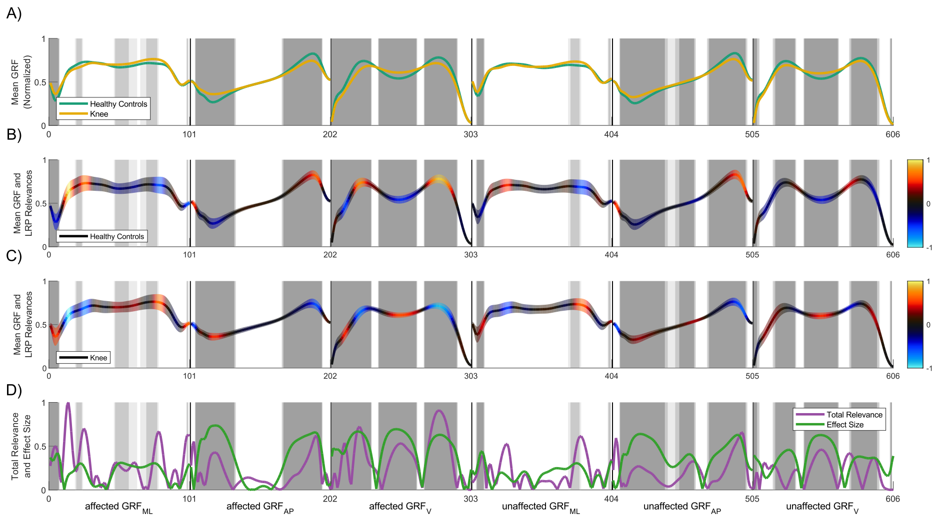

Fig. S9. Result overview for the classification of healthy controls (HC) and knee injury class (K) based on min-max normalized GRF signals using an SVM as classifier.

Classification Task: HC/A | Classification method: CNN

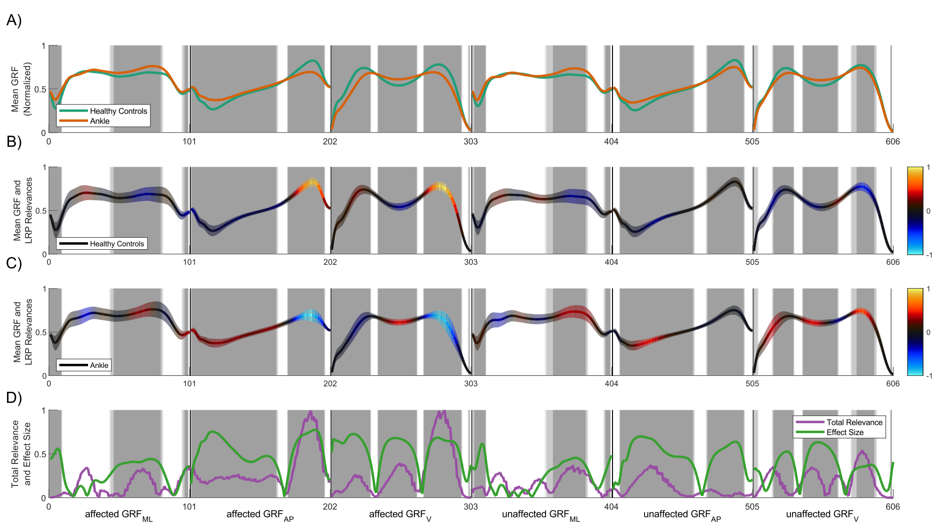

Fig. S10. Result overview for the classification of healthy controls (HC) and ankle injury class (A) based on min-max normalized GRF signals using a CNN as classifier.

Classification Task: HC/A | Classification method: MLP

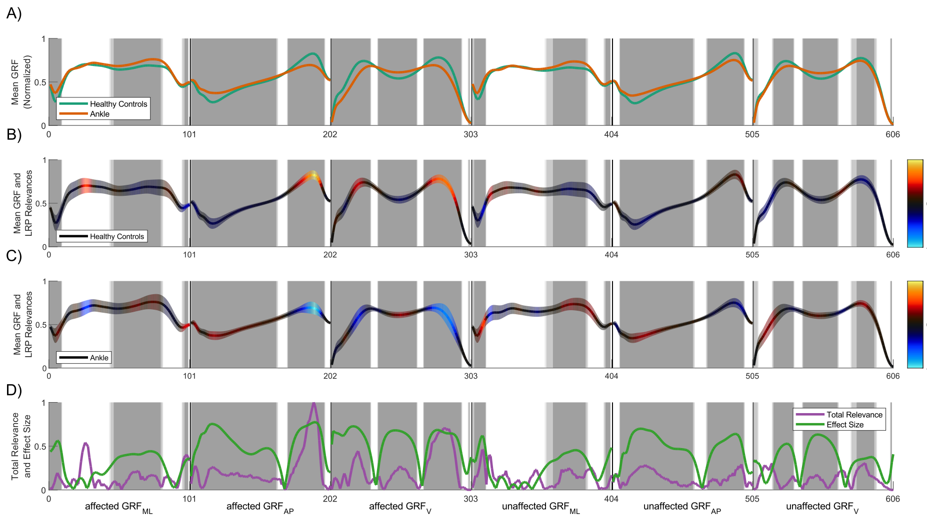

Fig. S11. Result overview for the classification of healthy controls (HC) and ankle injury class (A) based on min-max normalized GRF signals using an MLP as classifier.

Classification Task: HC/A | Classification method: SVM

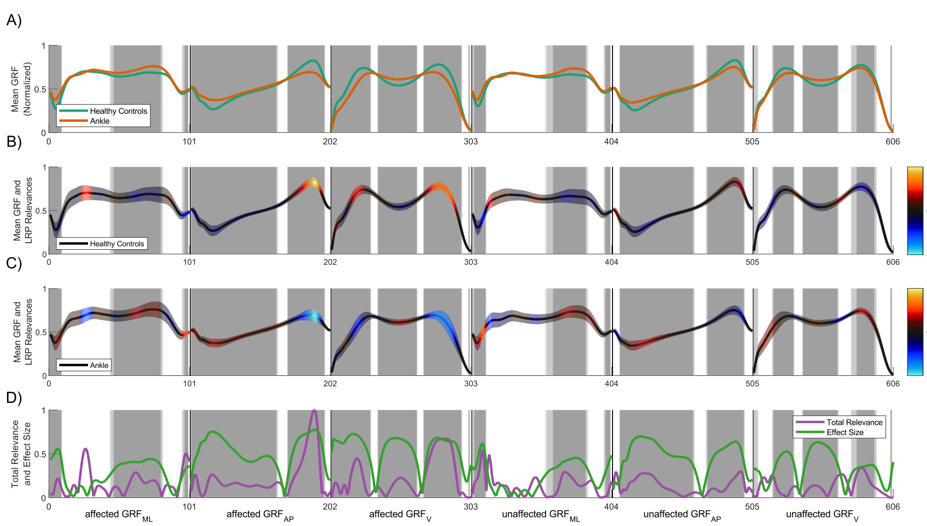

Fig. S12. Result overview for the classification of healthy controls (HC) and ankle injury class (A) based on min-max normalized GRF signals using an SVM as classifier.

EXPLAINABILITY RESULTS – NON-NORMALIZED DATA

Classification Task: HC/GD | Classification method: CNN

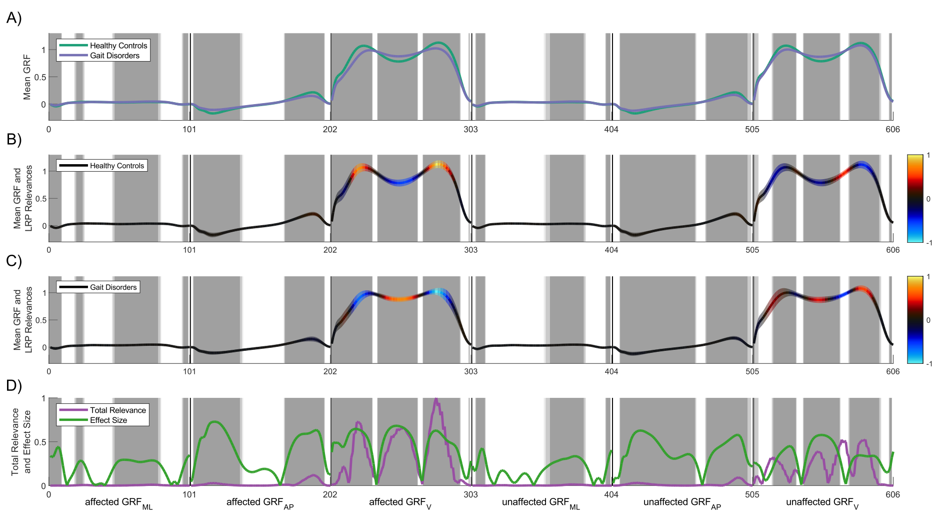

Fig. S13. Result overview for the classification of healthy controls and the aggregated class of all three gait disorders (HC/GD) based on non-normalized GRF signals using a CNN as classifier.

Classification Task: HC/GD | Classification method: MLP

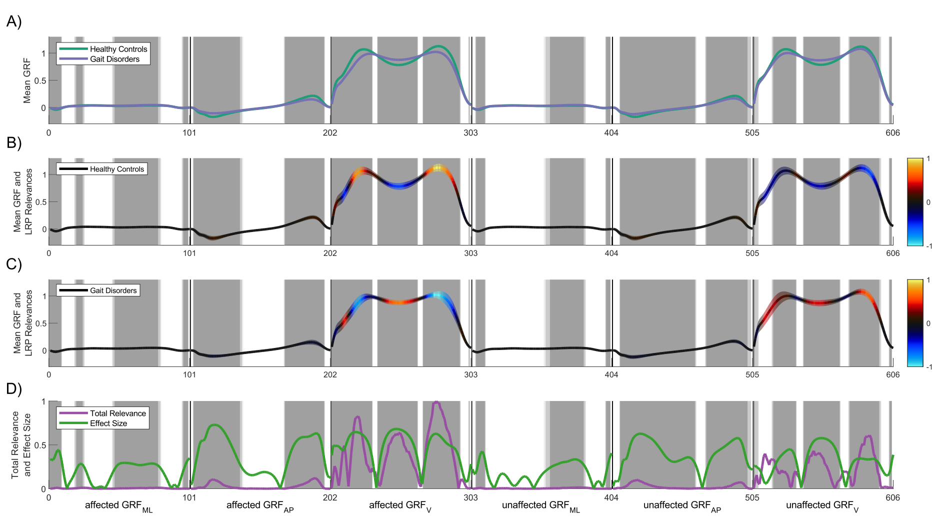

Fig. S14. Result overview for the classification of healthy controls and the aggregated class of all three gait disorders (HC/GD) based on non-normalized GRF signals using an MLP as classifier.

Classification Task: HC/GD | Classification method: SVM

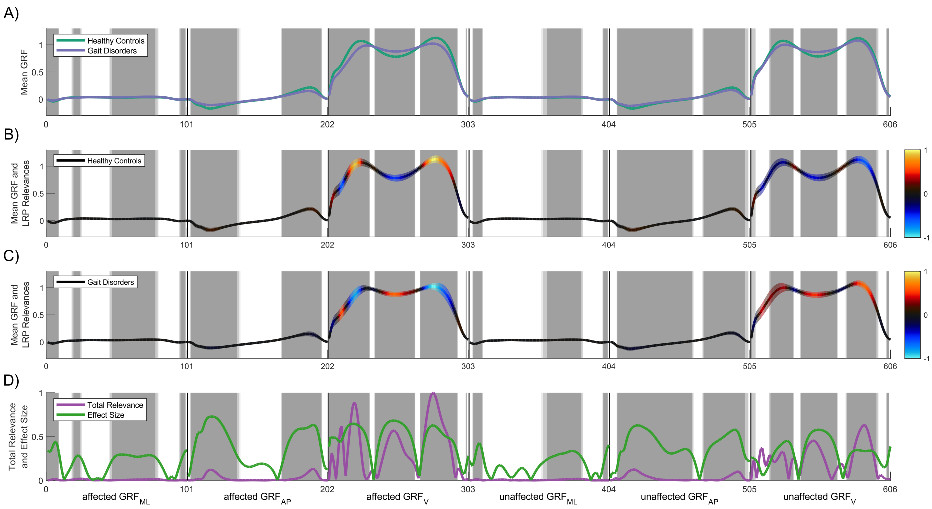

Fig. S15. Result overview for the classification of healthy controls and the aggregated class of all three gait disorders (HC/GD) based on non-normalized GRF signals using an SVM as classifier.

Classification Task: HC/H | Classification method: CNN

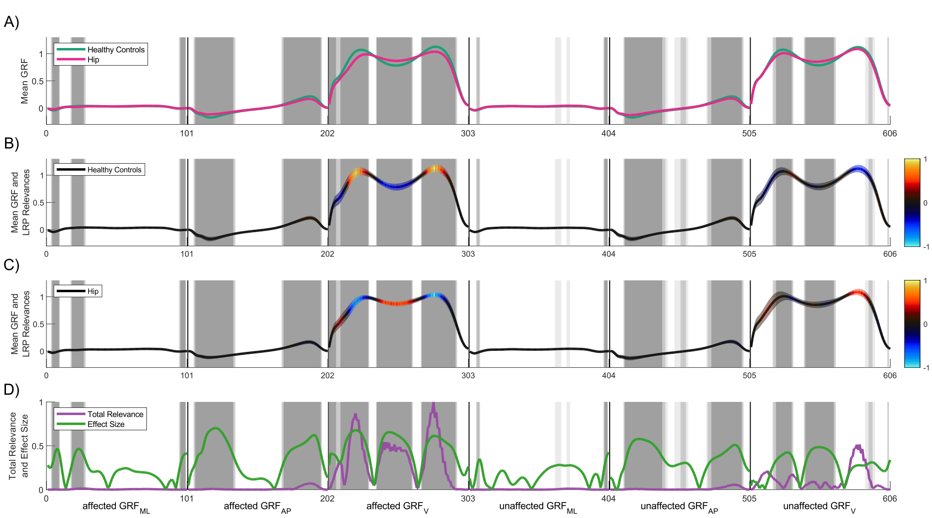

Fig. S16. Result overview for the classification of healthy controls (HC) and hip injury class (H) based on non-normalized GRF signals using a CNN as classifier.

Classification Task: HC/H | Classification method: MLP

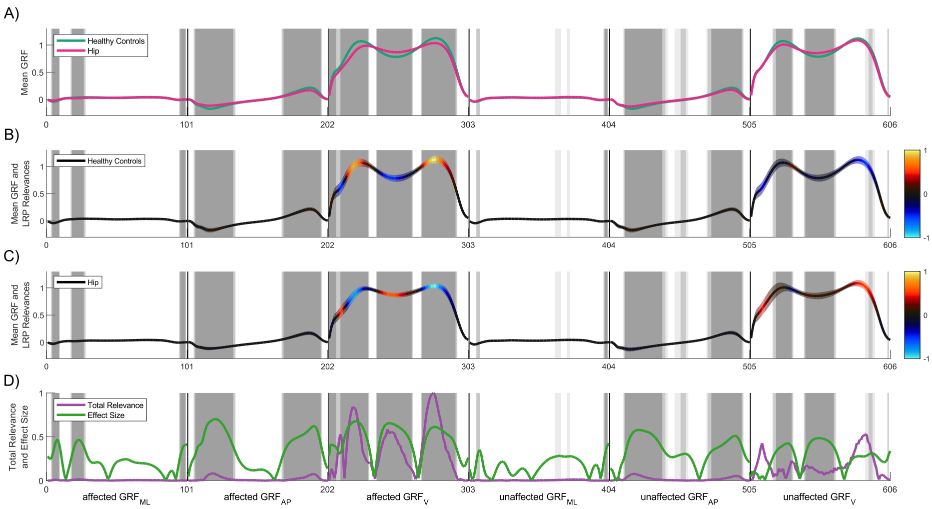

Fig. S17. Result overview for the classification of healthy controls (HC) and hip injury class (H) based on non-normalized GRF signals using an MLP as classifier.

Classification Task: HC/H | Classification method: SVM

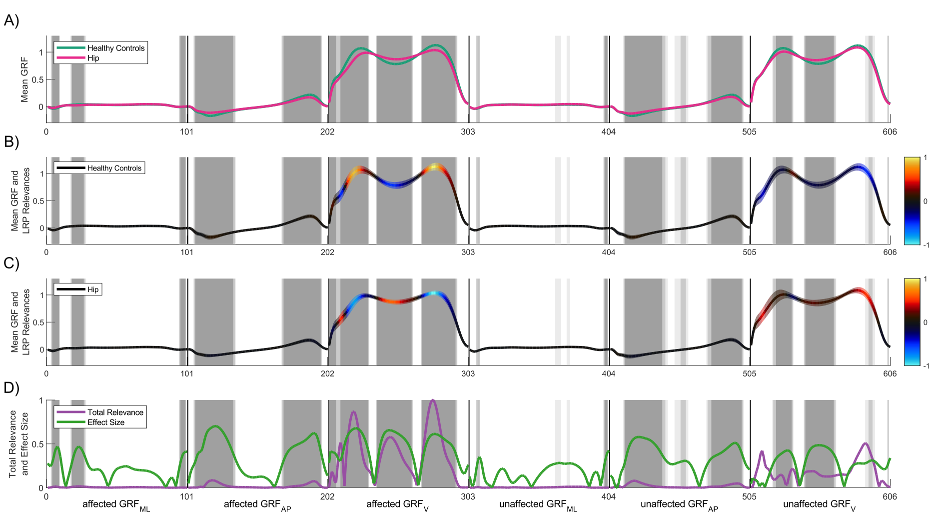

Fig. S18. Result overview for the classification of healthy controls (HC) and hip injury class (H) based on non-normalized GRF signals using an SVM as classifier.

Classification Task: HC/K | Classification method: CNN

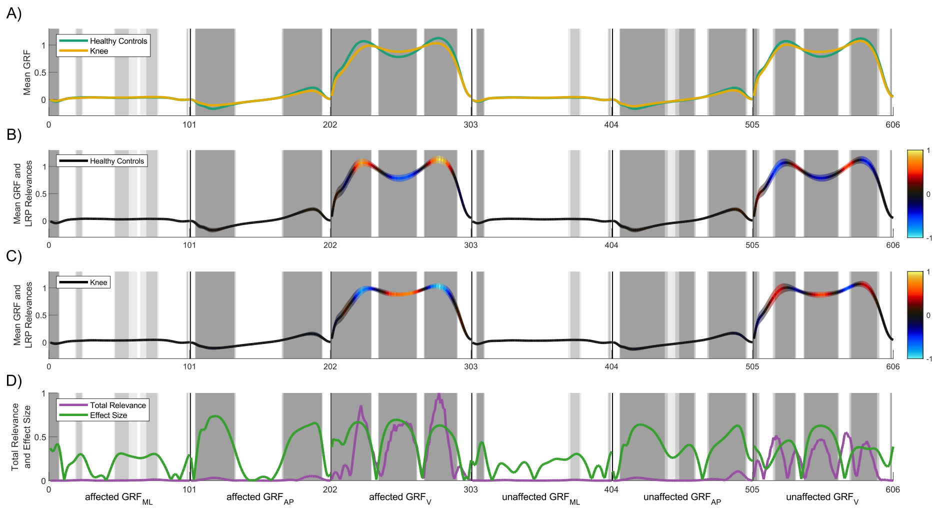

Fig. S19. Result overview for the classification of healthy controls (HC) and knee injury class (K) based on non-normalized GRF signals using a CNN as classifier.

Classification Task: HC/K | Classification method: MLP

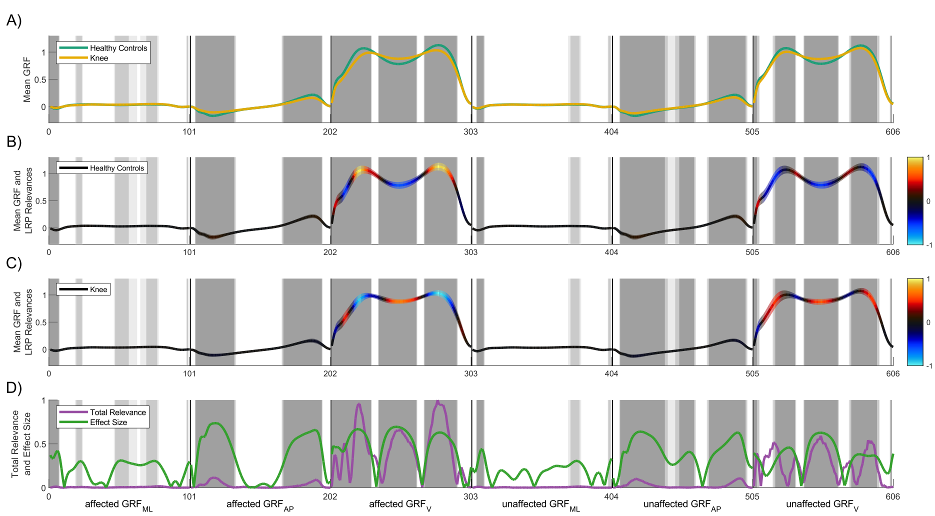

Fig. S20. Result overview for the classification of healthy controls (HC) and knee injury class (K) based on non-normalized GRF signals using an MLP as classifier.

Classification Task: HC/K | Classification method: SVM

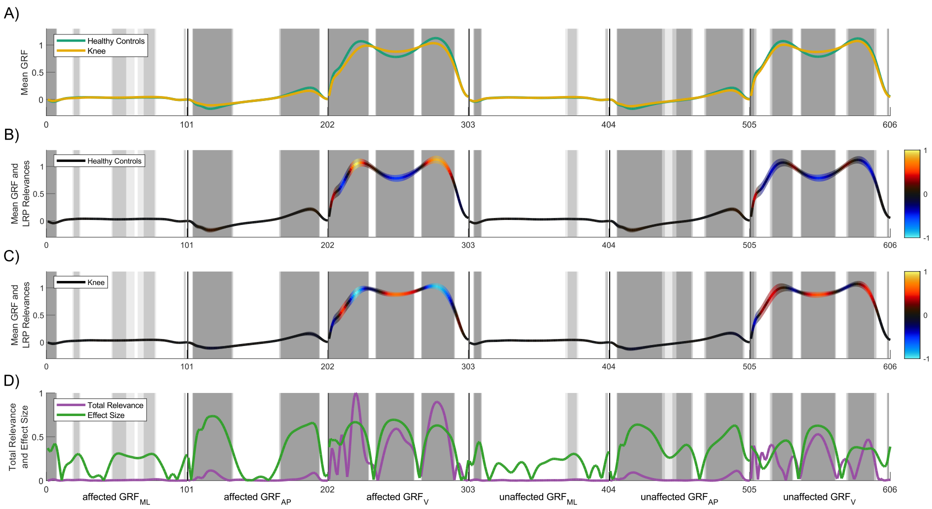

Fig. S21. Result overview for the classification of healthy controls (HC) and knee injury class (K) based on non-normalized GRF signals using an SVM as classifier.

Classification Task: HC/A | Classification method: CNN

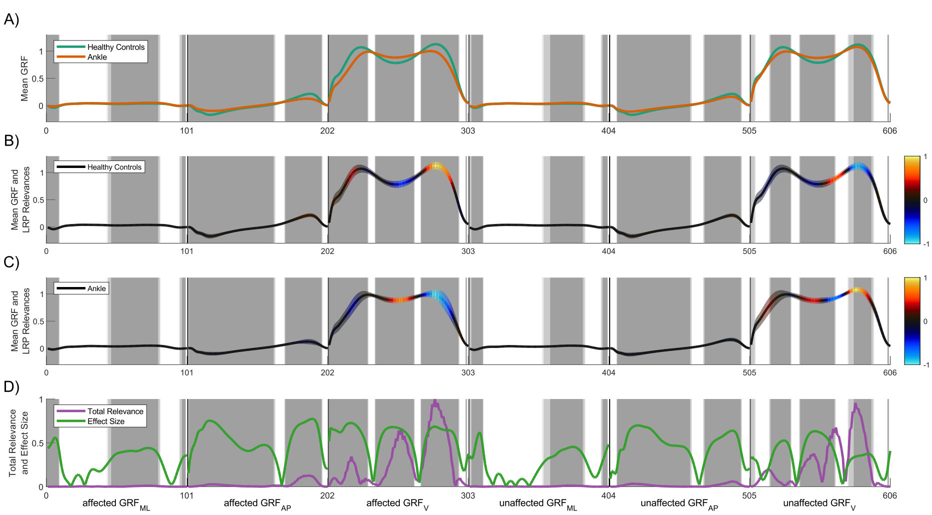

Fig. S22. Result overview for the classification of healthy controls (HC) and ankle injury class (A) based on non-normalized GRF signals using a CNN as classifier.

Classification Task: HC/A | Classification method: MLP

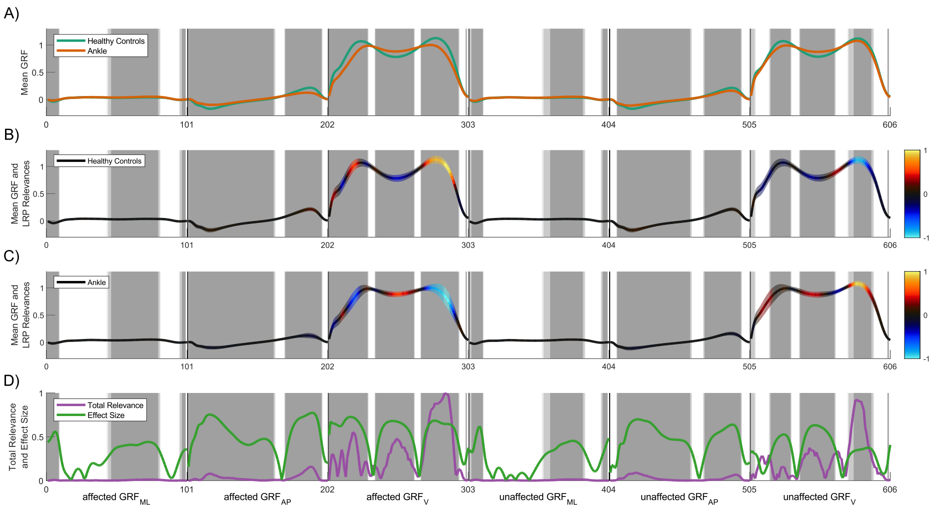

Fig. S23. Result overview for the classification of healthy controls (HC) and ankle injury class (A) based on non-normalized GRF signals using an MLP as classifier.

Classification Task: HC/A | Classification method: SVM

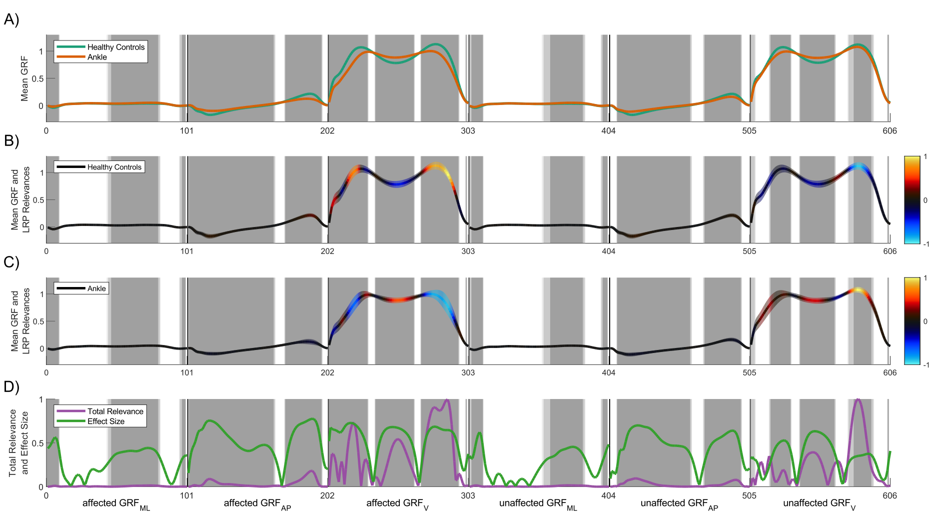

Fig. S24. Result overview for the classification of healthy controls (HC) and ankle injury class (A) based on non-normalized GRF signals using an SVM as classifier.