Exploring Dataset Bias and Scaling Techniques in Multi-Source Gait Biomechanics: An Explainable Machine Learning Approach

SOPHIE FLEISCHMANN and SIMON DIETZ, Machine Learning and Data Analytics Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany
JULIAN SHANBHAG, Engineering Design, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany
ANNIKA WUENSCH and MARLIES NITSCHKE, Machine Learning and Data Analytics Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany
JÖRG MIEHLING and SANDRO WARTZACK, Engineering Design, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany
SIGRID LEYENDECKER, Institute of Applied Dynamics, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany
BJOERN M. ESKOFIER, Machine Learning and Data Analytics Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany and Translational Digital Health Group, Institute of AI for Health, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany
ANNE D. KOELEWIJN, Machine Learning and Data Analytics Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany

Bjoern M. Eskofier and Anne D. Koelewijn contributed equally to this research.
This work was funded by the Sponsor Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—SFB 1483—Project-ID 442419336, EmpkinS.

Authors' Contact Information: Sophie Fleischmann (corresponding author), Machine Learning and Data Analytics Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany; e-mail: sophie.fl.fleischmann@fau.de; Simon Dietz, Machine Learning and Data Analytics Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany; e-mail: simon.j.dietz@fau.de; Julian Shanbhag, Engineering Design, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany; e-mail: shanbhag@mfk.fau.de; Annika Wuensch, Machine Learning and Data Analytics Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany; e-mail: annika.wuensch@fau.de; Marlies Nitschke, Machine Learning and Data Analytics Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany; e-mail: marlies.nitschke@fau.de; Jörg Miehling, Engineering Design, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany; e-mail: miehling@mfk.fau.de; Sandro Wartzack, Engineering Design, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany; e-mail: wartzack@mfk.fau.de; Sigrid Leyendecker, Institute of Applied Dynamics, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany; e-mail: sigrid.leyendecker@fau.de; Bjoern M. Eskofier, Machine Learning and Data Analytics Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany and Translational Digital Health Group, Institute of AI for Health, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany; e-mail: bjoern.eskofier@fau.de; Anne D. Koelewijn, Machine Learning and Data Analytics Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany; e-mail: anne.koelewijn@fau.de.

ACM 2157-6912/2024/12-ART20
https://doi.org/10.1145/3702646

Machine learning has become increasingly important in biomechanics. It allows to unveil hidden patterns from large and complex data, which leads to a more comprehensive understanding of biomechanical processes and deeper insights into human movement. However, machine learning models are often trained on a single dataset with a limited number of participants, which negatively affects their robustness and generalizability. Combining data from multiple existing sources provides an opportunity to overcome these limitations without spending more time on recruiting participants and recording new data. It is furthermore an opportunity for researchers who lack the financial requirements or laboratory equipment to conduct expensive motion capture studies themselves. At the same time, subtle interlaboratory differences can be problematic in an analysis due to the bias that they introduce. In our study, we investigated differences in motion capture datasets in the context of machine learning, for which we combined overground walking trials from four existing studies. Specifically, our goal was to examine whether a machine learning model was able to predict the original data source based on marker and GRF trajectories of single strides and how different scaling methods and pooling procedures affected the outcome. Layer-wise relevance propagation was applied to understand which factors were influential to distinguish the original data sources. We found that the model could predict the original data source with a very high accuracy (up to >99%), which decreased by about 15 percentage points when we scaled every dataset individually prior to pooling. However, none of the proposed scaling methods could fully remove the dataset bias. Layer-wise relevance propagation revealed that there was not only one single factor that differed between all datasets. Instead, every dataset had its unique characteristics that were picked up by the model. These variables differed between the scaling and pooling approaches but were mostly consistent between trials belonging to the same dataset. Our results show that motion capture data is sensitive even to small deviations in marker placement and experimental setup and that small inter-group differences should not be overinterpreted during data analysis, especially when the data was collected in different labs. Furthermore, we recommend scaling datasets individually prior to pooling them which led to the lowest accuracy. We want to raise awareness that differences in datasets always exist and are recognizable by machine learning models. Researchers should thus think about how these differences might affect their results when combining data from different studies.

CCS Concepts: • **Computing methodologies** → **Neural networks**; *Supervised learning by classification*; • **Applied computing** → *Life and medical sciences*;

Additional Key Words and Phrases: datasets, dataset combination, neural networks, explainable AI, scaling, biomechanics, motion capture, machine learning, LRP

ACM Reference format:
Sophie Fleischmann, Simon Dietz, Julian Shanbhag, Annika Wuensch, Marlies Nitschke, Jörg Miehling, Sandro Wartzack, Sigrid Leyendecker, Bjoern M. Eskofier, and Anne D. Koelewijn. 2024. Exploring Dataset Bias and Scaling Techniques in Multi-Source Gait Biomechanics: An Explainable Machine Learning Approach. *ACM Trans. Intell. Syst. Technol.* 16, 1, Article 20 (December 2024), 19 pages.
https://doi.org/10.1145/3702646

1 Introduction

The use of machine learning in biomechanics research has gained substantial popularity. It offers possibilities to analyze large and complex data, uncover hidden patterns, and deepen our understanding of human movement [21]. For example, machine learning models have been applied to characterize movement patterns [7], classify pathological gait [10], assess the risk of falls or injuries [31, 56], predict human motion [62] or gain insights into the control of movement [12].

However, in most studies in which machine learning is applied to human movement data, only a single dataset with few participants is used for model training and testing [21], which amplifies the risk of overfitting [17]. One reason for the small sample sizes in biomechanical studies is the high effort associated with data acquisition [33, 44]. Optical motion capture continues to be the benchmark method for recording motion data [52], but is time-consuming and expensive [14].

Pooling existing datasets can be an opportunity to solve the data scarcity problem without the need to spend additional time and effort on recording new data. This requires datasets to be shared publicly; however, the number of available datasets is still limited compared to the multitude of biomechanical studies conducted in the past years [37]. Nevertheless, more and more researchers have recognized the necessity and value of data sharing to create large-scale datasets [17, 21, 37, 54, 58, 61]. AddBiomechanics [58] is an example of a recent tool that aims to simplify the processing and exchange of motion data. Training a machine learning model on a pooled dataset that contains more and independent data can increase its accuracy and generalizability to new data [17]. At the same time, data from different labs will most likely differ due to variations in marker placement, measurement systems, or experimental protocols [8, 20], even if the same motion was recorded. These interlaboratory differences in biomechanics have been broadly investigated [8, 11, 20], however not in the context of machine learning. While data heterogeneity can improve model robustness, it might also negatively influence the performance of a machine learning model due to the bias that is introduced by the data source [19, 60]. In other disciplines like computer vision or systems biology, where large, open-source databases are already prevalent [45, 51] and merging datasets is more common [45], this problem is known as dataset bias [53] or batch effect [23]. It means that data in a specific dataset has unique characteristics stemming from factors like data collection procedures, which make them differentiable from other datasets of the same domain [4]. Torralba and Efris [53] investigated dataset bias in the context of object recognition using 12 often-used recognition datasets. Their classifier predicted the data source of random images from each dataset with an accuracy of 39%, which increased to 61% when using only images of a single class. The experiment showed that a machine learning model was able to differentiate between image data sources; however, no objective explanation was given which features of an image led to the correct classification [53] or what made the datasets unique.

Explanatory methods can help to understand the decision-making process of machine learning models. One popular method is **layer-wise relevance propagation (LRP)** [6]. It is more efficient compared to other algorithms such as LIME [43] or SHAP [38] and has shown to be suitable for time-series data [36, 55]. LRP computes a relevance score for each input feature of a neural network indicating its importance for the model decision. In biomechanics, LRP has been used to identify person-specific gait characteristics from joint angle and **ground reaction force (GRF)** trajectories [28] or electromyography data [1]. The technique has further been applied to explain kinematic differences [59] in running and to determine distinct [26] or overlapping [25] patterns among runners. LRP is a promising method to overcome the black-box character of machine learning models and gain insight into why a sample is classified into a certain class.

Furthermore, in machine learning, it is common to scale the input data before feeding it into the network. This speeds up the training process and reduces the impact of varying scales, which enhances the convergence of the model. However, while scaling is a standard practice in preprocessing biomechanical data for machine learning applications, there is no universally accepted procedure for how to scale biomechanical data for machine learning classification tasks [9]. It is also unknown how scaling affects the dataset bias when data from different sources is combined and to what extent scaling before or after pooling the datasets influences the outcome.

The purpose of this study was to explore dataset bias in biomechanical gait datasets and possibilities for combining data from existing studies. Similar to Torralba and Efros [53], we aimed to investigate whether a machine learning model could identify the original data source when we pooled motion capture data from multiple studies. However, we were not only interested in, but also why the model predicted that a trial stemmed from a specific dataset. Specifically, we applied LRP to analyze which factors in the datasets, which all measured overground walking, caused

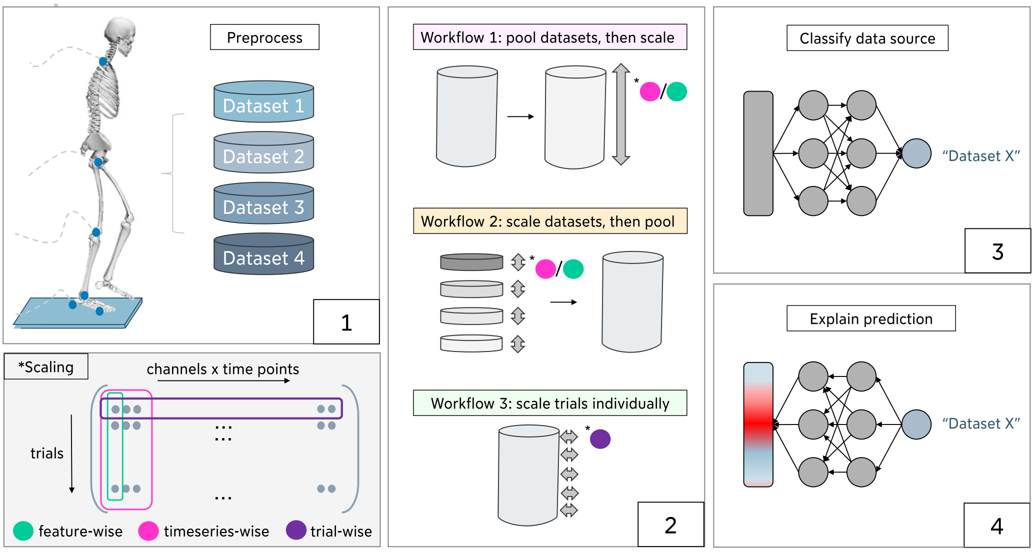


Fig. 1. High-level overview of the individual steps of the proposed method. (1) First, we extracted and preprocessed marker and GRF data from recordings of overground walking from four different datasets, resulting in 2,825 individual strides from 145 participants. (2) We compared three different data scaling methods and pooling WFs. (3) A neural network was used to classify the original data source of a stride. (4) We applied LRP to compute the importance of each input variable to the model prediction.

differences between the original data sources. Furthermore, we explored how different scaling techniques influenced the prediction and whether we could recommend a specific scaling technique that might reduce the dataset bias when data from different studies are combined.

This led to the following research questions:

- Can a machine learning model predict the original data source from marker and GRF force trajectories when we pool data from several datasets that all measure overground walking?
- Which variables in the data are crucial for the classification of the individual trials of a dataset?
- Can we reduce the dataset bias by applying different scaling techniques, and can we recommend a specific scaling technique when pooling datasets?

To answer these questions, we used data from four datasets and proposed different **workflows (WFs)** to pool and scale the data. We then trained neural networks to predict the data source and applied LRP to explain the results (Figure 1).

2 Methods

2.1 Datasets

We used gait data from three open source datasets [28, 30, 34, 39] and one internal dataset used in earlier research [15], which we named according to the respective first author. All datasets measured overground walking of healthy participants at multiple speeds and provided marker and GRF trajectories. However, the experimental protocol differed between the studies. For details on the data acquisition and the original purpose of the data, the reader is referred to the individual accompanying journal publications [15, 28, 35, 48].

We excluded data from children (age 0–17) as well as trials from the Moissenet dataset falling within the slowest speed range (0–0.4 ms^{-1}) because we considered the speed difference to all other

Table 1. Composition of the Pooled Dataset Used for This Study

Original dataset	Number of participants	Included walking speeds	Number of used trials
Dorschky [15]	9 (0 female, 9 male)	0.9–1.0 ms^{-1} 1.2–1.4 ms^{-1} 1.8–2.0 ms^{-1}	265
Horst [29]	57 (29 female, 28 male)	Self-selected speed	1,136
Moissenet [39]	50 (24 female, 26 male)	0.4–0.8 ms^{-1} 0.8–1.2 ms^{-1} Self-selected speed Fast self-selected speed	894
Lencioni [34]	29 (14 female, 15 male)	Self-selected speed Slow walking Fast walking	530
Pooled dataset	145 (67 female, 78 male)		2,825

datasets as too large. Furthermore, trials with clear measurement errors were discarded after a visual inspection. The remaining trials from all datasets were combined into one large gait dataset that was used for further analysis. Table 1 provides an overview of how the final, pooled dataset was composed.

2.2 Data Preparation

We extracted sagittal-plane marker and GRF data from all trials. Eleven markers were chosen, which were located at the same anatomical position in all original datasets: fifth metatarsal head (MT5), heel (HEE), lateral malleolus (ANKL), lateral femur epicondyle (KNE), and greater trochanter (GTRO) of both sides of the body as well as the acromion (SHO) on the side of the leading limb. As the Dorschky dataset provided only GRF data from one force plate, kinetic data from only one foot was used for all trials in the pooled dataset.

The knee and ankle markers in the Horst dataset, as well as the hip marker in the Lencioni dataset, had been removed before the dynamic trials. The trajectories of these markers were reconstructed in OpenSim [13, 49] using a three-dimensional model with 37 degrees of freedom based on the model by Hamner [22]. To do so, the model was scaled from the static trial and an inverse kinematic analysis was performed. For every time step, the position of the virtual marker corresponding to the missing experimental marker was extracted, and the resulting time series was then used as the experimental marker trajectory. This was possible because the original position of those markers on the body was known from the static trials.

The trajectories were filtered with a bidirectional second-order Butterworth filter with a 10 Hz cut-off frequency and cut to one gait cycle, starting and ending at **heel strike (HS)**. Each trial thus corresponded to one stride, which is why we use these terms interchangeably. For the Lencioni dataset, the data was already cut. The start of the stride was defined as the first sample with GRF data. However, they did not mention which GRF threshold they used nor how the end of the stride was defined, as they only used two force plates. For all other datasets, we determined HS using a vertical GRF threshold of 10N to match the method used in the Lencioni dataset. If no force plate data was available for a step, HS was defined as the time of the minimum vertical position of the heel marker. In the article, we refer to the trajectories of the leading leg and the following leg by "*_1*" and "*_2*," respectively. Every stride was resampled to 101 data points. The ANKL marker position at the first HS was used as the origin of all global coordinate systems to remove any dependency on the laboratory setup and allow for comparison between laboratories.

All trajectories were made dimensionless by dividing by body height or mass to enable a better comparison between subjects and to remove anthropometric influences [24]. Furthermore, all marker trajectories in anterior–posterior direction were expressed relative to the pelvis center, which was estimated as the middle between the left and right greater trochanter marker, in order to reduce the effect of different walking speeds. The 22 marker trajectories and 3 force trajectories (vertical and anterior–posterior GRF, ground reaction moment) of a trial were concatenated into one movement vector of length 2,525 (25 sequences times 101 data points). Eventually, all trials were arranged in a single data matrix $\boldsymbol{M} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_{2,525}] \in \mathbb{R}^{2,825x2,525}$, where every row corresponded to a different trial, and the number of columns equaled the number of data points that described one trial. We considered every time point as a feature and every marker or force trajectory as a separate channel.

2.3 Data Scaling and Pooling

Data normalization has an established role as a standard preprocessing step in biomechanical analyses. We tested three common scaling methods together with different data pooling WFs to investigate their effect on dataset bias (Figure 1). In feature-wise scaling, we individually normalized the elements of each feature vector $\boldsymbol{x_j} \in \mathbb{R}^{2,825}$ (for $j = 1, .., 2{,}525$) of our data matrix $\boldsymbol{M}$ to the interval [0, 1]. This is a standard technique in machine learning and can be expressed as

$$(x_j)_i' = \frac{(x_j)_i - min(\boldsymbol{x}_j)}{max(\boldsymbol{x}_j) - min(\boldsymbol{x}_j)} \qquad \forall i = 1, ..., 2{,}825 \quad \forall j = 1, ..., 2{,}525, \tag{1}$$

where $min(\boldsymbol{x}_j)$ and $max(\boldsymbol{x}_j)$ are the minimum and maximum values of vector $\boldsymbol{x}_j$.

In time series-wise scaling, we considered the fact that data points of the same channel trajectory were not independent but concatenated continuous waveforms. Instead of scaling the values of every feature vector individually, we expressed the data matrix as $\boldsymbol{M} = [\boldsymbol{C}_1, ..., \boldsymbol{C}_{25}]$ by summarizing the 101 feature vectors of each channel to one submatrix $\boldsymbol{C}_k \in \mathbb{R}^{2,825x101}$ (for $k = 1, .., 25$). The values of each submatrix $\boldsymbol{C}_k$ were then scaled to the interval [0,1], which can be expressed as

$$(C_k)_{il}' = \frac{(C_k)_{il} - min(\boldsymbol{C}_k)}{max(\boldsymbol{C}_k) - min(\boldsymbol{C}_k)} \qquad \forall k = 1, .., 25 \quad \forall i = 1, ..., 2{,}825 \quad \forall l = 1, .., 101. \tag{2}$$

In both methods, the values in the data matrix were scaled column-wise, meaning across trials or rows. In trial-wise scaling, the feature values were not scaled across all trials; instead, each submatrix was rewritten as $\boldsymbol{C}_k = [\boldsymbol{y}_1, ..., \boldsymbol{y}_{2,825}]^T$ and every channel trajectory $\boldsymbol{y}_i^T \in \mathbb{R}^{101}$ (for $i = 1, ..., 2{,}825$) of each trial was normalized to the interval [0, 1] using Equation (1). Normalizing each trial separately discards information, which can be advantageous if there is extraneous variability between samples [47], such as marker placement variability in our case.

In terms of the pooling WFs, we applied time series- and feature-wise scaling on two levels: Firstly, on our merged dataset as a whole (*WF1*), and secondly, separately on each original dataset prior to pooling (*WF2*) [5]. For trial-wise scaling, scaling before or after pooling yields the same outcome since each trial is considered individually. We, therefore, referred to this method as *WF3*. In any case, the scaling process was performed on the training set, and subsequently, the same scaling parameters were applied to scale the test and validation sets [44]. The combinations of normalization methods and pooling WFs eventually resulted in five different configurations.

2.4 Dataset Classification

The deployed classifier network architecture was a multi-layer perceptron with one hidden layer. This simple architecture has been proven to be sufficient and effective in classifying gait patterns based on time-continuous data. [25, 28]. The linear hidden layer was followed by a ReLU activation function, and the softmax function was applied to the output to obtain a probability distribution over the different datasets. The number of input and output nodes was determined from the data: 2,525 for the input layer (because one trial was described by $25 \cdot 101$ points) and 4 for the output layer (1 per dataset). The number of neurons in the hidden layer $H \in \{32, 64, 128, 256, 512, 1{,}024\}$ was set as a network hyperparameter. The network weights were initialized uniformly. We used the ADAM optimizer [32] and the cross entropy loss function to train the network. The batch size and learning rate were set as training hyperparameters and tuned together with the network hyperparameter using grid search. The search space was 1×10^{-5}, 1×10^{-4}, 1×10^{-3} for the learning rate and $\{8, 16, 32, 64, 128\}$ for the batch size. Prediction accuracies were reported over a stratified grouped five-fold cross-validation. In each fold, 25% of the training set was used for validation. The stratification ensured that the ratio of trials between the original datasets was preserved throughout the folds. We also made sure that trials of one person were either in the training, validation, or test set, as we wanted the model to learn unique characteristics of the dataset and not the walking patterns of individuals. The network was trained for a maximum of $1,000$ epochs and at the beginning of each epoch, the data was randomly shuffled before being divided into batches for training. If the validation loss decreased less than 5×10^{-3} for 30 epochs, training was stopped early. We trained a separate network for each of the five scaling and pooling configurations described before.

2.5 LRP

To explain why the network learned that a trial stemmed from a specific dataset, we applied LRP, an explanation technique that computes relevance scores for all input variables. LRP first runs a forward pass on the trained model. Then, the activation score of the output layer is propagated back layer by layer following the principle of relevance conservation until the input layer is reached [46]. In our study, we applied the ϵ-variant of LRP, which adds a small constant to the denominator to prevent numeric instabilities [40]. We chose an epsilon-value of 10^{-6}. For more details on LRP, we refer to the original article by Bach et al. [6], where a comprehensive explanation and analysis of the LRP algorithm can be found. The Zennit software package [3] was used for the implementation of LRP.

2.6 Evaluation

The classification performance of every network was evaluated by assessing the prediction accuracy and the corresponding confusion matrix. For any stride, LRP yielded a relevance pattern consisting of a relevance score for each of the 2,525 input variables indicating its importance for the prediction: positive values favored the decision for a certain dataset, while negative values could be seen as evidence against the class [46]. Since the interpretation of negative relevance scores is difficult in multiclass classification problems, we clipped the relevancies at zero and focused only on positive values [1]. Similar to [26], we only considered trials that were correctly assigned by the neural network. To enable a comparison between original data sources, all relevance patterns were normalized to their respective maximum [26] and averaged dataset-wise. The averaged relevancies were eventually rescaled to values ranging between zero and one, where values close to zero indicated the lowest, and values close to 1 the highest relevance. For better interpretability and comprehensibility, we visualized the class-specific relevance scores as heatmaps. To condense

Table 2. Prediction Accuracies of the Dataset Classification Task in Percentage, Reported as Mean and Standard Deviation of the Five-Fold Cross-Validation

Pooling WF	Scaling method		
	Feature-wise	Time series-wise	Trial-wise
WF1 (pool → scale)	99.3 ± 1.2	98.7 ± 1.7	–
WF2 (scale → pool)	83.6 ± 9.0	82.2 ± 5.2	–
WF3 (order not important)	–	–	99.5 ± 0.5

the information from the heatmap, we additionally displayed the overall relative contribution of each marker and of each time point for the dataset identification [25, 26]. This allows for the identification of key markers and critical moments within the gait cycle that have the highest impact on dataset identification, thereby providing deeper insights into the temporal and spatial dynamics that distinguish between different classes or conditions.

These average class heatmaps provided a global impression of the characteristics of each data source but no information about potential inter-trial differences. To determine whether the learned characteristics of trials within a dataset were consistent, we computed the relative channel importance for every trial by summing over the 101 relevance values per channel and normalizing the resulting values so that their sum equaled 1. This gave us, for each stride i, a vector $\boldsymbol{r}_i$ with 25 values (1 for each channel), which represented the relative relevance contribution of that respective channel for the classification of the stride. We then conducted a **principle component (PC)** analysis based on the 2,825 channel relevance vectors $\boldsymbol{r}_i$. The resulting eigenvectors spanned a new basis in the directions of the largest variance among the relevance vectors. Trials with similar relevancies were expected to lie close together when mapped onto this feature space, whereas trials were expected to lie apart when the model learned different characteristics. We visualized the result by plotting the first two PCs. The relative explained variance of each PC was a measure of how much this eigenvector contributed to the total variance in the data.

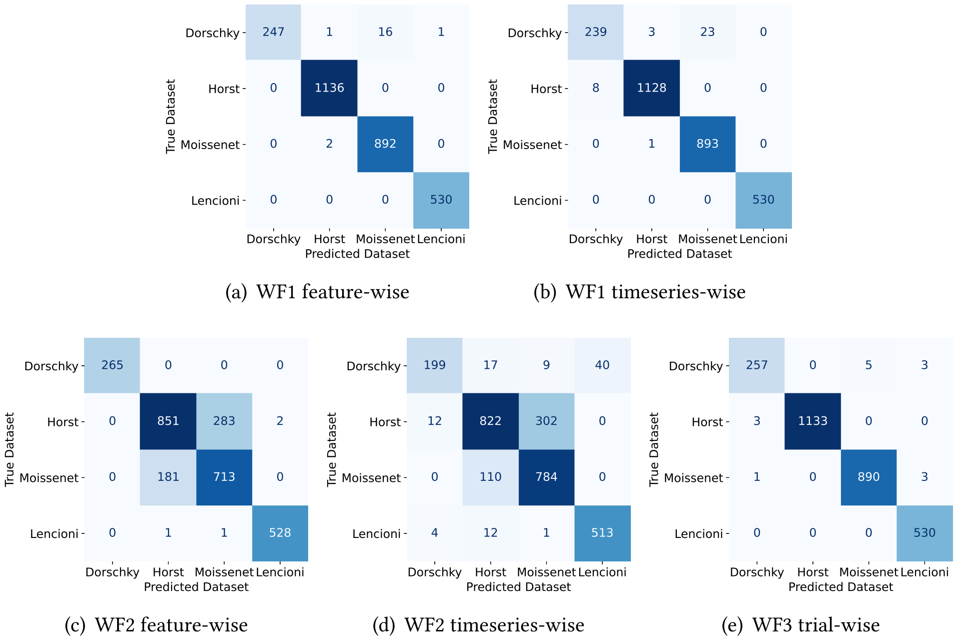
3 Results

3.1 Performance of Classification Models

The classification performance of the networks was consistently high with accuracies between 82% and 99.6% (Table 2). The model could distinguish trials from different datasets with a very high accuracy of above 98% when the data was first pooled and then scaled (WF1). Only a few trials from the Dorschky dataset and one trial (for time series-wise scaling) or two trials (for feature-wise scaling) from the Moissenet dataset were wrongly classified (Figure 2(a) and (b)). When the datasets were scaled prior to pooling (WF2), the accuracy decreased by about 15 percentage points. Here, only the Lencioni dataset could still be identified almost perfectly (Figure 2(c) and (d)). For WF3, during which every trial was scaled individually, the accuracy was again very high (99.5%).

3.2 Explaining the Model Predictions Using LRP

WF1 (Pool → Scale). The learned relevance patterns differed between the four datasets. For example, for the Dorschky dataset (Figure 3(a)), the vertical position of the knee markers was most relevant throughout the whole cycle, while other markers hardly contributed to the classification. In contrast, for the Lencioni dataset (Figure 3(b)), it was the vertical position of both heel markers that was most important over the complete cycle, accounting for about 40% of the complete relevance. Furthermore, some channels contributed strongly to the identification of a dataset but only at a



(a) WF1 feature-wise

(b) WF1 timeseries-wise

(c) WF2 feature-wise

(d) WF2 timeseries-wise

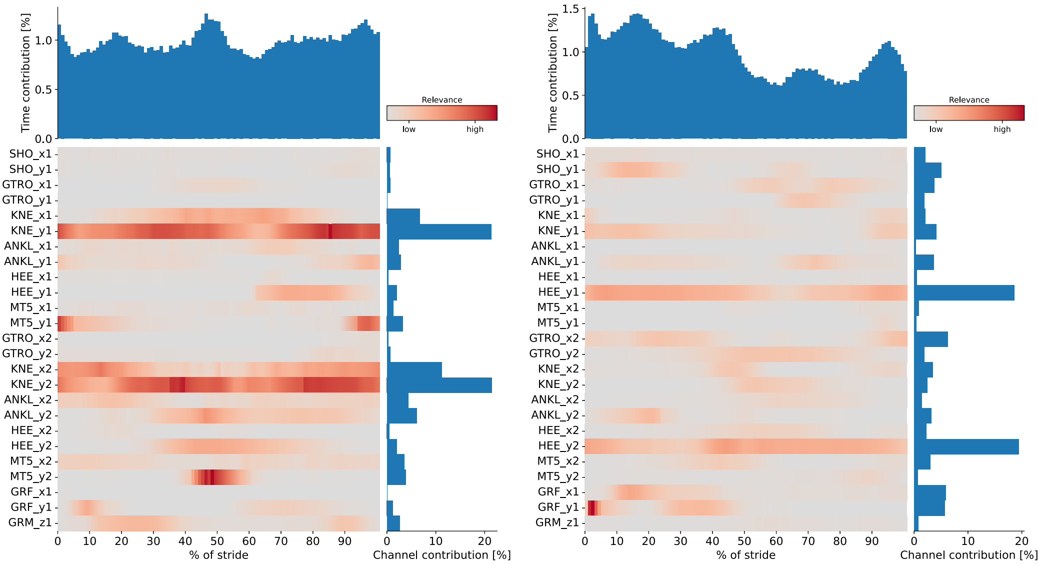
(e) WF3 trial-wise

Fig. 2. Confusion matrices for all five configurations.

specific period throughout the stride, such as the vertical MT5 marker in the Dorschky dataset around the HS of the respective foot or the vertical GRF in the Lencioni dataset at the beginning of the gait cycle. The relevancies were similar when scaling was done feature-wise. All heatmaps not depicted here can be found in the Supplemental Material.

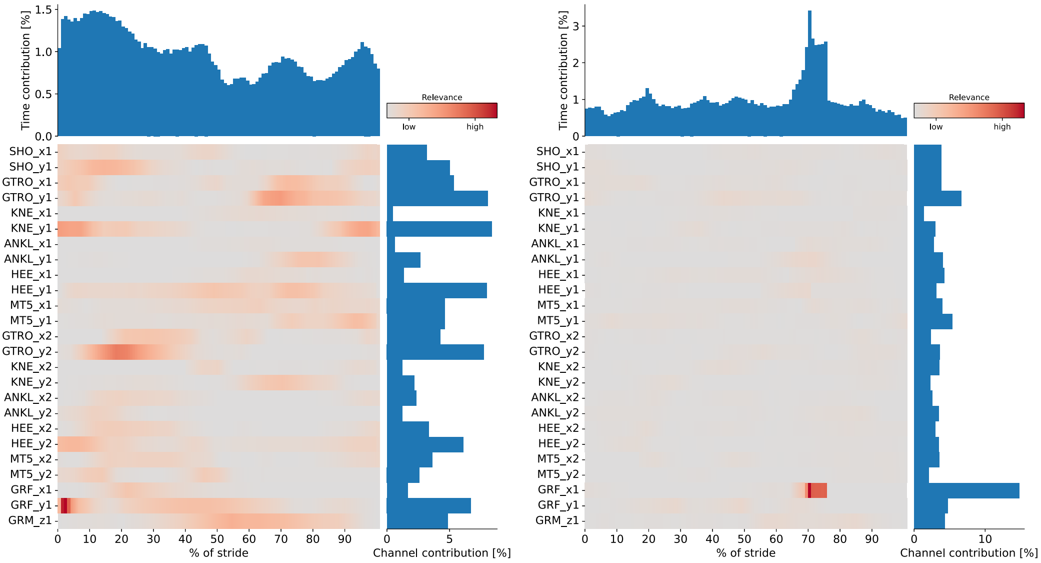
WF2 (Scale ⟶ Pool). Since the Lencioni dataset was still classified well in WF2 compared to the other datasets, we exemplarily show the relevance patterns of the Lencioni dataset in Figure 4 for both time series-wise and feature-wise scaling; all other heatmaps can be found in the Supplemental Material. The HEE markers were not as important as in WF1 (Figure 3(b)), and the relative relevance of each marker was generally lower, as the relevancies were distributed more evenly among the channels, which could also be observed for other datasets. As in WF1, not one single parameter differed between all datasets, but rather every dataset had certain unique characteristics. Other than in WF1, the learned characteristics of every dataset differed between time series-wise and feature-wise scaling. For example, for the Lencioni dataset, the relative influence of the vertical GRF after HS was preserved for feature-wise scaling (Figure 4(b), but not for time series-wise scaling (Figure 4(a)). Instead, the anterior–posterior GRF at around 70% of the gait cycle, which corresponds to the moment before toe-off, was most influential. This relevance peak around toe-off caused by the GRFs was observed for all datasets in feature-wise scaling but not in time series-wise scaling.

WF3 (Scale Trial-Wise). When every channel trajectory of each trial was scaled individually (Figure 5), the relevance heatmaps were generally sparser compared to scaling across all datasets in WF1 (Figure 3). Certain characteristics remained, such as the increased relevance of the vertical MT5 markers at the timing of the HSs for the Dorschky dataset (compare Figure 3(a) and Figure 5(a)). Other than in WF1, no marker was relevant throughout the complete stride though. Instead, we found that opposed distinct phases of the gait cycle were decisive across the different datasets, like the times shortly before HS (45% and 95% of the stride) for the Dorschky dataset (Figure 5(a)), the



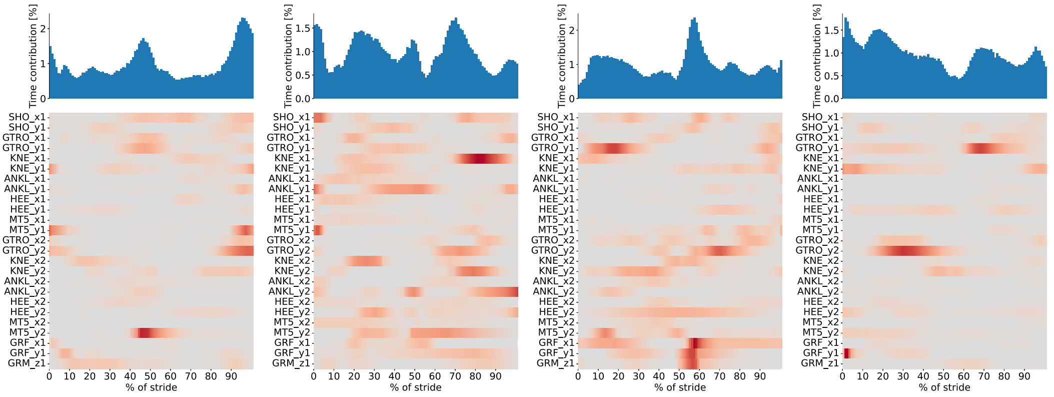
(a) Dorschky dataset (b) Lencioni dataset

Fig. 3. Class-specific relevancies averaged across the correctly classified trials of the respective dataset for WF1 and time series-wise scaling. In the center, red indicates a high feature relevance, while gray indicates a low feature relevance for the identification. The top part depicts the relative relevance of each time point of the stride, and the right part shows the relevance contribution of the individual channels.



(a) timerseries-wise scaling (b) feature-wise scaling

Fig. 4. Class-specific relevancies averaged across the correctly classified trials of the Lencioni dataset for WF2. In the center, red indicates a high feature relevance, while gray indicates a low feature relevance for the identification. The top part depicts the relative relevance of each time point of the stride, and the right part shows the relevance contribution of the individual channels.



(a) Dorschky dataset (b) Horst dataset (c) Moissenet dataset (d) Lencioni dataset

Fig. 5. Class-specific relevancies averaged across the correctly classified trials of the respective datasets for WF3. In the center, red indicates a high feature relevance, while gray indicates a low feature relevance for the identification. The top part depicts the relative relevance of each time point of the stride.

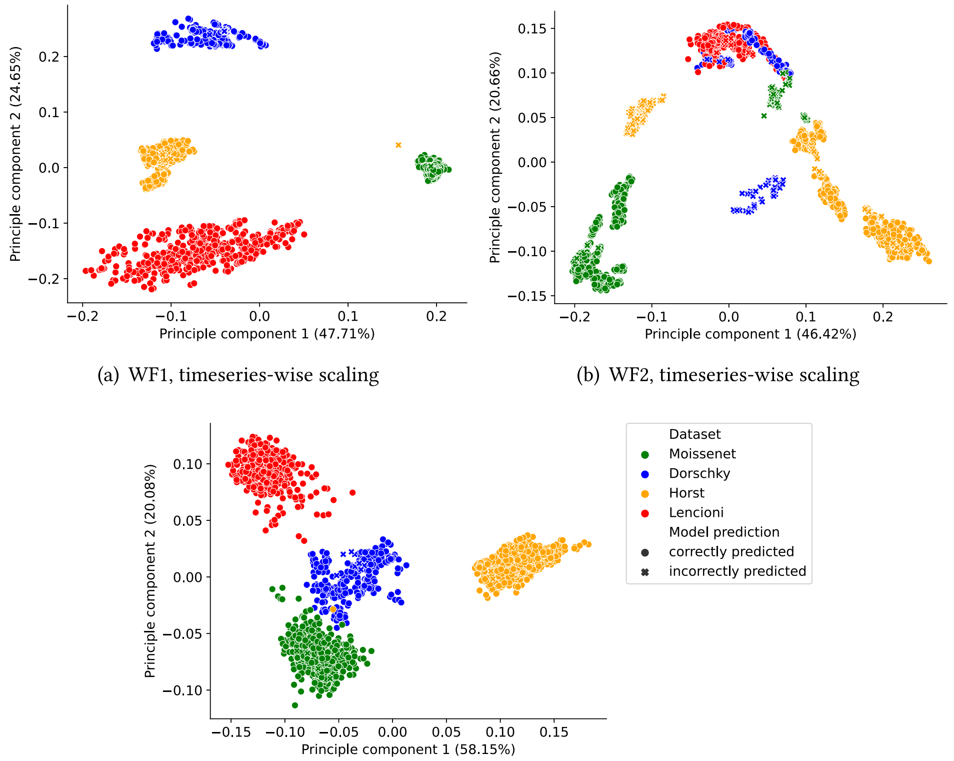
time right after HS and before toe off (50%–60% of stride) for the Moissenet dataset (Figure 5(c)) or the beginning of swing phases (15%–30% and 65%–80% of stride) for the Horst (Figure 5(b)) and Lencioni (Figure 5(d)) dataset.

Inter-Trial Consistency. For WF1 (Figure 6(a)), as well as for WF3 (Figure 6(c)), trials of the same data source lay close together in the PC space, while trials of different sources were clearly separated. For WF1, the first PCs separated the Moissenet dataset from the rest, while differences between the other three datasets could rather be seen in the second PC. In contrast, for WF3, the Horst dataset was the one to be separated from the others by the first PC. For WF2 (Figure 6(b)), especially the Dorschky and Lencioni dataset had overlapping regions. Furthermore, trials of the same dataset did not always lie together. Especially for the Horst and Moissenet dataset, the trials formed several smaller clusters. For all datasets, the trial clusters that lay far away from the majority of trials of the same dataset were trials that were wrongly classified by the model. In all cases, the first two PCs explained between 66% and 78% of the variance in the computed relevance patterns. The plots not shown can be found in the Supplemental Material.

4 Discussion

This work aimed to investigate the presence of dataset bias in motion capture studies and explore the possibilities of combining data from multiple studies. The goal of our study was threefold: firstly, we investigated whether a machine learning model could predict the original data source of an individual stride when data from different studies that all measured overground walking were pooled, which was confirmed. Secondly, we used LRP to explain how the model knew that a trial stemmed from a specific dataset. Here, we found that the learned attributes depended on the scaling method and that every dataset had unique characteristics, which were picked up by the model. Lastly, we compared different pooling WFs and common scaling methods to learn their effect on the prediction accuracy and the learned features, which showed us that it is harder for the model to predict the original data source when every dataset is scaled separately before pooling.

Our results demonstrated that a machine learning model could predict the original dataset with a high accuracy when combining data from different studies. This indicates that, although the data measured the same task, each dataset had distinct characteristics that allowed the model to



(a) WF1, timeseries-wise scaling

(b) WF2, timeseries-wise scaling

(c) WF3, trial-wise scaling

Fig. 6. Trials projected onto the first two PCs of the PC space, which was computed from the relative channel relevance vectors across all trials for the respective configuration. The percentage of each PC is the explained variance, which states how much of the total variance in the data is explained by the PC.

distinguish the sources. The result was in accordance with findings from other disciplines. Similar experiments that detected dataset bias in the image or neuroimaging datasets resulted in a 61% [53] or 71.5% [57] accuracy in predicting the original source, respectively. Remarkably, our model exhibited even higher accuracy rates, reaching up to 99%. This suggests that dataset characteristics in motion capture datasets seem to be especially pronounced and distinctive.

Differences in motion capture data from different labs are not necessarily bad. Training a machine learning model on data from different sources can enhance its robustness and generalizability compared to training on a single, small dataset. However, especially for unsupervised learning, the differences might influence the outcome. Our results show that motion capture data is sensitive even to small deviations, and we should not over-interpret small differences between groups when we analyze data, especially when the data was collected in different labs. Furthermore, with this study, we could get a better understanding of dataset bias in motion capture datasets and show that the commonly applied preprocessing strategies are insufficient to overcome it. We are convinced that data sharing and combining datasets is a great opportunity for the biomechanics community, but we want to sensitize researchers to think about how dataset bias could affect a model or an analysis when combining multiple datasets.

For the classification, it hardly made a difference if we scaled feature-wise or time series-wise (>2% difference in accuracy). In contrast, scaling all datasets independently prior to pooling (WF2) decreased the classification accuracy by approximately 15 percentage points compared to scaling across the pooled dataset (WF1), meaning that it was harder for the model to identify the original dataset. This finding shows that dataset bias affects the scaling outcome more when scaling happens after combining the data. One reason for this is that information about the relative differences in marker placement between datasets is present when pooling the unscaled data in WF1 but gets removed in WF2 when every marker trajectory of each dataset is scaled to the same range before pooling. Surprisingly, the accuracy was also above 99% when every time series of every trial was scaled individually, as this way of scaling also removes the information about differences in marker placement between datasets and trials. This shows that there are other hidden patterns in the data, which are picked up by the model.

Even though the scaling and pooling WFs could influence the model's ability to distinguish between data sources, none of the configurations that we tested could fully remove dataset bias when combining motion capture gait datasets. In this work, we intentionally limited our investigation to scaling approaches, given that scaling is a standard preprocessing step in the application of machine learning within the field of biomechanics. Future work should investigate more complex approaches to effectively remove the dataset bias. One way to mitigate performance degradation in downstream tasks caused by distribution shifts between biomechanical datasets is **multi-source domain adaptation (MDA)** [16, 63]. MDA seeks to learn domain-invariant latent features across multiple source domains, enabling the development of a model that generalizes well and is robust against domain shifts, thereby ensuring consistent performance on the target domain. A further approach could be to use the insights from LRP, which identified the most critical features for dataset identification, to guide the application of general adversarial networks or diffusion models aimed at overcoming dataset bias.

The characteristics that the model learned were clearly influenced by the WF and scaling method and to a certain extend be traced back to the experiments. For example, in WF1, when we first pooled and then scaled the data, marker placement played a decisive role in differentiating the datasets. The consistently high relevance of distinct markers throughout the stride indicates a systematic difference in the placement of these markers compared to the other datasets. However, the shorter relevance periods that occurred only during certain times of the gait cycle must emerge from other factors like processing methods. One example is the high relevance of the vertical GRF during the first three time points in the Lencioni dataset. The Lencioni dataset was already precut, allegedly starting at the first sample where a vertical GRF was detected. However, the vertical GRF at the first time point was often higher compared to the other datasets, potentially resulting from a higher GRF threshold than the 10N threshold that we used, which is what the model learned. Another factor that we identified as potentially relevant was the experimental design. The increased importance of the vertical MT5 markers around HS in the Dorschky dataset is likely to be attributed to the participants wearing shoes during the study instead of walking barefoot, possibly leading to a decrease in plantarflexion at HS [44]. Furthermore, the walking speed conditions influenced the model prediction. We noticed an increased relevance of the horizontal knee marker in the Horst dataset during the swing phase. The Horst dataset was the only one in which participants were not asked to walk at different predefined ranges. Fast walking increases the range of motion and step length [18]. In our case, this increased step length is reflected in a larger anterior–posterior distance between the knee joint and the pelvis center, resulting in higher x-values for the knee joint. As there were no slow walking trials in the dataset, we assume that the model learned this as a unique feature of the dataset.

In WF2, where we scaled every dataset before pooling, marker placement was less important as no marker was relevant throughout the complete stride as in WF1. However, the influence of the processing methods and experimental design seemed to persist or even increase in some cases. The fact that in the averaged plots, with a few exceptions, the contribution of the individual channels to the classification was more balanced can be an indication that each dataset now had more than one unique characteristic or that not one, but a combination of marker trajectories made it identifiable. The large influence of the anterior–posterior GRF in feature-wise scaling was connected to individual trials where this force becomes slightly negative before toe-off, which influences the scaling at this time point. It is thus advisable to use time series-wise scaling instead of feature-wise scaling when including GRFs. Since the heatmaps showed a high influence of the GRF for the model prediction in WF2, we repeated the experiment with only marker trajectories, which did not lead to a notably worse model performance though.

In WF3, where we scaled every channel trajectory of every trial separately, no information about the relative marker placement across datasets or trials or about the amplitude of movements compared to each other was retained. The only information from the trials is how a marker moves relative to its maximum and minimum position during the stride. We found that the model could still recognized characteristics such as the reduced plantarflexion at initial contact in the Dorschky dataset or the potentially different GRF threshold at the beginning of the gait cycle in the Lencioni dataset. It is, however, difficult for this WF to determine the origins of what the model learned. Nevertheless, the knowledge of high-contribution features can be beneficial for subsequent tasks even if it is not completely known why these features contributed more than others.

The PC analysis revealed that for WF1 and WF3, the learned characteristics were consistent for trials of the same dataset, even across folds, and differed between trials of different datasets. For WF2, there was more of an overlap between the Dorschky and Lencioni datasets, which is in accordance with the decreased prediction accuracy. The confusion matrices show that in addition, many trials of the Horst and Moissenet datasets were confounded, which is not represented in the scatter plot. However, the two axes only represent around 67% of the explained variance, and these datasets have overlaps in another dimension that is not depicted. For WF2, it is also noticeable that trials of the same dataset are in some cases again distributed into smaller clusters. This supports the theory that each dataset had more than one characteristic which was learned by the model.

One reason why we chose LRP as the explainable ML algorithm and an MLP with one hidden layer as our network architecture was to align with other biomechanical studies with the same input data type [25, 28]. However, LRP is also highly efficient compared to other explainable ML techniques [27]. In our case, the latency of a single backward pass was around 8 ms on an NVIDIA GeForce RTX 2080 Ti GPU, which allowed for a fast evaluation. The simple MLP model contributed to the computational efficiency of our approach, even with varying numbers of hidden neurons. The **floating point operations (FLOPs)** of the network scaled linearly with the hidden neurons and reached a maximum of around 2,590k FLOPs for the largest considered network. This is comparably small relative to modern computing capabilities, underscoring the scalability and efficiency of our network's design. Another aspect that highlights the scalability of our approach is that during training, multiple numbers of hidden neurons could achieve a comparable validation loss. Despite the higher complexity, we additionally tested a deeper MLP architecture with two hidden layers. However, the prediction results exhibited minimal change, with a deviation of 1.4 percentage points for WF2 and feature-wise scaling and a maximum deviation of 0.2 percentage points for all other configurations. Therefore, we refrained from a more extensive analysis of deeper networks.

Normalizing all variables to the participants' height or weight during preprocessing ensured that our experiment was not biased by variations in anthropometric characteristics between datasets. Furthermore, we ensured that the model did not learn the unique gait pattern of an individual by testing only on data from persons that were not used for training.

To validate our classification approach, we conducted a verification experiment by selecting data solely from a single study, randomly assigning labels, and training the model to predict these labels.

As anticipated, the results demonstrated a low accuracy of approximately 22%, which is comparable to random guessing. This outcome confirmed the effectiveness of our classification model and the presence of unique dataset characteristics. To validate the faithfulness of the predicted relevancies, we reran the experiment while omitting high-contribution markers or features, which should lead to a decreased prediction accuracy. When we omitted the knee markers, which had shown to be most decisive for the Dorschky dataset in WF1, the recall for this dataset decreased from 90.2% to 80.8%. In contrast, omitting non-contribution markers of a dataset had the opposite effect. For instance, excluding the trochanter markers in WF1 increased the recall for the Dorschky dataset to 95.1%. Additionally, when we masked the 10% of features with the highest importance score of each trial [2, 41], we observed a drop in accuracy for all configurations, with a maximum of 20.1 percentage points for WF2 and feature-wise scaling. This supports the effectiveness of LRP.

The structure in which the data was provided was quite heterogeneous. Data were stored not only in different file formats but also in folder structures, coordinate systems, marker names, and sampling frequencies, which varied between studies. Preprocessing was necessary to make the data comparable. Assessing every dataset individually is however time-consuming, which might hinder researchers from using data from multiple studies. Having a standardized way of storing and providing motion capture data would facilitate access and encourage the use of open-source data.

For the study, we had to virtually reconstruct some markers for the Horst and the Lencioni dataset using inverse kinematics, which might introduce a small error. However, when we ran the experiment without these markers, we observed similarly high accuracies, showing that our procedure did not introduce the dataset differences. Furthermore, our study was restricted to a core set of eleven markers. Future work should include a larger, full-body markerset. With this, it could, for example, be investigated which anatomical areas (e.g., upper or lower body) contribute most to the dataset bias. We performed our study on marker and GRF data, as it is the first and most basic output from motion capture experiments. Future studies should explore how the prediction accuracy changes when the classification is done based on the kinematics and kinetics of walking. However, other variations could be introduced when different processing methods or models are used for the computation. It would also be interesting to investigate which dataset differences are derivable from features gained from musculoskeletal simulations, such as muscle activation patterns or joint forces.

5 Conclusion

We showed that a machine learning model was able to predict the original data source from marker and GRF trajectories of overground walking when we combined data from different studies, from which we conclude that there are detectable differences between similar datasets from different sources. We presented a comparison between different scaling methods and pooling WFs and showed that the pooling WF had a higher influence on the prediction accuracy than the scaling technique and that scaling prior to pooling should be favored. However, none of the methods could fully remove the dataset bias. We applied LRP to understand how the model recognized that a trial stemmed from a specific dataset and to determine the most influential differences. We found that there was not one single factor that differed between all datasets, but that instead, every dataset had its unique characteristics that stemmed from factors like marker placement or experimental design. Furthermore, the learned characteristics differed between the scaling and pooling approaches. With our study, we wanted to raise awareness about differences in motion capture datasets. Combining datasets has many advantages, but we recommend keeping in mind that dataset bias exists and that this might influence the results.

References

[1] Jeroen Aeles, Fabian Horst, Sebastian Lapuschkin, Lilian Lacourpaille, and François Hug. 2021. Revealing the unique features of each individual's muscle activation signatures. *Journal of the Royal Society Interface* 18, 174 (2021), 20200770. DOI: https://doi.org/10.1098/rsif.2020.0770

[2] David Alvarez-Melis and Tommi S. Jaakkola. 2018. Towards robust interpretability with self-explaining neural networks. arXiv:1806.07538. Retrieved from https://arxiv.org/abs/1806.07538

[3] Christopher J. Anders, David Neumann, Wojciech Samek, Klaus-Robert Müller, and Sebastian Lapuschkin. 2023. Software for dataset-wide XAI: From local explanations to global insights with Zennit, CoRelAy, and ViRelAy. arXiv:2106.13200v2. Retrieved from https://arxiv.org/abs/2106.13200v2

[4] Boqing Gong, Fei Sha, and Kristen Graumann. 2021. Overcoming dataset bias: An unsupervised domain adaptation approach. In *Proceedings of the NIPS Workshop on Large Scale Visual Recognition and Retrieval*.

[5] B enjamin R. Babcock, Astrid Kosters, Junkai Yang, Mackenzie L. White, and Eliver E. B. Ghosn. 2021. Data matrix normalization and merging strategies minimize batch-specific systemic variation in scRNA-Seq data. bioRxiv. DOI: https://doi.org/10.1101/2021.08.18.456898

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

[7] Rezaul Begg and Joarder Kamruzzaman. 2005. A machine learning approach for automated recognition of movement patterns using basic, kinetic and kinematic gait data. *Journal of Biomechanics* 38, 3 (2005), 401–408. DOI: https://doi.org/10.1016/j.jbiomech.2004.05.002

[8] Maria G. Benedetti, Andrea Merlo, and Alberto Leardini. 2013. Inter-laboratory consistency of gait analysis measurements. *Gait & Posture* 38, 4 (2013), 934–939. DOI: https://doi.org/10.1016/j.gaitpost.2013.04.022

[9] 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. DOI: https://doi.org/10.3389/fbioe.2020.00260

[10] Helber A. Carvajal-Castaño, Juan D. Lemos-Duque, and Juan R. Orozco-Arroyave. 2022. Effective detection of abnormal gait patterns in Parkinson's disease patients using kinematics, nonlinear, and stability gait features. *Human Movement Science* 81 (2022), 102891. DOI: https://doi.org/10.1016/j.humov.2021.102891

[11] Jesse M. Charlton, Trevor B. Birmingham, Kristyn M. Leitch, and Michael A. Hunt. 2021. Knee-specific gait biomechanics are reliable when collected in multiple laboratories by independent raters. *Journal of Biomechanics* 115 (2021), 110182. DOI: https://doi.org/10.1016/j.jbiomech.2020.110182

[12] Olivier Codol, Jonathan A. Michaels, Mehrdad Kashefi, J. Andrew Pruszynski, and Paul L. Gribble. 2024. MotorNet: A Python toolbox for controlling differentiable biomechanical effectors with artificial neural networks. *eLife* 12, RP88591 (2024). DOI: https://doi.org/10.7554/eLife.88591.4

[13] Scott L. Delp, Frank C. Anderson, Allison S. Arnold, Peter Loan, Ayman Habib, Chand T. John, Eran Guendelman, and Darryl G. Thelen. 2007. OpenSim: Open-source software to create and analyze dynamic simulations of movement. *IEEE Transactions on Biomedical Engineering* 54, 11 (2007), 1940–1950. DOI: https://doi.org/10.1109/TBME.2007.901024

[14] Eva Dorschky, Marlies Nitschke, Christine F. Martindale, Antonie J. van den Bogert, Anne D. Koelewijn, and Bjoern M. Eskofier. 2020. CNN-based estimation of sagittal plane walking and running biomechanics from measured and simulated inertial sensor data. *Frontiers in Bioengineering and Biotechnology* 8 (2020), 604. DOI: https://doi.org/10.3389/fbioe.2020.00604

[15] Eva Dorschky, Marlies Nitschke, Ann-Kristin Seifer, Antonie J. van den Bogert, and Bjoern M. Eskofier. 2019. Estimation of gait kinematics and kinetics from inertial sensor data using optimal control of musculoskeletal models. *Journal of Biomechanics* 95 (2019), 109278. DOI: https://doi.org/10.1016/j.jbiomech.2019.07.022

[16] Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R. Arabnia. 2021. A brief review of domain adaptation. In *Advances in Data Science and Information Engineering*. Robert Stahlbock, Gary M. Weiss, Mahmoud Abou-Nasr, Cheng-Ying Yang, Hamid R. Arabnia, and Leonidas Deligiannidis (Eds.), Springer International Publishing, Cham, 877–894.

[17] Reed Ferber, Sean T. Osis, Jennifer L. Hicks, and Scott L. Delp. 2016. Gait biomechanics in the era of data science. *Journal of Biomechanics* 49, 16 (2016), 3759–3761. DOI: https://doi.org/10.1016/j.jbiomech.2016.10.033

[18] Claudiane A. Fukuchi, R. K. Fukuchi, and Marcos Duarte. 2019. Effects of walking speed on gait biomechanics in healthy participants: A systematic review and meta-analysis. *Systematic Reviews* 8 (2019), 153. DOI: https://doi.org/10.1186/s13643-019-1063-z

[19] Wilson Wen Bin Goh, Chern Han Yong, and Limsoon Wong. 2022. Are batch effects still relevant in the age of big data? *Trends in Biotechnology* 40, 9 (2022), 1029–1040. DOI: https://doi.org/10.1016/j.tibtech.2022.02.005

[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. DOI: https://doi.org/10.1016/j.gaitpost.2008.10.060

[21] 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. DOI: https://doi.org/10.1016/j.jbiomech.2018.09.009

[22] Samuel R. Hamner, Ajay Seth, and Scott L. Delp. 2010. Muscle contributions to propulsion and support during running. *Journal of Biomechanics* 43, 14 (2010), 2709–2716. DOI: https://doi.org/10.1016/j.jbiomech.2010.06.025

[23] Lukas Heumos, Anna C. Schaar, Christopher Lance, Anastasia Litinetskaya, Felix Drost, Luke Zappia, Malte D. Lücken, Daniel C. Strobl, Juan Henao, Fabiola Curion, Herbert B. Schiller, and Fabian J. Theis. 2023. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics* 24 (2023), 550–572. DOI: https://doi.org/10.1038/s41576-023-00586-w

[24] At L. Hof. 1996. Scaling gait data to body size. *Gait & Posture* 4, 3 (1996), 222–223. DOI: https://doi.org/10.1016/0966-6362(95)01057-2

[25] Fabian Hoitz, Laura Fraeulin, Vinzenz von Tscharner, Daniela Ohlendorf, Benno M. Nigg, and Christian Maurer-Grubinger. 2021. Isolating the unique and generic movement characteristics of highly trained runners. *Sensors* 21, 21 (2021), 7145. DOI: https://doi.org/10.3390/s21217145

[26] Fabian Hoitz, Vinzenz von Tscharner, Jennifer Baltich, and Benno M. Nigg. 2021. Individuality decoded by running patterns: Movement characteristics that determine the uniqueness of human running. *PLoS ONE* 16, 4 (2021), e0249657. DOI: https://doi.org/10.1371/journal.pone.0249657

[27] Andreas Holzinger, Anna Saranti, Christoph Molnar, Przemyslaw Biecek, and Wojciech Samek. 2022. *Explainable AI Methods–A Brief Overview*. Springer International Publishing, Cham, 13–38. DOI: https://doi.org/10.1007/978-3-031-04083-2_2

[28] 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 (2019), 2391. DOI: https://doi.org/10.1038/s41598-019-38748-8

[29] Fabian Horst, Sebastian Lapuschkin, Samek Samek Wojciech, Klaus-Robert Müller, and Wolfgang I Schöllhorn. 2019. A public dataset of overground walking kinetics and full-body kinematics in healthy adult individuals. Mendeley Data. V3. DOI: https://doi.org/10.17632/svx74xcrjr.3

[30] Fabian Horst, Djordje Slijepcevic, Matthias Zeppelzauer, Anna-Maria Raberger, Sebastian Lapuschkin, Wojciech Samek, Wolfgang I. Schöllhorn, Christian Breiteneder, and Brian Horsak. 2020. Explaining automated gender classification of human gait. *Gait & Posture* 81, Suppl. 1 (2020), 159–160. DOI: https://doi.org/10.1016/j.gaitpost.2020.07.114

[31] Susanne Jauhiainen, Jukka-Pekka Kauppi, Mari Leppänen, Kati Pasanen, Jari Parkkari, Tommi Vasankari, Pekka Kannus, and Sami Äyrämö. 2021. New machine learning approach for detection of injury risk factors in young team sport athletes. *International Journal of Sports Medicine* 42, 2 (2021), 175–182. DOI: https://doi.org/10.1055/a-1231-5304

[32] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. arXiv:1412.6980. Retrieved from https://arxiv.org/abs/1412.6980

[33] Duane Knudson. 2017. Confidence crisis of results in biomechanics research. *Sports Biomechanics* 16, 4 (2017), 425–433. DOI: https://doi.org/10.1080/14763141.2016.1246603

[34] Tiziana Lencioni, Ilaria Carpinella, Marco Rabuffetti, Alberto Marzegan, and Alberto Ferrari. 2019. Human kinematic, kinetic and EMG data during level walking, toe/heel-walking, stairs ascending/descending. figshare. Collection. DOI: https://doi.org/10.6084/m9.figshare.c.4494755.v1

[35] Tiziana Lencioni, Ilaria Carpinella, Marco Rabuffetti, Alberto Marzegan, and Maurizio Ferrarin. 2019. Human kinematic, kinetic and EMG data during different walking and stair ascending and descending tasks. *Scientific Data* 6 (2019), 309. DOI: https://doi.org/10.1038/s41597-019-0323-z

[36] Christoffer Loeffler, Wei-Cheng Lai, Bjoern Eskofier, Dario Zanca, Lukas Schmidt, and Christopher Mutschler. 2023. Don't get me wrong: How to apply deep visual interpretations to time series. arXiv:2203.07861. Retrieved from https://arxiv.org/abs/2203.07861

[37] Wan Shi Low, Chow Khuen Chan, Joon Huang Chuah, Yee Kai Tee, Yan Chai Hum, Maheza Irna Mohd Salim, and Khin Wee Lai. 2022. A review of machine learning network in human motion biomechanics. *Journal of Grid Computing* 20, 4 (2022), 37 pages. DOI: https://doi.org/10.1007/s10723-021-09595-7

[38] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS '17)*. Curran Associates Inc., 4768–4777.

[39] Florent Moissenet and Céline Schreiber. 2019. A multimodal dataset of human gait at different walking speeds established on injury-free adult participants. V8. figshare. DOI: https://doi.org/10.6084/m9.figshare.7734767

[40] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. 2019. *Layer-Wise Relevance Propagation: An Overview*. Springer International Publishing, Springer Cham, 193–209. DOI: https://doi.org/10.1007/978-3-030-28954-6_10

[41] Hyeonyeong Nam, Jun-Mo Kim, WooHyeok Choi, Soyeon Bak, and Tae-Eui Kam. 2023. The effects of layer-wise relevance propagation-based feature selection for EEG classification: a comparative study on multiple datasets. *Frontiers in Human Neuroscience* 17 (2023), 1–12. DOI: https://doi.org/10.3389/fnhum.2023.1205881

[42] Soham Raste, Rahul Singh, Joel Vaughan, and Vijayan N. Nair. 2022. Quantifying inherent randomness in machine learning algorithms. arXiv:2206.12353. Retrieved from https://arxiv.org/abs/2206.12353

[43] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "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 (KDD '16).* ACM, New York, NY, 1135–1144. DOI: https://doi.org/10.1145/2939672.2939778

[44] Mark A. Robinson, Jos Vanrenterghem, and Todd C. Pataky. 2021. Sample size estimation for biomechanical waveforms: Current practice, recommendations and a comparison to discrete power analysis. *Journal of Biomechanics* 122 (2021), 110451. DOI: https://doi.org/10.1016/j.jbiomech.2021.110451

[45] Yeonjae Ryu, Geun Hee Han, Eunsoo Jung, and Daehee Hwang. 2023. Integration of single-cell RNA-Seq datasets: A review of computational methods. *Molecules and Cells* 46, 2 (2023), 106–119. DOI: https://doi.org/10.14348/molcells.2023.0009

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

[47] Warren S. Sarle. 2023. comp.ai.neural-nets FAQ, Part 2 of 7: Learning. Retrieved from http://www.faqs.org/faqs/ai-faq/neural-nets/part2/

[48] Céline Schreiber and Florent Moissenet. 2019. A multimodal dataset of human gait at different walking speeds established on injury-free adult participants. *Scientific Data* 6, 111 (2019), 1–7. DOI: https://doi.org/10.1038/s41597-019-0124-4

[49] Ajay Seth, Jennifer L. Hicks, T. K. Uchida, Atif Habib, Christopher L. Dembia, James J. Dunne, C. F. Ong, M. S. DeMers, A. Rajagopal, M. Millard, et al. 2018. OpenSim: Simulating musculoskeletal dynamics and neuromuscular control to study human and animal movement. *PLoS Computational Biology* 14, 7 (2018), e1006223. DOI: https://doi.org/10.1371/journal.pcbi.1006223

[50] Felix Stief. 2018. *Variations of Marker Sets and Models for Standard Gait Analysis*. Springer International Publishing, Cham, 509–526. DOI: https://doi.org/10.1007/978-3-319-14418-4_26

[51] Tatiana Tommasi, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. 2017. *A Deeper Look at Dataset Bias*. Springer, Cham, 37–55. DOI: https://doi.org/10.1007/978-3-319-58347-1_2

[52] Matt Topley and James G. Richards. 2020. A comparison of currently available optoelectronic motion capture systems. *Journal of Biomechanics* 106 (2020), 109820. DOI: https://doi.org/10.1016/j.jbiomech.2020.109820

[53] Antonio Torralba and Alexei A. Efros. 2011. Unbiased look at dataset bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '11)*, 1521–1528. DOI: https://doi.org/10.1109/CVPR.2011.5995347

[54] Scott D. Uhlrich, Thomas K. Uchida, Marissa R. Lee, and Scott L. Delp. 2023. Ten steps to becoming a musculoskeletal simulation expert: A half-century of progress and outlook for the future. *Journal of Biomechanics* 154 (2023), 111623. DOI: https://doi.org/10.1016/j.jbiomech.2023.111623

[55] Ihsan Ullah, Andre Rios, Vaibhav Gala, and Susan Mckeever. 2022. Explaining deep learning models for tabular data using layer-wise relevance propagation. *Applied Sciences* 12, 1 (2022), 1–20. DOI: https://doi.org/10.3390/app12010136

[56] Martin Ullrich, Nils Roth, Arne Küderle, Robert Richer, Till Gladow, Heiko Gaßner, Franz Marxreiter, Jochen Klucken, Bjoern M. Eskofier, and Felix Kluge. 2023. Fall risk prediction in Parkinson's disease using real-world inertial sensor gait data. *IEEE Journal of Biomedical and Health Informatics* 27, 1 (2023), 319–328. DOI: https://doi.org/10.1109/JBHI.2022.3215921

[57] Christian Wachinger, Anna Rieckmann, and Sebastian Pölsterl. 2021. Detect and correct bias in multi-site neuroimaging datasets. *Medical Image Analysis* 67 (2021), 101879. DOI: https://doi.org/10.1016/j.media.2020.101879

[58] K eenon Werling, Michael Raitor, Jon Stingel, Jennifer L. Hicks, Steve Collins, Scott L. Delp, and C. Karen Liu. 2022. Rapid bilevel optimization to concurrently solve musculoskeletal scaling, marker registration, and inverse kinematic problems for human motion reconstruction. bioRxiv 2022.08.22.504896. [*bioRxiv.*] DOI: https://doi.org/10.1101/2022.08.22.504896

[59] Datao Xu, Wenjing Quan, Huiyu Zhou, Dong Sun, Julien S. Baker, and Yaodong Gu. 2022. Explaining the differences of gait patterns between high and low-mileage runners with machine learning. *Scientific Reports* 12 (2022), 2981. DOI: https://doi.org/10.1038/s41598-022-07054-1

[60] Haleh Yasrebi, Peter Sperisen, Viviane Praz, and Philipp Bucher. 2009. Can survival prediction be improved by merging gene expression data sets? *PLoS ONE* 4, 10 (2009), e7431. DOI: https://doi.org/10.1371/journal.pone.0007431

[61] Matteo Zago, Ana Francisca Rozin Kleiner, and Peter Andreas Federolf. 2020. Editorial: Machine learning approaches to human movement analysis. *Frontiers in Bioengineering and Biotechnology* 8 (2020), 638793. DOI: https://doi.org/10.3389/fbioe.2020.638793

[62] Abdelrahman Zaroug, Alessandro Garofolini, Daniel T. H. Lai, Kurt Mudie, and Rezaul Begg. 2021. Prediction of gait trajectories based on the long short term memory neural networks. *PLoS ONE* 16, 8 (2021), e0255597. DOI: https://doi.org/10.1371/journal.pone.0255597

[63] Sicheng Zhao, Bo Li, Colorado Reed, Pengfei Xu, and Kurt Keutzer. 2020. Multi-source domain adaptation in the deep learning era: A systematic survey. arXiv:2002.12169. Retrieved from https://arxiv.org/abs/2002.12169

Received 13 November 2023; revised 2 September 2024; accepted 14 October 2024