Computers in Biology and Medicine 184 (2025) 109338

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Explaining deep learning models for age-related gait classification based on acceleration time series

Xiaoping Zheng [a], Egbert Otten [a], Michiel F. Reneman [b], Claudine JC. Lamoth [a,*]

[a] University of Groningen, University Medical Center Groningen, Department of Human Movement Sciences, 9713 AV, Groningen, the Netherlands
[b] University of Groningen, University Medical Center Groningen, Department of Rehabilitation Medicine, 9751 ND, Groningen, the Netherlands

ARTICLE INFO

Keywords:
Accelerometers
Healthy ageing
Deep learning
Gait analysis
Machine learning
Explainable artificially intelligence
SHAP

ABSTRACT

Background: Gait analysis holds significant importance in monitoring daily health, particularly among older adults. Advancements in sensor technology enable the capture of movement in real-life environments and generate big data. Machine learning, notably deep learning (DL), shows promise to use these big data in gait analysis. However, the inherent black-box nature of these models poses challenges for their clinical application. This study aims to enhance transparency in DL-based gait classification for aged-related gait patterns using Explainable Artificial Intelligence, such as SHapley Additive exPlanations (SHAP).

Methods: In this cross-sectional study, a total of 244 participants, comprising 129 adults and 115 older adults (age>65), were included. They performed a 3-min walking task while accelerometers were affixed to the lumbar segment L3. DL models, convolutional neural network (CNN) and gated recurrent unit (GRU), were trained using 1-stride and 8-stride accelerations, respectively, to classify adult and older adult groups. SHAP was employed to explain the models' predictions.

Results: CNN achieved a satisfactory performance with an accuracy of 81.4 % and an AUC of 0.89, and GRU demonstrated promising results with an accuracy of 84.5 % and an AUC of 0.94. SHAP analysis revealed that both CNN and GRU assigned higher SHAP values to the data from vertical and walking directions, particularly emphasizing data around heel contact, spanning from the terminal swing to loading response phases. Furthermore, SHAP values indicated that GRU did not treat every stride equally.

Conclusion: CNN accurately distinguished between adults and older adults based on the characteristics of a single stride's data. GRU achieved accurate classification by considering the relationships and subtle differences between strides. In both models, data around heel contact emerged as most critical, suggesting differences in acceleration and deceleration patterns during walking between different age groups.

1. Introduction

Gait analysis plays a significant role in monitoring the quality of life, particularly among older individuals, since maintaining mobility and independence in later years is essential [1,2]. Gait performance can provide insights into the control and coordination of various systems, such as the neuromusculoskeletal system and nervous system [3,4]. Aging, as a continuous process, is often associated with the loss of muscle mass, decreased bone density, and declining nerve function, which can result in an altered gait pattern [5].

With the development of miniaturization of sensors (e.g., accelerometers), modern movement tracking systems can provide vast amounts of reliable data about human movements [6], allowing for the diagnosis, monitoring, and rehabilitation of gait patterns in daily living environments [4,7]. Given the high variability, dimensionality, non-linear interactions, and temporal dependencies of the data collected during walking, traditional statistical approaches have limited capabilities [8]. Therefore, machine learning (ML) approaches have gained importance in clinical gait analysis due to their ability to handle complex data.

ML has demonstrated promising results in clinical gait classification tasks. For example, Artificial Neural Network (ANN) achieved high accuracy (90 %) in classifying different age groups gait based on handcrafted gait outcomes [9], such as step length and step frequency which extracted from accelerometers. The extraction of these features involves manual design and selection by experts, including clinicians, rehabilitation specialists, and human movement scientists. This process requires

* Corresponding author.
E-mail address: c.j.c.lamoth@umcg.nl (C.JC. Lamoth).

https://doi.org/10.1016/j.compbiomed.2024.109338
Received 25 January 2024; Received in revised form 17 October 2024; Accepted 28 October 2024
Available online 12 November 2024

specialized knowledge and can be labor-intensive [10]. Deep learning (DL) can perform gait classification based on raw sensor signals and has demonstrated superior performance [7,11]. A recent study compared the classification performance of recurrent neural networks (bidirectional long short-term memory) and conventional ML (support vector machine (SVM) and linear regression) in classifying fallers and non-fallers in patients with multiple sclerosis [12]. The results of this study indicated that the DL approach outperformed conventional ML (area under the curve: 0.88 vs. 0.79 for SVM).

However, many ML models suffer from a lack of transparency and interpretability due to their black-box nature [13]. It is often unclear why a specific decision has been made, even though the mathematical principles underlying these methods are well-established and well-understood. In clinical settings, decisions directly impact patient care and outcomes. Healthcare regulations often require clear rationales for clinical decisions. The opacity of ML models challenges patient and clinician trust, significantly limiting these models' practical applications in clinical contexts. Furthermore, this lack of transparency does not comply with the requirements of the European General Data Protection Regulation (GDPR, EU 2016/679) [14], which mandates the explanation of the logic behind any automated decision-making process that significantly affects individuals. Apart from this, the black-box nature makes it impossible to know what the model has truly learned, consequently obstructing the potential for generating new knowledge and a better understanding of human gait movement.

To overcome these limitations, Explainable Artificial Intelligence (XAI) has gained attention in the field of medicine [15]. XAI is an approach aimed at revealing the reasoning behind a system's predictions and decisions, which becomes even more critical when handling sensitive and personal health data. XAI can be broadly categorized into two main categories based on the stage of use: 1) ante-hoc explainability; and 2) post-hoc explainability [16]. Ante-hoc explainability refers to simple models that are interpretable by design, such as linear regression models, decision trees, k-nearest neighbour models, and Bayesian models [17]. However, it is often assumed that ante-hoc explainable models do not achieve satisfactory performance; therefore, opaque models (such as DL) are frequently employed [17]. This leads to post-hoc explainability approaches, which can be used to explain a previously trained model or its prediction.

Layer-wise Relevance Propagation (LRP) [18,19], Local Interpretable Model-Agnostic Explanations (LIME) [20], and SHapley Additive exPlanations (SHAP) [4] are the popular post-hoc explainability approaches. LRP propagates relevance scores from the output layer back to the input layer to determine the relevance of each input variable to the output decision. LIME perturbs the original data to observe how it affects predictions and aims to provide interpretable and faithful explanations, but it suffers from instability. It has been reported that two very close input samples may get greatly varied explanations in a simulated setting [21]. Shapley values, derived from game theory [22], provide a fair distribution of the collective value generated by a group of collaborating agents. This method ensures that each agent's share is proportional to their unique contribution to the group's overall success. SHAP leverages these values to quantify the impact of each input variable on a specific prediction. With a robust theoretical foundation, SHAP ensures that the contribution of each input variable is fairly and efficiently distributed among all the input variables of the instance [23]. This efficiency property distinguishes SHAP from other approaches and suggests that it might be the only current approach that provides a full and fair explanation for the prediction of a ML model. This property may highlight the potential advantages of SHAP values in terms of fairness and legal compliance in certain situations.

The application of XAI approaches in DL-based clinical gait analysis is still in early stages. One study used LRP to explain a convolutional neural network (CNN) in classifying individual gait patterns based on one stride data collected by ground reaction forces and full-body joint angles [18]. In another study, CNN was used to classify the walking of healthy participants while performing four different dual tasks. The classification was based on data from no more than two strides, derived from ground reaction force and plastic optical fiber distributed sensors [19]. In this study, LRP was used to indicate which parts of the signal had the heaviest influence on the gait classification. The studies mentioned above were confined to the analysis of no more than two gait strides data which were collected using force plates. Each of these studies employed a CNN model for classification which excels at capturing local temporal and spatial features. However, such methodological constraints may limit the detection of long-term dependent changes in gait, which provide information about postural control ability [24].

The present study aimed to improve the transparency of DL-based gait classification, with acceleration time-series obtained during a 3-min walking task. The goal was to differentiate between the gait patterns of adults and older adults. More specifically, we: 1) employed a CNN and a gated recurrent unit (GRU, designed to learn long-term dependent features); 2) utilized the SHAP approach to indicate the importance of the input signal in classification for both models. This study could contribute to improving the transparency and interpretability of DL-based gait analysis and potentially lead to better clinical decision-making.

2. Methods

In this cross-sectional study, the overview of data acquisition and analysis is presented in Fig. 1, with a CNN being used as the example. The data were collected during a 3-min walking task (Fig. 1(a)). Stride data from all participants underwent preprocessing before being employed to train the CNN model which aimed to classify participants into adult and older adult groups (Fig. 1(b)). To interpret the CNN model, the SHAP approach was applied. The SHAP values were visualized using a colour spectrum to illustrate the contributions of input data to the classification process (Fig. 1(c)). In this representation, deeper red indicates a higher contribution.

2.1. Participants, equipment, and data collection

The dataset consisted of 386 participants, which were derived by merging data from existing datasets [25–29]. The datasets included participants who: (1) could walk safely for 3 min without assistance or a walking aid, and (2) did not have mobility disabilities caused by pain or neurological or orthopedic conditions affecting either leg that would prevent them from walking for 3 min without a walking aid. The Mini-Mental State Examination (MMSE) was used to evaluate cognitive impairment in older adults. The MMSE is a 30-point questionnaire, with scores of 24 or higher indicating normal cognition. In this study, participants with MMSE scores lower than 24 (n = 104) were excluded. Additionally, participants with insufficient walking data (n = 38, having fewer than 10 segments of 8 consecutive strides) were also excluded. Consequently, the remaining participants (n = 244) were categorized into two groups: adults (ages 18–65, n = 129) and older adults (ages >65, n = 115). For the adult group, the mean age was 38.3 years (SD: 15.4), with mean height and weight measurements of 175.8 cm (SD: 9.0) and 75.4 kg (SD: 11.4), respectively. In the older adult group, the mean age was 76.7 years (SD: 5.9), the mean height was 166.7 cm (SD: 8.4), and the mean weight was 72.0 kg (SD: 11.9). During the study, participants were instructed to walk at a comfortable speed for 3 min while wearing an accelerometer (iPod, dynamic range ±3 g; Dynaport, dynamic range ±8 g; or ActiGraph, dynamic range ±6 g) fixed to a belt near the lumbar segment L3. Device placement followed a standardized procedure outlined in a manual, ensuring consistent positioning and orientation of all devices and sensors. When assuming a standing and upright position, the orientation of the axes was as follows: the X-axis pointed toward the ground (representing the vertical direction, V), the Y-axis faced the walking direction (indicating the anteroposterior

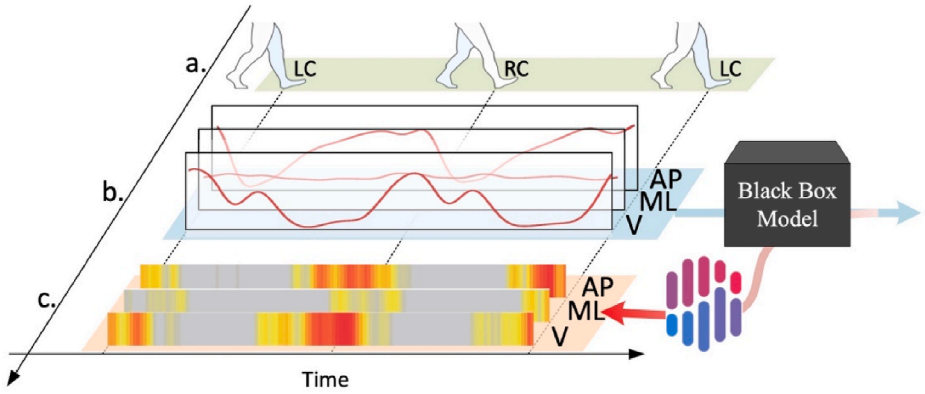

Fig. 1. Overview of data acquisition and analysis of CNN. (a): walking data collection; (b): preprocessing stride data and training CNN based on one stride data; (c) interpreting the CNN model by SHAP, deeper red colour represents a higher contribution to the classification process. CNN: convolutional neural network; LC: left contact; RC: right contact; V: vertical direction; ML: medio-lateral direction; AP: anterior-posterior direction.

direction, AP), and the Z-axis was perpendicular to the walking direction, extending from the patient's left to right (representing the mediolateral direction, ML). The sampling frequency was 100 Hz.

The studies were conducted between 2008 and 2022 and were approved by the Medical Ethical Committee of the University Medical Centre Groningen and the Medical Ethical Committee of the Slotervaart Hospital. All participants provided written informed consent in accordance with the Declaration of Helsinki.

2.2. Data preprocessing and data splitting

The sensor data were first checked to ensure the correct placement of the sensor. If incorrect placement was detected, the data were rotated to the correct orientation. To remove high-frequency noise, a second-order Butterworth low-pass filter with a cut-off frequency of 10 Hz was employed [30]. The resulting filtered signal was normalized to a range of −1 to 1 using min-max normalization. A stride is defined as the period from the initial left heel contact to the right heel contact and back to the subsequent left heel contact. It comprises two steps. Heel contact was detected based on the peaks of both AP- and V-axis acceleration data which correspond to the impact of the heel on the ground [31]. The left or right foot was determined by the values in the acceleration of ML direction. During the walking, the sensor sways with the movement of the body in the ML directions, and left heel contacts show higher readings in the ML direction than right contacts. To align the starting and ending timing of different strides, the data from each stride were interpolated to a uniform length of 128 samples. The segments, each with a length of 128, were employed in the CNN model, and for each participant, the initial 80 segments were chosen. In the case of the GRU model, 8 consecutive segments (stride) were merged into a singular segment comprising 1024 samples, yielding a total of 10 segments for each participant. The adult and older adult participants were randomly and proportionally divided into training, testing, and validation sets at a ratio of 146:49:49. Their corresponding data were used as training, testing, and validation data set for the CNN model and the GRU model, respectively.

2.3. Classifiers

CNN and GRU were utilized in this study because of excellent capacity of CNN for local special and temporal feature extraction and the outperformance of GRU in learning long-term dependent features [10]. A previous study with the same dataset, has shown that CNN can classify young and older adults based on one stride of data, while GRU benefits from a longer dataset (eight strides) [11].

The interpolated one-stride segments for CNN and interpolated eight-stride segments for GRU were organized in x-, y-, and z-axis order, to generate a single signal data with 3 channels (128*3 and 1024*3 respectively). The optimal hyperparameters for both the CNN and GRU models were tuned by Bayesian Optimization (BO) [32]. Unlike conventional techniques such as randomized search cross-validation, BO considers the prior performance of the hyperparameters and updates them to achieve better performance. This allows BO to find the global optimum with a minimum number of steps. For each model, 15 parameter combinations were tested. Detailed information about the hyperparameter space settings can be referenced in Table 1, while the learning rate, which was also optimized using BO, was configured within the range of [1e-5, 1e-2].

2.4. Evaluation

The assessment of classification performance was conducted using widely recognized evaluation measures such as accuracy, recall (sensitivity), precision, and F1 score (the harmonic mean of sensitivity and precision). Receiver operating characteristic (ROC) curves were generated and the area under the curve (AUC) was calculated as well.

To ensure transparency and reproducibility of the findings, we have made the project repository publicly accessible at https://github.com/xzheng93/Explainable_DL. The repository contains the source code, dataset, log files of experiments.

2.5. SHAP

The SHAP approach [23] was used to explain the prediction of a signal segment x in the given model f (CNN or GRU) based on the SHapley values from coalitional game theory. The original input

Table 1
Hyperparameters space for CNN and GRU.

	Layer		CNN	GRU
Input			128X3	1024X3
Stack (Stack number: [1,3])	Deep layer	Layer name	Conv1D	GRU
		Unit/Filter	[2, 768]	
		Kernel size	[1,15]	–
		Activation	"ReLU"	"tanh"
	Batch Normalization		–	
	Pooling		–	
	Dropout	Rate	Nan or [0.1, 0.9]	
	Dense	Unit	[2, 768]	
	Flatten		–	
	Dropout	Rate	Nan or [0.1, 0.9]	
Output	Dense		2 units and "softmax" activation	

CNN: convolutional neural network; GRU: gate recurrent unit; ReLU: rectified linear unit; tanh: hyperbolic tangent function.

segment x was mapped through the function $h_x(z')$ to get the input for the SHAP explanation $g(\bullet)$. $z' \in \{0,1\}^N$, where N is the number of features (data points or sets of data point in the signal segment) of x and, 0 and 1 mean the absence or presence of features in x. Applying to the model $f : f(h_x(z'))$, the SHAP explanation [23] can be defined as:

$$g(z') = f(h_x(z')) = \varnothing_0 + \sum_{i=1}^{N} \varnothing_i z_i' \tag{1}$$

where $\varnothing_i$ is the SHapley value of a feature i in the segment x, $\varnothing_0$ is the expatiation of $f(\overline{x})$ and $\overline{x}$ is the averaged of the input segment x.

The definition of SHapley values $\varnothing_i$ is as follows:

$$\varnothing_i(g) = \frac{1}{|N|!} \sum_{\{i\} \in s \text{ and } s \subseteq N} (|s|-1)!(|N|-|s|)![g(s) - g(s-\{i\})] \tag{2}$$

where s is the segment which data features i is present. $|*|$ represents the length of a segment *, excluding absent features. The definition of SHapley value make it satisfies the efficiency, symmetry, dummy, and additivity properties.

The efficiency property can be represented as:

$$\sum_{i \in N} \varnothing_i = g(x) \tag{3}$$

The sum of the SHapley values of all separated features equals the value of the coalition of all the features (the whole signal segment). Therefore, all the gain is distributed among the segments.

The symmetry property means that if the contributions of two features i and j are equal, they will contribute equally to all possible coalitions. This can be represented as, if $\varnothing_i = \varnothing_j$, then

$$g(s \cup \{i\}) = g(s \cup \{j\}) \tag{4}$$

where $s \subseteq N$ and $\{i.j\} \notin s$.

The dummy property means if a feature i does not change the predicted value:

$$g(s \cup \{i\}) = g(s) \tag{5}$$

Then, its SHapley value $\varnothing_i$ equals to 0.

Regarding the additivity property, if a coalition game is combined by two gain functions g' and g'', the SHapley values are additive:

$$\varnothing_i(g' + g'') = \varnothing_i(g') + \varnothing_i(g'') \tag{6}$$

The SHapley value is built based on a solid theory. The properties of SHapley value give the explanation a reasonable foundation and distinguish the SHAP from other methods such as LIME [33].

3. Results

The optimal hyperparameters and architecture for both CNN and GRU were determined through BO, and the results are presented in Figs. 2 and 3. In the case of CNN, the training dataset comprised 11680 (146*80) one-stride data (segments), the testing dataset included 3920 (49*80) segments, and the validation dataset contained 3920 (49*80) one-stride segments. For GRU, the training dataset consisted of 1460 eight-stride segments (146*10), the testing dataset encompassed 490 eight-stride segments (49*10), and the validation dataset comprised 490 eight-stride segments (49*10).

The CNN architecture included three 1D convolutional layers followed by batch normalization, max-pooling, and dropout layers. The first convolutional layer consisted of 88 filters with a kernel size of 13, while the second convolutional layer comprised 336 filters with a kernel size of 5. The third convolutional layer contained only 2 filters with a kernel size of 1. The dropout rates were 0.3, 0.6, and 0 for the first, second, and third dropout layers, respectively. A dense layer with 74 units, a fully connected layer, and a dropout layer with a rate of 0.5 were also incorporated into the architecture. Finally, a softmax activation function was employed for classification. Adam optimization was used, and the optimal learning rate of 0.0015 was discovered via BO. The model was trained for 150 epochs with early stopping based on the validation accuracy with a patience of 20 epochs.

Fig. 3 graphically illustrates the architecture and optimal hyperparameters of the proposed GRU model. The GRU architecture included three GRU layers, each followed by batch normalization, max-pooling, and dropout layers for regularization. The first GRU layer had 666 filters, the second had 438 filters, and the third had 2 filters. The three dropout layers had a rate of 0.5, 0.7, and 0, respectively. A dense layer with 676 units and a dropout layer with a rate of 0.1 were also incorporated into the architecture. The final layer of the GRU consisted of a fully connected layer followed by a softmax activation function for classification. A graphical representation of the GRU architecture is shown in Fig. 3. The Adam optimization was employed with a learning rate of 0.0003. The same training epoch setting and early stopping as for CNN were used.

Based on the testing data, the classification performance of the CNN is summarized in Table 2, achieving an accuracy of 81.4 %, precision of 82.7 %, recall of 76.3 %, F1-score of 79.3 %, and an AUC of 0.89. Detailed classification results for the CNN model, including the confusion matrix and ROC curve, are presented in Fig. 4. In the adult group, 85.9 % of data samples were correctly classified, while 14.1 % were incorrectly classified as older adults. In the older adult group, 76.3 % of the samples were correctly classified, while 23.7 % were incorrectly classified as adults.

Table 3 displays the classification performance of the GRU, with

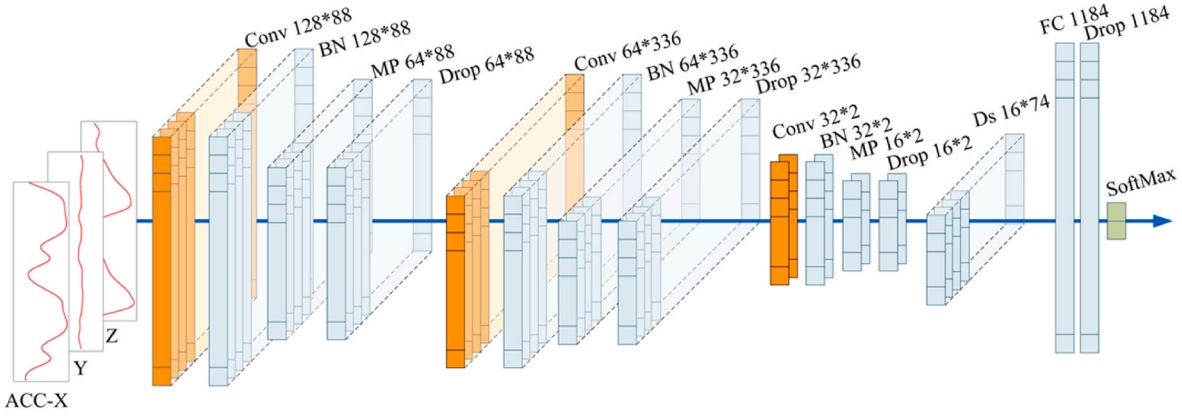

Fig. 2. The architecture and optimal hyperparameter of CNN. ACC: acceleration; CNN: convolutional neural network; Conv: 1-dimension convolutional layer (in orange); BN: batch normalization layer; MP: max-pooling layer; Drop: dropout layer; Ds: dense layer; SoftMax: softmax activation (in green).

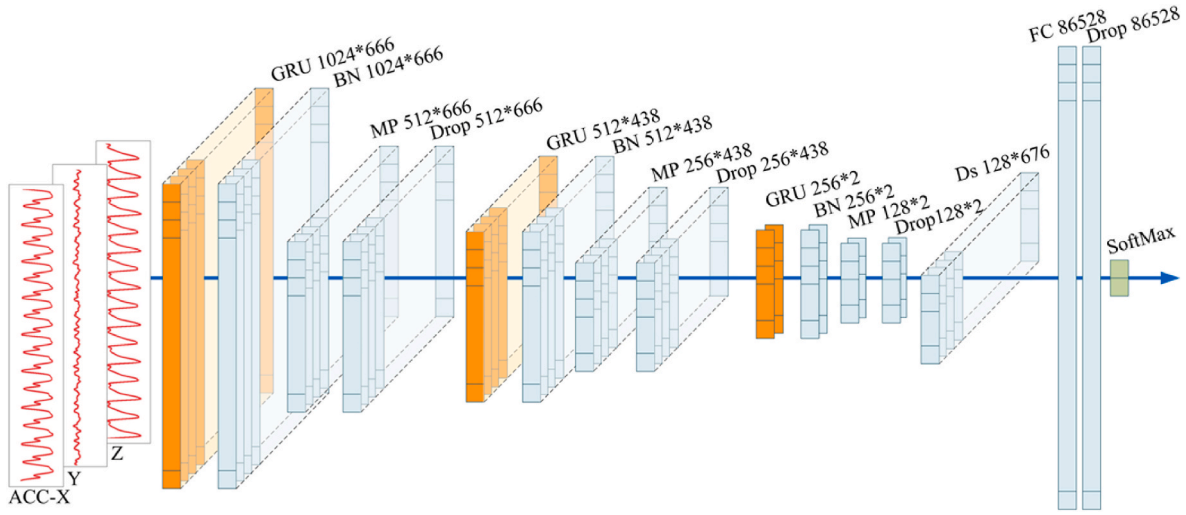

Fig. 3. The architecture and optimal hyperparameter of GRU. ACC: acceleration; GRU: gate recurrent unit layer (in orange); BN: batch normalization layer; MP: max-pooling layer; Drop: dropout layer; Ds: dense layer; SoftMax: softmax activation (in green).

Table 2
The performance metrics of CNN.

	Accuracy	Precision	Recall	F1-score	AUC
CNN	81.4 %	82.7 %	76.3 %	79.3 %	0.89

CNN: convolutional neural network; AUC: area under the curve.

further details provided in Fig. 5. The GRU model achieved an accuracy of 84.5 %, precision of 79.4 %, recall of 90.4 %, F1-score of 84.6 %, and an AUC of 0.94. The confusion matrix in Fig. 5(a) illustrates that 79.2 % of adults and 90.4 % of older adults were correctly classified.

After the evaluation, the mean absolute SHAP values for the testing data of both CNN and GRU models were computed and visualized in Fig. 6. In this representation, a deeper red colour signifies a greater contribution to the classification process.

In Fig. 6(a), an abundance of red colour is observed in the V and AP directions, particularly around heel contact. A similar pattern is evident in Fig. 6(b), with a prevalence of red colour in the V and AP directions, centered around heel contact event. Notably, Fig. 6(b) reveals that not all gait cycles are equally significant, as some exhibit a higher degree of red colour, indicating greater importance in the classification process.

4. Discussion

The primary aim of this study is to increase the transparency of non-linear DL models in gait analysis. To achieve this, state-of-the-art DL models, specifically CNN and GRU, were explained using a cutting-edge

Table 3
The performance metrics of GRU.

	Accuracy	Precision	Recall	F1-score	AUC
GRU	84.5 %	79.4 %	90.4 %	84.6 %	0.94

GRU: gate recurrent unit layer; AUC: area under the curve.

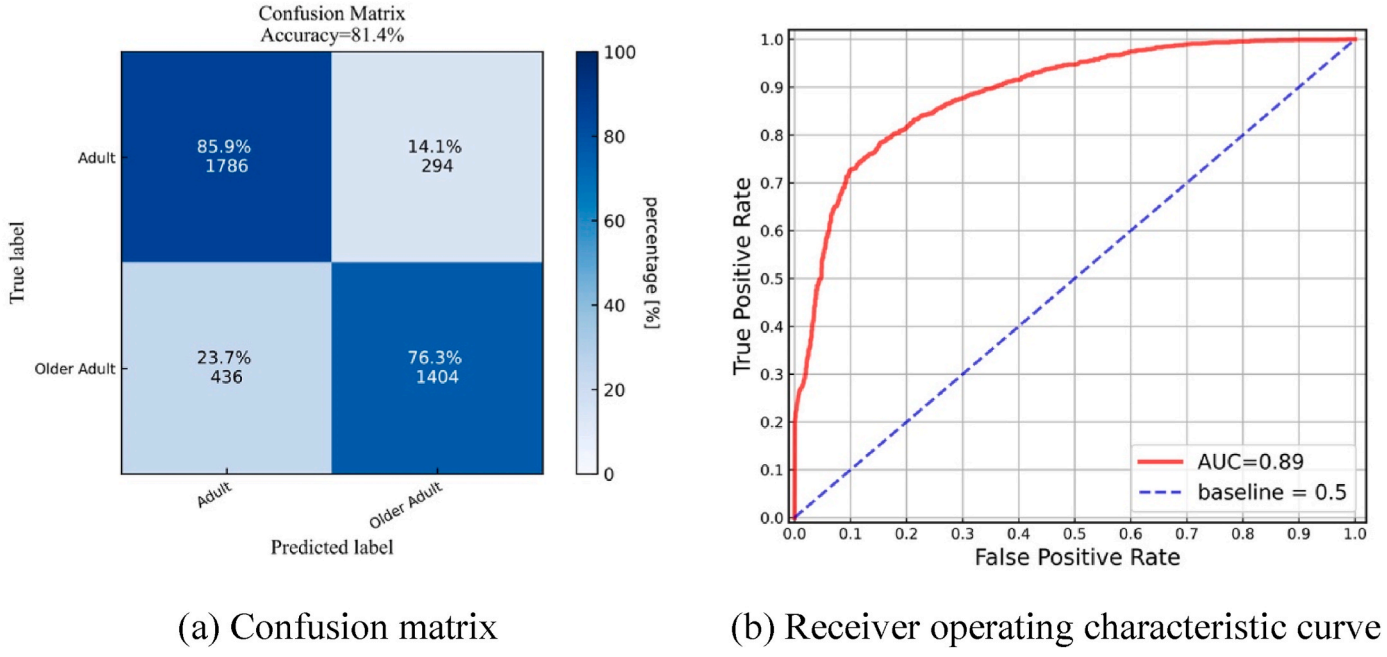

Fig. 4. (a) Confusion matrix and (b) Receiver operating characteristic curve for CNN. CNN: convolutional neural network; AUC: area under the curve.

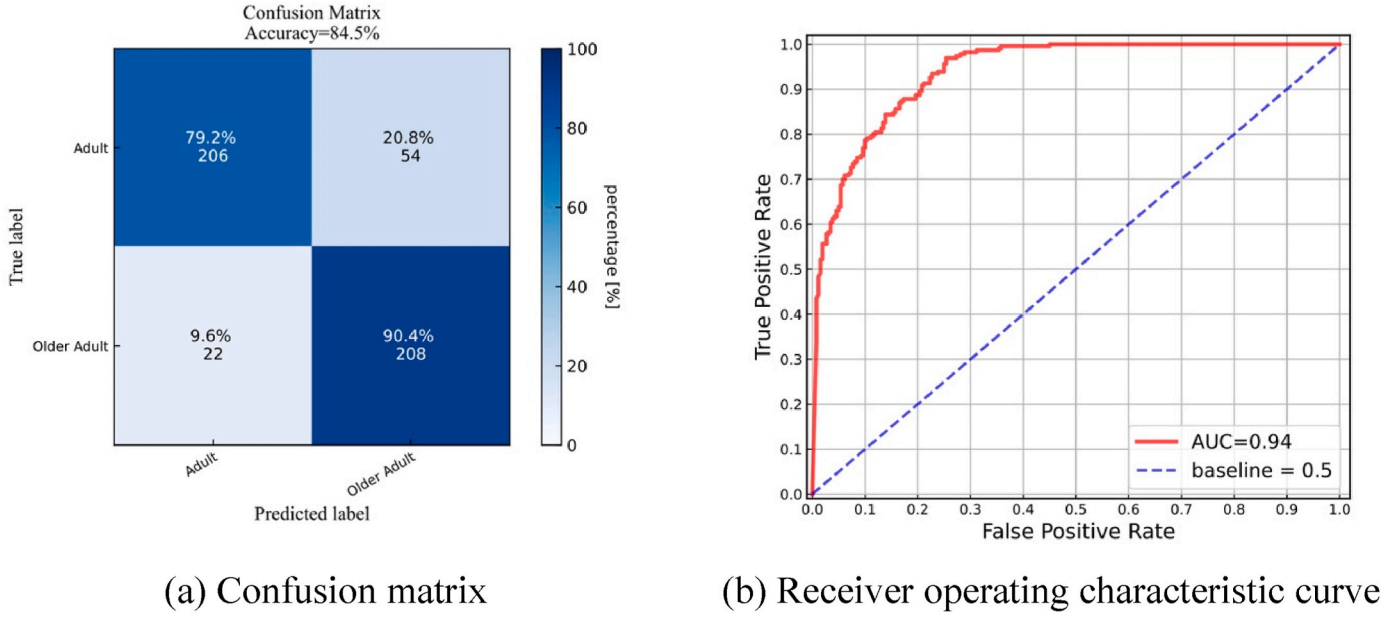

Fig. 5. (a) Confusion matrix and (b) Receiver operating characteristic curve for GRU. GRU: gate recurrent unit layer; AUC: area under the curve.

XAI approach, known as SHAP. These models are applied to classify individuals into two distinct groups: adults and older adults, based on acceleration time series data collected during a 3 min walking. The results indicate that the CNN model achieved satisfying classification performance, with an accuracy of 81.4 % and an AUC of 0.89, despite being trained on data from one stride. The GRU model demonstrated promising classification capabilities, achieving an accuracy of 84.5 % and an AUC of 0.94, utilizing eight-stride data. To understand and interpret the proposed DL models in the context of gait classification, the SHAP approach was employed. The SHAP values shed light on the models' decision-making processes, revealing a predominant reliance on acceleration data from the AP and V directions, rather than the ML direction, for the classification task. Specifically, data surrounding gait events such as heel contact in the AP and V directions emerged as the most influential inputs contributing to the differentiation between adults and older adults.

In this study, CNN and GRU were employed. CNN is renowned for its exceptional capacity to extract local spatial-temporal features. It has the potential to capture time-independent gait features, such as root mean square (indicative of gait intensity), rhythm (reflecting the proportion of stance and swing phases), and harmonic index (measuring the smoothness of the acceleration curve). These features have previously been successfully utilized in characterizing age-related gait differences in other studies [9,34]. The promising accuracy achieved by CNN suggests that even single stride data contains rich information that can effectively distinguish age-related gait patterns. The SHAP values underscore that data corresponding to the heel contact event play an important role in this discrimination. On the other hand, GRU is designed to capture both short- and long-term dependent features. It may learn time-dependent gait features that can reflect the intricate relationships and subtle differences between gait cycles. Time-dependent gait features, including regularity, variability, local stability (as measured by the largest Lyapunov exponent), gait symmetry (using the symmetry index), and complexity (evaluated through sample entropy), are crucial in age-related gait classification. These features offer insights into changes in postural control that arise due to aging [35,36]. The SHAP values presented in Fig. 6 (b) highlight that not all gait cycles were treated equally by GRU, as some gait cycles exhibit higher SHAP values. This observation suggests that GRU takes the relationships and slight variations between gait cycles into account when classifying individuals into the adult and older adult groups. Additionally, these findings inform model selection for specific gait problems. For example, CNN could effectively recognize gait characteristics that typically occur within one or two strides, such as asymmetric step length [37]. GRU might be more beneficial when focusing on extended time-series data to detect time-dependent changes occurring over multiple strides, like fatigue-related gait patterns [38].

Aging is an ongoing process often accompanied by a gradual decline in balance [39]. Changes within the neuromusculoskeletal system, such as the loss of muscle fibers and reduced muscle force production, can result in diminished muscle strength and flexibility [40]. These alterations may impact functionality and contribute to reduced mobility. Furthermore, the sensory systems are critical for effective postural control, including the visual, vestibular, and proprioceptive systems, tend to deteriorate with age, further affecting one's balance [41]. However, the SHAP results from this study indicate that DL models predominantly rely on data from the AP and V directions, instead of the ML direction, which is more closely associated with balance capacity. This observation aligns with a recent study that supports our findings, emphasizing the significance of dynamic gait parameters in the AP and V directions for classifying age-related gait patterns [9]. These parameters include Root Mean Square in AP and V directions, Lyapunov Exponent in the V direction, step regularity in the V direction, Cross Entropy in both V and ML directions, and gait speed when utilizing artificial neural networks for classification [9]. Given that the study [9] also utilized only one accelerometer, it is possible that the similar results may be attributed to the limited sensitivity of a single accelerometer in detecting balance-related postural control information. An additional explanation for the limited contribution of ML direction data to the classification process could be attributed to the simplicity of the task undertaken in this study. Participants engaged in a 3-min walking task within a clean and well-lit hallway, walking at their preferred pace without any perturbations. Consequently, this task may not effectively capture the variations in balance capacity between adults and older adults which may explain why ML-direction data did not play as a significant role in the classification process as AP and V direction data.

The colour spectrum of SHAP values in Fig. 6 not only reveals which axes contribute more to the classification process but also identifies specific gait events that play a crucial role. It shows that data spanning from the terminal swing to the loading response phase consistently yield higher SHAP values, particularly around the event of heel contact. The terminal swing is a phase before the heel contacts the ground. It involves

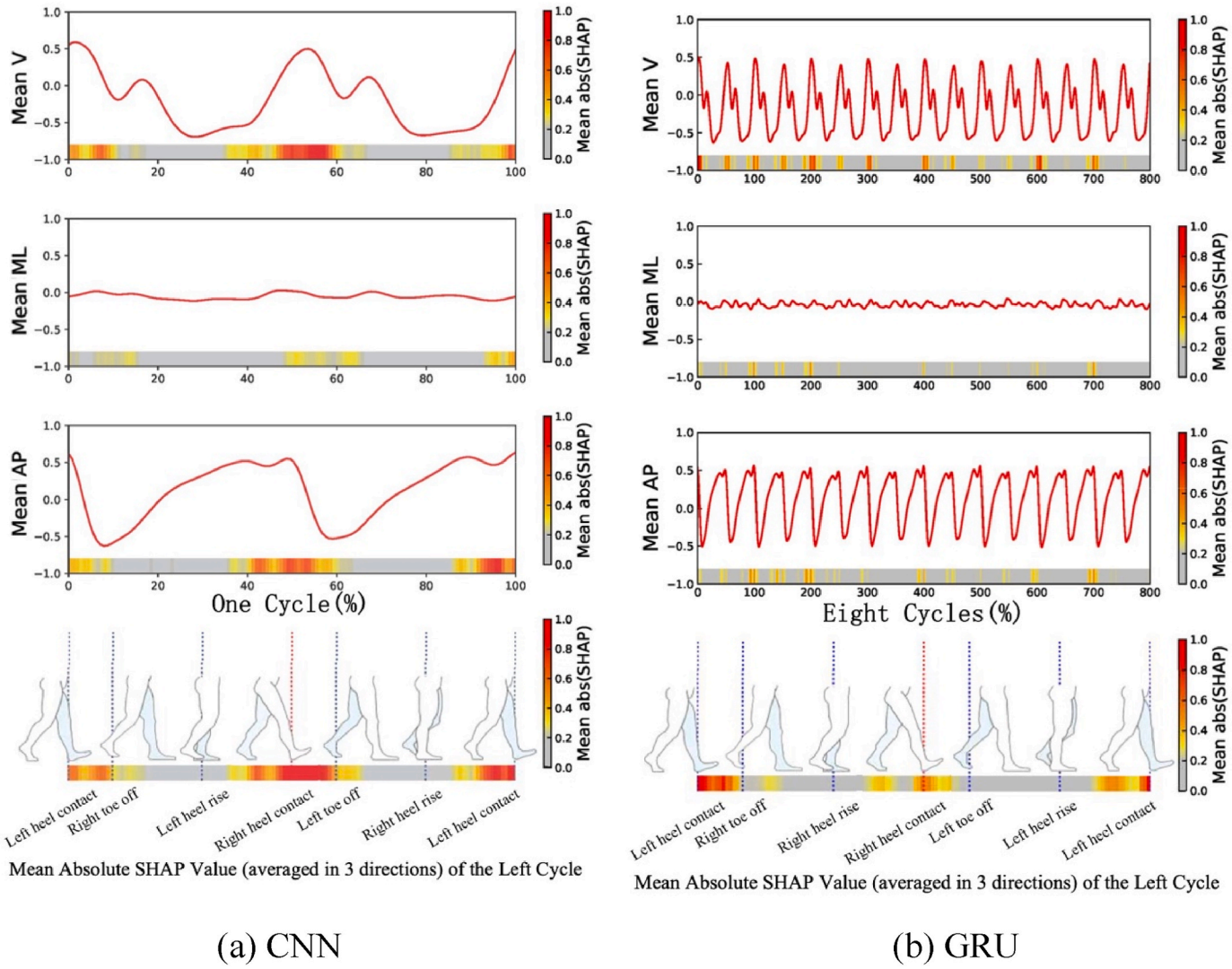

Fig. 6. SHAP values results of (a) CNN and (b) GRU. CNN: convolutional neural network; GRU: gate recurrent unit; V: vertical direction; AP: anteroposterior direction; ML: mediolateral direction; SHAP: SHapley Additive exPlanations. The acceleration data in each panel were normalized for both amplitude (ranging from −1 to 1) and time (using left heel contact as the reference point).

the final preparations of the leg and foot for ground contact. For example, muscles around the ankle and knee are activated to ensure that the limb is prepared to provide the necessary stability during the heel contact and loading response [42]. During this phase, the leg starts to slow down its forward swing to prevent excessive force upon heel strike. The acceleration, during this phase, in the walking direction (AP) has no obvious increase. The loading response phase starts with the initial contact of the heel with the ground and ends with toe-off of the opposite limb. After the heel contact, the body undergoes shock absorption and weight acceptance as the body's weight is transferred onto the stance limb [43]. During this phase, the acceleration data in the AP direction reach the maximum around the heel contact, and sharply decrease to reach the minimum around the toe-off event. The acceleration data around the terminal swing phase effectively represent how an individual prepares for deceleration, the moment of heel contact illustrates the process of deceleration, and the toe-off event signifies the initiation of acceleration [43]. Apart from this, from heel contact to the opposite toe-off is the double support phase of gait. During this phase, subjects transfer their weight from one leg to the other. In older adults, a longer double support time has been reported [44], potentially indicating a functional adaptation to a more unstable gait pattern [45]. This difference in gait dynamics may contribute to the discrimination of gait patterns in DL models.

The acceleration disparities observed by SHAP in these gait events/phases may indicate that adults and older adults exhibited different deceleration and acceleration patterns during walking. It can be attributed to the changes in kinematic and kinetic factors associated with aging which were observed by previous studies. Research has indicated that older adults tend to exhibit a reduced knee extension angle [46] and moment [47] at the point of heel contact, which can be closely linked to weaker muscles, such as the quadriceps [48,49]. These changes may result in a reduced absorption force in the knee joint [50]. These alterations can lead to compensatory gait adjustments aimed at alleviating joint discomfort or stiffness, particularly in the hip and knee joints. Compared to adults, older adults exhibit limited capabilities in limb advancement during the push-off period. Research shows that, during the loading response phase, older adults often demonstrate reduced hip extension and moment [47,51–53], particularly during the toe-off phase. This reduction may be indicative of decreased power in the hip extensors [52,54,55], which could imply weakness in swinging and kicking the lower limbs to generate forward propulsive force while walking. Furthermore, older adults tend to exhibit decreased independent movement of the subtendons [56] and reduced plantar flexor moments [53], contributing to lower propulsive power at the ankle [57].

The current study has insufficient data to examine whether these kinematic and kinetic factors are responsible for the distinct acceleration and deceleration patterns. Further studies are necessary in this regard.

SHAP has often been used to enhance transparency in CML-based gait analysis [4,58], while its application in DL-based gait analysis is still in its early stages, with LRP being primarily used [18,19]. One study attempted to compare the explainability of SHAP (for CML models) and LRP (for DL models) in gait analysis [59]. It reported that although the interpretations for LRP-based DL models were similar to those of SHAP-based CML models, LRP's reliability was lower due to variability in results across repeated experiments. However, this conclusion should be approached cautiously, given the low accuracy of the DL results and the differences in input/models. Our study opted to use SHAP instead of LRP due to its strong mathematical foundation in game theory, as opposed to LRP, which computes relevance between input and output based on back propagation. While this study suggested that SHAP might offer superior explanations, providing definitive proof is challenging due to the lack of clear metrics for assessing the trustworthiness of explanations [58]. Therefore, a fair comparison between different XAI approaches should be a focus in future research.

This study marks an initial step in applying XAI to gait analysis. It enhances the transparency of DL models by offering insights into the data used for predictions. In the future, this technology may help users, such as patients and clinicians, trust ML models, thereby facilitating the implementation of AI in clinical settings. However, XAI, such as the SHAP approach, provides explainability results based on input and model output data. Alterations in input signals can yield divergent outputs, and these changes may be influenced not only by aging but also by independent parameters, such as sensor brands. This study utilized three different types of accelerometers which have different dynamic ranges and accuracy. To minimize potential biases introduced by these independent parameters in prediction explanations, signal amplitude normalization was applied. Nevertheless, it is important to note that while these techniques help mitigate bias, they may also inadvertently eliminate valuable information, such as gait intensity, since the maximum and minimum sensor readings for each data segment are constrained to 1 and -1. To standardize the input for the DL models, gait cycle normalization was applied and walking speed information was excluded. However, a prior study has shown that gait cycle normalization has only a marginal impact on the classification performance of age-related gait patterns [11]. Notably, XAI provides explanations based on correlations and associations within the data rather than revealing causal relationships. Consequently, explanations offered by XAI may not always align with human intuition or domain expertise. These disparities can offer novel insights, or lead to misunderstandings or mistrust, particularly in critical domains like healthcare. Unfortunately, a ground truth for evaluating the quality of XAI explanations remains absent. Hence, the explainability results should be interpreted cautiously. Future studies should employ more extensive sensor systems, such as motion capture, to thoroughly examine the validity of SHAP explanations.

5. Conclusion

The present study enhances the transparency and interpretability of the proposed DL in gait analysis by incorporating the SHAP approach. The results demonstrate that CNN can accurately distinguish between adults and older adults based on data from a single stride. The key factors contributing to this classification were the accelerations around the heel contact in the AP and V directions. GRU also exhibited promising classification performance, leveraging data from eight consecutive strides. The SHAP results from GRU suggest that it may capture the relationships and subtle variations between gait cycles, particularly the accelerations around heel contact in the AP and V directions. These findings imply that adults and older adults exhibit distinct acceleration and deceleration patterns during 3 min of walking.

This study marks an initial step in applying XAI to gait analysis. It enhances the transparency of DL models by offering insights into the data used for predictions. In the future, this technology may increase user confidence—including patients and clinicians—in ML models, thus paving the way for AI implementation in clinical settings.

Summary table

What is already known on the topic.

- Machine learning has been successfully applied in classifying gait patterns.
- Deep learning improves classification performance, though it increases model complexity.
- The "black-box" nature of machine learning, particularly deep learning, limits its clinical application.

What has this study added to the body of knowledge?

- Two distinct deep learning models were applied to successfully classify age-related gait patterns.
- Explainable artificial intelligence was utilized to increase the transparency of the classification process.
- The approach developed in this study shows potential for broad application in various gait classifications and extends to other domains.

CRediT authorship contribution statement

Xiaoping Zheng: Writing – review & editing, Writing – original draft, Methodology, Data curation, Conceptualization. **Egbert Otten:** Writing – review & editing, Supervision, Conceptualization. **Michiel F. Reneman:** Writing – review & editing, Supervision, Conceptualization. **Claudine JC. Lamoth:** Writing – review & editing, Supervision, Data curation, Conceptualization.

Data availability

The data and source code of this study are accessible at https://github.com/xzheng93/Explainable_DL.

Funding

XZ was supported by the China Scholarship Council-University of Groningen Scholarship [Grant No.201906410084].

Declaration of competing interest

XZ was supported by the China Scholarship Council-University of Groningen Scholarship [Grant No.201906410084]. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Hanley, C. Silke, J. Murphy, Community-based health efforts for the prevention of falls in the elderly, Clin. Interv. Aging (2011) 19–25.
[2] T.B. Aderinola, T. Connie, T.S. Ong, W.-C. Yau, A.B.J. Teoh, Learning age from gait: a survey, IEEE Access 9 (2021) 100352–100368.
[3] M. Burnfield, Gait analysis: normal and pathological function, J. Sports Sci. Med. 9 (2010) 353.
[4] X. Zheng, M.F. Reneman, J.A. Echeita, R.H.S. Preuper, H. Kruitbosch, E. Otten, C. J. Lamoth, Association between central sensitization and gait in chronic low back pain: insights from a machine learning approach, Comput. Biol. Med. 144 (2022) 105329.
[5] M. Intriago, G. Maldonado, R. Guerrero, O. Messina, C. Rios, Bone mass loss and Sarcopenia in Ecuadorian patients, Journal of Aging Research (2020) 2020.

[6] D. Kobsar, J.M. Charlton, C.T. Tse, J.-F. Esculier, A. Graffos, N.M. Krowchuk, D. Thatcher, M.A. Hunt, Validity and reliability of wearable inertial sensors in healthy adult walking: a systematic review and meta-analysis, J. NeuroEng. Rehabil. 17 (2020) 1–21.
[7] L. Xiang, Y. Gu, Z. Gao, P. Yu, V. Shim, A. Wang, J. Fernandez, Integrating an LSTM framework for predicting ankle joint biomechanics during gait using inertial sensors, Comput. Biol. Med. (2024) 108016.
[8] T. Chau, A review of analytical techniques for gait data. Part 1: fuzzy, statistical and fractal methods, Gait Posture 13 (2001) 49–66.
[9] Y. Zhou, R. Romijnders, C. Hansen, J. van Campen, W. Maetzler, T. Hortobagyi, C. J.C. Lamoth, The detection of age groups by dynamic gait outcomes using machine learning approaches, Sci. Rep. 10 (2020).
[10] F. Gu, M.-H. Chung, M. Chignell, S. Valaee, B. Zhou, X. Liu, A Survey on Deep Learning for Human Activity Recognition, vol. 54, ACM Computing Surveys (CSUR), 2021, pp. 1–34.
[11] X. Zheng, E. Wilhelm, M.F. Reneman, E. Otten, C.J. Lamoth, Age-related Gait Patterns Classification Using Deep Learning Based on Time-series Data from One Accelerometer, TechRxiv (2023).
[12] B.M. Meyer, L.J. Tulipani, R.D. Gurchiek, D.A. Allen, L. Adamowicz, D. Larie, A. J. Solomon, N. Cheney, R.S. McGinnis, Wearables and deep learning classify fall risk from gait in multiple sclerosis, IEEE journal of biomedical and health informatics 25 (2020) 1824–1831.
[13] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160.
[14] 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, 2016.
[15] S. Ali, F. Akhlaq, A.S. Imran, Z. Kastrati, S.M. Daudpota, M. Moosa, The enlightening role of explainable artificial intelligence in medical & healthcare domains: a systematic literature review, Comput. Biol. Med. (2023) 107555.
[16] T. Speith, A review of taxonomies of explainable artificial intelligence (XAI) methods, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 2239–2250.
[17] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115.
[18] F. Horst, S. Lapuschkin, W. Samek, K.-R. Müller, W.I. Schöllhorn, Explaining the unique nature of individual gait patterns with deep learning, Sci. Rep. 9 (2019) 1–13.
[19] A.S. Alharthi, A.J. Casson, K.B. Ozanyan, Spatiotemporal analysis by deep learning of gait signatures from floor sensors, IEEE Sensor. J. 21 (2021) 16904–16914.
[20] C. Dindorf, W. Teufl, B. Taetz, G. Bleser, M. Fröhlich, Interpretability of input representations for gait classification in patients after total hip arthroplasty, Sensors 20 (2020) 4385.
[21] D. Alvarez-Melis, T.S. Jaakkola, On the robustness of interpretability methods, in: arXiv Preprint arXiv:1806.08049, 2018.
[22] E. Winter, The Shapley Value, Handbook of Game Theory with Economic Applications, vol. 3, 2002, pp. 2025–2054.
[23] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017).
[24] D. Borah, S. Wadhwa, U. Singh, S.L. Yadav, M. Bhattacharjee, V. Sindhu, Age related changes in postural stability, Indian J. Physiol. Pharmacol. 51 (2007) 395–404.
[25] L. Kikkert, N. Vuillerme, J.P. van Campen, B.A. Appels, T. Hortobagyi, C.J. Lamoth, Gait characteristics and their discriminative power in geriatric patients with and without cognitive impairment, J. NeuroEng. Rehabil. 14 (2017).
[26] C.J. Lamoth, F.J. van Deudekom, J.P. van Campen, B.A. Appels, O.J. de Vries, M. Pijnappels, Gait stability and variability measures show effects of impaired cognition and dual tasking in frail people, J. NeuroEng. Rehabil. 8 (2011) 1–9.
[27] N.M. Kosse, S. Caljouw, D. Vervoort, N. Vuillerme, C.J. Lamoth, Validity and reliability of gait and postural control analysis using the tri-axial accelerometer of the iPod touch, Ann. Biomed. Eng. 43 (2015) 1935–1946.
[28] M.H. de Groot, H.C. van der Jagt-Willems, J.P. van Campen, W.F. Lems, J. H. Beijnen, C.J. Lamoth, A flexed posture in elderly patients is associated with impairments in postural control during walking, Gait Posture 39 (2014) 767–772.
[29] T. Ijmker, C.J. Lamoth, Gait and cognition: the relationship between gait stability and variability with executive function in persons with and without dementia, Gait Posture 35 (2012) 126–130.
[30] C. Buckley, M.E. Micó-Amigo, M. Dunne-Willows, A. Godfrey, A. Hickey, S. Lord, L. Rochester, S. Del Din, S.A. Moore, Gait asymmetry post-stroke: determining valid and reliable methods using a single accelerometer located on the trunk, Sensors 20 (2019) 37.
[31] W. Zijlstra, A.L. Hof, Assessment of spatio-temporal gait parameters from trunk accelerations during human walking, Gait Posture 18 (2003) 1–10.
[32] J. Wu, X.-Y. Chen, H. Zhang, L.-D. Xiong, H. Lei, S.-H. Deng, Hyperparameter optimization for machine learning models based on Bayesian optimization, Journal of Electronic Science and Technology 17 (2019) 26–40.
[33] K. Aas, M. Jullum, A. Løland, Explaining individual predictions when features are dependent: more accurate approximations to Shapley values, Artif. Intell. 298 (2021) 103502.
[34] N.M. Kosse, N. Vuillerme, T. Hortobagyi, C.J.C. Lamoth, Multiple gait parameters derived from iPod accelerometry predict age-related gait changes, Gait Posture 46 (2016) 112–117.
[35] K.K. Patterson, N.K. Nadkarni, S.E. Black, W.E. McIlroy, Gait symmetry and velocity differ in their relationship to age, Gait Posture 35 (2012) 590–594.
[36] D. Kobsar, C. Olson, R. Paranjape, T. Hadjistavropoulos, J.M. Barden, Evaluation of age-related differences in the stride-to-stride fluctuations, regularity and symmetry of gait using a waist-mounted tri-axial accelerometer, Gait Posture 39 (2014) 553–557.
[37] S. Viteckova, P. Kutilek, Z. Svoboda, R. Krupicka, J. Kauler, Z. Szabo, Gait symmetry measures: a review of current and prospective methods, Biomed. Signal Process Control 42 (2018) 89–100.
[38] F.A. Barbieri, P.C.R. Dos Santos, E. Lirani-Silva, R. Vitório, L.T.B. Gobbi, J.H. Van Diëen, Systematic review of the effects of fatigue on spatiotemporal gait parameters, J. Back Musculoskelet. Rehabil. 26 (2013) 125–131.
[39] R.W. Bohannon, P.A. Larkin, A.C. Cook, J. Gear, J. Singer, Decrease in timed balance test scores with aging, Phys. Ther. 64 (1984) 1067–1070.
[40] C.A. Laughton, M. Slavin, K. Katdare, L. Nolan, J.F. Bean, D.C. Kerrigan, E. Phillips, L.A. Lipsitz, J.J. Collins, Aging, muscle activity, and balance control: physiologic changes associated with balance impairment, Gait Posture 18 (2003) 101–108.
[41] T. Coelho, Á. Fernandes, R. Santos, C. Paúl, L. Fernandes, Quality of standing balance in community-dwelling elderly: age-related differences in single and dual task conditions, Arch. Gerontol. Geriatr. 67 (2016) 34–39.
[42] A. Schmitz, A. Silder, B. Heiderscheit, J. Mahoney, D.G. Thelen, Differences in lower-extremity muscular activation during walking between healthy older and young adults, J. Electromyogr. Kinesiol. 19 (2009) 1085–1091.
[43] A. Kharb, V. Saini, Y. Jain, S. Dhiman, A review of gait cycle and its parameters, IJCEM International Journal of Computational Engineering & Management 13 (2011) 78–83.
[44] B.E. Maki, Gait changes in older adults: predictors of falls or indicators of fear? J. Am. Geriatr. Soc. 45 (1997) 313–320.
[45] E.A. Ihlen, O. Sletvold, T. Goihl, P.B. Wik, B. Vereijken, J. Helbostad, Older adults have unstable gait kinematics during weight transfer, J. Biomech. 45 (2012) 1559–1565.
[46] E. Chung, S.-H. Lee, H.-J. Lee, Y.-H. Kim, Comparative study of young-old and old-old people using functional evaluation, gait characteristics, and cardiopulmonary metabolic energy consumption, BMC Geriatr. 23 (2023) 1–11.
[47] D.C. Kerrigan, M.K. Todd, U. Della Croce, L.A. Lipsitz, J.J. Collins, Biomechanical gait alterations independent of speed in the healthy elderly: evidence for specific limiting impairments, Arch. Phys. Med. Rehabil. 79 (1998) 317–322.
[48] B. Holm, M.T. Kristensen, J. Bencke, H. Husted, H. Kehlet, T. Bandholm, Loss of knee-extension strength is related to knee swelling after total knee arthroplasty, Arch. Phys. Med. Rehabil. 91 (2010) 1770–1776.
[49] J.W. Kwon, S.M. Son, N.K. Lee, Changes of kinematic parameters of lower extremities with gait speed: a 3D motion analysis study, J. Phys. Ther. Sci. 27 (2015) 477–479.
[50] M. Bendall, E. Bassey, M. Pearson, Factors affecting walking speed of elderly people, Age Ageing 18 (1989) 327–332.
[51] C.F. Oliveira, E.R. Vieira, F.M. Machado Sousa, J.P. Vilas-Boas, Kinematic changes during prolonged fast-walking in old and young adults, Front. Med. 4 (2017) 207.
[52] W.S. Kim, E.Y. Kim, Comparing self-selected speed walking of the elderly with self-selected slow, moderate, and fast speed walking of young adults, Annals of rehabilitation medicine 38 (2014) 101–108.
[53] S.-u. Ko, J.M. Hausdorff, L. Ferrucci, Age-associated differences in the gait pattern changes of older adults during fast-speed and fatigue conditions: results from the Baltimore longitudinal study of ageing, Age Ageing 39 (2010) 688–694.
[54] D.A. Winter, Biomechanics and motor control of human gait: normal, elderly and pathological (1991).
[55] P. Morfis, M. Gkaraveli, Effects of aging on biomechanical gait parameters in the healthy elderly and the risk of falling, Journal of Research & Practice on the Musculoskeletal System (JRPMS) (2021) 5.
[56] J.R. Franz, D.G. Thelen, Imaging and simulation of Achilles tendon dynamics: implications for walking performance in the elderly, J. Biomech. 49 (2016) 1403–1410.
[57] K. Rasske, J.R. Franz, Aging effects on the Achilles tendon moment arm during walking, J. Biomech. 77 (2018) 34–39.
[58] J.-K. Kim, M.-N. Bae, K.B. Lee, S.G. Hong, Identification of patients with sarcopenia using gait parameters based on inertial sensors, Sensors 21 (2021) 1786.
[59] J.-K. Kim, M.-N. Bae, K. Lee, J.-C. Kim, S.G. Hong, Explainable artificial intelligence and wearable sensor-based gait analysis to identify patients with osteopenia and sarcopenia in daily life, Biosensors 12 (2022) 167.