Sansgiri, Sailee ORCID logoORCID: https://orcid.org/0000-0003-1010-6872, Matikainen-Tervola, Emmi, Rantakokko, Merja, Finni, Taija, Rantalainen, Timo and Cronin, Neil J ORCID logoORCID: https://orcid.org/0000-0002-5332-1188 (2026) From treadmill to outdoor overground walking: Enhancing ground contact timing detection for older adults using transfer learning. Experimental Gerontology, 215. art:113056. doi:10.1016/j.exger.2026.113056 (In Press)

Official URL: https://doi.org/10.1016/j.exger.2026.113056
DOI: http://dx.doi.org/10.1016/j.exger.2026.113056
EPrint URI: https://eprints.glos.ac.uk/id/eprint/15847

Disclaimer

The University of Gloucestershire has obtained warranties from all depositors as to their title in the material deposited and as to their right to deposit such material.

The University of Gloucestershire makes no representation or warranties of commercial utility, title, or fitness for a particular purpose or any other warranty, express or implied in respect of any material deposited.

The University of Gloucestershire makes no representation that the use of the materials will not infringe any patent, copyright, trademark or other property or proprietary rights.

The University of Gloucestershire accepts no liability for any infringement of intellectual property rights in any material deposited but will remove such material from public view pending investigation in the event of an allegation of any such infringement.

Experimental Gerontology 215 (2026) 113056

Contents lists available at ScienceDirect

Experimental Gerontology

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

From treadmill to outdoor overground walking: Enhancing ground contact timing detection for older adults using transfer learning☆

Sailee Sansgiri [a,*], Emmi Matikainen-Tervola [a,b], Merja Rantakokko [a,b,c], Taija Finni [a], Timo Rantalainen [a], Neil J. Cronin [a,d]

[a] *Faculty of Sport and Health Sciences, University of Jyväskylä, Jyväskylä, 40700, Finland*
[b] *Institute of Rehabilitation, JAMK University of Applied Sciences, Jyväskylä, 40100, Finland*
[c] *The Wellbeing services county of Central Finland, Jyväskylä, 40620, Finland*
[d] *School of Education and Science, University of Gloucestershire, Gloucester, GL50 2RH, United Kingdom*

ARTICLE INFO

Section Editor: Christiaan Leeuwenburgh

Keywords:
Ground contact timing
Deep learning
Transfer learning
Walking
Treadmill
Outdoor environment
Aging

ABSTRACT

Identification of ground contact timings (GCT) is critical for monitoring mobility in older adults. Laboratory methods are precise but limited to controlled environments, restricting their applicability in real-world settings. Treadmills allow extended measurements but fail to reflect the variability of overground walking. We evaluated the performance of deep learning models trained on treadmill data from young adults and their generalizability to treadmill and outdoor walking in older adults. We also explored transfer learning to enhance predictions by fine-tuning models with older adults' treadmill and outdoor walking data. Foot-mounted inertial measurement unit (IMU) walking data was collected from 20 young adults on treadmills and 26 older adults on treadmills and outdoor level, incline, and decline terrains. Ground truth GCTs were derived using pressure insoles (young adults) and manually-annotated motion capture (older adults). A fully connected neural network, a convolutional neural network (CNN), and a bidirectional long short-term memory network were trained on IMU data. Transfer learning was applied incrementally by fine-tuning the best-performing model with older adults' data. Model performance was evaluated on unseen outdoor data from 6 participants using F1-score and mean absolute error (MAE). The CNN achieved the highest F1-scores (0.9864 — treadmill, 0.9637 — outdoor level, 0.9538 — incline, and 0.9029 — decline walking) and the lowest MAE. Fine-tuning improved treadmill F1-scores up to n=10, while outdoor level scores plateaued at n=5. Decline walking showed poorer performance, highlighting the need for advanced modeling strategies. These findings underscore the potential of transfer learning for real-world mobility monitoring.

1. Introduction

As the population ages, the ability to monitor mobility in older adults has become increasingly important. Assessing prolonged gait in natural environments serves as a valuable indicator of health, especially in older populations. To analyze stride-to-stride variability between and within individuals, accurate identification of ground contact timings (GCT) is essential. GCT, defined as the time between initial contact (IC) and toe-off (TO) of the same foot is an important temporal parameter. High variability of step-to-step GCT can be an indicator of fall risk and cognitive decline in aging populations (Ruiz-Ruiz et al., 2021). Furthermore, large variation in GCT between the left and right limbs can be an indicator of gait asymmetry and imbalance (Plotnik et al., 2013). Thus, continuous and accurate measurement of GCT is critical in both clinical and everyday settings to monitor mobility.

Traditional methods of overground gait analysis, such as optoelectronic motion capture systems and pressure-sensitive walkways offer precise measurements but are limited by their high cost and reliance on controlled environments like laboratories. GCT is typically identified either through force plates, or through manual identification by experts, which can be time-consuming, labor-intensive and subjective (Bruening and Ridge, 2014). Laboratory environments also restrict the number of overground steps that can realistically be measured (Hansen et al., 2002). These setups also do not reflect the everyday walking conditions typically encountered by older adults,

☆ This article is part of a Special issue entitled: 'Aging & Technology' published in Experimental Gerontology.

* Corresponding author.

E-mail addresses: sailee.s.sansgiri@jyu.fi (S. Sansgiri), emmi.matikainen-tervola@jamk.fi (E. Matikainen-Tervola), merja.h.rantakokko@jyu.fi (M. Rantakokko), taija.m.juutinen@jyu.fi (T. Finni), timo.rantalainen@jyu.fi (T. Rantalainen), neil.j.cronin@jyu.fi (N.J. Cronin).

https://doi.org/10.1016/j.exger.2026.113056
Received 13 December 2024; Received in revised form 15 December 2025; Accepted 26 January 2026
Available online 5 February 2026

which can vary significantly depending on factors such as terrain and balance stability (Schmitt et al., 2021; Hillel et al., 2019).

Treadmills provide an attractive alternative for in-laboratory measurements to analyze prolonged walking. However, treadmills constrain walking speed, so treadmill gait does not fully represent the complexities of overground walking. For example, several studies have reported slower preferred walking speeds, higher double support times and shorter stride lengths on a treadmill compared to overground walking (Schmitt et al., 2021; Renggli et al., 2020; Aartolahti, 2024). When walking in natural environments, environmental variability such as different surfaces and slopes, irregular walking speeds, and stability challenges all affect gait patterns and increase gait variability (Renggli et al., 2020). These factors make automation of gait analysis in natural environments challenging.

One possible solution is to implement machine learning (ML) models, which facilitate automation by deriving non-linear relationships from data. Combining ML models with data from mobile sensors such as inertial measurement units (IMU) could further help reduce reliance on laboratory environments. However, a known caveat of ML models is that they can only make predictions on data similar to their training data (Mundt, 2023). Hence, while ML models trained on treadmill data perform well in controlled settings, they often struggle when applied to overground walking. This may be particularly evident with older adults, who exhibit more gait variability (Kang and Dingwell, 2008). Furthermore, ML models are reliant on the quantity of data, and this might be difficult to collect for smaller gait labs due to limitations in participant recruitment or resources. Beyond healthy older adults, IMU data combined with ML techniques have also been applied in clinical populations with reduced mobility. For example, ML models trained on IMU-derived gait features have been used to detect freezing of gait in Parkinson's disease with high accuracy (Yang et al., 2024), to predict rehabilitation outcomes and walking independence after stroke (O'Brien et al., 2024), and to classify gait impairments associated with neurological and musculoskeletal conditions (Trabassi et al., 2022). These studies highlight the potential of IMU-based ML approaches to provide clinically meaningful predictions, but also emphasize the challenge of generalizing models across different populations and environments.

Transfer learning, a ML technique, offers a promising solution to bridge this gap. This technique enables a model trained in one domain (e.g., treadmill walking in young adults) to adapt to a related domain (e.g., overground walking in older adults) using a small amount of data from the target environment (Torrey and Shavlik, 2010). This strategy leverages the generalizability of base-level features learned during the initial training phase. In recent previous research, transfer learning has been implemented in gait analysis for classification of gait pathologies using 2-D video camera data (Verlekar et al., 2018), classification of hip osteoarthritis using kinematic trajectories from 3-D motion capture data (Pantonial and Simic, 2024), detection of hemiplegic and diplegic gait using IMU data (Pandit et al., 2019), and prediction of ground reaction forces using 3-D motion capture data (Avdan et al., 2023). These studies demonstrated that transfer learning effectively reduces the amount of data required from the target domain while maintaining high predictive accuracy.

While gait variability increases with age due to physiological changes, there are some common features of gait that are present across all age groups, including periodicity, heel strikes and toe-offs. Therefore, models trained on treadmill data from young adults can effectively learn these foundational features and serve as an effective starting point for adaptation to older adults' walking patterns. An advantage of this approach is the relative ease of collecting large-scale treadmill data from young adults, who are often more readily available for participation in research studies. By initially training on treadmill data from younger adults and incrementally fine-tuning the model with data from older adults, this method reduces the need for extensive data collection from the latter group. Additionally, by combining older adults' treadmill data with overground walking data during fine-tuning, the model can learn both age-related gait variability and the unique challenges of natural walking environments.

In this study, we addressed the following research questions:

1. How well do deep learning models trained to detect GCT on treadmill walking data in young adults perform on outdoor walking data from older adults?

2. How does incremental fine-tuning with treadmill walking data from older adults improve the performance of these models on overground walking data from older adults?

3. Can transfer learning be used to improve predictive accuracy in detecting outdoor overground GCT in older adults?

By answering these questions, we aim to advance automated, real-world gait analysis for older adults, supporting earlier detection of mobility impairments and more personalized interventions. Although our evaluation focuses on a relatively mobile older cohort, the approach lays the groundwork for future studies in broader and clinically diverse populations.

2. Methods

2.1. Participants and data collection

This study utilized retrospective, cross-sectional data collected from three previous studies and focused on two cohorts — young adults (for model training) and older adults (for fine-tuning the models and model testing and evaluation) (Rantakokko et al., 2024; Matikainen-Tervola et al., 2024). Detailed descriptions of both cohorts, measurement procedures and ground truth annotations are available in Appendix A.

From the young cohort (n=20, Mean age: 27.6 ± 4.41 years, 15 females), data from foot-mounted IMUs were collected on a treadmill at speeds from 3.5 km/h to 6 km/h. Ground truth labels were obtained using in-shoe pressure insoles.

For the older cohort (n=26, mean age: 76 ± 5.2 years, 17 females), foot-mounted IMU data were collected from 26 participants on a level treadmill at self-reported comfortable walking speeds, and outdoors while walking on level ground, at an incline (+5°) and at a decline (−5°) at self-selected walking speeds. Ground truth were obtained by manually annotating opto-electronic data (for treadmill measurements) or video camera data (for outdoor measurements).

2.2. Data processing

Data pre-processing for both cohorts was performed in MATLAB (R2023b, MathWorks, USA) using custom scripts. These scripts generated Excel files for each measurement trial. Each file contained, in columns, the raw, unfiltered 3-D accelerometer and gyroscope data from each foot, corresponding timestamps, and annotated ground contact respectively. Gait events were encoded as: 0 — no gait event, 1 — left ground contact, 2 — right ground contact. The decision to retain unfiltered signals was consistent with approaches reported by a previous study (Jlassi and Dixon, 2024). IMU signals were normalized using min–max scaling on the recommendations of our previous study (Sansgiri et al., 2024). In addition, no detrending was applied in order to preserve the original signal characteristics and to expose the deep learning models to realistic variability present in raw IMU data.

IMU data from the younger cohort were down-sampled from 120 Hz to 100 Hz to match the sampling frequency of the pressure insoles. In older adults, IMU data collected at 200 Hz was first down-sampled to 100 Hz before converting into Excel sheets. Downsampling was performed using MATLAB's downsample function. Temporal alignment between IMU signals and ground-contact event timestamps was verified, and no additional downsampling compensation was required. The Excel sheets were then subsequently used for ML analysis. Samples of IMU data and the corresponding ground contact from each condition are available at github.com/saileesansgiri/transferlearning_gct.

Subsequently, participants from the older cohort were randomly assigned to either the fine-tuning cohort or the test/validation cohort using a 77%–23% split (20 participants for fine-tuning and 6 participants for testing). This ensured that each participant's data was exclusively present in one set to avoid biasing the models.

2.3. Implementation and evaluation of ML models

To address the first research question, a t-distributed Stochastic Neighbor Embedding (t-SNE) analysis was performed to evaluate the similarity between the walking conditions and the potential for generalization across age groups (Van der Maaten and Hinton, 2008). The t-SNE analysis included data from four scenarios: young adults walking on a treadmill, older adults walking on a treadmill, older adults walking overground on flat terrain, and older adults walking overground on inclined or declined terrain. From the entire dataset, samples of 300 rows (3 s) were created and flattened. A total of 500 random samples from each walking condition (e.g., overground level walking) were selected for t-SNE analysis. This technique reduces the higher-dimensional IMU signals (12 × T, where T is the length of the time-series input) into a dimensionless 2-D plot. The resulting plot reveals clusters or patterns, offering insight into whether treadmill data—especially from younger participants—resembles overground data from older participants. Similar patterns would suggest that models trained on treadmill data from younger individuals may generalize well to older adults' overground walking, while distinct separations between datasets would highlight the need for further fine-tuning or adaptations.

To enhance the clarity of the visualization, frequency-domain analysis using the Fast Fourier Transform (FFT) was applied to the original IMU signals. Features such as the mean and maximum magnitude of the FFT and the dominant frequency component were extracted for each sample. These FFT-based features are effective at capturing the consistent periodicity of walking. Please note that the FFT features were only used for visualization and were not provided as input to the models, wherein only raw IMU data was used as input.

Three deep learning models were then trained on data from younger adults to detect GCT: a Fully Connected Feed-Forward Neural Network (FCNN), a Convolutional Neural Network (CNN) and a Bi-directional Long Short-Term Memory (BiLSTM) network. The models were trained on treadmill data collected at 0° incline from 20 participants, with a total of 23900 steps. Each training sample consisted of 200 rows of 3-D accelerometer and gyroscope signals, corresponding to 2 s of data, extracted using a sliding window approach. The windows were cut randomly to prevent overfitting. A multi-head approach was implemented in the final layer of each model, where the network bifurcated into three distinct nodes, each corresponding to an event: 0 — no contact, 1 — left contact, and 2 — right contact.

The best set of hyperparameters for each model were determined on the basis of the highest average five-fold F1-score (see B.1). To assess model robustness, we additionally performed nested cross-validation (5 outer folds, 4 inner folds). The inner loop tuned a focused set of hyperparameters around the previously selected settings, and the outer loop reports per-fold metrics and mean ± SD for F1, micro/macro precision, and micro/macro recall (see B.2). The models were then tested on level treadmill, outdoor level, incline and decline walking from the older cohort to evaluate their generalizability.

Out of these three models, we selected one model for subsequent transfer learning. This was determined by the highest F1-score and lowest mean absolute error (MAE) on the test data of older adults (consisting of 4258 steps on treadmill, 2256 steps on level overground and 2888 steps on overground incline/decline). Transfer learning was implemented in two steps to evaluate its effectiveness in improving the model's performance on older adults' overground walking data.

In the first step, fine-tuning was performed incrementally using treadmill data from older adults. Subsets of treadmill data (n=1, 5, 10, 15, 20 participants) were added sequentially to the training set to evaluate how the inclusion of increasing amounts of treadmill data influenced the model's performance. For each subset, random subsets of participants were selected across multiple iterations to ensure robust evaluation. Specifically, 15 iterations were conducted for n=1, 10 iterations for n=5, 5 iterations for n=10, and 3 iterations for n=15. The earlier layers of the pre-trained model, which was initially trained on treadmill data from younger adults, were frozen to preserve the foundational features learned during the initial training phase. Fine-tuning was applied to the later layers of the model to adapt it to the dynamics of treadmill walking from older adults. After fine-tuning on each subset of data, the model's performance was evaluated on unseen overground walking data from the test set. F1-score and MAE were used as performance metrics to assess how well the model generalized to these unseen conditions. This step aimed to determine whether the addition of treadmill data from older adults improved the model's ability to generalize to overground walking.

In the second step, fine-tuning was extended by incrementally adding overground level walking data from older adults to the training set. As in Step 1, data were added in subsets to assess the impact of increasing amounts of overground data on model performance. This incremental approach made it possible to determine the point at which adding more overground data no longer resulted in substantial performance gains. Fig. 1 describes a schematic flowchart of the process.

The results from Steps 1 and 2 were compared to evaluate the relative contributions of treadmill and overground data to improving the F1-score, MAE and RMSE, as well as to identify the minimal amount of additional data required to achieve optimal performance. All models were implemented in Python3 with the Pytorch library (Paszke et al., 2019).

2.4. Performance metrics

The metrics used for evaluation of model performance included:

1. F1-score: the harmonic mean of precision and recall, where precision is the ratio of true positive predictions to the total predicted positives, and recall is the ratio of true positive predictions to the total actual positives.

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Both micro- and macro-averaged precision and recall were computed. Micro averaging aggregates contributions of all classes (0, 1, 2 labels) to calculate metrics globally. Macro averaging calculates metrics independently for each class and then averages them.

2. Accuracy: Accuracy is calculated as the ratio of correctly predicted labels to the total number of labels.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

3. Mean absolute error (MAE) Assesses the average magnitude of errors between predicted and actual GCT.
4. Root mean squared error (RMSE): A measure of the differences between predicted and actual values.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

MAE and RMSE were computed along with the classical ML metrics because while the latter provide information about the

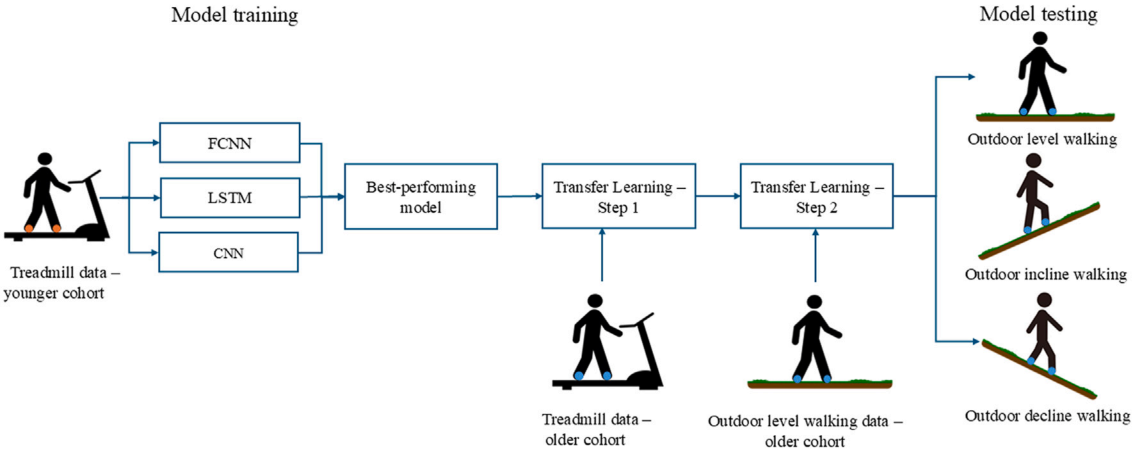
Model training
Model testing
FCNN
LSTM
CNN
Best-performing model
Transfer Learning – Step 1
Transfer Learning – Step 2
Treadmill data – younger cohort
Treadmill data – older cohort
Outdoor level walking data – older cohort
Outdoor level walking
Outdoor incline walking
Outdoor decline walking

Fig. 1. Schematic flowchart of the methodology. FCNN: Fully-connected neural network, BiLSTM: Bidirectional Long short-term memory network and CNN: Convolutional neural network.

one-to-one mapping of classes, they do not capture whether the temporal structure of the ground contact sequence was preserved. MAE and RMSE therefore quantify the magnitude of timing errors in milliseconds, providing complementary information on temporal accuracy.

3. Results

3.1. t-SNE analysis

Fig. 2 visualizes the t-SNE analysis on (a) raw IMU data and (b) FFT features of the IMU signals. In Fig. 2a, two clear clusters are formed with distinct separation between the data from the younger cohort and the older cohort, but there are overlapping points between the treadmill, overground level and overground incline/decline conditions from the older cohort. In Fig. 2b, there are three distinct clusters formed with the treadmill data from the younger cohort, treadmill data from the older cohort and a third cluster with overlapping points between the level outdoor and incline/decline data from the older cohort.

3.2. Selection of the best-performing ML model

Table 1 lists the performance metrics of the deep learning models trained on treadmill data from the younger cohort and tested on all test datasets from the older cohort. See Appendix C for micro- and macro-average precision and recall values.

As the CNN model had the highest F1-scores and lowest MAE across all 4 conditions, it was selected for transfer learning.

3.3. Transfer learning — iterative addition of treadmill walking data from the older cohort

Fig. 3 describes how the F1-score changed with iterative addition of treadmill data from the older cohort. As can be observed from the figure, the F1-scores started plateauing at n=10. At n=10, the mean F1-score were 0.9864 ± 0.0010 for treadmill walking, 0.9558 ± 0.0019 for outdoor level walking, 0.9463 ± 0.0021 for outdoor incline walking and 0.9029 ± 0.0012 for outdoor decline walking.

Hence, a random iteration of the fine-tuned model at n=10 was selected for further fine-tuning with overground data. See Appendix D for the mean (± 95% confidence interval (CI)) of the performance metrics for all subsets.

3.4. Transfer learning — iterative addition of overground level walking data from the older cohort

Fig. 4 describes the effect of iteratively adding outdoor level walking data from the older cohort on the F1-scores. As can be seen from the figure, the F1-score values started plateauing at n=5. The mean F1-scores at n=5 were 0.9637 ± 0.0023 for outdoor level walking, 0.9538 ± 0.0017 for outdoor incline walking and for 0.9122 ± 0.0021 outdoor decline walking. See Appendix E for the mean (± 95% CI) performance metrics for all subsets.

4. Discussion

The goal of this study evaluate how deep learning models trained to predict GCT on treadmill walking in a cohort of young adults performed on treadmill walking and outdoor overground walking data from older adults. Additionally, we aimed to assess whether sequential transfer learning with additive fine-tuning data from treadmill and overground level walking data could enhance predictive accuracy in outdoor overground level, incline and decline walking.

Our outcomes suggest that the CNN model was overall the most efficient in adapting learned features to data from older adults. Sequential transfer learning improved predictive accuracy for overground level and incline walking conditions, as evidenced by the increased F1-scores as fine-tuning data was incrementally added. However, the improvements for overground decline walking were less pronounced, suggesting that transfer learning with level treadmill or outdoor data used in this study were less effective in learning the features required to effectively predict GCT in downhill walking. Furthermore, the results across conditions demonstrated that F1-scores plateaued around n=10 iterations for treadmill data and n=5 for overground data, indicating that this is likely the optimal amount of fine-tuning data required to maximize performance while avoiding diminishing returns.

4.1. t-SNE analysis

The t-SNE analysis provided insights into the similarities and differences between walking conditions and age groups, as visualized in the plots. In the first t-SNE plot, raw IMU data was directly used. This data likely contained noise introduced by mechanical vibrations of the treadmill during treadmill walking and signal disruptions from wifi connectivity during outdoor overground walking. This noise may have contributed to the lack of clear separations between the treadmill, outdoor level and outdoor incline/decline data from the older cohort, as the noise obscured the underlying gait patterns.

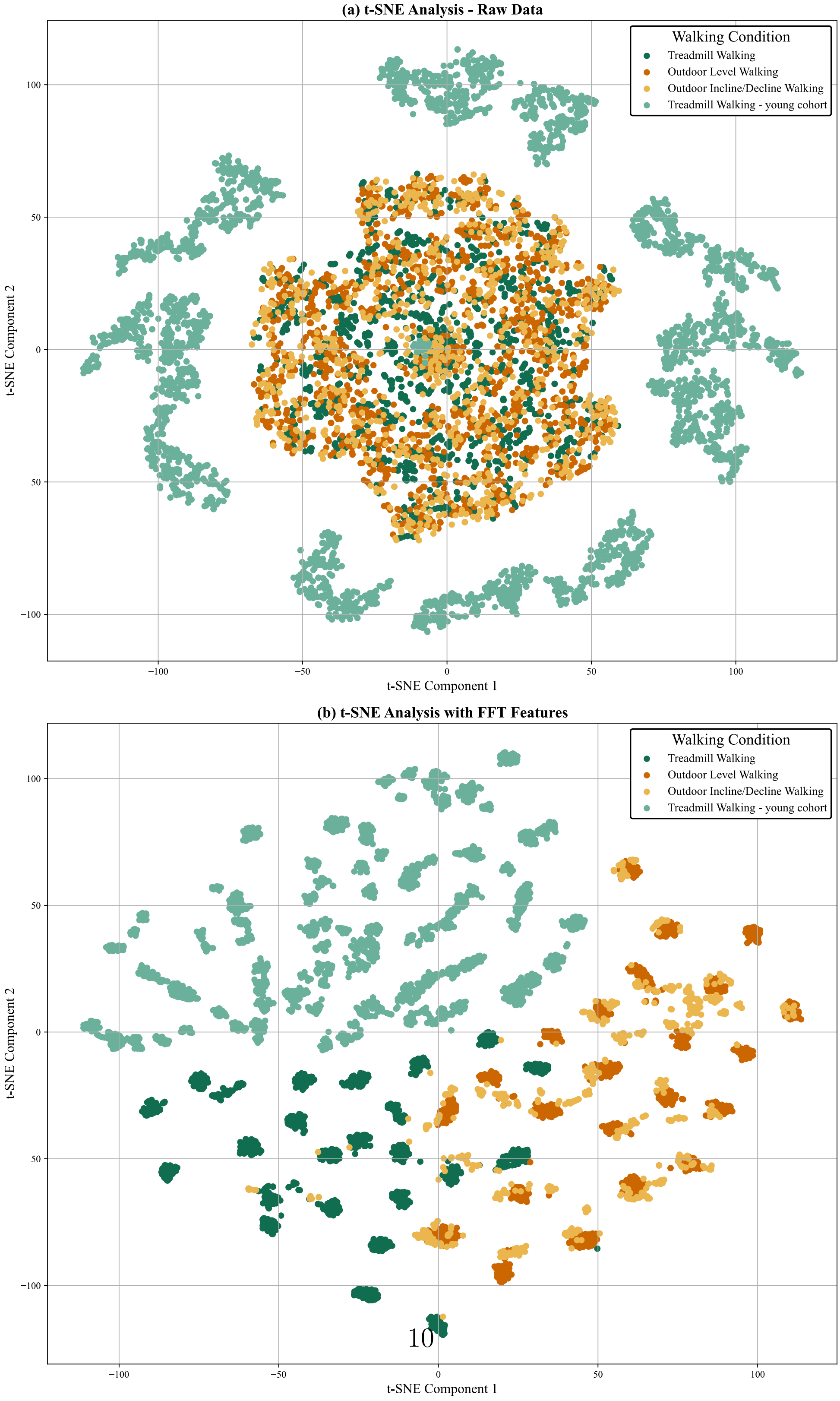

Fig. 2. t-SNE analysis of (a) raw IMU data and (b) FFT features of the IMU signals. Units are dimensionless.

Applying FFT to the IMU data transformed the noisy time-series signals into the frequency domain, effectively isolating the periodicity and rhythmic patterns inherent to walking. This allowed for clearer differentiation between the walking conditions. Hence, treadmill walking from both cohorts formed well-separated clusters, highlighting distinct frequency-domain characteristics of treadmill walking between these two age groups. However, outdoor walking showed less separability, indicating similar gait features across the three conditions. Some overlap was also observed between treadmill and outdoor walking by older adults, particularly on flat terrain. This overlap indicates that while

Table 1
F1-score, accuracy, MAE and RMSE of FCNN, BiLSTM and CNN on treadmill walking, outdoor level walking, outdoor incline walking, outdoor decline walking and mean values from all conditions. FCNN: fully connected neural network, BiLSTM: bidirectional long short term memory network, CNN: Convolutional neural network.

Model	F1-score	Accuracy	RMSE (ms)	MAE (ms)
1. Treadmill walking				
FCNN	0.9238	0.9242	2.7529	0.7578
BiLSTM	0.8508	0.8508	3.7658	1.4922
CNN	0.9498	0.9503	2.2284	0.4966
2. Outdoor level walking				
FCNN	0.9289	0.9298	2.6485	0.7014
BiLSTM	0.8468	0.8414	3.8554	1.5859
CNN	0.9323	0.9335	2.5783	0.6647
3. Outdoor incline walking				
FCNN	0.9205	0.9213	2.6205	0.6867
BiLSTM	0.8334	0.8277	4.0239	1.7228
CNN	0.9307	0.9320	2.6073	0.6798
4. Outdoor decline walking				
FCNN	0.8702	0.8723	3.5737	1.2771
BiLSTM	0.7587	0.7504	4.9282	2.4961
CNN	0.8808	0.8827	3.4233	1.1726
5. Mean values from all conditions				
FCNN	0.9134	0.9144	2.8989	0.8558
BiLSTM	0.8234	0.8176	4.1433	1.8243
CNN	0.9234	0.9247	2.7093	0.7534

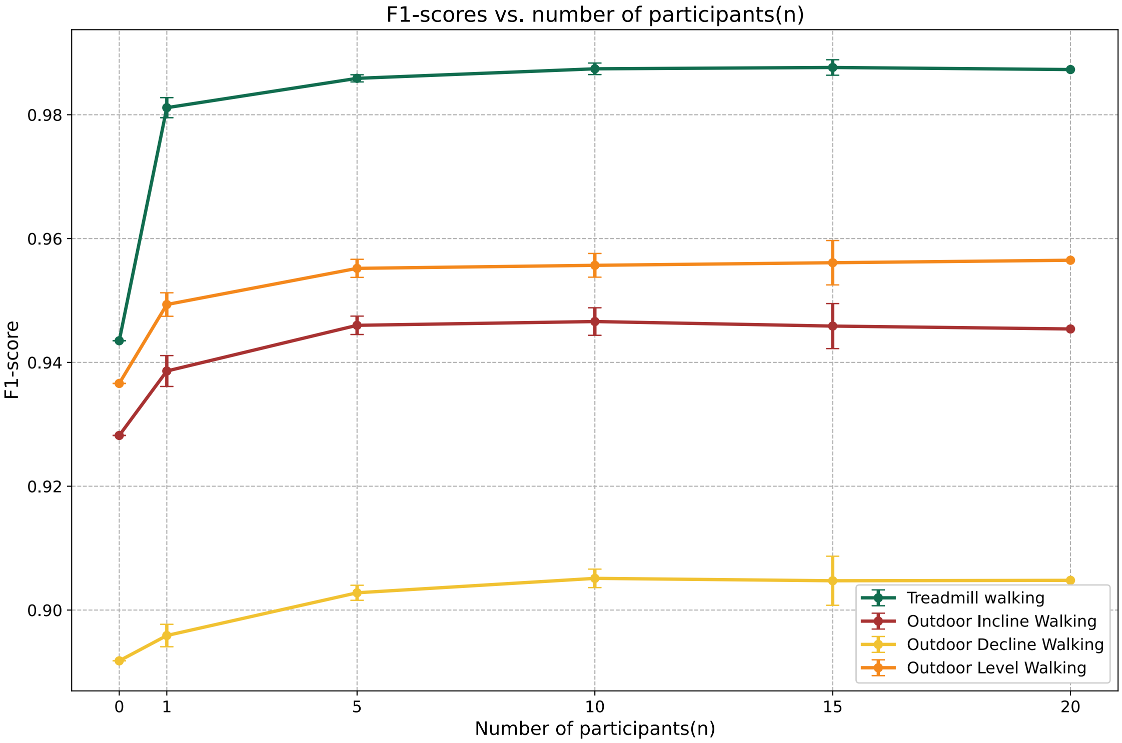

Fig. 3. A plot of the mean F1-scores (± 95% CI) versus number of participants with iterative addition of treadmill data from the older cohort.

there are unique aspects to overground walking due to environmental and biomechanical variability, there are also shared features that models trained on treadmill data could potentially leverage.

While no prior studies have performed t-SNE or other dimensionality reduction techniques specifically to explore differences between treadmill and overground walking, these techniques have previously been applied in related contexts. Specifically, principal component analysis (PCA) has been used in previous research to differentiate between the ground reaction forces patterns of male and female runners (Yu et al., 2021), identify variations in running biomechanics between novice and experienced runners (Jiang et al., 2023) and to identify the differences in joint kinematics between the paretic and non-paretic legs in hemiplegic patients (Milovanović and Popović, 2012). t-SNE analysis has been previously used to visualize the similarity between real and augmented gait data generated using generative adversarial networks (Wang et al., 2021).

4.2. Evaluation of model performance on data from the older cohort

The CNNs trained on treadmill walking from the younger cohort demonstrated the highest F1-scores and the lowest MAE consistently

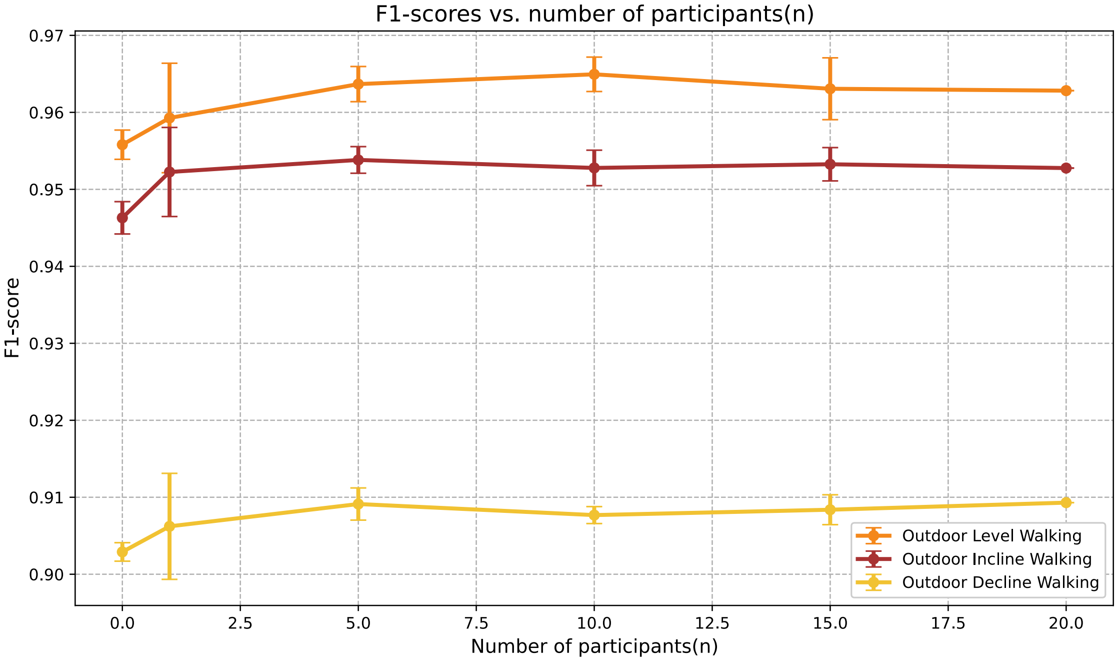

Fig. 4. A plot of F1-scores versus number of participants with iterative addition of outdoor level data from the older cohort.

on all test datasets. The implemented BiLSTM, on the other hand, demonstrated the lowest metrics. These findings are consistent with those from our previous study (Sansgiri et al., 2024), wherein BiLSTMs trained on treadmill data from young adults failed to generalize to unseen overground data. This could be due to many reasons. Firstly, given the small size of the test dataset, the BiLSTM may have struggled to generalize effectively as they are prone to overfitting. On the other hand, CNNs are good at identifying localized spatial patterns in the data. Since a gait cycle is repetitive and periodic, the CNNs could have been more efficient at capturing these generalizable features without overfitting on the training data. Both FCNN and CNN architectures are also generally more robust to noise compared to BiLSTMs. Since our training data were unfiltered due to recommendations from previous research (Jlassi and Dixon, 2024), this could also be an additional reason why these two models fared better than the BiLSTM. Future work should also explore other neural architectures to determine their efficacy in predicting features in unseen gait data.

4.3. Effect of incrementally adding fine-tuning data from treadmill walking and outdoor walking

In the first fine-tuning step, the F1-scores for all conditions increased upon adding treadmill data from the older cohort, but plateaued at n = 10. The performance remained consistently high at n = 15 and 20, suggesting that fine-tuning with data from 10 participants was sufficient for the model to generalize to the unseen treadmill and overground data. The F1-scores for downhill walking were consistently lower than the other three conditions, suggesting that the underlying data characteristics from downhill walking are different from the other three conditions.

When outdoor level data were incrementally added to the fine-tuned model, the model showed a slower improvement in the F1-score compared to the first scenario. Here, the F1-scores plateaued at n = 5, suggesting that the model benefitted from early fine-tuning, but adding more data provided no additional benefit. In the first step, the F1-scores for treadmill walking plateaued at about 0.986 with the addition of treadmill data, whereas the addition of outdoor level data plateaued at about 0.964. As expected, adding more data generally improved model performance, but the main value of this analysis was in identifying the point at which performance stabilized (n = 10 for treadmill data, n = 5 for overground data). This provides practical guidance on the minimal amount of fine-tuning data required to achieve stable performance.

It is also important to note that while the reported F1-scores are comparable to other studies investigating gait event detection in older adults (Kim et al., 2024; Sharifi Renani et al., 2020; Skvortsov et al., 2023), a possible reason for the relatively lower score could be that outdoor data has greater variation in stride lengths and times due to factors such as uneven surfaces and more noise because of environmental factors. These factors could affect the model's ability to achieve performance comparable to treadmill walking. As in the first step, the values for downhill walking remained lower than the other conditions.

To better understand the differences in the underlying data characteristics across outdoor walking conditions, we performed a PCA on the outdoor test cohort (Fig. 5). As can be seen in the plot, outdoor level and incline walking share more overlapping regions while outdoor decline walking shows a more distinct and dispersed cluster, indicating a greater variability in the underlying features. This separation could be attributed to the unique biomechanical demands of downhill walking, such as the need for increased braking forces to control descent along with shorter stride lengths, reduced step times, and higher variability in step placement. Hence, for future studies replicating this research, specialized modeling approaches need to be developed to better handle the dynamics of decline walking.

4.4. Limitations

This study has several limitations to acknowledge. Firstly, the older adult cohort included in this study were healthy and mobile, demonstrating mean walking speeds of 3.9 km/h on the treadmill and 5.4 km/h overground (Matikainen-Tervola et al., 2024), which are similar to the walking speeds observed in young adults (Kim and Kim, 2014). As a result, there was limited variability between the two cohorts, potentially reducing the challenges of transfer learning. Future studies should aim to include a wider range of functional abilities within the older adult group to better evaluate the robustness of transfer learning approaches.

Secondly, the test cohort was small, consisting of six participants, and a fixed test set was used due to the retrospective nature of the data.

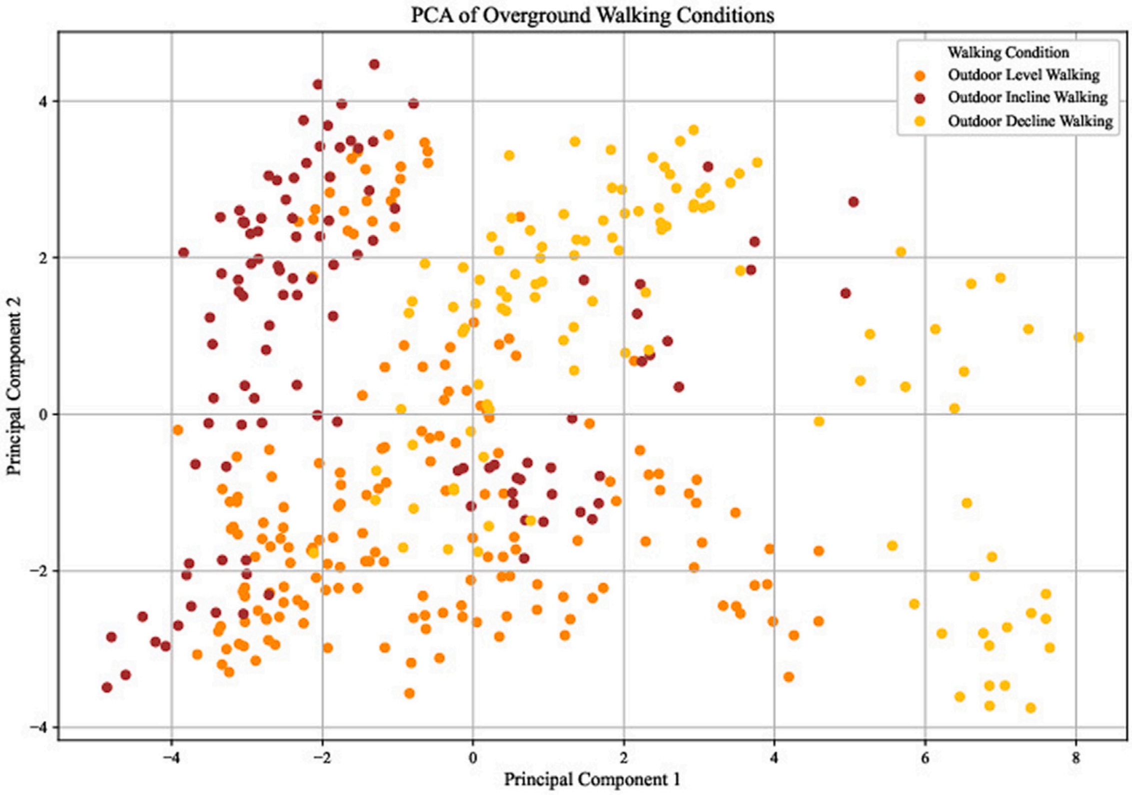

Fig. 5. PCA of the outdoor walking data. Units are dimensionless.

While this allowed for detailed analysis within the scope of the study, it limits the generalizability of the findings to the broader population of older adults. A larger test sample size and/or testing on an independent dataset would be necessary to confirm the effectiveness of these models.

Thirdly, the use-case of this study — predicting GCT — is a relatively simpler research question, and hence, we caution against directly generalizing these findings to more complex problems such as prediction of step lengths or joint angles without further research.

Lastly, this study did not compare GCT predictions to those obtained using traditional signal processing methods. While this comparison could provide additional insights, this choice is justified by our research question, which focused on evaluating the potential of transfer learning to adapt models trained on young adults' data to data from older adults.

4.5. Conclusion

In conclusion, this study provided insights into the practical application of transfer learning in scenarios where data collection from target populations is limited.

Our findings highlighted the efficacy of CNNs in adapting to unseen datasets. Sequential fine-tuning improved predictive accuracy for treadmill and overground walking. Notably, the model struggled to generalize to downhill walking, suggesting that unique biomechanical demands, such as increased braking forces and variability in step placement, require specialized modeling approaches.

Future work should focus on addressing the limitations identified in this study by incorporating a more diverse older adult cohort with a wider range of functional abilities, as well as increasing the test sample size to improve generalizability. Implementing domain adaptation techniques such as adversarial networks should be explored to potentially improve model performance for challenging conditions like downhill walking. By addressing these areas, future studies could advance robust and scalable models capable of bridging the gap between controlled laboratory settings and real-world applications.

CRediT authorship contribution statement

Sailee Sansgiri: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Emmi Matikainen-Tervola:** Writing – review & editing, Writing – original draft, Methodology, Data curation. **Merja Rantakokko:** Writing – review & editing, Writing – original draft, Methodology. **Taija Finni:** Writing – review & editing, Writing – original draft, Supervision. **Timo Rantalainen:** Writing – review & editing, Writing – original draft, Supervision, Conceptualization. **Neil J. Cronin:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization.

Declaration of funding sources

Data collection in the older cohort was supported by Research Council of Finland (grant number 330185 to Merja Rantakokko). The financial sponsors had no role in the design, implementation, analyses or reporting of the results.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, author Sailee Sansgiri used ChatGPT 4o to rephrase and shorten the abstract, methods and discussion sections of the manuscript. After using this tool, all authors reviewed and edited the content as needed, and take full responsibility for the content of the published article.

Declaration of competing interest

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.

Table B.2
Selected hyperparameters. FCNN: fully connected neural network, BiLSTM: bidirectional long short term memory network, CNN: Convolutional neural network.

Hyperparameters tested	Best hyperparameters
FCNN	
Batch Size	32
Learning Rate	0.001
Number of Layers	4
Hidden Dimensions	64, 128, 64, 64
Dropout Rate	0.1
Number of Epochs	50
Optimization Algorithm	adam
Loss Function	Binary Cross-Entropy with Logits Loss
Activation Function	relu
Mean 5-fold F1-score	0.9868
Mean precision	0.9872
Mean recall	0.9864
BiLSTM	
Batch Size	32
Learning Rate	0.005
Number of LSTM Layers	3
Number of Fully Connected Layers	1
Hidden Size	128, 128
Dropout Rate	0.3
Number of Epochs	50
Optimization Algorithm	adam
Loss Function	Focal Loss
Activation Function	relu
Mean 5-fold F1-score	0.9940
Mean precision	0.9943
Mean recall	0.9937
CNN	
Batch Size	32
Learning Rate	0.001
Number of Convolutional Layers	3
Number of Fully Connected Layers	2
Convolutional Filters	128
CNN Hidden Size	256
Kernel Size	3
Dropout Rate	0.2
Number of Epochs	50
Weight Decay	1e−5
Optimization Algorithm	adam
Loss Function	Binary Cross-Entropy with Logits Loss
Activation Function	relu
Mean 5-fold F1-score	0.9911
Mean precision	0.9915
Mean recall	0.9892

Appendix A. Data collection

A.1. Data collection in the younger cohort

Data were collected from a sample of 27 young, healthy adults (Mean age: 27.6 ± 4.41 years, 15 females) without any self-reported history of cardiovascular diseases or recent injuries. Out of these, 20 random participants were assigned to the training cohort, and their data were used for training the ML models. Participants wore comfortable sports clothing and their own running shoes during the measurement. Participants walked on a level treadmill (Gymstick Walking Pad Pro, 0.44 × 1.20 m) at speeds from 3.5 km/h to 6 km/h. The speeds were increased at increments of 0.5 km/h after 100 continuous steps (per leg) were collected at each speed.

Their gait was simultaneously measured using two different wearable gait measurement systems: 2 IMUs attached on top of the shoes' midfoot (Movella DOT, Movella Technologies B.V., Netherlands) sampled at 120 Hz, and in-shoe pressure insoles with inbuilt IMUs (OpenGO sensor insoles, Moticon DE, Germany) sampled at 100 Hz. Data collected from all devices were transmitted via Bluetooth 5.0 to a smartphone and later manually transferred to a computer.

Before each measurement, the participant was asked to jump twice on the spot. These jumps resulted in two distinct peaks in the vertical acceleration signals of the foot-mounted IMUs and the inbuilt IMUs of the in-shoe insoles, and the timings of these peaks were used to synchronize the 3 signals. Ground truth GCT annotations were obtained from the vertical ground reaction force (vGRF) from the insoles. A threshold of 30N was set on the vGRF of each foot, with values exceeding 30N indicating ground contact.

A.2. Data collection in the older cohort

Participants included 26 ambulatory adults aged 69–92 (mean age: 76 ± 5.2 years, 17 females), without any severe sensory deficits, memory impairments or neurological conditions (Mini-Mental State Examination (MMSE) score < 22), and who could walk at least 1 km without assistive devices or orthotics. Participants were instructed to wear dark-colored, tight-fitting, comfortable sports clothing and their own sports shoes. 3-D accelerometer and gyroscope data (sampled at 200 Hz) were collected from 2 IMUs (NGIMU, x-ion, UK) mounted on top of each shoe, near the midfoot. Data collection occurred in two settings: indoor treadmill walking and outdoor overground walking. During treadmill trials, participants walked indoors on a level treadmill. Each treadmill walking trial lasted for three minutes, preceded by a

Table B.3
Results of nested cross-validation. FCNN: fully connected neural network, BiLSTM: bidirectional long short term memory network, CNN: Convolutional neural network.

Outer fold (n)	F1-score	Precision (micro)	Recall (micro)	Precision (macro)	Recall (macro)
			FCNN		
1	0.9884	0.9884	0.9884	0.9859	0.9877
2	0.9887	0.9887	0.9887	0.9863	0.9880
3	0.9888	0.9888	0.9888	0.9862	0.9884
4	0.9888	0.9888	0.9888	0.9866	0.9879
5	0.9881	0.9881	0.9881	0.9850	0.9879
Mean (± SD)	0.9885 (±0.0003)	0.9885 (±0.0003)	0.9885 (±0.0003)	0.9860 (±0.0006)	0.9880 (±0.0002)
			BiLSTM		
1	0.9895	0.9908	0.9883	0.9859	0.9859
2	0.9897	0.9915	0.9878	0.9861	0.9861
3	0.9904	0.9916	0.9892	0.9870	0.9870
4	0.9901	0.9706	0.9896	0.9867	0.9867
5	0.9881	0.9894	0.9867	0.9839	0.9839
Mean (± SD)	0.9896 ± 0.0008	0.9908± 0.0008	0.9883± 0.0010	0.9859 ± 0.0011	0.9859 ± 0.0011
			CNN		
1	0.9845	0.9845	0.9845	0.9808	0.9838
2	0.9846	0.9846	0.9846	0.9809	0.9841
3	0.9848	0.9848	0.9848	0.9821	0.9833
4	0.9849	0.9849	0.9849	0.9814	0.9842
5	0.9841	0.9841	0.9841	0.9804	0.9834
Mean (± SD)	0.9846 ± 0.0003	0.9846 ± 0.0003	0.9846 ± 0.0003	0.9811 ± 0.0006	0.9838 ± 0.0004

Table B.4
Full hyperparameter grids explored during nested cross-validation (inner $k = 4$, outer $k = 5$).

Model	Hyperparameter grid
BiLSTM	Hidden size: $\{128, 256\}$; Number of LSTM layers: $\{2, 3\}$; Bidirectional: $\{\text{True}, \text{False}\}$; Number of fully connected layers: $\{1, 2\}$; Dropout rate: $\{0.1, 0.3\}$; Learning rate: $\{0.001, 0.0005\}$; Batch size: $\{64, 128\}$; Loss: $\{\text{BCE with Logits}, \text{Focal}\}$; Focal loss parameters (if used): $\alpha \in \{0.25, 0.5\}$, $\gamma \in \{1.0, 2.0\}$.
CNN	Hidden size: $\{128, 256\}$; Convolutional filters: $\{64, 128\}$; Kernel size: $\{3, 5\}$; Number of convolutional layers: $\{2, 3\}$; Number of fully connected layers: $\{1, 2\}$; Dropout rate: $\{0.1, 0.3, 0.5\}$; Learning rate: $\{0.001, 0.0005\}$; Batch size: $\{16, 32, 64\}$; Weight decay: $\{1e-4, 1e-5\}$;
FCNN	Number of fully connected layers: $\{2, 3, 4\}$; Hidden layer widths: 64–128 units per layer; Dropout rate: $\{0.1, 0.2, 0.3\}$; Learning rate: $\{0.001, 0.005\}$; Batch size: $\{32, 64\}$; Batch normalization: $\{\text{True}, \text{False}\}$;

Table C.5
Micro and macro precision and recall of FCNN, BiLSTM and CNN on treadmill walking, outdoor level walking, outdoor incline walking, outdoor decline walking and mean values from all conditions. FCNN: fully connected neural network, BiLSTM: bidirectional long short term memory network, CNN: Convolutional neural network.

Model	Precision (micro)	Recall (micro)	Precision (macro)	Recall (macro)
1. Treadmill walking				
FCNN	0.9268	0.9208	0.9118	0.9008
BiLSTM	0.8568	0.8449	0.8368	0.8199
CNN	0.9538	0.9458	0.9418	0.9298
2. Outdoor level walking				
FCNN	0.9319	0.9260	0.9169	0.9060
BiLSTM	0.8528	0.8410	0.8328	0.8160
CNN	0.9353	0.9293	0.9233	0.9133
3. Outdoor incline walking				
FCNN	0.9235	0.9176	0.9085	0.8976
BiLSTM	0.8394	0.8277	0.8194	0.8027
CNN	0.9337	0.9277	0.9217	0.9117
4. Outdoor decline walking				
FCNN	0.8732	0.8673	0.8582	0.8473
BiLSTM	0.7707	0.7468	0.7507	0.7218
CNN	0.8858	0.8759	0.8738	0.8599
5. Mean values from all conditions				
FCNN	0.9139	0.9080	0.8999	0.8879
BiLSTM	0.8300	0.8151	0.8100	0.7901
CNN	0.9271	0.9197	0.9151	0.9037

one-to-five-minute familiarization period at the beginning of the measurement. Treadmill speed was adjusted in 0.5 km/h increments until the participant indicated that the speed was comfortable. Participants were permitted to use the treadmill rails for support as needed.

Opto-electronic data were collected concurrently using an 16-camera Vicon motion capture system (Vicon Vero, Oxford, UK), sampling at 200 Hz. Reflective markers were placed on anatomical landmarks following the Vicon lower-body Plug-in Gait model (Vicon, 2002) with an additional hallux marker on each foot. Note that the foot markers were placed on top of the participants' shoes, and the other lower body markers were placed on bony landmarks, either on skin or on top of clothing. GCT were manually annotated by researchers using marker trajectories from the heel and hallux markers. This served as the ground truth labels for ML model training and evaluation.

IC was defined at the preceding frame when the heel marker started moving backwards, or as the immediate next frame after the heel marker stopped moving forward at the beginning of the stance phase. TO was defined at the immediate frame preceding the frame where the hallux marker moved upwards, just at the beginning of the swing phase. Both events were annotated from the sagittal view. Mokka (Motion kinematic & kinetic analyzer, Biomechanical ToolKit, https://biomechanical-toolkit.github.io/mokka/) was used to analyze the trajectories of the markers.

For outdoor trials, participants walked on a track for six minutes. The track consisted of two 29-meter straight sections connected by semicircular turnings (70 m per lap). Data from the first and last minute of walking were annotated with ground truth labels and used for further analysis. Participants then performed five repetitions of uphill and downhill walking on a 20-meter straight asphalt path at a self-selected comfortable speed. A figure illustrating the outdoor route is available in a previous publication (Rantakokko et al., 2024).

Ground truth labels were obtained by manually annotating GCT from video recordings of a high-speed camera (sampling frequency 120 Hz) that was mounted on a stroller moving parallel to the participant. The time of the frame when the shoe touched the ground after the swing phase was used as the IC, and the time of the frame before the shoe left the ground, was used as the TO. Shotcut software (Meltytech LLC, Oceanside, CA) was used to analyze the step events from the videos.

Appendix B. Selection of the best hyperparameters

B.1. Determining the optimal hyperparameters

Table B.2 summarizes the best hyperparameters and the mean five-fold F1-score for all three models.

B.2. Nested cross-validation

While we determined the optimal set of hyperparameters through a 5-fold cross-validation, a nested cross-validation with 4 inner folds and 5 outer folds was also performed to reveal the stability and generalizability of model performance across different train–validation splits. In this setup, the inner loop was used to select the best hyperparameters within each outer fold, and the outer loop then provided an unbiased estimate of performance on held-out data. Since the primary aim of the nested CV was to examine the dispersion of results rather than to exhaustively optimize the models, a smaller hyperparameter grid was employed. Table B.3 lists the results of this nested cross-validation. Table B.4 lists the hyperparameter grid used for nested cross-validation.

Appendix C. Selection of the best-performing model — precision and recall values

Table C.5 lists the micro and macro precision and recall values of the deep learning models trained on treadmill data from the younger cohort and tested on all test datasets from the older cohort

Table D.6
Results of iteratively fine-tuning with treadmill data. CI: confidence interval.

Subset (n)	F1-score (mean ± 95% CI)	Accuracy (mean ± 95% CI)	MAE (ms) (mean ± 95% CI)	RMSE (ms) (mean ± 95% CI)
Treadmill Walking				
1 (15 iterations)	0.9790 ± 0.0016	0.9720 ± 0.0022	0.2798 ± 0.0224	1.6688 ± 0.0653
5 (10 iterations)	0.9844 ± 0.0012	0.9793 ± 0.0016	0.2066 ± 0.0164	1.4355 ± 0.0561
10 (5 iterations)	0.9864 ± 0.0010	0.9819 ± 0.0012	0.1810 ± 0.0125	1.3448 ± 0.0466
15 (3 iterations)	0.9863 ± 0.0003	0.9818 ± 0.0004	0.1824 ± 0.0037	1.3506 ± 0.0137
20 (1 iteration)	0.9862 ± 0.0000	0.9816 ± 0.0000	0.1838 ± 0.0000	1.3558 ± 0.0000
Overground Level Walking				
1 (15 iterations)	0.9583 ± 0.0020	0.9455 ± 0.0027	0.5449 ± 0.0269	2.3321 ± 0.0574
5 (10 iterations)	0.9541 ± 0.0016	0.9400 ± 0.0022	0.6001 ± 0.0217	2.4491 ± 0.0444
10 (5 iterations)	0.9558 ± 0.0019	0.9423 ± 0.0026	0.5771 ± 0.0255	2.4020 ± 0.0530
15 (3 iterations)	0.9566 ± 0.0018	0.9432 ± 0.0023	0.5677 ± 0.0235	2.3825 ± 0.0491
20 (1 iteration)	0.9580 ± 0.0000	0.9452 ± 0.0000	0.5483 ± 0.0000	2.3415 ± 0.0000
Overground Incline Walking				
1 (15 iterations)	0.9461 ± 0.0019	0.9294 ± 0.0027	0.7058 ± 0.0267	2.6552 ± 0.0504
5 (10 iterations)	0.9445 ± 0.0014	0.9273 ± 0.0018	0.7271 ± 0.0182	2.6960 ± 0.0339
10 (5 iterations)	0.9463 ± 0.0021	0.9296 ± 0.0027	0.7037 ± 0.0275	2.6525 ± 0.0518
15 (3 iterations)	0.9461 ± 0.0021	0.9293 ± 0.0029	0.7065 ± 0.0292	2.6580 ± 0.0548
20 (1 iteration)	0.9475 ± 0.0000	0.9313 ± 0.0000	0.6870 ± 0.0000	2.6211 ± 0.0000
Overground Decline Walking				
1 (15 iterations)	0.9046 ± 0.0015	0.8779 ± 0.0023	1.2212 ± 0.0230	3.4941 ± 0.0330
5 (10 iterations)	0.8998 ± 0.0020	0.8717 ± 0.0028	1.2832 ± 0.0276	3.5818 ± 0.0385
10 (5 iterations)	0.9029 ± 0.0012	0.8761 ± 0.0015	1.2388 ± 0.0151	3.5196 ± 0.0215
15 (3 iterations)	0.9036 ± 0.0019	0.8769 ± 0.0020	1.2307 ± 0.0203	3.5081 ± 0.0290
20 (1 iteration)	0.9062 ± 0.0000	0.8803 ± 0.0000	1.1973 ± 0.0000	3.4602 ± 0.0000

Appendix D. Transfer learning — iterative addition of treadmill walking data from the older cohort

Table D.6 lists the F1-scores, accuracy, MAE and RMSE (mean ± 95% CI) of iteratively fine-tuning the CNN model with treadmill data.

Table D.7 lists the micro and macro precision and recall values (mean ± 95% CI) of iteratively fine-tuning the CNN model with treadmill data.

Appendix E. Transfer learning — iterative addition of overground level walking data from the older cohort

Table E.8 lists the F1-scores, accuracy, MAE and RMSE (mean ± 95% CI) of iteratively fine-tuning the CNN model with overground level data.

Table D.7
Results of iteratively fine-tuning with treadmill data. CI: confidence interval.

Subset (n)	Precision (micro) (mean ± 95% CI)	Recall (micro) (mean ± 95% CI)	Precision (macro) (mean ± 95% CI)	Recall (macro) (mean ± 95% CI)
Treadmill Walking				
1 (15 iterations)	0.9718 ± 0.0026	0.9720 ± 0.0022	0.2798 ± 0.0224	1.6688 ± 0.0653
5 (10 iterations)	0.9772 ± 0.0016	0.9793 ± 0.0016	0.2066 ± 0.0164	1.4355 ± 0.0561
10 (5 iterations)	0.9792 ± 0.0011	0.9819 ± 0.0012	0.1810 ± 0.0125	1.3448 ± 0.0466
15 (3 iterations)	0.9783 ± 0.0036	0.9818 ± 0.0004	0.1824 ± 0.0037	1.3506 ± 0.0137
20 (1 iteration)	0.9786 ± 0.0000	0.9816 ± 0.0000	0.1838 ± 0.0000	1.3558 ± 0.0000
Overground Level Walking				
1 (15 iterations)	0.9583 ± 0.0020	0.9455 ± 0.0027	0.5449 ± 0.0269	2.3321 ± 0.0574
5 (10 iterations)	0.9541 ± 0.0016	0.9400 ± 0.0022	0.6001 ± 0.0217	2.4491 ± 0.0444
10 (5 iterations)	0.9558 ± 0.0019	0.9423 ± 0.0026	0.5771 ± 0.0255	2.4020 ± 0.0530
15 (3 iterations)	0.9566 ± 0.0018	0.9432 ± 0.0023	0.5677 ± 0.0235	2.3825 ± 0.0491
20 (1 iteration)	0.9580 ± 0.0000	0.9452 ± 0.0000	0.5483 ± 0.0000	2.3415 ± 0.0000
Overground Incline Walking				
1 (15 iterations)	0.9461 ± 0.0019	0.9294 ± 0.0027	0.7058 ± 0.0267	2.6552 ± 0.0504
5 (10 iterations)	0.9445 ± 0.0014	0.9273 ± 0.0018	0.7271 ± 0.0182	2.6960 ± 0.0339
10 (5 iterations)	0.9463 ± 0.0021	0.9296 ± 0.0027	0.7037 ± 0.0275	2.6525 ± 0.0518
15 (3 iterations)	0.9461 ± 0.0021	0.9293 ± 0.0029	0.7065 ± 0.0292	2.6580 ± 0.0548
20 (1 iteration)	0.9475 ± 0.0000	0.9313 ± 0.0000	0.6870 ± 0.0000	2.6211 ± 0.0000
Overground Decline Walking				
1 (15 iterations)	0.9046 ± 0.0015	0.8779 ± 0.0023	1.2212 ± 0.0230	3.4941 ± 0.0330
5 (10 iterations)	0.8998 ± 0.0020	0.8717 ± 0.0028	1.2832 ± 0.0276	3.5818 ± 0.0385
10 (5 iterations)	0.9029 ± 0.0012	0.8761 ± 0.0015	1.2388 ± 0.0151	3.5196 ± 0.0215
15 (3 iterations)	0.9036 ± 0.0019	0.8769 ± 0.0020	1.2307 ± 0.0203	3.5081 ± 0.0290
20 (1 iteration)	0.9062 ± 0.0000	0.8803 ± 0.0000	1.1973 ± 0.0000	3.4602 ± 0.0000

Table E.8
Results of iteratively fine-tuning with outdoor level data. CI: confidence interval.

Subset (n)	F1-score (mean ± 95% CI)	Accuracy (mean ± 95% CI)	MAE (ms) (mean ± 95% CI)	RMSE (ms) (mean ± 95% CI)
		Overground Level Walking		
1 (15 iterations)	0.9493 ± 0.0111	0.9455 ± 0.0027	0.5449 ± 0.0269	2.3321 ± 0.0574
5 (10 iterations)	0.9637 ± 0.0023	0.9535 ± 0.0030	0.4649 ± 0.0295	2.1544 ± 0.0668
10 (5 iterations)	0.9649 ± 0.0022	0.9542 ± 0.0029	0.4582 ± 0.0270	2.1396 ± 0.0650
15 (3 iterations)	0.9631 ± 0.0040	0.9529 ± 0.0052	0.4709 ± 0.0522	2.1698 ± 0.1197
20 (1 iteration)	0.9628 ± 0.0000	0.9526 ± 0.0000	0.4742 ± 0.0000	2.1777 ± 0.0000
		Overground Incline Walking		
1 (15 iterations)	0.9523 ± 0.0058	0.9380 ± 0.0075	0.6195 ± 0.0751	2.4764 ± 0.1435
5 (10 iterations)	0.9538 ± 0.0017	0.9416 ± 0.0046	0.5842 ± 0.0524	2.4395 ± 0.1135
10 (5 iterations)	0.9528 ± 0.0023	0.9393 ± 0.0029	0.6069 ± 0.0291	2.4632 ± 0.0594
15 (3 iterations)	0.9533 ± 0.0022	0.9400 ± 0.0027	0.6004 ± 0.0272	2.4502 ± 0.0554
20 (1 iteration)	0.9528 ± 0.0000	0.9393 ± 0.0000	0.6074 ± 0.0000	2.4645 ± 0.0000
		Overground Decline Walking		
1 (15 iterations)	0.9062 ± 0.0069	0.8809 ± 0.0088	1.1912 ± 0.0880	3.4445 ± 0.1242
5 (10 iterations)	0.9122 ± 0.0021	0.8893 ± 0.0027	1.1075 ± 0.0269	3.3274 ± 0.0406
10 (5 iterations)	0.9077 ± 0.0011	0.8839 ± 0.0013	1.1610 ± 0.0128	3.4073 ± 0.0188
15 (3 iterations)	0.9084 ± 0.0019	0.8850 ± 0.0026	1.1498 ± 0.0258	3.3909 ± 0.0380
20 (1 iteration)	0.9093 ± 0.0000	0.8861 ± 0.0000	1.1389 ± 0.0000	3.3747 ± 0.0000

Table E.9
Results of iteratively fine-tuning with overground level data. CI: confidence interval.

Subset (n)	Precision (micro) (mean ± 95% CI)	Recall (micro) (mean ± 95% CI)	Precision (macro) (mean ± 95% CI)	Recall (macro) (mean ± 95% CI)
Overground Level Walking				
1 (15 iterations)	0.9487 ± 0.0065	0.9487 ± 0.0065	0.9453 ± 0.0069	0.9436 ± 0.0073
5 (10 iterations)	0.9570 ± 0.0016	0.9570 ± 0.0016	0.9536 ± 0.0017	0.9533 ± 0.0017
10 (5 iterations)	0.9537 ± 0.0030	0.9537 ± 0.0030	0.9500 ± 0.0030	0.9498 ± 0.0036
15 (3 iterations)	0.9549 ± 0.0030	0.9549 ± 0.0030	0.9512 ± 0.0030	0.9512 ± 0.0035
20 (1 iteration)	0.9545 ± 0.0000	0.9545± 0.0000	0.9508 ± 0.0000	0.9508 ± 0.0000
Overground Incline Walking				
1 (15 iterations)	0.9371 ± 0.0075	0.9371 ± 0.0075	0.9297 ± 0.0079	0.9333 ± 0.0086
5 (10 iterations)	0.9480 ± 0.0018	0.9480 ± 0.0018	0.9409 ± 0.0020	0.9461 ± 0.0021
10 (5 iterations)	0.9462 ± 0.0035	0.9462 ± 0.0035	0.9393 ± 0.0034	0.9436 ± 0.0044
15 (3 iterations)	0.9470 ± 0.0051	0.9470 ± 0.0051	0.9403 ± 0.0052	0.9445 ± 0.0060
20 (1 iteration)	0.9479 ± 0.0000	0.9479 ± 0.0000	0.9409 ± 0.0000	0.9459 ± 0.0000
Overground Decline Walking				
1 (15 iterations)	0.8853 ± 0.0062	0.8853 ± 0.0062	0.8783 ± 0.0065	0.8762 ± 0.0069
5 (10 iterations)	0.8934 ± 0.0027	0.8934 ± 0.0027	0.8869 ± 0.0029	0.8847 ± 0.0028
10 (5 iterations)	0.8922 ± 0.0035	0.8922 ± 0.0035	0.8855 ± 0.0037	0.8836 ± 0.0039
15 (3 iterations)	0.8925 ± 0.0003	0.8925 ± 0.0003	0.8859 ± 0.0006	0.8842 ± 0.0003
20 (1 iteration)	0.8918 ± 0.0000	0.8918 ± 0.0000	0.8851 ± 0.0000	0.8834 ± 0.0000

Table E.9 lists the micro and macro precision and recall values (mean ± 95% CI) of iteratively fine-tuning the CNN model with overground level data.

References

Aartolahti, E., 2024. Comparison of spatiotemporal gait parameters in indoor and outdoor walking among older adults. Arch. Phys. Med. Rehabil. 105 (4), e85–e86.

Avdan, G., Onal, S., Rekabdar, B., 2023. Regression transfer learning for the prediction of three-dimensional ground reaction forces and joint moments during gait. Int. J. Biomed. Eng. Technol. 42 (4), 317–338.

Bruening, D.A., Ridge, S.T., 2014. Automated event detection algorithms in pathological gait. Gait Posture 39 (1), 472–477.

Hansen, A.H., Childress, D.S., Meier, M.R., 2002. A simple method for determination of gait events. J. Biomech. 35 (1), 135–138.

Hillel, I., Gazit, E., Nieuwboer, A., Avanzino, L., Rochester, L., Cereatti, A., Croce, U.D., Rikkert, M.O., Bloem, B.R., Pelosin, E., et al., 2019. Is every-day walking in older adults more analogous to dual-task walking or to usual walking? Elucidating the gaps between gait performance in the lab and during 24/7 monitoring. Eur. Rev. Aging Phys. Act. 16, 1–12.

Jiang, X., Xu, D., Fang, Y., Bíró, I., Baker, J.S., Gu, Y., 2023. PCA of running biomechanics after 5 km between novice and experienced runners. Bioengineering 10 (7), 876.

Jlassi, O., Dixon, P.C., 2024. The effect of time normalization and biomechanical signal processing techniques of ground reaction force curves on deep-learning model performance. J. Biomech. 168, 112116.

Kang, H.G., Dingwell, J.B., 2008. Separating the effects of age and walking speed on gait variability. Gait Posture 27 (4), 572–577.

Kim, W.S., Kim, E.Y., 2014. Comparing self-selected speed walking of the elderly with self-selected slow, moderate, and fast speed walking of young adults. Ann. Rehabil. Med. 38 (1), 101–108.

Kim, Y.K., Pai, S.G., Choi, J.O., Tan, K.Z., Gwerder, M., Frautschi, A., Taylor, W.R., Singh, N.B., 2024. Leveraging deep learning and wearables for automatically identifying gait event: Effects of age and location of sensors on the assessment of gait events. IEEE Sensors J.

Matikainen-Tervola, E., Cronin, N., Aartolahti, E., Sihvonen, S., Sansgiri, S., Finni, T., Mattila, O.-P., Rantakokko, M., 2024. Validity of IMU sensors for assessing features of walking in laboratory and outdoor environments among older adults. Gait Posture 114, 277–283.

Milovanović, I., Popović, D.B., 2012. Principal component analysis of gait kinematics data in acute and chronic stroke patients. Comput. Math. Methods Med. 2012 (1), 649743.

Mundt, M., 2023. Bridging the lab-to-field gap using machine learning: a narrative review. Sport. Biomech. 1–20.

O'Brien, M.K., Lanotte, F., Khazanchi, R., Shin, S.Y., Lieber, R.L., Ghaffari, R., Rogers, J.A., Jayaraman, A., 2024. Early prediction of poststroke rehabilitation outcomes using wearable sensors. Phys. Ther. 104 (2), pzad183.

Pandit, T., Nahane, H., Lade, D., Rao, V., 2019. Abnormal gait detection by classifying inertial sensor data using transfer learning. In: 2019 18th IEEE International Conference on Machine Learning and Applications. ICMLA, pp. 1444–1447.

Pantonial, R., Simic, M., 2024. Transfer learning method for the classification of hip osteoarthritis using kinematic gait parameters. Procedia Comput. Sci. 246, 4692–4701.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. 32.

Plotnik, M., Bartsch, R.P., Zeev, A., Giladi, N., Hausdorff, J.M., 2013. Effects of walking speed on asymmetry and bilateral coordination of gait. Gait Posture 38 (4), 864–869.

Rantakokko, M., Matikainen-Tervola, E., Aartolahti, E., Sihvonen, S., Chichaeva, J., Finni, T., Cronin, N., et al., 2024. Gait features in different environments contributing to participation in outdoor activities in old age (GaitAge): Protocol for an observational cross-sectional study. JMIR Res. Protoc. 13 (1), e52898.

Renggli, D., Graf, C., Tachatos, N., Singh, N., Meboldt, M., Taylor, W.R., Stieglitz, L., Schmid Daners, M., 2020. Wearable inertial measurement units for assessing gait in real-world environments. Front. Physiol. 11, 90.

Ruiz-Ruiz, L., Jimenez, A.R., Garcia-Villamil, G., Seco, F., 2021. Detecting fall risk and frailty in elders with inertial motion sensors: a survey of significant gait parameters. Sensors 21 (20), 6918.

Sansgiri, S., Mody, P., Vohlakari, K., Finni, T., Rantalainen, T., Cronin, N.J., 2024. Evaluating the transferability of machine learning models from treadmill walking to overground gait conditions. Available at SSRN 4923857.

Schmitt, A.C., Baudendistel, S.T., Lipat, A.L., White, T.A., Raffegeau, T.E., Hass, C.J., 2021. Walking indoors, outdoors, and on a treadmill: Gait differences in healthy young and older adults. Gait Posture 90, 468–474.

Sharifi Renani, M., Myers, C.A., Zandie, R., Mahoor, M.H., Davidson, B.S., Clary, C.W., 2020. Deep learning in gait parameter prediction for OA and TKA patients wearing IMU sensors. Sensors 20 (19), 5553.

Skvortsov, D., Chindilov, D., Painev, N., Rozov, A., 2023. Heel-strike and toe-off detection algorithm based on deep neural networks using shank-worn inertial sensors for clinical purpose. J. Sensors 2023 (1), 7538611.

Torrey, L., Shavlik, J., 2010. Transfer learning. In: Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques. IGI global, pp. 242–264.

Trabassi, D., Serrao, M., Varrecchia, T., Ranavolo, A., Coppola, G., De Icco, R., Tassorelli, C., Castiglia, S.F., 2022. Machine learning approach to support the detection of parkinson's disease in IMU-based gait analysis. Sensors 22 (10), 3700.

Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (11).

Verlekar, T.T., Correia, P.L., Soares, L.D., 2018. Using transfer learning for classification of gait pathologies. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine. BIBM, IEEE, pp. 2376–2381.

Vicon, 2002. Plug-in-Gait modelling instructions. Vicon manual, Vicon 612 motion systems. Oxford Metrics Ltd., Oxford, UK.

Wang, Y., Li, Z., Wang, X., Yu, H., Liao, W., Arifoglu, D., 2021. Human gait data augmentation and trajectory prediction for lower-limb rehabilitation robot control using GANs and attention mechanism. Machines 9 (12), 367.

Yang, P.K., Filtjens, B., Ginis, P., Goris, M., Nieuwboer, A., Gilat, M., Slaets, P., Vanrumste, B., 2024. Freezing of gait assessment with inertial measurement units and deep learning: effect of tasks, medication states, and stops. J. NeuroEng. Rehabil. 21 (1), 24.

Yu, L., Mei, Q., Xiang, L., Liu, W., Mohamad, N.I., István, B., Fernandez, J., Gu, Y., 2021. Principal component analysis of the running ground reaction forces with different speeds. Front. Bioeng. Biotechnol. 9, 629809.