scientific reports

OPEN

Human-centered evaluation of statistical parametric mapping and explainable machine learning for outlier detection in plantar pressure data

Carlo Dindorf[1]✉, Jonas Dully[1], Steven Simon[1], Dennis Perchthaler[1], Stephan Becker[1], Hannah Ehmann[2], Kjell Heitmann[2], Bernd Stetter[3], Christian Diers[2] & Michael Fröhlich[1]

Plantar pressure mapping is essential in clinical diagnostics and sports science, yet large heterogeneous datasets often contain outliers from technical errors or procedural inconsistencies. Statistical Parametric Mapping (SPM) provides interpretable analyses but is sensitive to alignment and its capacity for robust outlier detection remains unclear. This study compares an SPM approach with an explainable machine learning (ML) approach to establish transparent quality-control pipelines for plantar pressure datasets. Data from multiple centers were annotated by expert consensus and enriched with synthetic outliers resulting in 798 valid samples and 2000 outliers. We evaluated (i) a non-parametric, registration-dependent SPM approach and (ii) a convolutional neural network (CNN), explained using SHapley Additive exPlanations (SHAP). Performance was assessed via nested cross-validation; explanation quality via a semantic differential survey with domain experts. The ML model reached high accuracy and outperformed SPM, which misclassified clinically meaningful variations and missed true outliers (Matthews Correlation Coefficient: ML = 0.96 ± 0.01; SPM = 0.78 ± 0.02). Experts perceived both SPM and SHAP explanations as clear, useful, and trustworthy, though SPM was assessed less complex. These findings highlight the complementary potential of SPM and explainable ML as approaches for automated outlier detection in plantar pressure data, and underscore the importance of explainability in translating complex model outputs into interpretable insights that can effectively inform decision-making.

Keywords Explainable artificial intelligence (XAI), Deep learning, Human-centered design, Semantic differential, Clinical decision support, Biomechanics quality control

Plantar pressure mapping—capturing how vertical forces distribute across the foot during static (e.g., standing) or dynamic (e.g., walking, running) activities—has become indispensable in clinical diagnostics, sports science, and rehabilitation[1–3]. By revealing biomechanical irregularities in pressure profiles, the diagnosis, monitoring, and screening of conditions such as diabetic neuropathy[4], Parkinson's disease[5], and various musculoskeletal disorders[6–8], is supported. It is also an established approach to measure the biomechanical impact of medical aids, such as (knee) ankle-foot orthotics or insoles, to ensure positive clinical outcomes[9] and in the recent past to adapt orthotics to plantar pressure profiles[9,10]. Yet the accuracy of any downstream analysis hinges on data quality—and in practice, pressure datasets are often contaminated by outliers, or anomalies, which are data points that deviate from the expected pattern[11] (we use the term *outliers* to refer to technical errors that produce measurements without clinical relevance, and the term *anomalies* to describe deviations from a healthy foot caused by anatomical abnormalities, such as flat foot).

Variations in participant instruction, protocol adherence, or the use of different systems across multiple centers are causes for inconsistencies and increase the likelihood of outliers in plantar pressure data. Instructor-

[1]Department of Sports Science, RPTU University Kaiserslautern-Landau, Kaiserslautern, Germany. [2]DIERS International GmbH, Wiesbaden, Germany. [3]Institute of Sports and Sports Science, Karlsruhe Institute of Technology, Karlsruhe, Germany. ✉email: carlo.dindorf@rptu.de

or protocol-related issues may include incorrect guidance, inadequate monitoring, or failure to correct participant behavior (e.g., wearing unintended footwear or moving unexpectedly during trials). Technical factors include sensor malfunctions, algorithmic misalignment (e.g., on treadmills), premature truncation of pressure curves, environmental noise, or errors in automated foot identification and segmentation. If left unaddressed, such outliers can distort automated segmentation algorithms[12,13] and reduce the accuracy of classification pipelines[14,15], potentially leading to erroneous or lower-quality biomechanical insights. This issue is particularly critical in multicenter data collection, which is often required to overcome data sparsity, especially when developing automated machine learning (ML) pipelines[16]. In such settings, datasets can become large and heterogeneous, making manual data verification economically and logistically impractical.

One established approach for analyzing plantar pressure data is Statistical Parametric Mapping (SPM)[17]. Originally developed for the analysis of 3D neuroimaging data[18], SPM enables the statistical comparison of continuous spatial data by performing voxel- or point-wise variance analyses across pre-defined regions. In the context of plantar pressure, this enables researchers to identify areas of statistically significant deviation across the plantar surface rather than relying solely on summary metrics such as peak pressure alone[17]. Frameworks such as Personalized Analysis of Plantar Pressure Images (PAPPI)[19] leverage SPM by first constructing a normative model that accounts for individual demographic factors (e.g., age, weight, or foot size) and then applying SPM to highlight deviations from this normative reference. While PAPPI showcases the interpretability and individualized assessment strengths of SPM, it was not explicitly designed to isolate technical errors or procedural inconsistencies; instead, it flags any departure from the norm, independently whether pathological or spurious. This limitation underscores a critical requirement of SPM analyses: spatial alignment. To ensure that each pixel corresponds to the same anatomical region across subjects, plantar pressure data must be normalized for rotation, scale, and anatomical landmarks[1]. Without such alignment, statistically significant deviations may reflect misregistration rather than true biomechanical differences—for example, identical pixel locations could span both heel and midfoot areas in different participants, leading to misleading inferences.

ML methods offer a promising alternative, as they can potentially learn alignment invariances directly from the data, reducing or even eliminating the need for labor-intensive preprocessing steps[20]. In medical imaging, deep learning approaches have demonstrated high-precision anomaly detection[21,22]. However, it remains largely unexplored whether these methods can achieve similar success in plantar pressure data, particularly when trained on labeled examples of diverse outlier types.

Importantly, under Article 22 of the General Data Protection Regulation (GDPR), individuals subjected to automated decision-making have the right to obtain meaningful information about the logic underlying such decisions[23]. This requirement presents a significant challenge for ML models, which often operate as “black boxes” that do not readily provide interpretable explanations. To address this challenge, Explainable Artificial Intelligence (XAI) techniques have gained increasing importance, enabling researchers and practitioners to probe the internal decision-making processes of complex ML models[24]. Beyond facilitating model debugging, XAI methods support risk assessment, bias detection, regulatory compliance, and the development of end-user trust and acceptance[25].

Crucially, the value of an explanation extends beyond quantitative metrics (e.g., fidelity or completeness) to encompass human-centered attributes, including understandability, reliability, and the potential to inform subsequent actions[26,27]. Without systematic, human-in-the-loop evaluation, XAI outputs risk relegation to academic curiosities rather than serving as practical decision-support tools in clinical settings.

Despite recent advances in the field of explainable outlier detection[28], the integration of supervised outlier classification with XAI—along with human-centered assessment of explanatory outputs—has, to the best of the authors’ knowledge, not yet been explored in plantar pressure data. This constitutes a critical research gap, given the growing need for automated yet interpretable quality-control pipelines in clinical and sports analysis[29]. To address this gap, this study directly compares the more established SPM approach against a novel explainable ML approach. This investigation is structured around two core questions:

1. How do these approaches compare in terms of detection accuracy?
2. In an exploratory follow-up, how do human evaluators perceive and trust the explanations generated by each approach?

By addressing these questions, this study aims to inform strategies for refining data-cleaning protocols and guiding the development of real-time monitoring systems that can alert technicians to potential acquisition errors (e.g., prompting trial repetition or verifying foot-side annotations). Ultimately, these advancements are intended to enhance data quality and reliability in both research and practical diagnostic settings, specifically within the context of plantar pressure analysis.

Methods

Workflow overview

An overview of the workflow is presented in Fig. 1 and described in more detail below.

Participants and data acquisition

Participation in the study was restricted to individuals of legal age. All participants received detailed information about the study protocol and relevant data protection guidelines before giving their written informed consent. The study was conducted in accordance with the ethical standards set forth in the Declaration of Helsinki and received approval from the Ethics Committee of the University of Kaiserslautern-Landau (approval number: 55). Data were collected across several centers using two types of pressure measurement systems: resistive pressure sensor plates (RSscan Lab Ltd., Ipswich, England) and capacitive pressure sensor plates (Zebris Medical

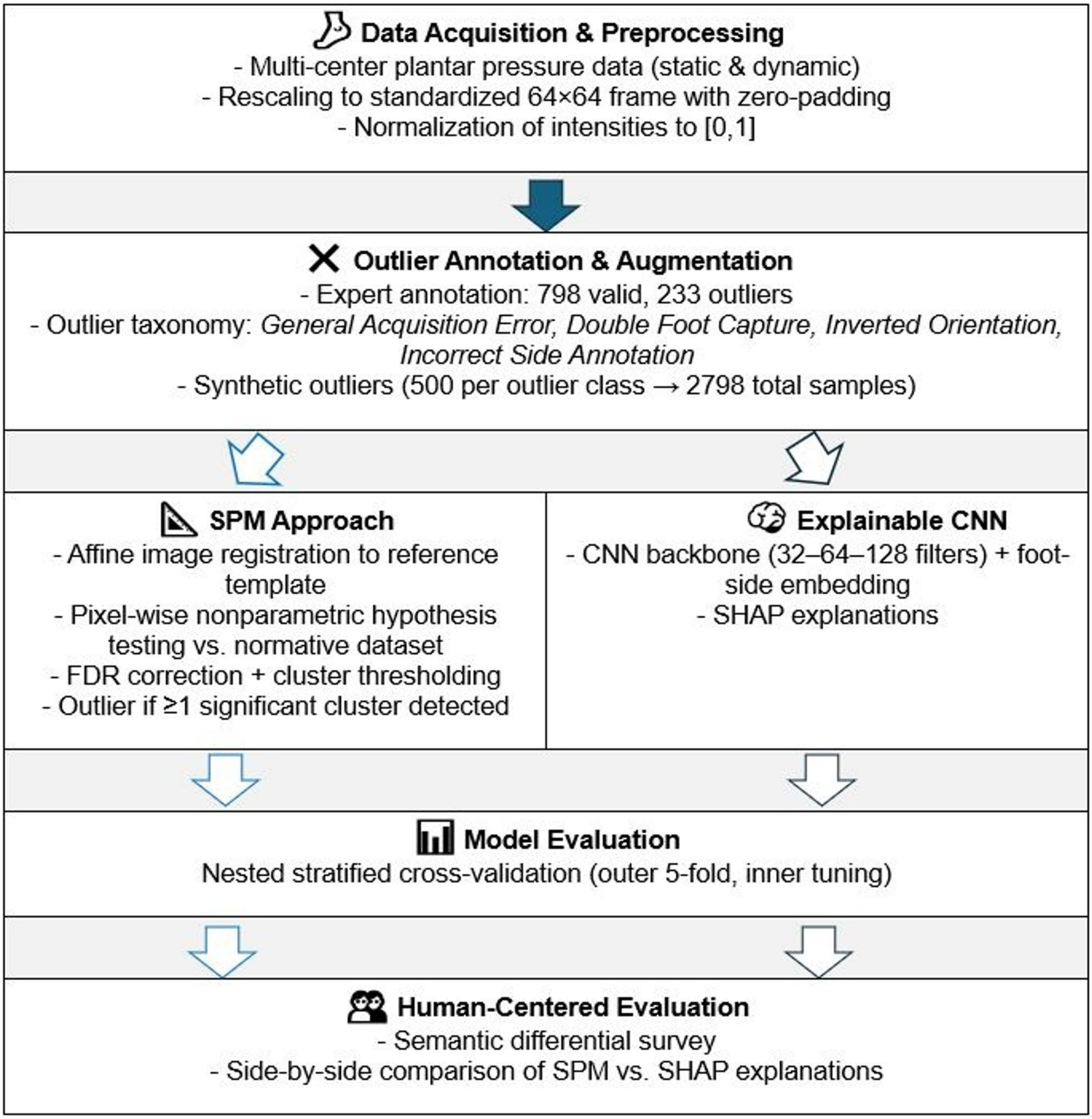

Data Acquisition & Preprocessing
- Multi-center plantar pressure data (static & dynamic)
- Rescaling to standardized 64×64 frame with zero-padding
- Normalization of intensities to [0,1]
Outlier Annotation & Augmentation
- Expert annotation: 798 valid, 233 outliers
- Outlier taxonomy: General Acquisition Error, Double Foot Capture, Inverted Orientation, Incorrect Side Annotation
- Synthetic outliers (500 per outlier class → 2798 total samples)
SPM Approach
- Affine image registration to reference template
- Pixel-wise nonparametric hypothesis testing vs. normative dataset
- FDR correction + cluster thresholding
- Outlier if ≥1 significant cluster detected
Explainable CNN
- CNN backbone (32–64–128 filters) + foot-side embedding
- SHAP explanations
Model Evaluation
Nested stratified cross-validation (outer 5-fold, inner tuning)
Human-Centered Evaluation
- Semantic differential survey
- Side-by-side comparison of SPM vs. SHAP explanations

Fig. 1. Overview of the study workflow, illustrating the comparison between Statistical Parametric Mapping (SPM) and a machine learning approach based on a Convolutional Neural Network (CNN), with model interpretation provided through SHapley Additive exPlanations (SHAP).

GmbH, Isny, Germany). Measurements were taken from both the left and right feet under static (53%) and dynamic (47%) conditions. In the static trials, participants stood barefoot in an upright position on the platform for 10 s, with data captured at 50 Hz. Following a 60-second habituation phase on a treadmill, dynamic trials consisted of three overground walking passes at each participant's self-selected pace, recorded at 100 Hz. For these dynamic measurements, stride-level plantar pressure profiles and peak pressure values were extracted, in line with common reporting practices[3]. Both static and dynamic profiles were subsequently combined for model development, which increases the sample size, enhances generalizability, and supports more robust model training. Duplicate datasets were identified and excluded before analysis.

Only data essential to the development and evaluation of the computational model were collected, while anthropometric and other descriptive variables were deliberately omitted. This choice reflected the principle of data minimization (GDPR Art. 5(1)(c)[23], requiring personal data to be "adequate, relevant, and limited to what is necessary" for the stated purpose.

Because the two pressure systems differed in spatial resolution and sensor geometry, preprocessing steps were applied to harmonize the datasets. First, all pressure distributions were rescaled to adjust for non-uniform sensor spacing. The resulting data were proportionally resized and embedded into a standardized 64 × 64 pixel grid, maintaining aspect ratios and avoiding distortion by applying zero-padding around the patterns. Finally, pressure intensities were normalized to the range [0,1] to reduce inter-subject variability and eliminate weight-dependent effects.

Outlier annotation and dataset augmentation

Outlier categories were defined by domain experts through systematic review of the dataset, considering both naturally occurring artifacts and recurrent sources of error in the initial dataset. The resulting taxonomy is presented in Table 1. The categorization was guided not only by technical accuracy but also by the practical relevance of providing automated feedback to end users regarding data integrity. Specifically, recordings classified as *General Acquisition Errors* were consolidated into a single category, because such trials are irreparably flawed (e.g., incomplete foot contact, trials performed with footwear, or corrupted sensor output) and require re-acquisition rather than post hoc correction. In contrast, the remaining categories capture systematic but potentially correctable errors—such as mislabeling of foot laterality or inverted foot orientation—that could be addressed through post-processing.

If a sample was identified as having both an *Incorrect Side Annotation* label and another outlier category, the *Incorrect Side Annotation* was considered the lowest priority and was superseded by the more critical label. This method was implemented to ensure that each sample was assigned to the most practically relevant outlier category, which aids in subsequent actions, such as deciding whether re-recording the data is required.

Three domain experts focused on identifying and labeling outliers and valid (inliers) samples. The annotation process was carried out collaboratively. Each sample was independently reviewed by at least one expert, after which labels were cross-verified to ensure consensus across raters. This process yielded a curated dataset of 1,031 samples from 703 subjects, consisting of 798 valid samples and 233 outlier samples (*General Acquisition Error*: 124; *Double Foot Capture*: 29; *Inverted Orientation*: 38; *Incorrect Side Annotation*: 42), with approximately equal representation of left and right feet (~ 50% each). To improve model robustness and allow for systematic evaluation, the dataset was further augmented with synthetically generated outliers, ensuring that each outlier category contained 500 samples. Parameter choices for generating these artificial samples were developed in close collaboration with the domain experts who originally annotated the data. Experts iteratively reviewed prototype examples to verify that the transformations produced realistic characteristics matching those observed in true acquisition errors. This synthetic balancing strategy aligns with findings from related fields, where artificially equalized outlier classes have been shown to enhance model performance[30]. Four types of artificial outliers were created:

- General Acquisition Error (Label 1): This category mimics incomplete foot contact caused by early or late stance capture or by missing regional pressure. Random samples from the valid inlier set were spatially cropped to remove either the forefoot or heel. For the missing forefoot condition, the distal 50% of the image (± 5% random variation per sample) was zeroed. For the missing heel condition, the proximal 65% of the image (± 5% random variation per sample) was removed. Original laterality labels were retained.
- Double Foot Capture (Label 2): To simulate erroneous recordings showing both feet simultaneously, pairs of valid left and right foot samples from the same subject folder were combined. Each individual foot image was downscaled to 32 × 32 pixels and then inserted into diagonally opposite quadrants of a 64 × 64 canvas (e.g., left foot in the upper-left quadrant and right foot in the lower-right quadrant, or the reverse configuration selected at random). To increase variability and avoid overly regular composite structures, each downscaled foot image was additionally shifted by zero to two pixels in a randomly chosen direction (left, right, upward, or downward) before placement. Laterality annotations are intrinsically undefined for such merged images and were therefore assigned at random.
- Inverted Orientation (Label 3): A random subset of valid inlier samples was vertically flipped, producing images equivalent to a 180° rotation. Original laterality annotations were retained.
- Incorrect Side Annotation (Label 4): A random sample of valid inlier images was selected, and their left/right annotations were inverted while the underlying pressure maps remained unchanged.

The fidelity of these synthetic outliers was evaluated by the three experts, who confirmed that the artificially generated samples closely reproduced the biomechanical and acquisition-related characteristics of genuine outliers within each category. After augmentation, the final dataset comprised 2798 samples in total, reflecting 798 inliers and 2000 outliers.

Statistical parametric mapping (SPM) approach

Plantar pressure registration

A critical prerequisite for the SPM approach is that all images be precisely aligned to ensure that each pixel corresponds to the same anatomical foot region across all subjects, thereby preventing misalignment from introducing misleading statistical inferences[17]. To address this, a plantar pressure registration pipeline was implemented using a similarity-based optimization method, which has previously been shown to be effective for plantar pressure alignment[31].

To ensure spatial consistency, each raw input was registered to a single, pre-defined prototypical reference pressure distribution, generated separately for the left and right foot. This reference acts as a template for alignment. The registration process employed an affine transformation, which corrects for variations in rotation, translation, and scaling[32]. The optimal transformation parameters (angle, shift, and zoom) were found by minimizing a mean squared error (MSE) loss function between the transformed input and the corresponding prototypical reference, following established image registration procedures[33]. The optimization was performed

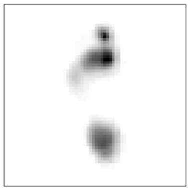
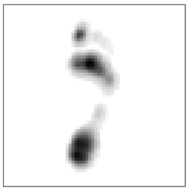
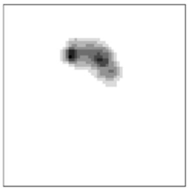
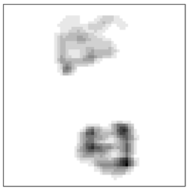
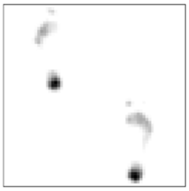
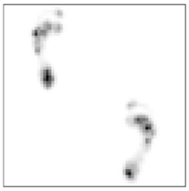
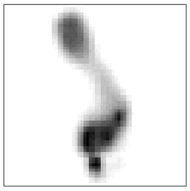
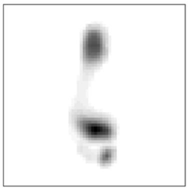
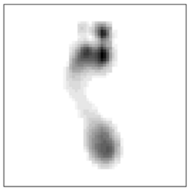
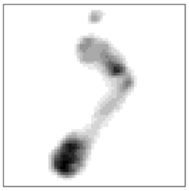
Label	Example		Description
Valid Sample **Label 0**	 Left Side	 Right Side	**Correct plantar pressure recordings without measurement or detection errors** Both feet are properly identified, and side labels (left vs. right) are accurate. Importantly, atypical or pathological plantar pressure patterns (e.g., due to gait abnormalities) are not classified as outliers if acquisition quality is intact.
General Acquisition Error **Label 1**	 Right Side	 Left Side	**Recordings with severe acquisition artifacts** These include incomplete foot contact (e.g., only forefoot or heel captured), trials performed with footwear instead of barefoot, or corrupted sensor output (e.g., motion blur, hardware malfunction).
Double Foot Capture **Label 2**	 Right Side	 Right Side	**Failure to separate left and right plantar pressure distribution** Trials in which both feet are in a single frame due to failed automated segmentation of individual footprints.
Inverted Orientation **Label 3**	 Left Side	 Right Side	**Non-standard orientation of plantar pressure maps** Plantar pressure maps with non-standard orientation, where the forefoot is not aligned upwards, e.g. due to incorrect foot placement on the pressure plate or erroneous test leader instructions.
Incorrect Side Annotation **Label 4**	 Right Side	 Left Side	**Incorrect side labeling** Plantar pressure distribution is valid, and acquisition quality is intact, but metadata incorrectly labels the side of the foot, e.g., right foot recorded but annotated as left.

Table 1. Overview of the data categories included in the dataset.

using the L-BFGS-B algorithm[34], ensuring that the transformations were constrained within realistic bounds. Figure 2 shows a visual example of the registration results.

To quantitatively verify the registration accuracy, the spatial overlap between each registered plantar pressure map and the corresponding reference was calculated by binarizing the images (pressure > 0 → 1, background → 0) and computing the Intersection over Union (IoU). Across the dataset, a mean IoU of 0.77 ± 0.11 on the valid samples was observed, indicating that the registration pipeline achieves reliable spatial correspondence for subsequent SPM analyses.

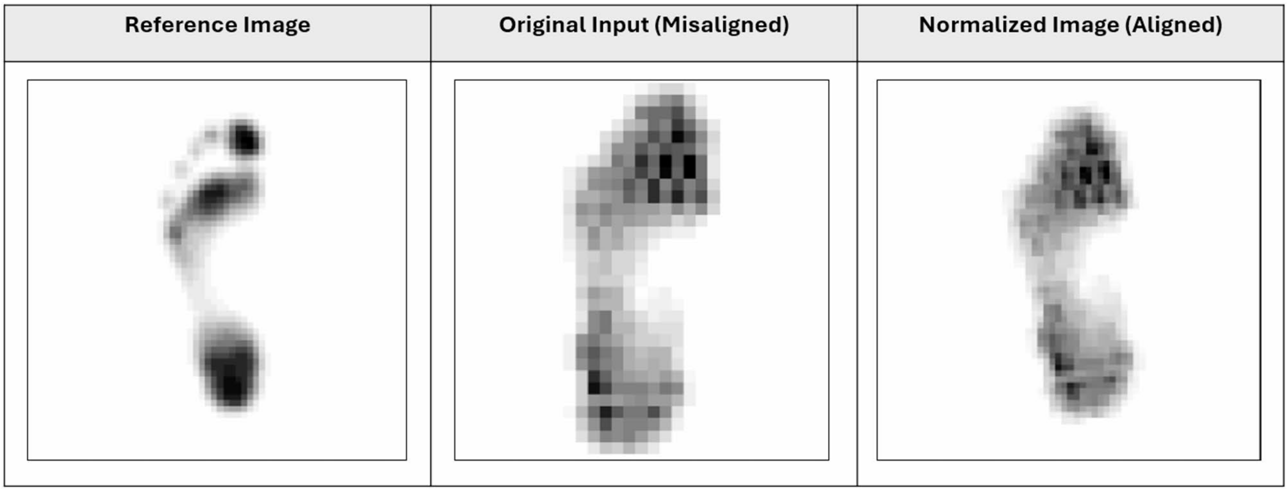


Fig. 2. Exemplary plantar pressure registration results.

Statistical analysis

We implemented a non-parametric SPM approach tailored for plantar pressure data, using cluster-based permutation testing to identify deviant (outlier) pressure maps. This approach extends the widely used methodology of non-parametric SPM[35,36] to the specific task of technical error or procedural inconsistency detection in plantar pressure data. The analysis is grounded in a normative modeling framework, where each test sample is statistically compared against a reference distribution derived from a cohort of non-outlier (healthy) plantar pressure maps. Because left and right feet differ in anatomy and loading patterns, analyses were conducted independently for each laterality, preventing anatomical misalignment even after spatial normalization.

Our methodology proceeds in two main steps:

1. Pixel-wise non-parametric testing: For each pixel in a test sample, we compute a two-tailed empirical p-value by comparing its pressure intensity against the empirical distribution at the same pixel across the normative reference cohort. This is achieved via a rank-based permutation approach[36], which is robust to non-Gaussian distributions and does not rely on parametric assumptions. The resulting p-map represents the probability of observing such an extreme pressure value under the normative model.
2. Cluster-based multiple comparison correction: To control the large number of simultaneous pixelwise tests, we employed a cluster-based permutation procedure[35]. First, clusters of contiguous suprathreshold pixels were identified using an uncorrected cluster-forming threshold (α_forming). Next, a null distribution of maximum cluster sizes was constructed by repeatedly permuting the normative data ($n = 1{,}000$ permutations) and recomputing the p-map, thereby quantifying the cluster sizes expected under the null hypothesis. Finally, only clusters whose size exceeded the (1 − α_FWE) percentile of the null distribution (here, α_FWE = 0.05) were retained as significant. This procedure controls the family-wise error rate at the cluster level and, by considering contiguous clusters rather than individual pixels, indirectly accounts for spatial correlations across neighboring pixels.

To increase robustness, we additionally required clusters to meet a minimum cluster size (min_cluster) criterion, which served as an initial filter before running the full permutation-based correction. The two key parameters of this—α_forming (cluster-forming threshold) and min_cluster (minimum cluster size)—were tuned in a nested cross-validation scheme (see Sect. *Evaluation and calculations*). Specifically, a randomized search was performed in the inner loop of the cross-validation, where α_forming was sampled uniformly from the range 0.01–0.05, and min_cluster from the range 0–30 pixels. Candidate parameter sets were evaluated on the validation folds using the F1-score (binary inlier vs. outlier classification), and the best-performing combination was carried forward to the outer cross-validation loop for final evaluation.

Machine learning approach

For the ML approach, the original (unregistered) plantar pressure data were used, as previous research has shown that deep learning models can effectively handle spatial misalignments in plantar pressure data[12]. A Convolutional Neural Network (CNN) classifier designed to process the plantar pressure data along with additional categorical metadata (foot lateral label) was implemented[37,38]. Parameter search was performed manually by evaluating the model's performance on the validation sets. We also evaluated transfer learning approaches using pre-trained image classification models (e.g., ResNet) fine-tuned on the plantar pressure data; however, these did not generalize well due to structural differences between plantar pressure images and the original training images, which is also supported by other research[39]. Starting with a more complex architecture, the model's complexity was gradually reduced until further reductions led to worse validation performance.

The resulting CNN backbone consists of three sequential convolutional blocks with increasing filter sizes (32, 64, 128). Each block is composed of two convolutional layers with a 3 × 3 kernel, followed by batch normalization, a ReLU activation function, and a 2 × 2 max-pooling layer. Dropout layers (Dropout2D) were included after each max-pooling operation to mitigate overfitting. The flattened output of the convolutional layers is concatenated with a learned embedding vector from a categorical embedding layer. This layer converts the integer-encoded side labels into an 8-dimensional vector representation. The combined vector is then passed to a classifier head comprising two fully connected layers with batch normalization and dropout. The final layer outputs the class probabilities.

The plantar pressure samples were normalized using the mean and standard deviation calculated exclusively from the training data of each cross-validation fold (see Sect. *Evaluation and calculations*) to prevent data leakage. The categorical variable was encoded as an integer and passed directly to the embedding layer. The data were structured using a custom dataset class, with class-balanced sampling to address label imbalance during training.

The model was trained using a weighted cross-entropy loss, where class weights were inversely proportional to their frequency in the training set. The Adam optimizer with a learning rate of 0.001 was used for weight updates. Training was managed with an early stopping mechanism, which halted the process if the validation loss did not improve for ten consecutive epochs, and the model with the lowest validation loss was retained.

To understand the key plantar presser regions that contributed to the CNN's predictions, we employed the XAI method SHapley Additive exPlanations (SHAP)[40]. This method explains the output of a ML model as a sum of contributions from each input feature, providing local interpretability. We used a Deep SHAP explainer, which is tailored for deep learning models. The explainer was trained on a background dataset of a random subset of 100 plantar pressure samples and their corresponding metadata from the training set. For each test sample, the SHAP explainer calculated the contribution of each sample pixel to the final prediction.

Evaluation and calculations

To ensure an unbiased assessment of the two approaches, we employed nested stratified cross-validation. The nested structure is critical for preventing data leakage and ensuring that the final performance metrics accurately reflect the models' generalization ability on unseen data[41]. The same data partitions were used for both approaches, enabling a direct and fair comparison of their performance. An outer 5-fold stratified cross-validation was used to partition the dataset into a training/validation set and a held-out test set. Within each outer training/validation fold, an inner stratified shuffle split (80/20 ratio) was applied to create a dedicated training set and a validation set for hyperparameter tuning. For both the SPM and ML approaches, the grouped data structure was respected, ensuring that all data from a single subject (e.g., both left and right foot) were confined to a single partition. This is important to mitigate the risk of artificially inflated performance due to anatomical or measurement similarities within a subject.

The best-performing model configuration—identified by its performance on the validation set during the inner loop—was then evaluated on the completely independent, held-out test set. This process was repeated for each fold of the outer loop. The final model's performance was assessed using the Matthews Correlation Coefficient (MCC) and F1-score, which are robust metrics for evaluating models in the presence of class imbalance[42,43]. To assess the influence of the synthetic outlier samples, these metrics were also computed exclusively on the real test data, with the synthetically generated outliers omitted from testing. Thanks to the grouped cross-validation scheme, which splits data by subject, synthetic outliers were generated only from the training and validation data within each fold, ensuring strict separation from the test set. To ensure a fair comparison, the multi-class predictions from the ML approach were additionally post-processed into a binary classification output, analogous to the SPM-inspired predictions. In addition, confusion matrices were generated to visualize class-wise prediction accuracy and to identify potential sources of bias. All modeling, training, and evaluation procedures were implemented in Python using PyTorch[44], scikit-learn[45], and SciPy[46], while visualizations were generated with Matplotlib[47] and Seaborn[48]. Computations were performed on a desktop equipped with an 11th Gen Intel Core i7-11800 H CPU, 16 GB of RAM, a 512 GB SSD, and an NVIDIA GeForce RTX 3070 Laptop GPU (8 GB).

Human-centered results evaluation

Visual representation

To provide a comprehensive understanding of the models' decision-making processes, a side-by-side visualization of the outputs from the SPM and the ML approach is provided. For each sample analyzed, we generated a three-panel figure. The leftmost panel displays the original grayscale plantar pressure. The central panel presents the output of the non-parametric SPM approach. Here, the original pressure is shown with a green contour line overlaying regions that were identified as statistically significant outliers according to the approach. This highlights the specific foot regions where pressure values deviate substantially from the normative, valid plantar pressure dataset.

The rightmost panel presents the explanation of the CNN model's predictions using SHAP values. It overlays a heatmap on the original plantar pressure, highlighting pixels that positively or negatively contributed to the model's output. This provides a visual representation of the most influential plantar pressure regions underlying the predicted classification. To enhance clarity, SHAP values below 20% of the maximum absolute value were omitted, and the resulting map was smoothed using a bilateral filter.

Semantic differential survey

For this exploratory part of the study, we recruited 16 participants (9 male, 7 female) with expertise in biomedical data analysis and plantar pressure assessment. Post-hoc power calculations were performed using G*Power (version 3.1)[49], indicating a power of ≈ 0.83 for large effects ($d \approx 0.8$) at $\alpha = 0.05$, and ≈ 0.50 when applying a Bonferroni correction for the comparisons. All participants held at least a university degree in sports science, biomechanics, or a health-related field, and reported extensive prior experience with plantar pressure data.

The participants were provided with written, standardized explanations on how to interpret the explanations provided. The presentation to participants and data collection were carried out using the digital survey platform LimeSurvey (LimeSurvey GmbH, Hamburg, Germany). The estimated duration for completing the entire survey was approximately 25 min.

To explore how end-users perceive the quality of explanations generated we employed a semantic differential survey. A semantic differential is a well-established psychometric technique in which respondents rate a concept along bipolar adjective scales, thereby yielding quantitative measures of subjective impressions[50].

Drawing on prior work in human-centered XAI and established criteria for evaluating explanation quality, we selected eight key attribute pairs that capture core dimensions relevant to users' understanding and trust in AI explanations[26,27,51]. These pairs were selected to evaluate two key aspects of explanations: their cognitive processing and their perceived utility. Each of the following adjective pairs was presented on a 7-point Likert scale, with the two extremes representing the poles of the pair:

- Understandability (Understandable – Unintelligible).
- Correctness (Correct – Incorrect).
- Trustworthiness (Trustworthy – Suspicious).
- Usefulness (Useful – Useless).
- Clarity (Clear – Unclear).
- Completeness (Complete – Incomplete).
- Simplicity (Simple – Complex).
- Relevance (Relevant – Irrelevant).

The left-to-right order of the attributes (e.g., "Correct-Incorrect" vs. "Incorrect-Correct") was randomized for each participant to minimize bias. Each approach was evaluated separately using the semantic differential scale, and only correctly classified samples were shown. This ensured that participants assessed the quality of the explanations themselves, not the model's prediction accuracy. Including misclassified cases introduced confounding effects during the pilot phase, as participants tended to judge the explanation in light of the error, making direct comparisons between approaches difficult. Given this and the exploratory nature of the study, we deliberately restricted the evaluation to correctly classified samples to maintain clarity and comparability.

To manage the workload for each participant, we selected a random subset of ten generated explanations to be evaluated by the participants. Each participant was presented with the *same* set of model predictions along with their corresponding explanations from both approaches, as described in Sect. *Visual representation*. This side-by-side presentation allowed for a direct evaluation of the relative strengths and weaknesses of each approach's explanations on the same task. To ensure a fair comparison, we removed the additional level of detail provided by the ML model, which not only indicated whether a sample was predicted as an outlier but also specified the type of outlier. This adjustment was made to avoid bias resulting from differences in the amount of information conveyed by the labels. Furthermore, we labeled the approaches A and B to prevent any bias that could arise from participants knowing which approach was used.

To assess potential statistical differences between the approaches for each semantic differential attribute, we applied the Wilcoxon signed-rank test as a non-parametric paired test. This choice accounts for the ordinal nature of the semantic differential data and its non-normal distribution. Since each subject rated ten images per approach, we first computed the median rating across the images for each subject and attribute. These median values were then used as the paired data in the Wilcoxon test. The test was applied separately for each of the eight semantic differential attributes using data from all participants. To control for multiple comparisons across all attributes, p-values were adjusted using the Bonferroni correction, considering the total number of tested attributes. The significance threshold was set at $\alpha = 0.05$.To assess the perceived consistency of the explanations, a question was posed to participants for each sample after they had evaluated both approaches. Using a 5-point Likert scale, we asked participants to rate their agreement with the following statement: "Both approaches A and B highlight similar features and reasoning behind the model's classification." This question was designed to gauge the experts' perception of explanation alignment between the two approaches. Finally, for each sample presented, participants were asked, "Which approach would you personally prefer?" Response options included: Approach A, approach B, a combination of both (as presented in the survey), or neither.

Results

Classification results

Results are summarized in Table 2. Overall, the ML approach outperformed the SPM approach. A detailed single-case analysis of the misclassified samples revealed that, for both approaches, valid samples incorrectly identified as outliers (false positives) were predominantly feet exhibiting pathological characteristics (e.g., hallux valgus, hammer toe, claw toe, flatfoot). Further analysis of the false negative samples of the SPM approach indicates that samples from the outlier class 4 (*Incorrect Side Annotation*) were most often misclassified as valid ($n = 109$), followed by classes 1 (*General Acquisition Error*; $n = 49$) and class 3 (*Inverted Orientation*; $n = 29$).

A confusion matrix for the ML approach is shown in Fig. 3. The lowest label accuracy was observed for the *General Acquisition Error* class. Interestingly, samples belonging to this class were most frequently misclassified as valid samples. For the *Inverted Orientation* class, the primary source of error was misclassification as *General Acquisition Error*.

	SPM approach		ML approach	
Confusion matrix	**754 (754)**	44 (44)	**783 (783)**	15 (15)
	237 (27)	**1763 (206)**	30 (4)	**1970 (229)**
MCC (min; max)	0.76; 0.81 (0.74; 0.83)		0.95; 0.98 (0.92; 0.98)	
MCC (mean ± std)	0.78 ± 0.02 (0.81 ± 0.03)		0.96 ± 0.01 (0.95 ± 0.01)	
F1-score (min; max)	0.92; 0.94 (0.86; 0.93)		0.98; 1.00 (0.94; 0.99)	
F1-score (mean ± std)	0.93 ± 0.01 (0.88 ± 0.03)		0.99 ± 0.00 (0.96 ± 0.02)	

Table 2. Results for the held-out test sets across all cross-validation folds, comparing the SPM and machine learning (ML) approaches. For comparability, predictions of the multiclass ML approach were reduced to a binary classification of outlier vs. non-outlier. The confusion matrices show actual classes on the rows and predicted classes on the columns, where the top row corresponds to valid samples and the bottom row corresponds to outliers. Correct predictions (true positives and true negatives) are highlighted in bold. Metrics include the minimum (min) and maximum (max) values, as well as the mean ± standard deviation (std) for the Matthews Correlation Coefficient (MCC) and F1-score (F1). Values shown in brackets represent performance calculated exclusively on the real test data, excluding the synthetically generated outliers, which were used only during training and validation.

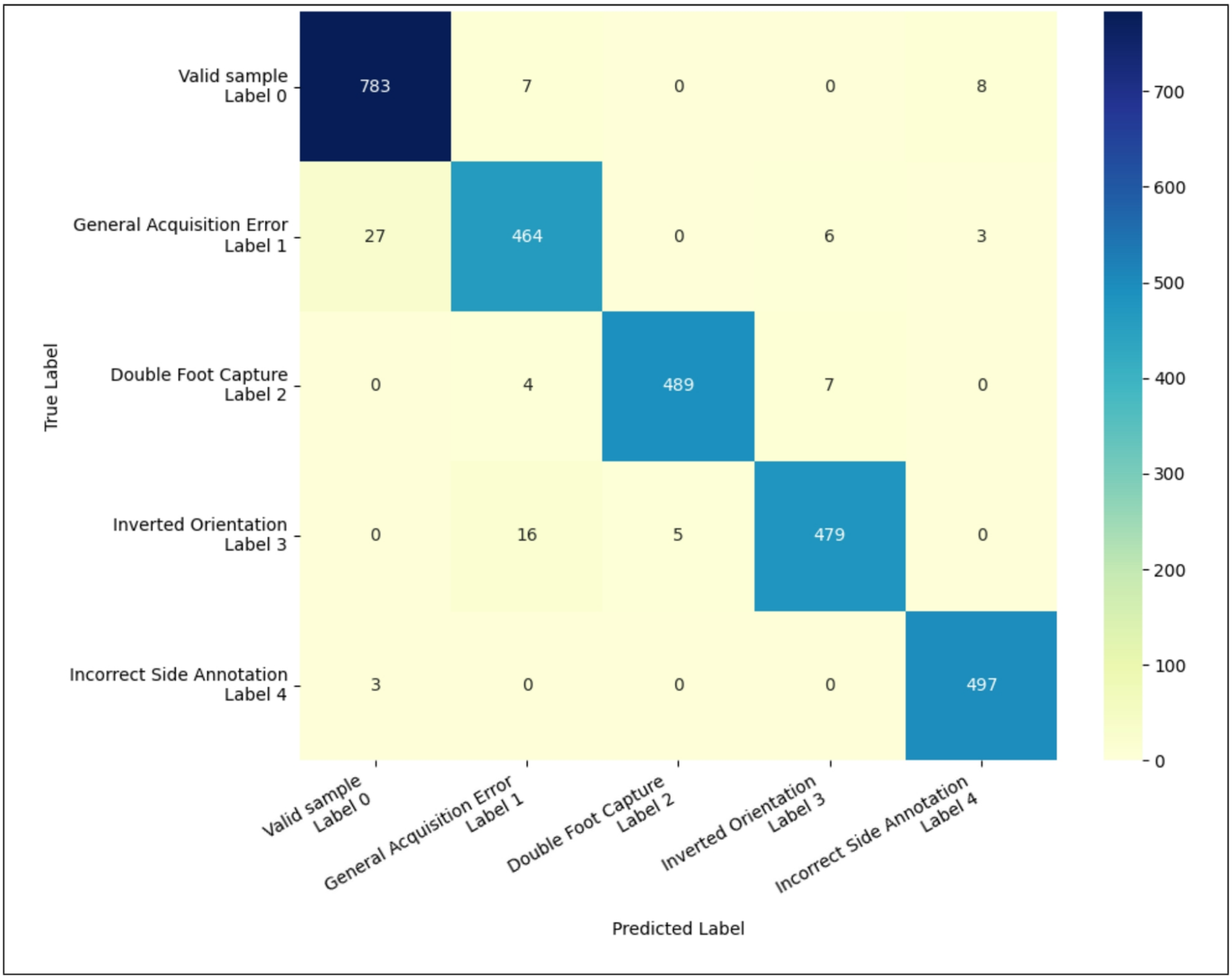


Fig. 3. Class-specific confusion matrices for the machine learning (ML) approach on the test set.

Semantic differential results

Figure 4 presents exemplary cases of generated explanations using both approaches. While the SPM-approach highlights clusters with statistically significant deviations from the valid dataset, the ML-approach with SHAP explanations highlights areas that contributed to or against the decision.

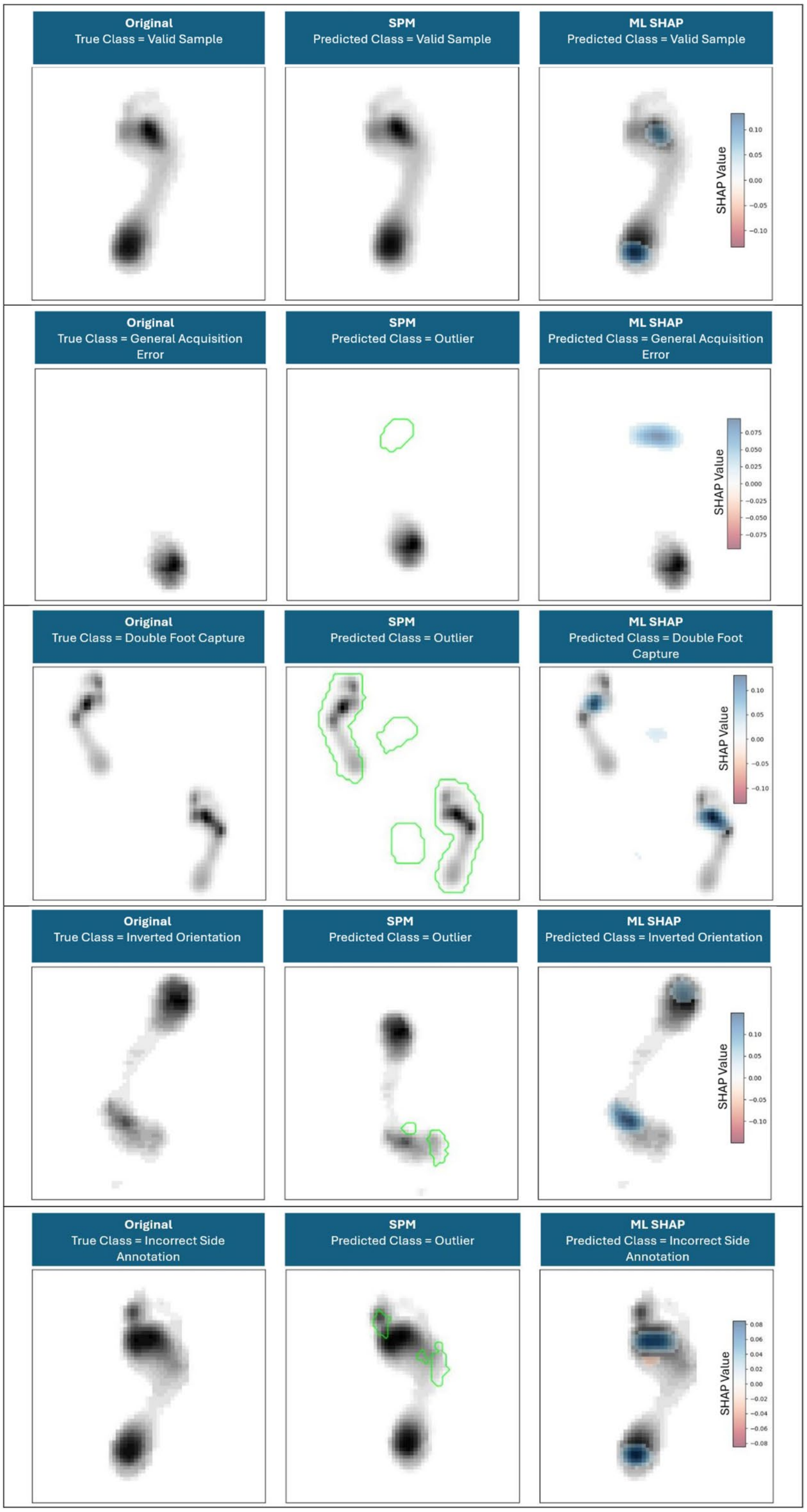


Fig. 4. Side-by-side visualization comparing outputs from the Statistical Parametric Mapping (SPM) approach and the machine learning (ML) approach for each data category (see Table 1). Each row shows: (left) the original grayscale plantar pressure, (middle) the SPM output with green contours marking regions identified as statistically significant outliers relative to a normative dataset, and (right) the ML explanation using SHAP values, where a heatmap overlays the original pressure distribution to highlight plantar pressure regions with the strongest positive or negative contributions to the model's prediction. Regions colored blue represent pixels that positively contribute to the model's prediction for the classified label, while regions colored red indicate pixels that push the prediction away from that label.

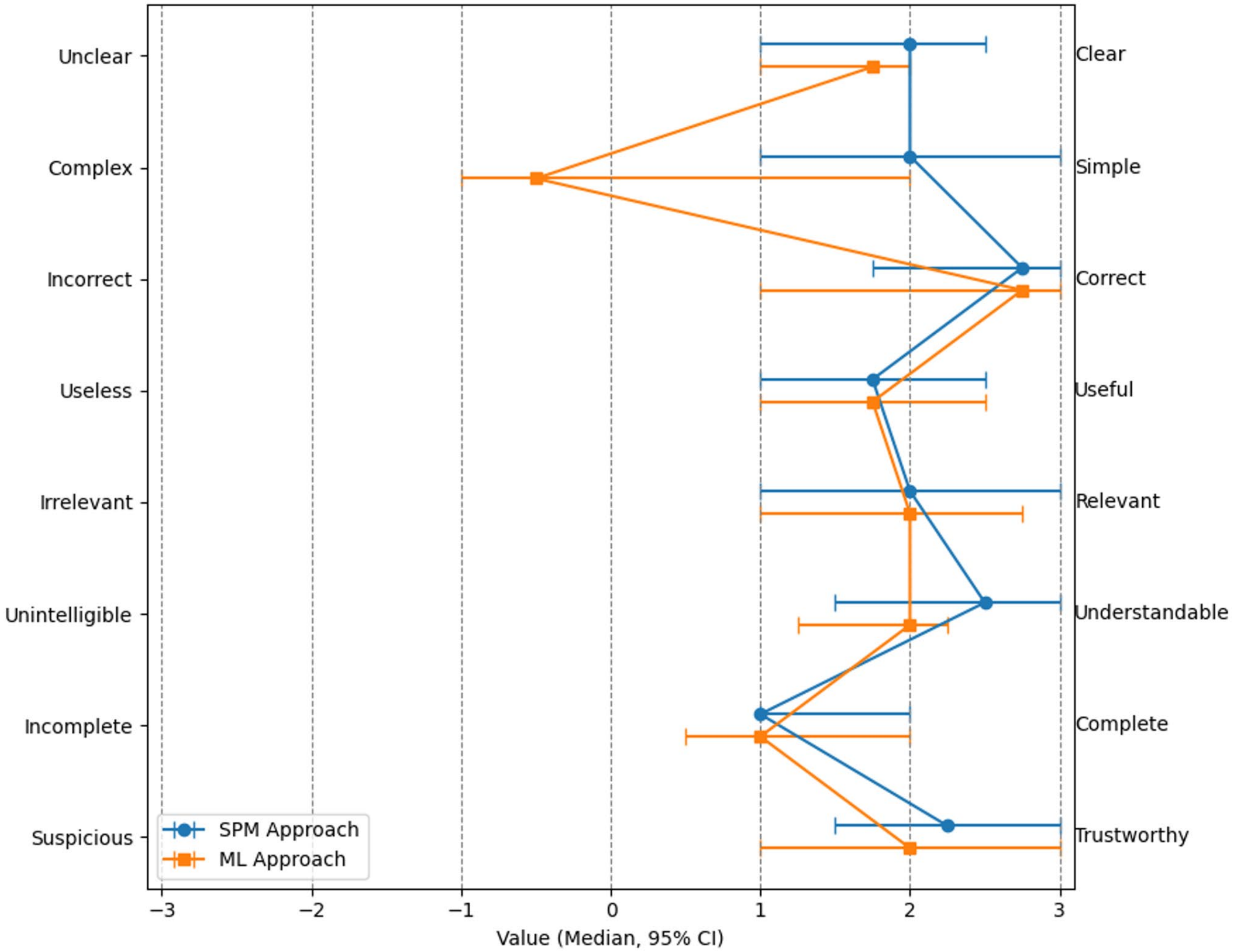


Fig. 5. Results of the semantic differential survey, shown separately for the SPM (blue) and ML (orange) approaches. For each subject, ratings across the 10 images per approach were summarized using the median, and these subject-level medians were then used to calculate overall median values across all participants for each attribute. Confidence intervals (95% CI) were estimated using bootstrapping with 5000 resamples of the subject-level medians. For clarity, positively connoted attributes are placed on the right side.

The human evaluation of the explanations is presented in Fig. 5. Overall, both approaches were rated positively, being perceived as clear, correct, useful, relevant, understandable, trustworthy, and relatively complete. A notable descriptive difference emerged with respect to simplicity: SPM was rated as simpler, whereas the ML approach was considered more complex and exhibited greater variability in participants' ratings. However, no statistically significant differences between the two approaches for any attribute in the semantic differential were observed ($p > 0.05$).

Experts rated the similarity between the SPM and ML approaches on a Likert scale, yielding a median score of 3.75 (median absolute deviation = 0.25). This indicates a relatively high level of perceived agreement between the two approaches. Regarding subjective preferences, 43.48% of the expert ratings favored the ML approach with SHAP explanations, followed by the SPM approach (34.78%) and the side-by-side presentation of both approaches (17.39%). Only 4.35% of ratings indicated no preference for any of the presented approaches.

Discussion

Both the SPM and ML approaches were able to detect outliers within plantar pressure data, but the ML approach outperformed the SPM approach in all evaluated metrics. Specifically, the ML model achieved an F1 score of 0.99 ± 0.00 and an MCC of 0.96 ± 0.01, compared to F1 = 0.93 ± 0.01 and MCC = 0.78 ± 0.02 for the SPM approach (research question 1). The dataset contained both real and synthetically generated outliers; however, performance estimates remained nearly unchanged when synthetic outliers were excluded from the test sets, suggesting that their inclusion did not bias the evaluation. This aligns with the nature of the task: the model learns to detect systematic structural or procedural abnormalities, such as missing regions or inverted orientation, rather than memorizing subject-specific plantar pressure characteristics. Consequently, the synthetic outliers appear to provide a practical means of ensuring sufficient representation of rare error types.

These results demonstrate the superior performance of the ML approach for outlier detection in this dataset. The observed differences in accuracy likely stem from the fundamentally different ways in which each approach interprets deviations. Within our dataset, clinically relevant but non-anomalous cases exist that exhibit substantial pixel-wise differences due to underlying pathological conditions or biomechanical adaptations. While these variations are clinically meaningful, they do not constitute technical errors or procedural inconsistencies. SPM is highly sensitive to localized, pixel-level variations and relies on precise normalization for rotation, scale, and anatomical landmarks to ensure that corresponding pixels accurately represent the same plantar region across subjects[17]. Although our proposed alignment procedure performed well visually, residual misalignments may have disproportionately affected SPM's classification performance.

This limitation is especially pronounced for pathological yet valid plantar pressure samples, where alignment with a normative reference plantar pressure distribution is challenging or even infeasible. For example, optimally registering a foot with Hallux Valgus to a normative template is nearly impossible, as either the hallux protrudes beyond the reference region or pressure voids appear where tissue is normally present. As a result, pathologic inliers were more often misclassified as outliers with the SPM approach due to their pixel-wise deviations from normative plantar pressure patterns. Regarding false negatives in the SPM approach, these largely arose from samples in outlier class 4 (*Incorrect Side Annotation*). This outcome may reflect the fact that lateral feet are relatively similar, and lateral differences are often localized and subtle. As a result, even when the lateral side is incorrectly labeled, the SPM approach may still interpret the sample as conforming to the normative dataset for that side, leading to misclassification.

The lowest label accuracy for the ML approach was observed for the outlier class *General Acquisition Error*, with 5.4% of the samples predominantly being wrongly classified as valid. A possible explanation is the class's inherent heterogeneity—ranging from subjects wearing shoes to instances with only partial plantar pressure data. Though combining these diverse characteristics into a single class was intended to boost sample size, it may have inadvertently compromised accuracy. Similar problems have been documented in other domains, such as medical imaging, where hidden stratification—arising from unrecognized heterogeneity within a class—has significantly reduced model performance[52]. Consequently, as the sample size of this outlier category grows, subdividing it into more homogeneous subclasses may enable finer-grained classification and improve outlier detection performance.

Overall, the two approaches were rated similarly on most attributes of the semantic differential, being generally rated positively across the dimensions, including clarity, correctness, usefulness, relevance, understandability, trustworthiness, and perceived completeness (research question 2). While the aggregated ratings suggest that experts generally found the explanations aligned with domain knowledge, individual cases reveal occasional differences in ratings, showing that full agreement with expert logic was not achieved for every sample. Nonetheless, the generally high usefulness ratings imply that both approaches could support expert understanding of the classification process, potentially facilitating interpretation of how specific features or regions contribute to model decisions. Taken together, these findings provide preliminary evidence that both approaches generate explanations that are interpretable and relevant from an expert perspective, though further investigation is needed to confirm the extent and robustness of this alignment.

The observed difference in perceived complexity was descriptive rather than statistically significant: although ratings exhibited high variability, SPM explanations were, on average, considered simpler. This descriptive trend aligns with expectations, as SHAP provides fine-grained, pixel-level attributions that detail how individual features contribute to predictions, whereas the SPM approach highlights only significant clusters of pixels, offering a less detailed but more immediately interpretable representation. The substantial variability in ratings, particularly for the ML approach using SHAP, likely contributed to the absence of statistical significance, reflecting the subjective nature of participants' perceptions of complexity.

The integration of ML-based and statistically driven explanations has been proposed as a promising avenue in XAI research, particularly in domains such as biomechanics where interpretability and trust are critical[16,53–55]. Previous findings suggest that the optimal XAI approach must be adapted to the user's context[56]. Consequently, providing both statistical and ML explanations allows end users to choose the representation that best fits their background and task requirements. Interestingly, our findings suggest that experts did not primarily value the combined presentation of SPM and ML explanations. Instead, the ML approach with SHAP explanations received the highest preference, followed by the SPM approach, while the side-by-side presentation of both approaches was less frequently favored. This indicates that, although SPM and ML operate at different levels of abstraction—feature-level versus group-level—the added value of presenting both simultaneously may not be as compelling to domain experts. At the same time, experts rated the overall similarity between the two approaches as relatively high, suggesting that despite methodological differences, both approaches were perceived as largely consistent. This perceived alignment may explain why experts felt comfortable selecting a single preferred approach rather than relying on a dual-validation perspective.

Finally, the necessity of XAI in the current study's task requires careful consideration. Although the classification task in this study is relatively straightforward for human experts, it is time-consuming, making automation valuable. In such scenarios, XAI primarily supports compliance with regulatory frameworks, fosters trust in automated systems, and facilitates human-in-the-loop monitoring[57]. It can also highlight cases where models fail, enhancing the robustness and reliability of ML-assisted decision-making. In our current evaluation, we focused on instances where SPM and SHAP explanations agreed with ground-truth labels. A practical workflow might prioritize ML predictions for decision-making, given their higher accuracy, while using SPM outputs as supporting explanations when the two approaches align.

This study has several limitations, which also suggest promising directions for future research. While our ML approach proved highly effective, simpler methods—such as expert-defined rules—could be explored for identifying specific outlier categories (e.g., multiple feet or upside-down feet). However, heuristic approaches tend to be less robust when handling highly abnormal foot shapes, as they may not adequately account for complex or anomalous data patterns[58]. The SPM approach is computationally intensive due to registration and voxel-wise analyses (~ 0.5–1 s per image), whereas the ML approach enables fast inference (~ 5–10 ms per image on GPU), making it better suitable for potential real-time applications. Hyperparameter search for the ML model was performed manually, guided by validation performance. While this approach yielded very good results on key metrics, it may limit the reproducibility and generalizability of the model, as more systematic optimization methods (e.g., automated search or ablation studies) could potentially uncover architectures that perform equally well or better. In the context of inference-statistics–based outlier detection, future work could investigate advanced biomechanical alignment techniques, such as deformable image registration[59]. These approaches may better accommodate anatomical variability in pathological feet, improving registration quality and enhancing the robustness of SPM-based outlier detection. Moving beyond pixel-wise comparisons, aggregating plantar pressure values over anatomically or functionally meaningful regions (e.g., heel, metatarsal heads) could also provide more clinically meaningful and functional metrics for comparison with normative datasets[60]. Our dataset included both real and artificially generated outlier samples. Although experts confirmed that the synthetic cases closely resembled realistic outliers, some residual bias cannot be ruled out. Future research could leverage generative AI to produce even more realistic artificial outliers, building on recent work demonstrating its capability to generate accurate biomechanical data and thus enhancing the utility of ML predictions in biomechanics[61–64].

Although explanations for misclassified samples are essential for evaluating the full utility of XAI systems, they constitute a separate research question beyond the scope of this exploratory study and should be addressed in future work. Moreover, our study focused exclusively on SHAP. Future work could systematically compare multiple XAI methods, assessing both their technical interpretability and perceived usefulness from the perspective of human experts. Expert ratings in our semantic differential analysis may have been influenced by differences in explanatory depth. Since standardized instruments for evaluating XAI explanations are still underdeveloped, our study represents an initial exploratory effort. Given the sample size, post hoc power analysis indicates moderate sensitivity for detecting large effects, whereas smaller or medium effects are unlikely to be detected. Consistent with the exploratory aim of this investigation, the focus on practically meaningful, large effects is statistically appropriate. Moreover, most studies employing the semantic differential technique are fundamentally descriptive rather than inferential[50], which further supports the methodological decisions regarding sample size and observed statistical trends. Consequently, the observed patterns offer valuable preliminary insights, while future, adequately powered investigations with larger samples will be required to confirm subtle effect sizes. Future research could also refine evaluation protocols by expanding the set of assessment attributes and applying factor analysis to capture latent dimensions of user perception.

Conclusion

This study demonstrates that the statistical SPM analysis and ML modeling are highly promising approaches for detecting and categorizing technical errors and procedural inconsistencies in plantar pressure data, thereby enabling automated and targeted guidance for addressing outliers. The results underscore that advancing artificial intelligence in biomechanics requires not only evaluating predictive performance but also understanding how users perceive model explanations. By applying semantic differential analysis to assess user perceptions of SPM and ML explanations, this study provides a first step toward developing human-centered tools that evaluate interpretability and practical usefulness, highlighting the need for further research in this direction.

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Received: 25 September 2025; Accepted: 22 December 2025
Published online: 11 January 2026

References

1. Kirtley, C. Clinical gait analysis: Theory and practice. Edinburgh and New York: Elsevier. ISBN: 9780702036712 (2006).
2. Mann, R., Malisoux, L., Urhausen, A., Meijer, K. & Theisen, D. Plantar pressure measurements and running-related injury: A systematic review of methods and possible associations. *Gait Posture.* **47**, 1–9. https://doi.org/10.1016/j.gaitpost.2016.03.016 (2016).
3. Arzehgar, A. et al. An overview of plantar pressure distribution measurements and its applications in health and medicine. *Gait Posture.* **117**, 235–244. https://doi.org/10.1016/j.gaitpost.2024.12.022 (2025).
4. Cao, Z. et al. Characteristics of Plantar Pressure Distribution in Diabetes with or without Diabetic Peripheral Neuropathy and Peripheral Arterial Disease. *J. Healthc. Eng.* 2437831. https://doi.org/10.1155/2022/2437831 (2022).
5. Zou, Y. F. et al. Clinical utility of plantar pressure measurements as screening in patients with Parkinson disease with and without freezing of gait history. *Arch. Phys. Med. Rehabil.* **104**, 1091–1098. https://doi.org/10.1016/j.apmr.2023.02.019 (2023).
6. Buldt, A. K. et al. Foot posture is associated with plantar pressure during gait: A comparison of normal, planus and cavus feet. *Gait Posture.* **62**, 235–240. https://doi.org/10.1016/j.gaitpost.2018.03.005 (2018).
7. Hofmann, U. K. et al. Transfer of plantar pressure from the medial to the central forefoot in patients with hallux valgus. *BMC Musculoskelet. Disord.* **20**. https://doi.org/10.1186/s12891-019-2531-2 (2019).
8. da Silveira, G. E. et al. The Effects of Short- and Long-Term Spinal Brace Use with and without Exercise on Spine, Balance, and Gait in Adolescents with Idiopathic Scoliosis. *Medicina* **58**. https://doi.org/10.3390/medicina58081024 (2022).
9. Güven, E., Çıtaker, S. & Alsancak, S. The effect of orthotics on plantar pressure in children with infantile tibia vara (Blount's disease). *Sci. Rep.* **13**, 2875. https://doi.org/10.1038/s41598-023-30066-4 (2023).
10. Ma, M., Song, Q. & Liu, H. The effect of personalized orthopedic insoles on plantar pressure during running in subtle cavus foot. *Front. Bioeng. Biotechnol.* **12**, 1343001. https://doi.org/10.3389/fbioe.2024.1343001 (2024).

11. Zhang, C., Pan, S., Qi, Y. & Yang, Y. A footprint extraction and recognition algorithm based on plantar pressure. *TS* **36**, 419–424. https://doi.org/10.18280/ts.360506 (2019).
12. Dindorf, C. et al. Toward automated plantar pressure analysis: machine learning-based segmentation and key point detection across multicenter data. *Front. Bioeng. Biotechnol.* **13**, 1579072. https://doi.org/10.3389/fbioe.2025.1579072 (2025).
13. Wang, D. et al. Deep-segmentation of plantar pressure images incorporating fully convolutional neural networks. *Biocybernetics Biomedical Eng.* **40**, 546–558. https://doi.org/10.1016/j.bbe.2020.01.004 (2020).
14. Chae, J., Kang, Y. J. & Noh, Y. A Deep-Learning approach for Foot-Type classification using heterogeneous pressure data. *Sens*. **20**. https://doi.org/10.3390/s20164481 (2020).
15. Han, J. et al. Plantar pressure image classification employing residual-network model-based conditional generative adversarial networks: a comparison of normal, planus, and Talipes equinovarus feet. *Soft Comput.* **27**, 1763–1782. https://doi.org/10.1007/s00500-021-06073-w (2023).
16. Stetter, B. J., Stein, T. Machine Learning in Biomechanics: Enhancing Human Movement Analysis. In: *Artificial Intelligence in Sports, Movement, and Health.* (eds. Dindorf, C., Bartaguiz, E., Gassmann, F. & Fröhlich, M.) (Springer, Cham, 2024). https://doi.org/10.1007/978-3-031-67256-9_9.
17. Pataky, T. C. & Goulermas, J. Y. Pedobarographic statistical parametric mapping (pSPM): a pixel-level approach to foot pressure image analysis. *J. Biomech.* **41**, 2136–2143. https://doi.org/10.1016/j.jbiomech.2008.04.034 (2008).
18. Friston, K. J. et al. Statistical parametric maps in functional imaging: A general linear approach. *Hum. Brain. Mapp.* **2**, 189–210. https://doi.org/10.1002/hbm.460020402 (1994).
19. Booth, B. G. et al. Personalized analysis of plantar pressure images using statistical modelling and parametric mapping. *PloS One*. **15**, e0229685. https://doi.org/10.1371/journal.pone.0229685 (2020).
20. Vieira, S., Pinaya, W. H. L. & Mechelli, A. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: methods and applications. *Neurosci. Biobehav. Rev.* **74**, 58–75. https://doi.org/10.1016/j.neubiorev.2017.01.002 (2017).
21. Tschuchnig, M. E. & Gadermayr, M. Anomaly Detection in Medical Imaging - A Mini Review. *International Data Science Conference*, 33–38. https://doi.org/10.1007/978-3-658-36295-9_5 (2022).
22. Baur, C., Wiestler, B., Albarqouni, S. & Navab, N. Deep Autoencoding Models for Unsupervised Anomaly Segmentation in Brain MR Images. In: Brainlesion: *Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2018. Lecture Notes in Computer Science*, Vol. 11383. (eds. Crimi, A. et al.) (Springer, Cham, 2019). https://doi.org/10.1007/978-3-030-11723-8_16.
23. European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) 95/46/EC (General Data Protection Regulation). 119th ed. https://eur-lex.europa.eu/eli/reg/2016/679/oj (2016).
24. Varshney, K. R. Trustworthy machine learning and artificial intelligence. *XRDS* **25**, 26–29. https://doi.org/10.1145/3313109 (2019).
25. Arya, V. et al. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. https://doi.org/10.48550/arXiv.1909.03012 (2019).
26. Doshi-Velez, F. & Kim, B. Towards a rigorous science of interpretable machine learning. https://doi.org/10.48550/arXiv.1702.08608 (2017).
27. Mohseni, S., Zarei, N. & Ragan, E. D. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Trans. Interact. Intell. Syst.* **11**, 1–45. https://doi.org/10.1145/3387166 (2021).
28. Sejr, J. H. & Schneider-Kamp, A. Explainable outlier detection: What, for Whom and Why? *Mach. Learn. Appl.* **6**, 100172. https://doi.org/10.1016/j.mlwa.2021.100172 (2021).
29. Dindorf, C. et al. Machine Learning in Biomechanics: Key Applications and Limitations in Walking, Running and Sports Movements. In: *Artificial Intelligence, Optimization, and Data Sciences in Sports. Springer Optimization and Its Applications*, vol 218. (eds. Blondin, M. J., Fister Jr., I. & Pardalos, P. M.) (Springer, Cham, 2025). https://doi.org/10.1007/978-3-031-76047-1_4.
30. Dina, A. S., Siddique, A. B. & Manivannan, D. Effect of balancing data using synthetic data on the performance of machine learning classifiers for intrusion detection in computer networks. *IEEE Access.* **10**, 96731–96747. https://doi.org/10.1109/ACCESS.2022.3205337 (2022).
31. Oliveira, F. P. M. & Tavares, J. M. R. S. Novel framework for registration of pedobarographic image data. *Med. Biol. Eng. Comput.* **49**, 313–323. https://doi.org/10.1007/s11517-010-0700-4 (2011).
32. Goshtasby, A. A. 2-D and 3-D image registration for medical, remote sensing, and industrial applications. (John Wiley & Sons, Inc., New Jersey, 2005) https://doi.org/10.1002/0471724270.
33. Zitová, B. & Flusser, J. Image registration methods: a survey. *Image Vis. Comput.* **21**, 977–1000. https://doi.org/10.1016/S0262-8856(03)00137-9 (2003).
34. Byrd, R. H., Lu, P., Nocedal, J. & Zhu, C. A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.* **16**, 1190–1208. https://doi.org/10.1137/0916069 (1995).
35. Maris, E. & Oostenveld, R. Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods*. **164**, 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024 (2007).
36. Oostenveld, R., Fries, P., Maris, E. & Schoffelen, J. M. FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. Intell. Neurosci.* **2011** (156869). https://doi.org/10.1155/2011/156869 (2011).
37. Litjens, G. et al. A survey on deep learning in medical image analysis. *Med. Image. Anal.* **42**, 60–88. https://doi.org/10.1016/j.media.2017.07.005 (2017).
38. Wang, X. IEEE, et al. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly Supervised Classification and Localization of Common Thorax Diseases. *IEEE/CVF CVPR* 3462–3471. https://doi.org/10.1109/CVPR.2017.369 (2017).
39. Tartaglione, E. et al. COVID-19 from CHEST X-Ray with deep learning: A hurdles race with small data. *Int. J. Environ. Res. Public Health*. **17**. https://doi.org/10.3390/ijerph17186933 (2020).
40. Lundberg, S. & Lee, S. I. A Unified Approach to Interpreting Model Predictions. https://doi.org/10.48550/arXiv.1705.07874 (2017).
41. Sasse, L. et al. Overview of leakage scenarios in supervised machine learning. *J. Big Data*. **12**. https://doi.org/10.1186/s40537-025-01193-8 (2025).
42. Chicco, D. & Jurman, G. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Min.* **16**. https://doi.org/10.1186/s13040-023-00322-4 (2023).
43. Branco, P., Torgo, L. & Ribeiro, R. A survey of predictive modelling under imbalanced distributions. https://doi.org/10.48550/arXiv.1505.01658 (2015).
44. Paszke, A. et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. https://doi.org/10.48550/arXiv.1912.01703 (2019).
45. Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
46. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in python. *Nat. Methods.* **17**, 261–272. https://doi.org/10.1038/s41592-019-0686-2 (2020).
47. Hunter, J. D. & Matplotlib A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95. https://doi.org/10.1109/MCSE.2007.55 (2007).
48. Waskom, M. Seaborn: statistical data visualization. *JOSS* **6**, 3021. https://doi.org/10.21105/joss.03021 (2021).
49. Faul, F., Erdfelder, E., Buchner, A. & Lang, A. G. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. *Behav. Res. Methods.* **41**, 1149–1160. https://doi.org/10.3758/BRM.41.4.1149 (2009).

50. Osgood, C. E., Suci, G. J. & Tannenbaum, P. H. *The Measurement of Meaning* (University of Illinois Press, 1978).
51. Kim, J., Maathuis, H. & Sent, D. Human-centered evaluation of explainable AI applications: a systematic review. *Front. Artif. Intell.* **7**, 1456486. https://doi.org/10.3389/frai.2024.1456486 (2024).
52. Oakden-Rayner, L., Dunnmon, J., Carneiro, G. & Ré, C. Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. *Proceedings of the ACM Conference on Health, Inference, and Learning 2020*, 151–159. https://doi.org/10.1145/3368555.3384468 (2020).
53. Slijepcevic, D. et al. Explaining machine learning models for clinical gait analysis. *ACM Trans. Comput. Healthc.* **3**, 1–27. https://doi.org/10.1145/3474121 (2022).
54. Dindorf, C., Teufl, W., Taetz, B., Bleser, G. & Fröhlich, M. Interpretability of Input Representations for Gait Classification in Patients after Total Hip Arthroplasty. *Sensors.* **20**. https://doi.org/10.3390/s20164385 (2020).
55. Kokkotis, C. et al. Leveraging explainable machine learning to identify gait Biomechanical parameters associated with anterior cruciate ligament injury. *Sci. Rep.* **12**, 6647. https://doi.org/10.1038/s41598-022-10666-2 (2022).
56. Cugny, R., Aligon, J., Chevalier, M., Roman Jimenez, G. & Teste, O. Association for Computing Machinery. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. (eds. Al Hasan, M. & Xiong, L.) 315–324. https://doi.org/10.1145/3511808.3557247 (2022).
57. Kieseberg, P., Schantl, J., Frühwirt, P., Weippl, E. & Holzinger, A. Witnesses for the Doctor in the Loop. In: *Brain Informatics and Health. BIH 2015. Lecture Notes in Computer Science*, vol 9250. (eds. Guo, Y. et al.) (Springer, Cham, 2015). https://doi.org/10.1007/978-3-319-23344-4_36.
58. Kanno, Y. Simple heuristic for data-driven computational elasticity with material data involving noise and outliers: a local robust regression approach. *Japan J. Indust Appl. Math.* **35**, 1085–1101. https://doi.org/10.1007/s13160-018-0323-y (2018).
59. Sotiras, A., Davatzikos, C. & Paragios, N. Deformable medical image registration: a survey. *IEEE Trans. Med. Imaging.* **32**, 1153–1190. https://doi.org/10.1109/TMI.2013.2265603 (2013).
60. Bai, X. et al. Plantar pressure classification and feature extraction based on multiple fusion algorithms. *Sci. Rep.* **15**, 13274. https://doi.org/10.1038/s41598-025-96440-6 (2025).
61. Bicer, M., Phillips, A. T. M., Melis, A., McGregor, A. H. & Modenese, L. Generative deep learning applied to biomechanics: A new augmentation technique for motion capture datasets. *J. Biomech.* **144**, 111301. https://doi.org/10.1016/j.jbiomech.2022.111301 (2022).
62. Dindorf, C. et al. Enhancing Biomechanical machine learning with limited data: generating realistic synthetic posture data using generative artificial intelligence. *Front. Bioeng. Biotechnol.* **12**, 1350135. https://doi.org/10.3389/fbioe.2024.1350135 (2024).
63. Halmich, C., Höschler, L., Schranz, C. & Borgelt, C. Data augmentation of time-series data in human movement biomechanics: A scoping review. *PloS One.* **20**, e0327038. https://doi.org/10.1371/journal.pone.0327038 (2025).
64. Kárason, H., Ritrovato, P., Maffulli, N. & Tortorella, F. Generative Data Augmentation of Human Biomechanics. In: *Image Analysis and Processing - ICIAP 2023 Workshops. ICIAP 2023. Lecture Notes in Computer Science, vol 14365.* (eds. Foresti, G. L., Fusiello, A. & Hancock, E.) (Springer, Cham, 2024). https://doi.org/10.1007/978-3-031-51023-6_40.

Author contributions

CaD: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review and editing. JD: Conceptualization, Validation, Writing – original draft, Writing – review and editing. SS: Writing – original draft, Writing – review and editing. DP: Conceptualization, Investigation, Writing – review and editing. SB: Writing – original draft, Writing – review and editing. HE: Conceptualization, Investigation, Writing – review and editing. KH: Conceptualization, Software. BS: Conceptualization, Writing – review and editing. ChD: Conceptualization, Investigation, Writing – review and editing. MF: Funding acquisition, Project administration, Supervision, Writing – review and editing.

Funding

Open Access funding enabled and organized by Projekt DEAL. The authors declare that financial support was received for the research and/or publication of this article. This research was supported by the Central Innovation Program for Small and Medium-Sized Enterprises (Zentrales Innovations Program Mittelstand, ZIM) of the German Federal Ministry for Economic Affairs and Climate Action under Grant numbers 16KN113027 and KK5209402NK4.

Declarations

Competing interests

Authors HE, KH, and ChD were employed by DIERS International GmbH. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Additional information

Correspondence and requests for materials should be addressed to C.D.

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

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