Trustworthy Visual Analytics in Clinical Gait Analysis: A Case Study for Patients with Cerebral Palsy

Alexander Rind*†
Inst. of Creative\Media/Technologies
St. Pölten Univ. of Applied Sciences

Djordje Slijepčević*†
Inst. of Creative\Media/Technologies
St. Pölten Univ. of Applied Sciences

Matthias Zeppelzauer†
Inst. of Creative\Media/Technologies
St. Pölten Univ. of Applied Sciences

Fabian Unglaube‡
Orthopaedic Hospital Vienna-Speising

Andreas Kranzl‡
Orthopaedic Hospital Vienna-Speising

Brian Horsak†
CDHSI & Inst. of Health Sciences
St. Pölten Univ. of Applied Sciences

ABSTRACT

Three-dimensional clinical gait analysis is essential for selecting optimal treatment interventions for patients with cerebral palsy (CP), but generates a large amount of time series data. For the automated analysis of these data, machine learning approaches yield promising results. However, due to their black-box nature, such approaches are often mistrusted by clinicians. We propose gaitXplorer, a visual analytics approach for the classification of CP-related gait patterns that integrates Grad-CAM, a well-established explainable artificial intelligence algorithm, for explanations of machine learning classifications. Regions of high relevance for classification are highlighted in the interactive visual interface. The approach is evaluated in a case study with two clinical gait experts. They inspected the explanations for a sample of eight patients using the visual interface and expressed which relevance scores they found trustworthy and which they found suspicious. Overall, the clinicians gave positive feedback on the approach as it allowed them a better understanding of which regions in the data were relevant for the classification.

Index Terms: Human-centered computing—Visualization—Visualization application domains—Visual analytics; Computing methodologies—Machine learning; Applied computing—Life and medical sciences

1 INTRODUCTION

Impairments in our ability to walk pose a major threat to participation in social activities and the labor market, and are therefore closely linked to quality of life. In children, one of the most common causes of physical disability is cerebral palsy (CP). It is diagnosed approximately in 2.5 per 1,000 births in developed countries [27] and constitutes a group of neurological disorders associated with varying symptoms such as tremors, muscle weakness, muscle stiffness, and spasticity [31]. These symptoms affect the child's motor functions, including the ability to walk. The most common causes for CP are brain lesions that occur shortly before or after birth. These lesions can result in musculoskeletal impairments, which develop and worsen throughout childhood and adolescence [18]. To provide the best possible care for children with CP, a comprehensive *physical examination and quantification of the patients' gait pattern* have become essential tools in the treatment of CP. *Three-dimensional gait analysis* is the gold standard for quantifying human movement and determining how musculoskeletal impairments affect a child's ability to walk, both in clinical and research settings [44].

Clinical gait analyses produce a vast amount of data with high-dimensionality, temporal dependencies, strong variability, non-linear relationships, and inter-correlations among variables [9]. This is especially true for patients with CP, who often exhibit high variability in their gait pattern, which can significantly complicate the interpretation of clinical gait analysis data [25]. The diversity in gait patterns and clinical representations has led researchers in the past to develop automated gait classification methods that can aid in diagnosis, clinical decision making, and communication [29]. These methods enable the classification of gait patterns into clinically meaningful groups, which can be distinguished from each other based on a set of defined (biomechanical) variables [15]. One of the most frequently used clinical classification schemes for CP is the one proposed by Rodda et al. [33]. Briefly explained, it allows to distinguish CP-related gait patterns in the sagittal plane such as drop foot, true equinus, jump knee, apparent equinus, and crouch gait depending on whether one or both sides of the body are affected (i.e., spastic hemiplegia vs. spastic diplegia). Accurate and objective classification of these patterns is critical, as this information is the basis on which clinicians make decisions about optimal treatment interventions. To support clinicians in diagnosis of CP-related gait patterns, great efforts have been undertaken in the past to develop automated classification algorithms [10, 13, 16, 29, 34, 45]. While these methods are promising for supporting clinicians in their daily routines, most of the developed approaches, however, employ complex classification methods which have a black-box character, making it impossible to understand why the automated algorithm made a particular classification.

Clinical experts have extensive experience with CP-related gait patterns from patient observation and the diagnostic literature, but they cannot effectively explain the functioning of a given black-box model. Transparency is, however, essential in the medical field. Its absence leads to a lack of trust in the algorithm, as it is not clear whether the model makes its predictions based on clinically relevant features or based on bias in the data (e.g., due to walking speed differences, marker misplacement, or other data collection issues). Thus, automated classification based on machine learning (ML) has not been widely used in clinical practice. To overcome this problem, this work combines explainable artificial intelligence (XAI) [1, 20] with visual analytics [2, 23, 40].

We present *gaitXplorer,* an explainability-enriched visual analytics approach for the classification of gait patterns, and demonstrate it in a case study with gait data of children with CP. We exploit the potential of modern ML techniques in combination with visualizations of clinical time series. To make the trained ML models more transparent, our approach highlights the regions in the patients' gait data that are particularly relevant for the prediction of the model. In addition, our approach allows for the comparison of

*Alexander Rind and Djordje Slijepčević equally contributed to this paper and are both to be regarded as first authors.
†e-mail: {firstname}.{lastname}@fhstp.ac.at
‡e-mail: {firstname}.{lastname}@oss.at

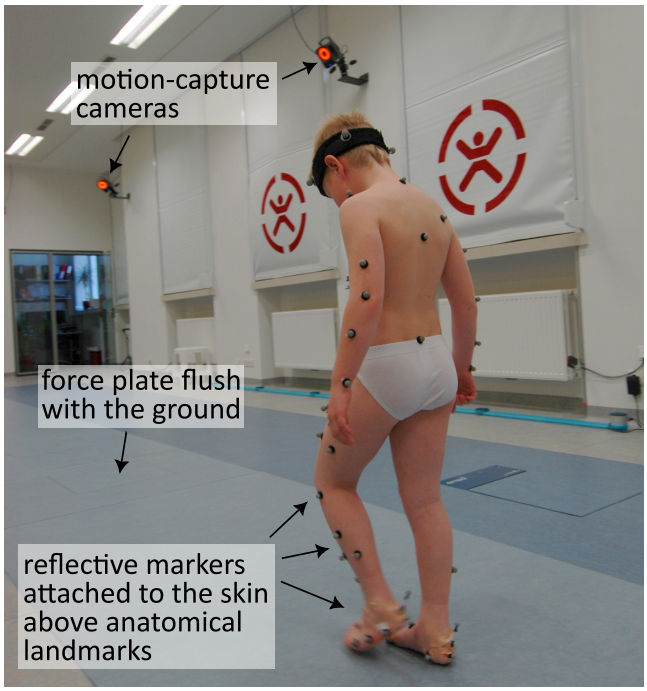

Figure 1: A typical gait analysis scenario. The patient is equipped with a marker set and walks up and down an approximately 10-meter walkway. A set of motion-capture cameras tracks the three-dimensional trajectories of the markers and force plates embedded flush with the floor measure the external ground reaction forces (GRFs). (Photo © Orthopaedic Hospital Vienna-Speising)

patient time series data to a variety of subgroups in the dataset, allowing clinicians to examine the entirety of their dataset that is used to train ML models. The developed visual analytics approach enables clinical experts to engage in joint human-machine reasoning and to compare the grounding of the automated classification methods with their expert knowledge. We conducted this case study in the course of a long-term collaboration with clinical gait experts from the Orthopaedic Hospital Vienna-Speising in Austria. It provides a unique opportunity to study explainability-enriched visual analytics approaches in a real-world setting and explore the effects it has on the trust of clinical experts in automated gait classification.

This work is structured as follows: After the Related Work in Sect. 2, we characterize and abstract the application domain of clinical gait analysis in Sect. 3. In Sect. 4, we describe our visual analytics approach and its underlying design decisions. Sect. 5 reports on the evaluation of the case study with our clinical collaboration partners on gait data from eight patients with CP. Sect. 6 discusses the clinicians' feedback, limitations, and future work and Sect. 7 provides a brief summary and conclusion of the presented work.

2 Related Work

Due to the ability to analyze a large amount of gait data in a cost-effective, fast, and objective manner, there is a growing interest in the use of ML in the field of gait analysis [17, 19]. ML methods have been successfully applied to patients with, for example, stroke [12], Parkinson's disease [42], osteoarthritis [26], or various functional gait disorders [37, 38]. Automatic classification of CP-related gait patterns, in particular, is a frequently addressed topic in the literature [29]. Ferrari et al. [16] compared multilayer perceptrons (MLP), support vector machines (SVM), and recurrent neural networks (RNNs) in a classification task with four self-defined CP-related gait patterns, and the latter showed the best performance. Zhang and Ma [45] examined seven machine learning algorithms (e.g., MLP, SVM, Random Forest (RF)) for the classification of CP-related gait patterns as defined by Rodda et al. [33], with MLPs performing best. For the same task, Darbandi et al. [13] used a fuzzy algorithm to translate the expert knowledge into rules and to perform fuzzy clustering. Sangeux, Rodda, and Graham [34] presented the plantarflexor-knee extension index, a non-ML-based approach for the same classification task. The index represents the distance of the patient's ankle and knee kinematics in mid-stance from normative data [34]. Chia et al. [10] proposed a decision support system consisting of two models using physical examination and kinematic data to identify CP-related impairments and surgical recommendations. They employed a stratified RF and leveraged feature importance to provide an explanation for the prediction.

Besides being relatively accurate, complex ML models have a limitation, namely their black-box character [1]. Thus, it is difficult to determine what an ML model has learned from the data or why certain decisions are made. As a result, even well-functioning ML models are rarely used in clinical practice [20].

To overcome this problem, Slijepčević, Horst et al. [36] proposed several XAI approaches based on Layer-wise Relevance Propagation (LRP) [4] to explain the functioning of ML models trained to classify different functional gait disorders. Dindorf et al. [14] utilized Local Interpretable Model-Agnostic Explanations (LIME) [32] to explain an ML model distinguishing between healthy controls and patients after total hip arthroplasty.

Interestingly, in the field of visualization and visual analytics, there is a relatively small body of literature that is focused on clinical gait analysis and related topics. In particular, the KAVAGait [41] approach supports exploring gait data in a clinical setting and interactively externalizing expert knowledge about gait pattern classifications. However, it utilizes only spatiotemporal parameters (e.g., walking speed). Three approaches visualize motion data of the upper extremities for rehabilitation: The visualization approach by Krekel et al. [24] shows kinematic data of shoulder and arm joints in a combination of three-dimensional views and time series plots. The Motion Browser [8] provides a time series visualization to analyze muscle activity patterns in an approach that integrates it with motion data and video information. The NE-Motion [11] approach supports assessing movement impairment and compensation of the upper extremities caused by stroke. It visualizes the relationships between joint angle data based on a graph learning method. In a veterinary setting, the FuryExplorer [43] approach visualizes motion capture data of horses for analysis in lameness recovery. Several visual analytics approaches, i.e., MotionExplorer [6], GestureAnalyzer [21], and MotionFlow [22], support exploratory search for motion sequences in a hierarchically clustered motion tracking database. Bernard et al. [5] also presented a visual-interactive approach for the semi-supervised labeling of human motion capture data. Thus, we could not identify any previous work on visual analytics with three-dimensional motion capture data collected in a clinical gait laboratory.

3 Problem Characterization and Abstraction

In clinical and research settings, gait performance is most frequently quantified using an opto-electronic motion capture system (Figure 1). These are very similar to the hardware used in the film and animation industry. A set of retro-reflective spherical markers are attached to the skin above bony landmarks. The three-dimensional coordinates of these markers are tracked by the motion capture system. Time-synchronized with these trajectories, external ground reaction forces (GRFs) are measured using a set of force plates that are flush with the floor. These data can then be used to calculate variables such as kinematics (e.g., knee flexion-extension angle), kinetics (e.g., knee flexion moment), and spatiotemporal parameters (e.g., walking speed, step length, step time) by utilizing a biomechanical model and

inverse dynamic calculations. Most of the kinematic and kinetic data are calculated for all three anatomical planes. To allow comparison in the time domain to other patients or to normative data, the time series in gait analysis are time-normalized to one gait cycle (0-100% gait cycle), defined as the time interval between successive initial foot contacts of the same foot with the ground. The data are summarized in a visual gait report and used to inform clinicians during decision making.

For this case study, we focus on the time series data for the lower body. In combination with spatiotemporal parameters and various indicators of gait quality, the time series form the most important basis for diagnosis. The clinical experts work with a total of 58 time series (kinematics and kinetics). These consist of

- the joint angles (degrees) of the pelvis, hip, knee, and ankle in all three anatomical planes ($n = 12$),
- the foot progression angle and the angle between the floor and the bottom of the foot ($n = 2$),
- the joint moments (Nm/kg) of the hip, knee, and ankle in three anatomical planes ($n = 9$),
- the power (W/kg) for the hip, knee, and ankle joints ($n = 3$),
- and the GRF (N or % body weight) in three anatomical planes ($n = 3$)

for the left and right sides ($29 * 2 = 58$). While the time series are measured in different units and exhibit diverging value ranges, they still follow typical patterns which are meaningful to clinicians.

The primary task of the clinical experts involved in the present case study is medical decision making. For this purpose accurate and objective classification of gait patterns is a prerequisite to offer optimal treatment strategies to the patient. To assist them in the process of data analysis, there is a strong interest in integrating automated analysis algorithms into their workflows. The clinical experts have annotated a large training dataset based on anonymized patient records and are collaborating with us to explore the utility of ML models to classify gait patterns into clinically meaningful categories. For both scenarios, medical decision making and ML model evaluation, the clinical experts need a visual analytics system that

- shows the patient's biomechanical data in a familiar and efficient way,
- allows comparison of the patient's biomechanical data with the data of specific patient groups,
- automatically suggests a gait pattern classification, and
- provides a detailed explanation for this classification in terms of the input gait measurement data.

4 Design and Implementation

Our prototypical visual analytics approach *gaitXplorer* provides access to a list of patients that have been newly classified and should be inspected by domain experts to confirm or override the automated gait pattern classification. It consists of a Python backend for data management, ML, and explainability as well as a web-based frontend providing the interactive visual interface.

4.1 Data & Data Management

We used a retrospective dataset comprising data from 257 children (355 affected legs) with CP, who could walk independently at self-selected walking speed, but had clearly identifiable gait abnormalities. All gait abnormalities were categorized by a clinically well-established procedure [34] into four CP-specific common gait patterns that served as the ground truth, namely: **true equinus**, **jump gait**, **apparent equinus**, and **crouch gait**.

The data is managed in the Python-based backend, which provides data access to the biomechanical data of the patients to be classified and to the group averages and standard deviations for each gait classification via a REST API.

4.2 Explainable Machine Learning

For the ML pipeline, we used a subset of the available time series that clinical experts mainly focus on, i.e., joint angles in all three anatomical planes of the pelvis, hip, knee, and ankle (only sagittal and transverse plane), as well as all three GRFs. Each of these signals has 101 time points, since they are time-normalized to one gait cycle (0–100% gait cycle). During a patient's measurement session, multiple gait cycles are recorded. We used the averaged signals per session and body side to account for gait variability within a person. After min-max normalizing each signal we concatenated all signals. Thus, for each "affected" leg of a patient we obtained a 1x1414 input vector. The ML model was trained to predict the classes at the level of individual legs.

As ML method we utilized a Convolutional Neural Network (CNN) model comprised of four consecutive one-dimensional convolutional layers, a flatten layer, a fully-connected layer with 512 neurons, and a softmax output layer with four output neurons. To counteract potential overfitting during training, we used alpha dropout layers prior to the last two fully connected layers with an dropout rate of 0.05. We chose the scaled exponential linear unit (SELU) activation function for all layers. The convolutional layers had the following properties: 64 feature maps, a filter size of three, and a stride of two. We trained the model using the Adam optimizer and the categorical cross-entropy loss function.

To explain and examine the internal functioning of the trained model we integrated Grad-CAM [35], a explainability algorithm commonly used for visual data and adapted it to one-dimensional data. Grad-CAM provides explanations for a certain prediction based on the more abstract features learned in the last convolutional layer. The Grad-CAM explanation for an input sample highlights local regions in the input vector that are strongly relevant for the predicted class. This relevance (grounding) for a particular automated prediction is visualized in the interactive visual interface.

The implementation of the ML method and the Grad-CAM method was conducted within the software framework Python 3.7 (Python Software Foundation, USA) and TensorFlow 2.4 (Google Inc., USA).

4.3 Interactive Visual Interface

The interactive visual interface (Fig. 2) consists of a patient list in the top left, a configuration panel in the bottom left, and patient-specific information in the right, occupying about 75% of the screen.

The patient list (Fig. 2.a) shows the list of new patients with their automatically predicted gait pattern classification. Clinicians can use the list to navigate between patients. In the upper center (Fig. 2.b), the patient's identifier, the examination date, and their walking speed are shown. Below, clinicians can select a gait classification and possibly override the automated classification. Note that updating the dataset and the ML model with the newly labeled patients will be in the scope of future work.

Time series in detail The largest part of the screen is used to provide detailed information about the patient's time series (Fig. 2.d) based on the established gait report of our clinical collaborator. Each left/right pair of time series is displayed in a line plot with the common time axis (x-axis) from 0% to 100% gait cycle. The y-axis scales are displayed in the physical units of the respective time series (e.g., degree for angles). By default, they are scaled to an established value range that includes the values for a broad range of patients. Additionally, the line plots can be zoomed to the value range of the current patient. The colors blue and red are consistently used for the left and right body side. As within a patient's measurement session, multiple gait cycles are recorded, by default the averaged time series are visualized. However, it is possible to fade in the individual trials (gait cycles), i.e. each step the patient took during the examination (Fig. 3.a). Thin vertical support lines indicate events in the gait cycle when the foot and the opposite foot touch (initial contact) and leave

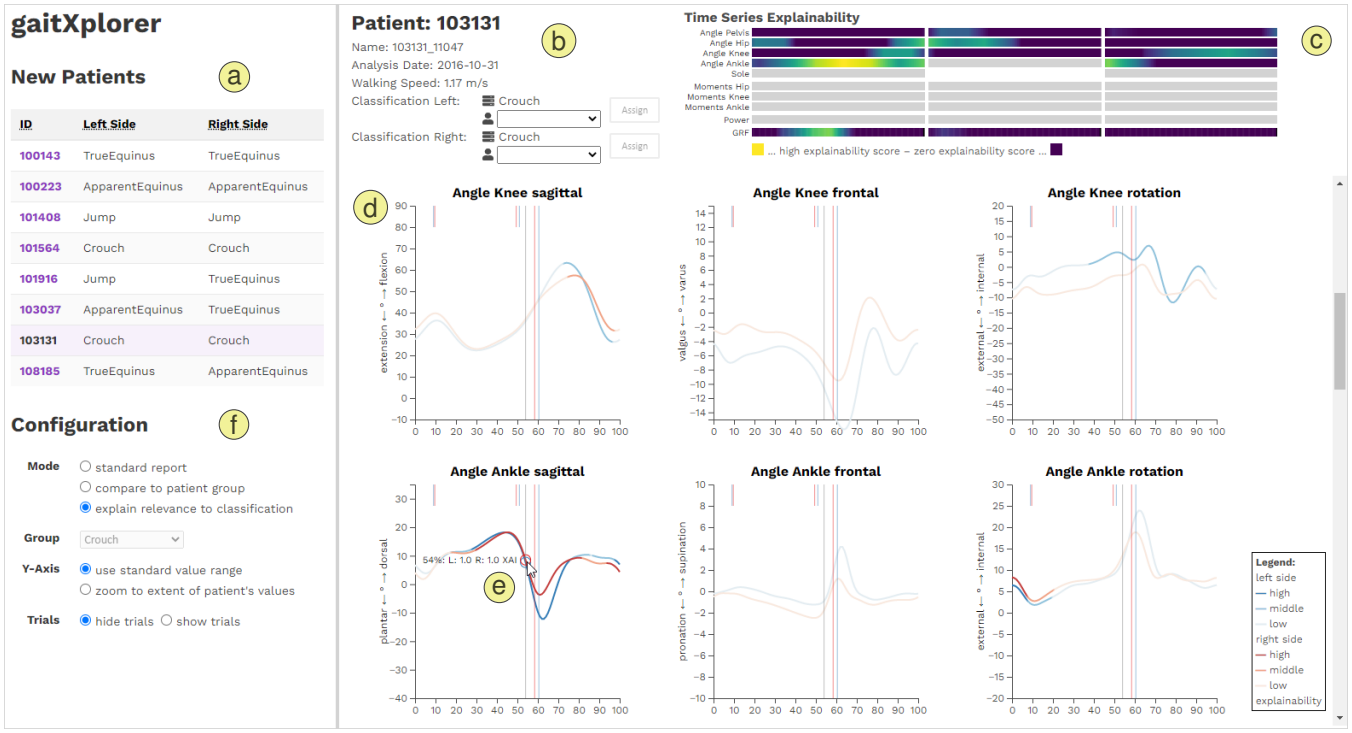

Figure 2: Visual interface showing a patient's data in explainability mode: (a) list of new patients that need to be classified; (b) master data of the current patient including controls to select a gait classification; (c) compact overview of the patient's time series with color encoding that indicates the relevance (grounding) for the automated prediction; (d) time series details as line plots with color intensity according to their relevance (discretized into three levels, i.e., high, middle, and low); (e) cursor displaying the relevance score at the current mouse position; (f) controls to switch between modes and change settings.

the ground (toe-off). The line plots are arranged in a matrix with three columns for the three anatomical planes and ten rows grouped by biomechanical variable (i.e., joint angles, moments, powers, and GRFs) and body part (e.g., knee, hip, and ankle). As stated above, the line plot idiom, the colors used, and the matrix arrangement are based on the established gait report in order to fit well with the clinicians' work practices. We refer to this configuration as the standard mode.

To integrate the annotated dataset and the ML model, we extended the line plots of the standard report with two novel configurations, the explainability mode and the group comparison mode.

The explainability mode (Fig. 3.c) changes the color intensity of the line plot segments based on the results of the Grad-CAM algorithm. For this, the relevance scores are first discretized to three ordinal bins (low $< 1/3$, $1/3 \leq$ middle $< 2/3$, high $\geq 2/3$) and then used as color gradient for the line plots. In order to maintain the blue/red color convention for body sides and consistent color intensity, colors from ColorBrewer's "RdBu" diverging color scheme [7] are used. We assume that binned relevance levels provide a sufficient level of detail for clinicians and this is in line with design guidelines from social sciences [28] that explanation information should be simple. In contrast, color gradients from continuous relevance scores are hard to comprehend and compare. We also experimented with various multi-hue color schemes from D3 but these made it harder to distinguish left and right side. Since color encoding of relevance scores was limited to three bins and a spatial encoding can allow for more perceptually effective reading of the explainability information, we prepared an alternative design that represents continuous relevance scores as superimposed area charts in the lower sixth of the line plot (Fig. 3.d). Furthermore, an interactive cursor (Fig. 2.e) prints the relevance scores for left and right at the current mouse position.

The group comparison mode (Fig. 3.b) superimposes aggregated information of the patient time series labeled with a certain gait classification. The average is shown as a purple line plot and the range of average plus-minus one standard deviation are shown as a purple band.

Time series overview Since the line plots require space for detailed inspection, it is not possible to show them all at once and scrolling is needed. To provide a compact overview of all time series and guide the clinician towards interesting parts of the report, we added a time series heatmap into the top right corner (Fig. 2.c). Each stripe represents a left/right pair of time series for the 29 relevant combinations of biomechanical variable, body part, and anatomical plane. The arrangement of time series is consistent with the matrix of line plots below and the established gait report. By clicking on a time series stripe, the detailed report scrolls to the corresponding line plot.

While the line plots show two time series for both legs, each stripe of the compact overview plot displays a single time series that should represent how interesting this line plot is. The calculation of this aggregated time series depends on the report mode: In the explainability mode, the overview plot presents the maximum of the left and right relevance score for each time point (Fig. 2.c). In the group comparison mode, the interesting time points have a large difference to the group average relative to the group's standard deviation. We calculate the maximum of the absolute values of the z-scores for the left and right side at each time point (Fig. 4). In the standard mode, asymmetry between the legs is assumed to be interesting [41], which we quantify as the difference between left and right relative to the total extent of values in this stripe.

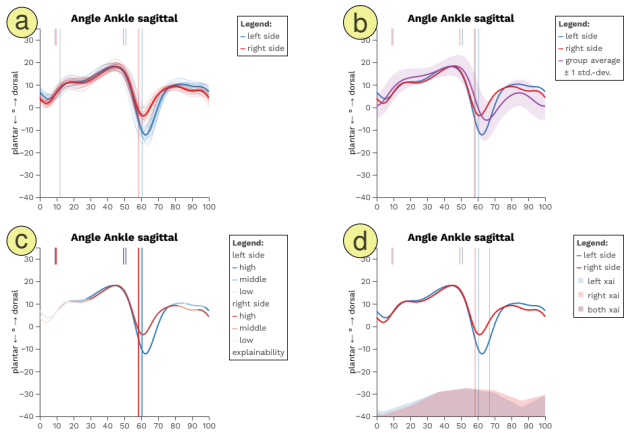

Figure 3: Line plots for detailed inspection of time series (sagittal ankle angles for patient 103131 of the "*crouch*" gait pattern): (a) standard mode showing the average (thick lines) and all trials (thin lines); (b) comparison to the average and standard deviation of all patients with crouch gait in the dataset; (c) relevance level (high, middle, or low) indicated by color intensity; (d) relevance scores shown as superimposed area chart in the lower part of the plot.

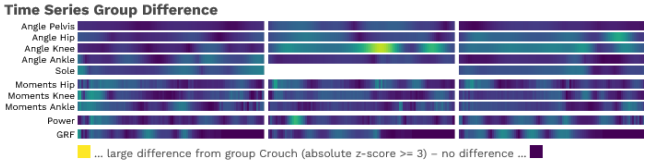

Figure 4: Overview plot showing group differences (comparing patient 103131 to the crouch gait class).

The heatmap uses the "viridis" sequential color scheme by van der Walt, Smith, and Firing [7] for all modes. An alternative design with distinct color schemes for each mode was shown and discussed in the case study.

In the bottom left part is a configuration panel (Fig. 2.f) to switch modes, select a patient group for comparison, zoom the y-axis, and show individual trials, i.e. each step recorded during the examination. The visual interface was implemented using TypeScript 4.3, D3 7.0, Svelte 3.49, Bulma 0.9, and Font Awesome 5.15 icons.

5 Case Study

We investigated the applicability of this visual analytics approach in a case study based on annotated patient data from our collaboration partner and collected their feedback in a series of focus group interviews. The case study builds upon a long-term, iterative collaboration between the involved clinical experts and the researchers, especially in the field of ML. The two clinical experts, who are also coauthors of this work, have formal education in sport science and training and over 29 years and seven years, respectively, of practical experience in clinical gait analysis.

As described in Sect. 4.2, the ML model was trained on a dataset comprising data from 257 children with CP (355 affected legs). To simulate new unlabeled data for demonstration in the focus groups, we drew a random subsample of eight patients so that two samples were available for each class (only the left legs were considered during the selection process). In addition, for this subsample we selected only samples that were correctly classified by the ML model. This information was of course not shared with the clinicians in advance.

For the design iteration described in the work at hand, we conducted three focus group interviews of about one hour over the video conference platform MS Teams. We demonstrated the visual interface over MS Teams' screen sharing feature based on the pair analytics method [3]. The clinical experts gave spoken inputs to navigate in the data visualization and the facilitator (i.e., the visual analytics expert) interacted with the tool accordingly. In these interviews we discussed the usability of the visual interface and how such a visual analytics tool can be integrated into the clinical workflow. In addition, the clinical experts received screenshots of the eight patients showing all data visualizations (cp. supplemental material) and were asked whether these explanations made a plausible or suspicious impression. We also discussed how the visual interface affects their trust in the automated classifications.

Visual Interface Design The clinicians found the visual interface clear and intuitive. They appreciated the reuse of the line plot style and matrix arrangement from their established layout.

Showing the explainability results in ordinal bins of high, middle, and low relevance level was regarded as sufficient and the color gradients were confirmed as readable. The clinicians emphasized the need for a color legend, which needs to clarify in particular how regions with a zero relevance score are handled. They pointed out that the color gradients are not readable when the two line plots for the left and right side overlap (e.g., Fig. 3.c). The interactive cursor showing the relevance score at the mouse position served as a workaround for such regions. The alternative design using area plots for relevance scores (Fig. 3.d) made a good first impression but on closer inspection the superimposed area plots with areas of mixed colors were hard to interpret. While discussing various further alternatives the clinicians expressed their preference for the color gradient rather than adding more visual elements to the screen.

A stable color scheme for the compact time series overview was preferred by the clinical experts over three color schemes for each mode. The clinicians expressed that familiarizing with different color schemes poses more burden than the concern of confusing the current mode. In addition, the selected mode can be easily recognized from the detailed line plots. They asked for the stripes to be larger, so that they would be easier to discern and click on. Thus, after the first interview the stripe height was increased from 5 to 10 pixels and tiny labels determining the biomechanical variable and body part were added. Six rows of time series, angle speeds and angle acceleration, were removed because they are not needed for analysis of patients with CP.

Explanations & Trust The two clinical experts independently inspected data visualizations of eight patients and assessed their trust in the automated classification based on the displayed relevance scores. All the visualizations can be found in the supplemental material and can be looked up using the 6-digit patient identifier.

Over the eight patients and 16 legs, the clinicians stated they would rather trust the explanations for both legs of two patients (101408, 101564) and for one leg of two patients (101916 left, 108185 right). For one patient the clinicians gave diverging answers (100223). In this case, both clinicians considered the relevant regions to be plausible, but one of the clinicians expected higher relevant regions in the sagittal knee angle. For the remaining legs (100143, 103037, 103131, 101916 right, 108185 left), the clinicians were distrusting the explanations (Table 1).

Summarizing the feedback, there were two types of suspicious patterns in the explanations: On the one hand, certain time series regions had a lower relevance score than the clinicians expected. For example, patient 103131's gait (Fig. 2) is classified as crouch gait which has a characteristic behavior in the sagittal knee angle, but that time series contains only low to middle relevance scores. On the other hand, the relevance score was high for time series regions

Table 1: Plausibility assessment of explanations.

patient id	left	right
100143	suspicious	suspicious
100223	diverging	diverging
101408	plausible	plausible
101564	plausible	plausible
101916	plausible	suspicious
103037	suspicious	suspicious
103131	suspicious	suspicious
108185	suspicious	plausible

that the clinicians did not regard as characteristic for the predicted class. Low relevance scores were more frequently criticized. Often too low relevance scores in some regions (e.g., sagittal knee or ankle angles) were accompanied by too high relevance scores in other (for clinicians unexpected) regions. Samples classified as true equinus exhibited often this suspicious pattern, e.g., both legs of patient 100143, classified as true equinus, had high relevance scores for the expected frontal knee angle but low relevance scores for the sagittal ankle angle. The clinicians expected relevant regions in the sagittal ankle angle, as this is the most important biomechanical variable for characterizing true equinus. However, for the true equinus class, the ML model identified greater differences in sagittal knee angle compared to the other CP-related gait patterns, which were sufficient for classification.

For all legs classified as jump gait (101408 both, 101916 left), the explanations were rated as plausible. This is notable as explanations for this class contained much larger regions of high relevance scores than explanations for other classes. Even though this is clearly visible in the compact time series overview plots with more yellow/ green regions, the clinicians did not consider this as a sign for distrust.

The clinicians expressed certain concerns about the ML model. In summary, confidence in the ML model was more affected by missing relevance scores for regions where they expected them than for regions that had high relevance scores but were not expected. Additionally, they emphasized that a visual interface that makes the algorithm more transparent and provides such detailed insights into its functioning is essential to gain trust in the use of ML-based classification in the clinical setting.

Integration in Workflows The clinicians stated that the group comparison mode is good and an interesting addition to their standard report. However, the most important addition was the inclusion of explainability and visualization thereof, which enabled the evaluation of the functioning of the automatic classification algorithm. The clinicians clearly stated that this should be a part of every clinical gait analysis that utilizes ML in order to achieve the best possible treatment outcome for the patient.

Reflecting on the analysis of the patient visualizations (cp. supplemental material), the senior clinician described his workflow as follows: First, he looked at the detailed line plots from top to bottom. Normally, he would use their standard report; here he took the explainability mode ignoring color gradients. Based on the time series data, he decided on a CP-related gait pattern. Finally, he looked at the relevance scores and compared them to his decision and expectations. He did not consult the group comparison mode because he is familiar with the time series for the different gait groups in the dataset. He neither used the compact time series overview, because he always reads a patient report from top to bottom.

6 Discussion & Future Work

An interesting observation from the focus group was that there was a certain discrepancy between clinicians' expectations and the signal regions actually used by the ML model to make its predictions (i.e., the visualized relevance scores). The clinicians expected the ML model to use all regions unique to the specific class (and thus regions where the samples of one class differ from the other classes), but actually the ML model seemed to use only a subset of all these regions. This is evident in the samples classified as true equinus, where the model used regions in the sagittal hip and knee angles, but the clinicians expected the model to use primarily the sagittal ankle angle, as this is considered the main biomechanical variable for this gait pattern. The expectation that relevant regions include all clinically relevant variables may also be fostered by clinicians' experience with methods such as statistical parametric mapping (SPM) [30], a method to identify statistically significant differences between two different patient groups, but which cannot be used for automated classification.

The explainability-enriched visual analytics approach still has limitations. First, there is no feedback mechanism that would allow the clinicians to externalize their domain knowledge on relevant regions and thus influence the training of the ML model. A natural evolution of this work will be to integrate such feedback into future iterations of this approach. Second, the Grad-CAM algorithm determines its explanations solely from the relevance of input data to the predicted classification. There is a broad landscape of alternative explanation approaches [1]. For example, counterfactual explanations could point out how much a gait pattern would need to change until it is classified differently. It will be a fruitful area for further work to investigate how clinicians react to counterfactual explanations or a combination with relevance explanations. Third, most patients in clinical practice suffer from two or more gait abnormalities affecting also the frontal and transversal plane (e.g., true equinus and valgus deformity). To incorporate overlapping gait abnormalities, a multi-label ML approach and set-visualization techniques can be incorporated. Finally, the case study evaluation has the limitation that the two domain experts were already involved in the conception of the visual analytics approach. Therefore, studies with additional clinicians are needed in the future. However, to broaden the user base, onboarding mechanisms [39] need to be provided. These mechanisms would support clinical experts in learning to use the visual analytics tool and to extract information from the visualizations.

7 Conclusion

In this work, we present *gaitXplorer*, a visual analytics approach for clinical gait analysis of patients with CP that we developed and evaluated with two clinical experts. While we focus on CP as a case study, the visual analytics approach is applicable to clinical gait analysis in general, given that labeled training data is available. Even beyond the clinical context, the results can be generalized to other visual analytics scenarios with multiple interrelated time series. Our approach can serve as a reference for the integration of an explainability algorithm into a ML-based visual analytics approach. Overall, this case study strengthens the idea that visual analytics approaches, which integrate an explainability algorithm and embed its explanations into the interactive visual interface, are essential to gain confidence in the use of ML-based classification in clinical settings.

Acknowledgments

This work was partly funded by the Austrian Research Promotion Agency (FFG, #866855), by the Austrian Science Fund (FWF): P33531-N, as well as by the Gesellschaft für Forschungsförderung NÖ (Research Promotion Agency of Lower Austria) and the Provincial Government of Lower Austria within IntelliGait3D (#FTI17-014) and within the Endowed Professorship for Applied Biomechanics and Rehabilitation Research (#SP19-004).

REFERENCES

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

[2] N. Andrienko, T. Lammarsch, G. Andrienko, G. Fuchs, D. Keim, S. Miksch, and A. Rind. Viewing visual analytics as model building. *Computer Graphics Forum*, 37(6):275–299, 2018. doi: 10/gdv9s7

[3] R. Arias-Hernandez, L. Kaastra, T. Green, and B. Fisher. Pair Analytics: Capturing reasoning processes in collaborative visual analytics. In *Proc. 2011 44th Hawaii International Conference on System Sciences*, pp. 1–10, 2011. doi: 10.1109/HICSS.2011.339

[4] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015. doi: 10.1371/journal.pone.0130140

[5] J. Bernard, E. Dobermann, A. Vögele, B. Krüger, J. Kohlhammer, and D. Fellner. Visual-interactive semi-supervised labeling of human motion capture data. *Electronic Imaging*, 29:34–45, 2017. doi: 10.2352/ISSN.2470-1173.2017.1.VDA-387

[6] J. Bernard, N. Wilhelm, B. Krüger, T. May, T. Schreck, and J. Kohlhammer. MotionExplorer: Exploratory search in human motion capture data based on hierarchical aggregation. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2257–2266, Dec. 2013. doi: 10.1109/TVCG.2013.178

[7] M. Bostock and al. d3-scale-chromatic: Sequential, diverging and categorical color scales. GitHub, 2021. `https://github.com/d3/d3-scale-chromatic`, accessed Jul 19, 2022.

[8] G. Y.-Y. Chan, L. G. Nonato, A. Chu, P. Raghavan, V. Aluru, and C. T. Silva. Motion Browser: Visualizing and understanding complex upper limb movement under obstetrical brachial plexus injuries. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):981–990, 2019. doi: 10.1109/TVCG.2019.2934280

[9] T. Chau. A review of analytical techniques for gait data. Part 2: neural network and wavelet methods. *Gait & Posture*, 13(2):102–120, 2001. doi: 10.1016/S0966-6362(00)00095-3

[10] K. Chia, I. Fischer, P. Thomason, H. K. Graham, and M. Sangeux. A decision support system to facilitate identification of musculoskeletal impairments and propose recommendations using gait analysis in children with cerebral palsy. *Frontiers in Bioengineering and Biotechnology*, 8, 2020. doi: 10.3389/fbioe.2020.529415

[11] R. C. Contreras, A. Parnandi, B. G. Coelho, C. Silva, H. Schambra, and L. G. Nonato. NE-Motion: Visual analysis of stroke patients using motion sensor networks. *Sensors*, 21(13):4482, Jan. 2021. doi: 10.3390/s21134482

[12] C. Cui, G.-B. Bian, Z.-G. Hou, J. Zhao, G. Su, H. Zhou, L. Peng, and W. Wang. Simultaneous recognition and assessment of post-stroke hemiparetic gait by fusing kinematic, kinetic, and electrophysiological data. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(4):856–864, 2018.

[13] H. Darbandi, M. Baniasad, S. Baghdadi, A. Khandan, A. Vafaee, and F. Farahmand. Automatic classification of gait patterns in children with cerebral palsy using fuzzy clustering method. *Clinical Biomechanics*, 73:189–194, 2020. doi: 10.1016/j.clinbiomech.2019.12.031

[14] C. Dindorf, W. Teufl, B. Taetz, G. Bleser, and M. Fröhlich. Interpretability of input representations for gait classification in patients after total hip arthroplasty. *Sensors*, 20(16):4385, 2020.

[15] F. Dobson, M. E. Morris, R. Baker, and H. K. Graham. Gait classification in children with cerebral palsy: A systematic review. *Gait & Posture*, 25(1):140–152, 2007. doi: 10/c4t784

[16] A. Ferrari, L. Bergamini, G. Guerzoni, S. Calderara, N. Bicocchi, G. Vitetta, C. Borghi, R. Neviani, and A. Ferrari. Gait-based diplegia classification using LSMT networks. *Journal of Healthcare Engineering*, 2019. doi: 10.1155/2019/3796898

[17] J. Figueiredo, C. P. Santos, and J. C. Moreno. Automatic recognition of gait patterns in human motor disorders using machine learning: A review. *Medical Engineering & Physics*, 53:1–12, 2018. doi: 10.1016/j.medengphy.2017.12.006

[18] H. K. Graham, P. Rosenbaum, N. Paneth, B. Dan, J.-P. Lin, D. L. Damiano, J. G. Becher, D. Gaebler-Spira, A. Colver, D. S. Reddihough, K. E. Crompton, and R. L. Lieber. Cerebral palsy. *Nature Reviews Disease Primers*, 2(1):1–25, 2016. doi: 10.1038/nrdp.2015.82

[19] E. Halilaj, A. Rajagopal, M. Fiterau, J. L. Hicks, T. J. Hastie, and S. L. Delp. Machine learning in human movement biomechanics: Best practices, common pitfalls, and new opportunities. *Journal of Biomechanics*, 81:1–11, 2018. doi: 10.1016/j.jbiomech.2018.09.009

[20] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell. What do we need to build explainable AI systems for the medical domain? *arXiv preprint arXiv:1712.09923*, 2017.

[21] S. Jang, N. Elmqvist, and K. Ramani. GestureAnalyzer: Visual analytics for pattern analysis of mid-air hand gestures. In *Proceedings of the 2nd ACM Symposium on Spatial User Interaction*, SUI '14, pp. 30–39, 2014. doi: 10.1145/2659766.2659772

[22] S. Jang, N. Elmqvist, and K. Ramani. MotionFlow: Visual Abstraction and Aggregation of Sequential Patterns in Human Motion Tracking Data. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):21–30, 2015. doi: 10.1109/TVCG.2015.2468292

[23] D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, eds. *Mastering the Information Age – Solving Problems with Visual Analytics*. Eurographics, Goslar, Germany, 2010.

[24] P. R. Krekel, E. R. Valstar, J. De Groot, F. H. Post, R. G. H. H. Nelissen, and C. P. Botha. Visual analysis of multi-joint kinematic data. *Computer Graphics Forum*, 29(3):1123–1132, 2010. doi: 10.1111/j.1467-8659.2009.01681.x

[25] G. S. Liptak and P. J. Accardo. Health and social outcomes of children with cerebral palsy. *The Journal of Pediatrics*, 145(2, Supplement):S36–S41, 2004. doi: 10.1016/j.jpeds.2004.05.021

[26] M. J. Long, E. Papi, L. D. Duffell, and A. H. McGregor. Predicting knee osteoarthritis risk in injured populations. *Clinical Biomechanics*, 47:87–95, 2017. doi: 10.1016/j.clinbiomech.2017.06.001

[27] S. McIntyre. The continually changing epidemiology of cerebral palsy. *Acta Paediatrica*, 107(3):374–375, 2018. doi: 10.1111/apa.14232

[28] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019. doi: 10.1016/j.artint.2018.07.007

[29] E. Papageorgiou, A. Nieuwenhuys, I. Vandekerckhove, A. Van Campenhout, E. Ortibus, and K. Desloovere. Systematic review on gait classifications in children with cerebral palsy: An update. *Gait & Posture*, 69:209–223, 2019. doi: 10/gh388q

[30] T. C. Pataky. Generalized n-dimensional biomechanical field analysis using statistical parametric mapping. *Journal of Biomechanics*, 43(10):1976–1982, July 2010. doi: 10.1016/j.jbiomech.2010.03.008

[31] N. Pérez and A. Rodríguez. Cerebral Palsy: Hope Through Research. Technical report, NIH NINDS, 2013.

[32] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016. doi: 10.1145/2939672.2939778

[33] J. Rodda and H. K. Graham. Classification of gait patterns in spastic hemiplegia and spastic diplegia: A basis for a management algorithm. *European Journal of Neurology*, 8(s5):98–108, 2001. doi: 10.1046/j.1468-1331.2001.00042.x

[34] M. Sangeux, J. Rodda, and H. K. Graham. Sagittal gait patterns in cerebral palsy: The plantarflexor–knee extension couple index. *Gait & Posture*, 41(2):586–591, 2015. doi: 10.1016/j.gaitpost.2014.12.019

[35] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proc. IEEE ICCV*, pp. 618–626, 2017.

[36] D. Slijepcevic, F. Horst, S. Lapuschkin, B. Horsak, A.-M. Raberger, A. Kranzl, W. Samek, C. Breiteneder, W. I. Schöllhorn, and M. Zeppelzauer. Explaining machine learning models for clinical gait analysis. *ACM Transactions on Computing for Healthcare*, 3(2):14:1–14:27, 2021. doi: 10.1145/3474121

[37] D. Slijepcevic, M. Zeppelzauer, A.-M. Gorgas, C. Schwab, M. Schüller, A. Baca, C. Breiteneder, and B. Horsak. Automatic classification of functional gait disorders. *IEEE journal of Biomedical and Health Informatics*, 22(5):1653–1661, 2017. doi: 10.1109/JBHI.2017.2785682

[38] D. Slijepcevic, M. Zeppelzauer, C. Schwab, A.-M. Raberger, C. Breiteneder, and B. Horsak. Input representations and classification strategies for automated human gait analysis. *Gait & Posture*, 76:198–203,

2020. doi: 10.1016/j.gaitpost.2019.10.021

[39] C. Stoiber, F. Grassinger, M. Pohl, H. Stitz, M. Streit, and W. Aigner. Visualization onboarding: Learning how to read and use visualizations. In *IEEE Workshop on Visualization for Communication*. Vancouver, BC, Canada, 2019. doi: 10.31219/osf.io/c38ab

[40] J. J. Thomas and K. A. Cook, eds. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE, Los Alamitos, CA, USA, 2005.

[41] M. Wagner, D. Slijepcevic, B. Horsak, A. Rind, M. Zeppelzauer, and W. Aigner. KAVAGait: Knowledge-assisted visual analytics for clinical gait analysis. *IEEE Transactions on Visualization and Computer Graphics*, 25(3):1528–1542, 2018. doi: 10.1109/TVCG.2017.2785271

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

[43] N. Wilhelm, A. Vögele, R. Zsoldos, T. Licka, B. Krüger, and J. Bernard. FuryExplorer: visual-interactive exploration of horse motion capture data. In *Proceedings Visualization and Data Analysis*, vol. 9397, p. 93970F. SPIE, 2015. doi: 10.1117/12.2080001

[44] T. A. L. Wren, G. E. Gorton, S. Ounpuu, and C. A. Tucker. Efficacy of clinical gait analysis: A systematic review. *Gait & Posture*, 34(2):149–153, 2011. doi: 10.1016/j.gaitpost.2011.03.027

[45] Y. Zhang and Y. Ma. Application of supervised machine learning algorithms in the classification of sagittal gait patterns of cerebral palsy children with spastic diplegia. *Computers in Biology and Medicine*, 106:33–39, 2019. doi: 10.1016/j.compbiomed.2019.01.009