

Deep Learning-Based Environmental and Terrain Classification for Enhanced Free-Living Fall Risk Assessment Using Wearable Video Data

Daniel Fischer^{1*}, Laura Meier¹, Stefan Koch², Thomas Braun³

¹ Department of Geriatric Medicine and Fall Prevention, Faculty of Medicine, University of Freiburg, Freiburg, Germany.

² Department of Wearable Video Data and Environmental Classification, Faculty of Medicine, Technical University of Munich, Munich, Germany.

³ Department of Free-Living Fall Risk Assessment, Faculty of Medicine, University of Kiel, Kiel, Germany.

*E-mail ✉ daniel.fischer@pharmazie.uni-freiburg.de

Abstract

A better understanding of mobility and gait patterns can substantially improve fall risk assessment. Recent progress in this field has been driven by wearable inertial measurement units (IMUs), which record temporal aspects of movement (such as step time) in both laboratory or clinical environments and during daily activities. Such data contribute to fall risk evaluations grounded in habitual behavior. However, comprehensive interpretation of IMU-derived mobility measures in free-living conditions is currently restricted by a lack of contextual awareness. For example, while variations in stride length can be quantified, the specific reasons for these changes—whether stemming from personal factors or external surroundings—frequently remain undetermined. To address this gap and achieve deeper insight into habitual fall risk via IMU-based metrics, wearable video cameras are proposed as a complementary tool. Nevertheless, manually analyzing video material is time-consuming and labor-intensive, requiring extended viewing and annotation of surroundings while also raising important privacy concerns. Therefore, robust automated artificial intelligence (AI) techniques, trained on varied datasets, are essential for reliable environment classification. The present work introduces a new dataset compiled from online video sources. It presents a deep learning system that employs chained convolutional neural networks to automatically detect environment type (indoor versus outdoor) and terrain characteristics (for example, carpet or grass). This dataset includes 146,624 video-based images (environment: 79,251; floor visible: 28,347; terrain: 39,026). After training, the individual classifiers reached F1-scores of ≥ 0.84 on an independent, manually annotated validation set (environment: 0.98; floor visible indoor: 0.86; floor visible outdoor: 0.96; terrain indoor: 0.84; terrain outdoor: 0.95). When applied to fresh data, the isolated networks achieved accuracies between 51 and 100%, while the full pipeline attained 45–90%. This project is ongoing, with continued optimization of the AI components to increase classification precision and enable automated contextualization of mobility and gait data for fall risk evaluation. Additional work will focus on collecting primary recordings from participants in their natural living environments to improve the dataset's diversity and representativeness.

Keywords: Computer vision, Ambulatory gait analysis, Wearables, Terrain classification, Convolutional neural networks

Introduction

Falls represent a leading contributor to morbidity and mortality among individuals with Parkinson's disease (PD), where approximately 60% of patients experience at

least one fall annually and 39% encounter recurrent falls [1]. A significant shortcoming of existing fall risk assessment methods lies in their limited capacity for real-world monitoring in remote or unrestricted living settings, coupled with the scarcity of objective, person-specific indicators [2]. Digital wearable technologies now offer a modern, individualized, and cost-effective solution for fall risk evaluation by enabling analysis of mobility under free-living conditions. This capability positions wearables as practical instruments for guiding strategies that promote safe and efficient movement to prevent falls [3].

Access this article online

<https://smerpub.com/>

Received: 31 March 2026; Accepted: 25 June 2026

Copyright CC BY-NC-SA 4.0

How to cite this article: Fischer D, Meier L, Koch S, Braun T. Deep Learning-Based Environmental and Terrain Classification for Enhanced Free-Living Fall Risk Assessment Using Wearable Video Data. J Med Sci Interdiscip Res. 2026;6(1):264-78. <https://doi.org/10.51847/TthjEjx3Ok>

Wearable inertial measurement units (IMUs) generally incorporate tri-axial accelerometers and gyroscopes that jointly deliver detailed information on movement and orientation, supporting the acquisition of high-resolution data suitable for in-depth mobility evaluation. When only one inertial sensor is employed, the device is typically designated as either an accelerometer or a gyroscope [4-6]. For instance, IMU signals can be gathered unobtrusively and discreetly [7-9] over extended durations (e.g., 7 days) to derive temporal and spatial gait features, as well as frequency and rotational parameters. These metrics offer valuable insights for fall prediction [10] and detecting gait disturbances such as freezing of gait [11]. At present, lengthy IMU recordings are commonly processed by aggregating all mobility episodes or by restricting analysis to bouts lasting ≥ 1 min or ≥ 2 min [12-14]. Nevertheless, IMUs alone typically fail to capture environmental context (e.g., indoor versus outdoor settings or surface characteristics), which limits the robustness of mobility assessments. Recognizing how environmental elements, particularly terrain, affect gait is especially critical in neurological conditions such as Parkinson's disease, where gait deviations arise from both internal disease-related factors and external influences such as high-friction surfaces. The absence of such contextual details therefore impedes comprehensive fall risk evaluation [15, 16]. Combining IMU data collected during routine mobility with simultaneous environmental context would enable more precise and impactful fall risk assessments. The approach described here supports thorough evaluation by distinguishing intrinsic from extrinsic contributors during everyday activities.

Contextual information can refine fall risk assessment in habitual settings by determining whether elevated instability detected by IMUs stems from intrinsic factors (e.g., basal ganglia dysfunction due to pathology) or extrinsic factors (e.g., uneven terrain). Although prior efforts have attempted to derive context directly from raw IMU signals at the sample level [17], these methods remain constrained in the depth of contextual detail they provide. Similarly, global positioning system (GPS)-enabled wearables have demonstrated value in monitoring outdoor movement [18]. However, this strategy cannot reliably reflect current terrain conditions, as GPS coordinates combined with map overlays (e.g., Google Maps) may be outdated or fail to identify specific hazards that increase fall likelihood. Without wearable cameras, it is not possible to definitively establish

whether fall risk factors are intrinsic or extrinsic [14]. For example, heightened step time variability is frequently observed in PD and may respond to pharmacological treatment [19]. Yet, relying solely on IMU mobility data cannot clarify whether such variability results from (a) an underlying neurological process requiring intervention adjustment (intrinsic) or (b) environmental or terrain characteristics, such as an irregular surface (extrinsic) [20].

Earlier studies involving video capture of mobility in natural environments have primarily relied on stationary cameras [14, 21]. Fixed-camera setups, however, confine data collection to predetermined sites and offer no coverage of atypical or edge-case walking locations [22]. To overcome these constraints, mobile solutions employing waist-mounted cameras oriented toward the ground have been suggested. While effective for terrain assessment, this configuration provides limited broader environmental context and depends on cumbersome, obtrusive equipment [23]. In contrast, first-person view (FPV) wearable cameras have been advocated to supply complete environmental context for mobility across diverse locations. Nonetheless, routine use of camera-based data presents practical difficulties, as researchers must manually review extensive video recordings to annotate environments and identify objects or obstacles. Ethical concerns also arise from capturing sensitive material, including financial documents, children, or private activities. The main benefit of the method proposed in this work is applying automated artificial intelligence (AI) techniques to interpret environmental context from video data. This enhances habitual fall risk assessment while safeguarding participant privacy. For example, AI systems can automatically review footage and remove or delete sensitive or irrelevant frames. Moreover, automated AI methods deliver more efficient contextual descriptions than manual review. This is further supported by comfortable, discreet wearable eye-tracking glasses, which have been used to capture environmental context [24]. Such devices supply precise information without the drawbacks of bulky waist-mounted cameras [23] or potentially imprecise satellite-based GPS data that may lack sufficient detail [9]. Successful deployment of AI in this domain requires training data that accurately reflects habitual conditions. In general, AI models depend on well-annotated ground-truth datasets to automate the interpretation of new inputs. To address these limitations, the current study introduces an innovative strategy that uses data mining

from social media platforms to build a dataset, avoiding the need for burdensome primary data collection from participants. Effective generalization of AI models necessitates training on a wide variety of images under differing conditions. Insufficient or unrepresentative training data can result in errors when models encounter novel scenarios, such as varying illumination. Collecting sufficiently diverse training examples that encompass a wide spectrum of environments and conditions remains a key challenge in creating reliable AI tools for environmental classification in free-living fall risk assessment [23]. Advancing automated contextual analysis for habitual fall risk evaluation requires clarifying how different flooring and terrain types influence fall probability in everyday mobility [25]. This underscores the value of developing AI solutions from foundational principles, encompassing both dataset construction and the evaluation of terrain effects on movement.

One solution is to use existing public datasets; however, identifying datasets that precisely match the required criteria—such as ground-level FPV head-mounted video—is difficult. An alternative is to extract relevant material from social media. These platforms contain vast quantities of video content recorded in everyday settings with diverse lighting and quality characteristics, which can be mined to generate a representative and heterogeneous dataset. This study introduces a biomedical computing application designed to harvest such data to create an automated, AI-based environment classifier. This classifier is intended to complement habitual fall risk assessment derived from IMU mobility data. Specifically, the approach draws upon video-sharing websites that provide thousands of hours of footage, including property tours, village walks, and urban explorations, many of which are available under CC-BY licensing. Overall, this methodology constitutes a meaningful advance in fall risk assessment by integrating data mining and artificial intelligence applied to FPV videos. It delivers definitive environmental context and automates terrain classification.

The following sections outline the development of (1) a data curation and mining tool and (2) a visualization and classification tool founded on multiple deep neural networks. These networks address distinct aspects of environmental recognition, including (a) differentiation between indoor and outdoor scenes and (b) identification of terrain types.

Program description

Creating novel datasets often depends on manual procedures that involve a time-intensive sequence of acquiring, downloading, and labeling or classifying video material and individual frames. Labeling is commonly performed by saving images into specifically named folders—for instance, placing frames depicting wood into a dedicated “wood” folder. Although essential for establishing the foundation of subsequent AI model development, this workflow is notably labor-intensive.

To address these challenges, the initial objective (Section 2.1) was to design an efficient pipeline by developing a specialized ground truth curation and annotation tool (GT-CAT). This tool accelerates video retrieval, downloading, and labeling processes. GT-CAT extracts individual frames from videos and uses a semi-automated mechanism to organize them into appropriate folders, thereby constructing a training dataset. The second objective centers on a visualization module that supports an automated AI framework. This framework leverages the prepared training dataset to classify novel, previously unseen video recordings (Section 2.2).

Data curation tool: Collection and annotation

The GT-CAT facilitates efficient video acquisition and annotation to support the classification of first-person view (FPV) terrain video frames. Although various annotation tools serve similar purposes, the unique requirements of transforming social media videos into terrain-classified frames justified creating this dedicated tool. Many existing solutions assume frames have already been extracted and uploaded, adding a manual step of segmenting videos into individual images. By contrast, GT-CAT accepts a web link, downloads the corresponding video, extracts all frames, and automatically crops the lower central portion of each image, assuming this region contains relevant terrain information. While some manual oversight remains necessary—such as assigning terrain images to their respective categories (e.g., wood, carpet, tile)—the tool substantially streamlines and accelerates data acquisition and annotation. This provides the AI system with high-quality examples for constructing a robust training dataset. All images incorporated into the training set were captured at an approximate height of 1.5 m, consistent with typical handheld camera use. To increase diversity, the dataset was further supplemented with frames captured from varying heights and perspectives.

Collection

The GT-CAT was developed using a Python-based representational state transfer application programming interface (REST API), along with hypertext markup language (HTML), cascading style sheets (CSS), and JavaScript. Upon receiving a valid social media link, the Python-based GT-CAT downloads the video to a designated local directory and records associated metadata (including title, channel name, and link) in a .CSV file. Next, an additional script processes the video using OpenCV's video capture functionality to extract individual frames, which are saved into an organized

folder structure for subsequent examination and labeling. A description of the GT-CAT is provided below (**Figures 1 and 2**):

- A text input field allows users to enter a social media link (**Figure 2a**), enabling straightforward pasting and confirmation. The tool delivers progressive feedback on processing status, such as notifications indicating completion of video cropping and frame extraction.
- Upon processing, the interface displays all extracted video frames for review and supports multi-selection to efficiently assign terrain labels (**Figure 2b**).

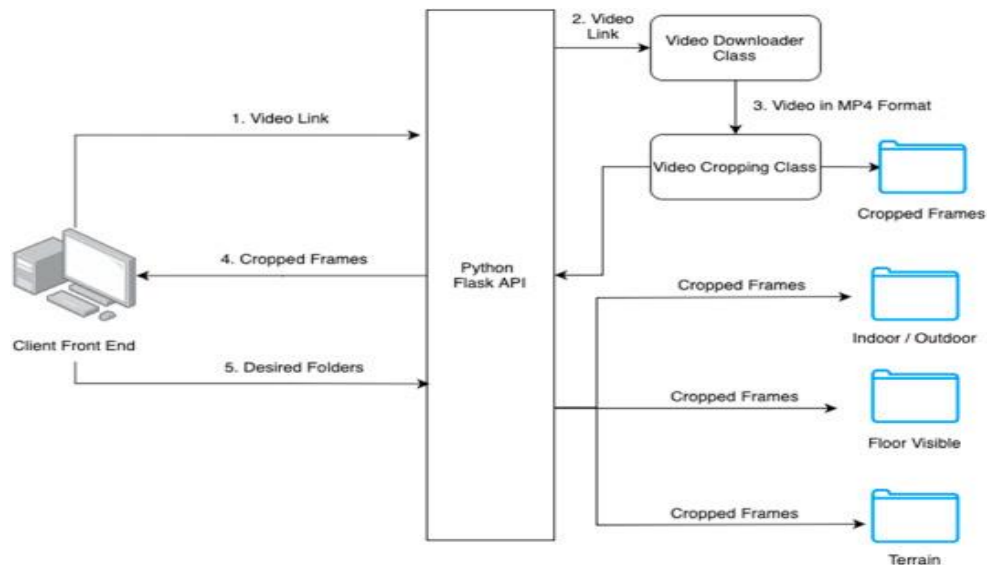
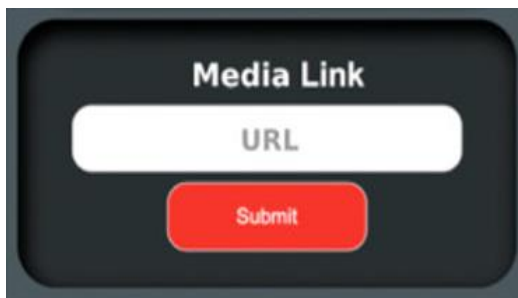
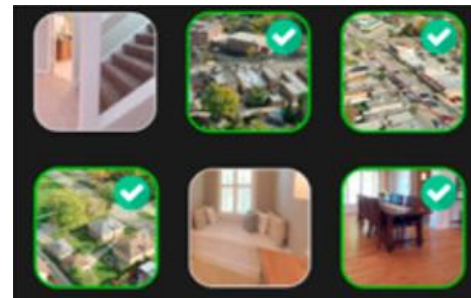


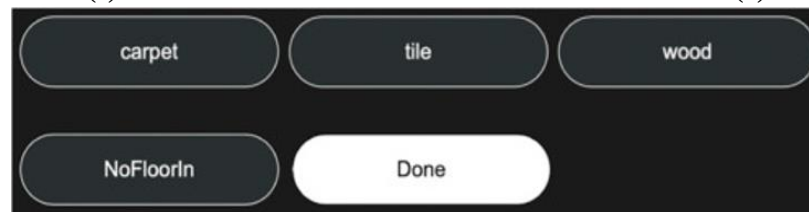
Figure 1. Architecture of the dataset curation tool based on a REST API.



(a)



(b)



(c)

Figure 2. Screenshot of the newly developed training data collection tool used for categorizing frames extracted from a YouTube video. (a) Input interface for specifying the target YouTube video link; (b) multi-selection interface for labeling multiple frames; (c) bottom toolbar enabling selection of visible terrain types for the chosen frames.

Annotation

After extracting and storing individual video frames, a 500×500 pixel (px) crop is taken from the bottom-center region of each image and saved. These cropped frames are then manually assigned to the appropriate terrain-specific folders (e.g., wood, tile). The procedure is as follows:

- The user reviews and selects images that clearly depict the target terrain, excluding any frames lacking visible terrain. The user then assigns the selected images to the correct terrain category via options presented on an action bar (**Figure 2c**). Once a terrain type is chosen, the GT-CAT automatically relocates the images to the corresponding pre-designated folder. In cases where multiple terrain classes appear within a single frame, priority is given to the most prominent class.
- An exception applies to the environment classifier (indoor/outdoor). For this particular model, the GT-CAT was not employed, as compiling its training data proved considerably less demanding. Unlike the other classifiers, the environment classifier operates on full-resolution

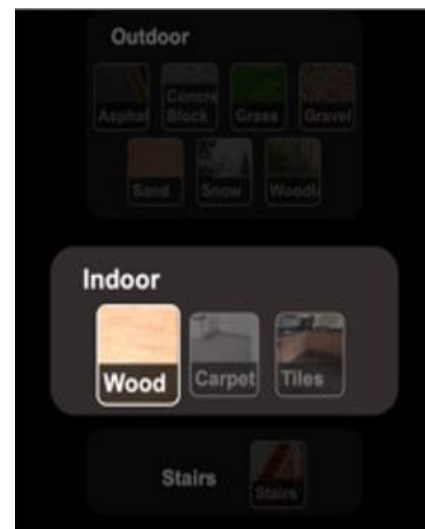
images (1280×720 px) rather than 500×500 px crops, eliminating the need for individual frame cropping. Utilizing complete images also simplifies categorization, as the overall scene content is immediately apparent and limited to two distinct classes (indoor or outdoor).

Visualization and classifier tool

A web-based visualization and classification tool was developed to demonstrate practical implementation (**Figure 3**) and establish proof of concept for clinical application. This interface allows users to interact with the model by uploading a video, which is then played back alongside a sidebar that displays the predicted terrain classification in real time. The underlying system architecture resides in the same directory as the Python-based REST API component (**Figure 4**). It processes the video by evaluating every 30 frames through the complete classification pipeline, determining both environment and terrain categories for each frame. Upon completion, the API returns classification results for all frames to the client interface in JSON format.



(a)



(b)

Figure 3. Example of the testing interface showing the classification toolbar (right side) that updates every 30 video frames with the current prediction. (a) Video playback of a novel, unseen indoor setting and (b) the classifier accurately recognizing the indoor environment along with a wood terrain or floor surface.

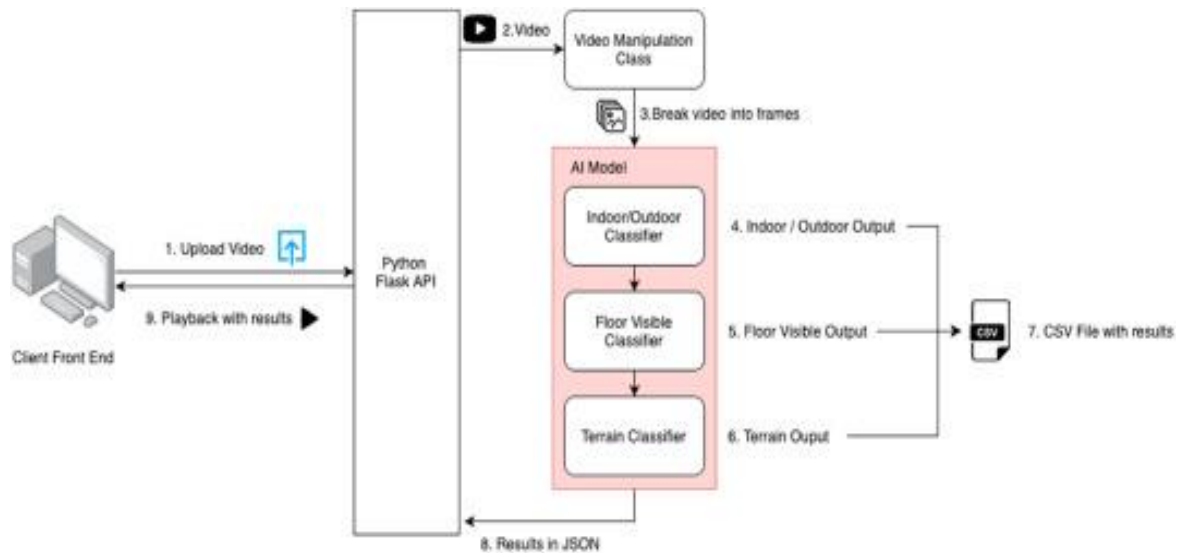


Figure 4. REST API architecture of the visualization and classification tool.

Computational methods

Preparing ground truth training datasets

Developing a robust training dataset required considerable diversity for each of the five classifiers (#1 to #5) that form the integrated system architecture (**Figure 5**). Each classifier was designed to handle variations in lighting conditions, surface textures, floor obstructions, and geometric patterns. Distinct collection strategies were therefore implemented to ensure adequate heterogeneity for each component:

- #1 Indoor and outdoor: In total, 20,000 images were gathered. The outdoor portion consisted of 10,000 frames extracted from extended videos of hiking paths, urban strolls, and village explorations recorded across different daylight hours and illumination levels, with an average video length of approximately 60 minutes. Indoor data collection was more difficult because generic searches yielded shorter clips that often featured promotional overlays obscuring the floor area. Nonetheless, we successfully assembled 10,000 images from office tours, property viewings, historic site recordings, and university residence walkthroughs, spanning multiple sites, times of day, and lighting conditions.

- #2 (indoor) and #3 (outdoor) floor visible: Images originally prepared for classifier #1 were repurposed for these models. A fresh dataset was necessary to represent both situations with clearly observable terrain and those without any visible floor. Utilizing the wide environmental range already present in classifier #1, frames were re-labeled according to terrain visibility, creating the categories “floor visible” and “floor not visible.” The identical approach was applied to generate the outdoor-specific floor visibility model (#3). The resulting frames for classifiers #2 and #3 were then reused as the basis for the subsequent terrain classifiers (#4 and #5).

- #4 (indoor) and #5 (outdoor) terrain: Every 500×500 px crop previously identified as containing visible floor underwent further subdivision according to the specific surface type shown (e.g., asphalt, grass). Constructing a balanced and varied dataset proved difficult because certain indoor categories, such as wood and carpet, appeared far more frequently than others, such as tiles. This imbalance was addressed by incorporating additional real estate videos from areas known for higher prevalence of tiled flooring, notably the United States.

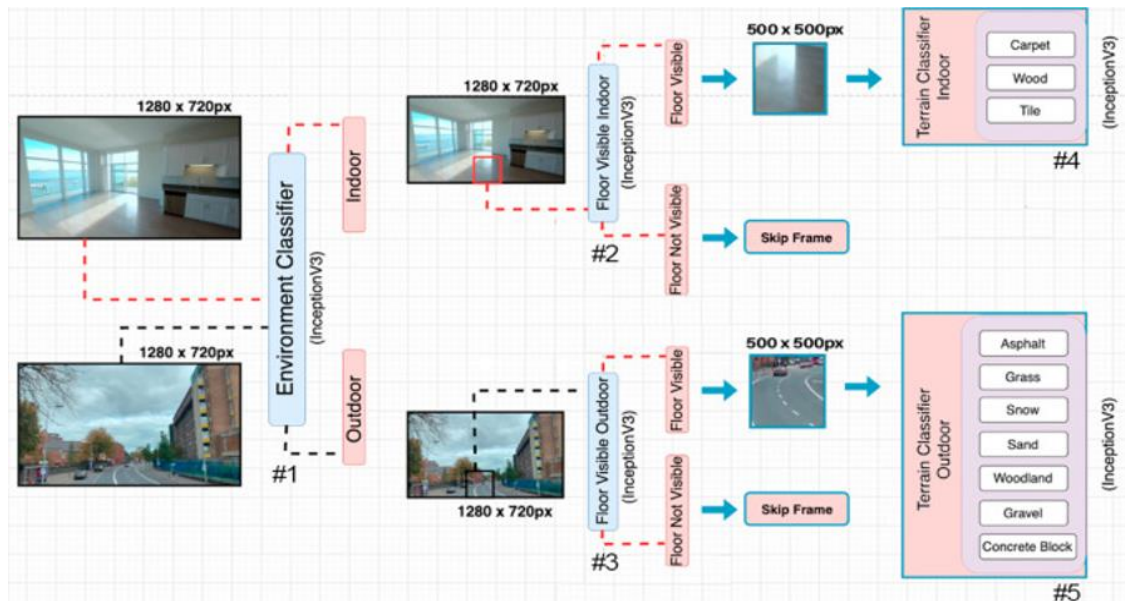


Figure 5. Full system architecture illustrating the sequential flow of classifiers #1 to #5. (#1) Evaluates a 1280×720 image to distinguish between indoor and outdoor scenes. (#2–#3) Examine the 500×500 central crop to determine terrain visibility and skip processing when absent. (#4–#5) Determine the precise terrain category within the 500×500 central crop.

AI methodology

The developed framework integrates a series of chained two-dimensional (2D) convolutional neural networks (CNNs). It incorporates five distinct deep learning models (#1 to #5), all implemented in Python 3.8.5 with the TensorFlow and Keras libraries (**Figure 5**). These CNN classifiers (**Figures 5a–5e**) sequentially determine the overall environment (indoor versus outdoor) before proceeding to detailed terrain recognition (e.g., wood, carpet, tile). In each case, we applied transfer learning, starting from ImageNet pre-trained weights and adding custom layers tailored to the task. The stepwise classification procedure is as follows:

- Classifier #1 receives input images sized $299 \times 299 \times 3$. It uses a CNN backbone with transfer learning, replacing the original final layer with a dense layer of 1024 neurons with ReLU activation, followed by a dropout layer (rate 0.2) and a single sigmoid-activated output neuron. Following environment classification, a 500×500 px central crop (starting at coordinates 0, 390) is extracted and supplied to the floor visibility model.
- Classifiers #2 and #3 share a comparable CNN design, including a 1024-neuron ReLU hidden layer, a dropout layer (rate 0.2), a second 1024-neuron ReLU layer, another dropout layer (rate 0.2), and a final softmax output layer with three neurons. When the floor visibility assessment determines that no terrain is present, the system skips terrain classification and maintains the most

recent terrain prediction. Frames confirmed to show visible floor advance to the relevant terrain classifier (#4 for indoor or #5 for outdoor).

- Classifiers #4 and #5 maintain an architecture nearly identical to #2 and #3, differing primarily in the number of output neurons to match the respective number of target classes. These models classify multiple indoor surfaces (e.g., carpet, wood, tile) and outdoor surfaces (e.g., asphalt, grass, sand). Terrain types were selected according to their frequency of occurrence in the sourced videos; categories with negligible or zero representation were excluded from the classification task.

Choosing the optimal CNN architecture

A variety of CNN architectures exist for image classification, each presenting different levels of computational demand and performance characteristics. Selection was guided by established, high-performing contemporary models, specifically InceptionV3 [26], InceptionResNetV2 [27], VGG-16 [28], and MobileNet [29]. These architectures were systematically assessed and compared using accuracy metrics derived from the training datasets to identify the most effective option. InceptionV3 consistently delivered the highest accuracy levels (**Table 1**) for environment classification, floor visibility detection, and terrain recognition. **Table 2** presents the outcomes of hyperparameter optimization

trials involving different numbers of hidden layers and learning rates. The optimal configuration combined a learning rate of 0.0001 with a single hidden layer. Although a learning rate of 0.0001 paired with two hidden layers yielded equivalent accuracy, the reduced

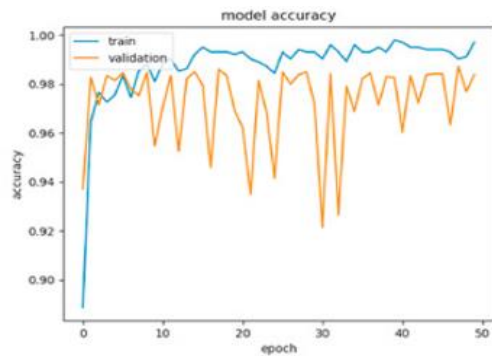
computational load associated with one hidden layer was preferred for overall efficiency. The number of training epochs was determined according to observed convergence patterns (**Figure 6**).

Table 1. Classification accuracies achieved by alternative architectures for (a) environment, (b) floor visibility, and (c) terrain tasks.

Deep learning architecture	Training epochs	Classifier #1Environment	Classifier #2Indoor Terrain	Classifier #3Outdoor Terrain	Classifier #4Indoor Floor Visibility	Classifier #5Outdoor Floor Visibility
InceptionV3	350	99%	91%	97%	91%	97%
InceptionResNetV2	350	99%	88%	97%	90%	97%
MobileNetV2	350	98%	87%	55%	89%	80%
VGG16	350	96%	78%	95%	84%	96%
ResNet-50	350	85%	42%	64%	44%	79%

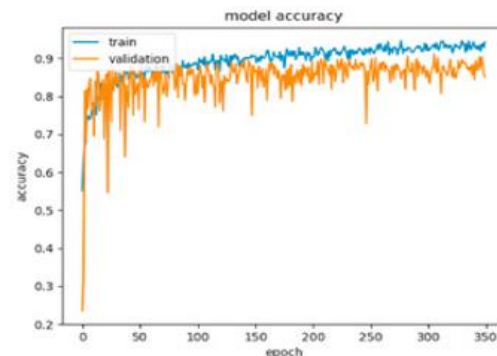
Table 2. Results of hyperparameter optimization for (a) environment, (b) floor visible, and (c) terrain classifiers.

Hidden layers	Learning rate	Epochs	Environment	Floor visible	Terrain
1	0.0001	50	99%	89%	84%
2	0.0001	50	99%	89%	84%
1	0.01	50	98%	75%	50%
2	0.01	50	97%	73%	43%
1	0.2	50	98%	70%	45%
2	0.2	50	98%	69%	40%



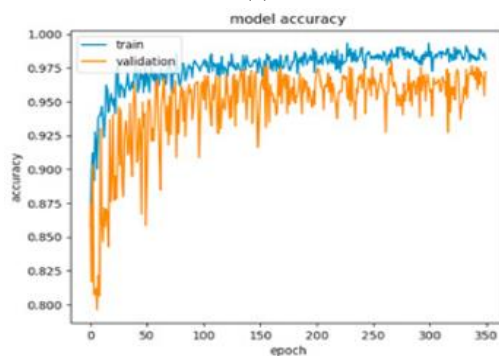
Environment Classifier (Indoor/Outdoor)

(a)

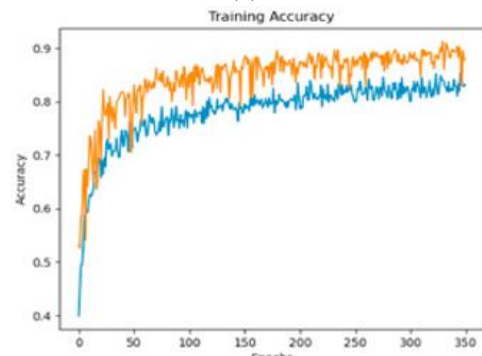


Indoor Floor Visible Classifier

(b)



Outdoor Floor Visible Classifier



Indoor Terrain Classifier

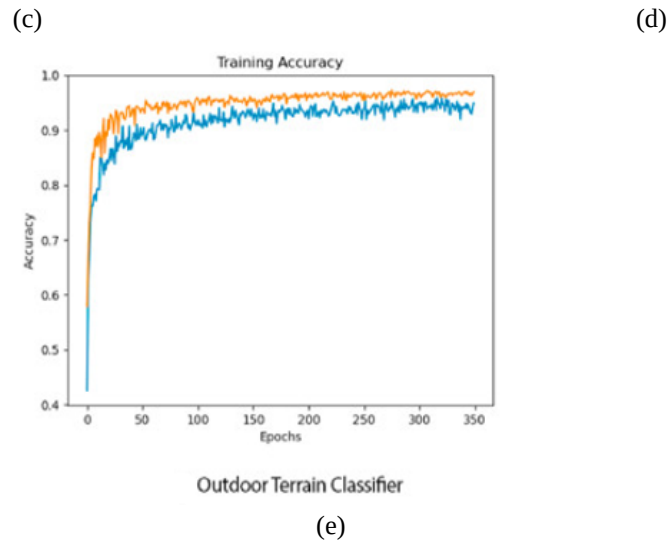


Figure 6. Convergence behavior of the trained models.

Training and preliminary testing

The selected CNN architecture (InceptionV3) was trained using the custom dataset assembled through the newly developed data curation tool from publicly available online videos. To balance computational efficiency with reliable performance, the system processes input videos by classifying every 30 frames. This approach accelerates analysis while maintaining effectiveness, given that substantial environmental shifts are unlikely to occur within such short intervals. Since the source videos were recorded at 30 frames per second (fps), this strategy results in classification of each second of footage.

The completed model underwent initial evaluation by assessing each constituent classifier against a preliminary validation dataset. A research team member recorded this dataset during a 10-minute walkthrough that included both indoor building spaces and outdoor campus areas.

Statistical approach

Model performance was validated by computing several key statistical measures. Precision was first determined to quantify the proportion of correct positive predictions. Recall (sensitivity) was then calculated, and both values were combined to derive the F1-score. This metric represents the harmonic mean of precision and recall, offering a balanced indicator of each model's overall effectiveness.

Program run

Manual labeling and contextual interpretation of first-person view (FPV) video recordings is highly time-

consuming. In benchmark evaluations conducted by the research team, a 10-minute video could be annotated manually (for example, using pen and paper) in roughly 20 minutes. In comparison, (a) employing only the data curation tool reduced annotation time to 10 minutes, while (b) full integration of the AI system further decreased this to approximately 5 minutes. On average, the proposed AI solution processed videos in roughly half their actual duration (55% of running time). Upon completion, the system generates a .CSV file containing the classification outcomes.

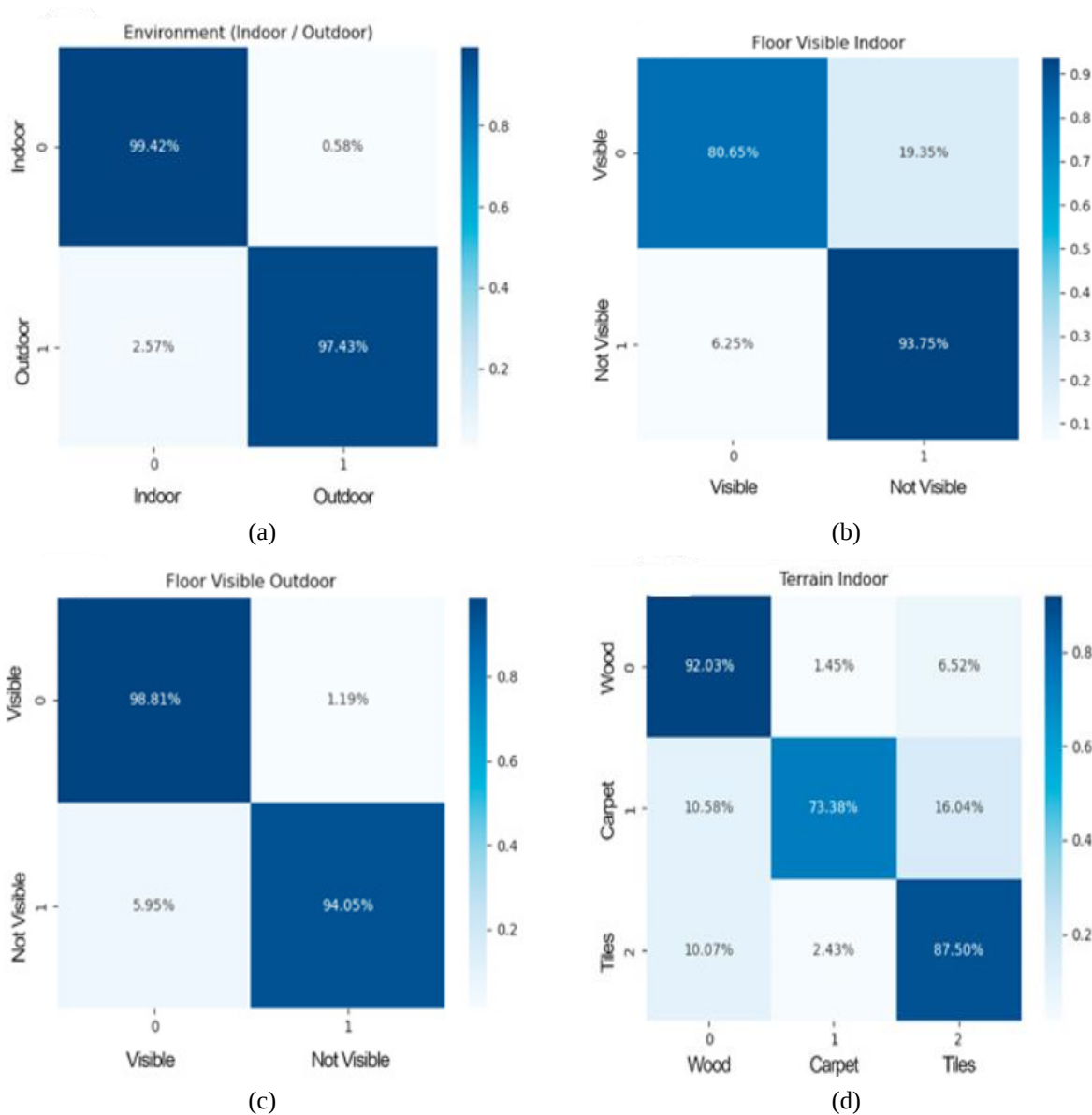
Classification

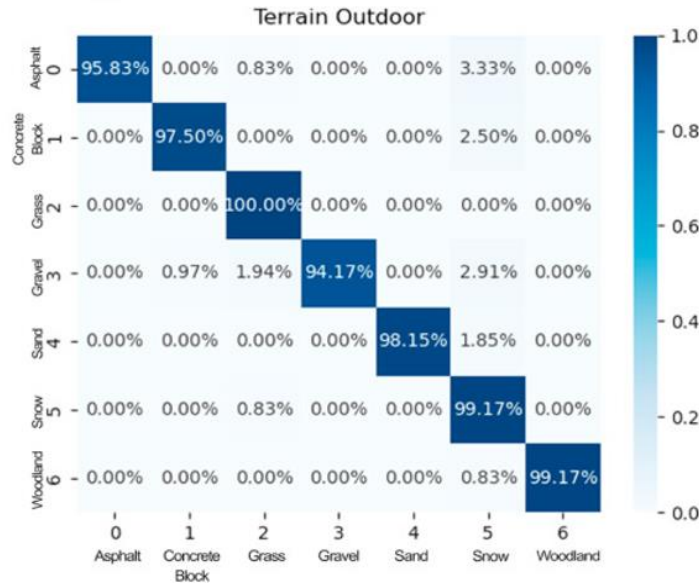
Evaluation on the validation dataset (**Table 3**) indicates that the developed AI framework offers a promising and accurate solution for terrain classification in video footage across varied environments. All models attained F1-scores of ≥ 0.84 when assessed on a held-out subset of previously unseen data, derived from an 80:20 training-validation split of the original dataset (**Figure 7**). The environment classifier (**Figure 7a**) demonstrated the strongest performance, reaching an F1-score of 0.98. This superior result is likely attributable to the extensive training data available (28,000 images spanning diverse scenes). The floor visibility models achieved F1-scores of 0.86 for the indoor variant (**Figure 7b**) and 0.96 for the outdoor variant (**Figure 7c**). The indoor floor visibility classifier proved most demanding, primarily due to a comparatively limited training set. The terrain classification models (**Figure 7d indoor and Figure 7e outdoor**) exhibited a similar pattern, with the outdoor

model attaining notably higher accuracy, again reflecting the benefit of richer and more varied training examples.

Table 3. Classification accuracy results.

Classification model	Precision	Sensitivity (Recall)	F1-measure
#1 Environmental Classification	0.97	0.99	0.98
#2 Indoor Floor Visibility Detection	0.92	0.81	0.86
#3 Outdoor Floor Visibility Detection	0.94	0.98	0.96
#4 Indoor Terrain Classification	0.84	0.84	0.84
#5 Outdoor Terrain Classification	0.95	0.95	0.95





(e)

Figure 7. Confusion matrices for (a) the indoor/outdoor classifier, (b) the indoor floor visibility classifier, (c) the outdoor floor visibility classifier, (d) the indoor terrain classifier, and (e) the outdoor terrain classifier.

Preliminary testing: New/unseen data

Initial evaluation of the AI system utilized recordings obtained in a university environment. Data were collected from (a) an older adult (OA: male, 72 years) who had experienced a stroke and (b) a younger adult (YA: male, 24 years). Ethical approval was obtained

from the Northumbria University Research Ethics Committee (REF: 21,603), and all participants provided informed written consent before participation. Participants wore two different cameras (OA: 720p; YA: 1080p) while performing brief 30-second walks in both indoor and outdoor settings (**Figure 8**).



(a)



(b)

Figure 8. Examples of challenging edge cases. Left: Although the participant is walking on a narrow path, a substantial portion of the frame shows grass, resulting in misclassification as grass. Right: As the participant approaches a revolving door, glass with significant glare obscures the actual terrain, causing the system to classify the frame as having no visible terrain.

The recorded environments featured multiple terrain types (e.g., grass, asphalt, concrete paving, and indoor tiles). Preliminary testing showed varying accuracy,

highlighting the difficulty of developing a universally robust terrain classifier when using footage from different camera specifications. While performance

remained strong on higher-quality recordings (YA), more demanding scenarios involving lower camera resolution posed significant challenges, as evidenced in **Table 4** (OA video 2). Similar limitations have been noted in

earlier studies employing waist-mounted cameras [23]. **Table 4** summarises classifier results when evaluated separately and as an integrated pipeline (**Figure 7**).

Table 4. Performance metrics across different 30-second video segments of previously unseen data.

Participant	Recording	Standalone CNN evaluation	
		Environment	Floor visibility
OA (720p)	Clip #1	100%	83%
	Clip #2	100%	93%
	Clip #3	97%	52%
YA (1080p)	Clip #1	100%	98%
	Clip #2	100%	98%
	Clip #3	100%	97%

Results and Discussion

This investigation established a new dataset by systematically extracting online video footage with a specialized custom annotation instrument. The study examined deep learning strategies based on sequential CNN architectures to enable automatic recognition of environmental settings (indoor versus outdoor) and surface types (for example, carpet or grass). Cross-validation results were favorable, and assessments on previously unseen data revealed specific challenging scenarios that merit additional scrutiny.

General environment: Indoor and outdoor

Construction of the indoor/outdoor classification model (**Figure 5a**), based on the InceptionV3 CNN framework, proved successful even with a comparatively limited set of full-resolution images captured in relatively straightforward scenes. This outcome highlights both the intrinsic strength of the selected architecture and the potential for achieving cutting-edge performance in basic indoor/outdoor scene differentiation once supplied with expanded training datasets. This is evident in **Figure 6e**, which shows the model reaching an F1-score of 0.98 on the validation set after only 10 epochs of training. Beyond the strong results of the indoor/outdoor classifier, greater volumes of training imagery are anticipated to boost F1-scores across the other models. In addition, the floor visibility and terrain classifiers show substantial performance gains with higher-resolution inputs. By comparison, models trained exclusively on lower-resolution 720p material showed unstable outcomes due to frequent edge cases. In contrast, higher-quality 1080p footage consistently achieved accuracies in the range of

70–90%, even when sequences included numerous edge cases such as motion-blurred frames.

Floor visible

The floor visibility model (**Figure 5b**) was specifically created to detect and isolate video frames that contain no discernible terrain surface. These frames typically arise in confined indoor spaces such as bathrooms or kitchens, where the body-worn camera (positioned at chest or waist level) records only overhead elements like cabinet tops or bathtub rims. Such non-informative frames compromise the precision of downstream analyses and unnecessarily increase processing demands and overall classification time. By deploying a dedicated model to filter out these irrelevant segments, the approach lowers computational load—since frames no longer need to pass through the more demanding terrain classifier—and improves the reliability of the entire pipeline. Consequently, the AI system gains improved capacity to differentiate between comparable indoor surfaces (such as tile versus textured carpet) by leveraging surrounding contextual cues, including the presence of objects like toilets or bathtubs.

Terrain

At present, the terrain classification model (**Figure 5c**) remains the most demanding component for improving F1-scores, largely because its multiclass design inherently adds layers of complexity. Evaluation on the validation set requires the model to handle highly variable images with irregular terrain perspectives caused by sudden body movements, motion blur, or off-center positioning. In contrast, the floor visibility task constitutes a simpler binary decision. Restricting

classification to the floor area within each frame by excluding peripheral regions proved crucial for boosting accuracy, as it prevents extraneous objects or obstacles from influencing predictions. Without this step, the model might, for instance, incorrectly assign a tile label when it detects toilet-like shapes elsewhere in the image.

Mode of availability of software

The software is openly available at no cost and generates environmental-context classifications from video inputs. These outputs can support more informed contextual gait analysis, particularly when integrated with IMU sensor recordings. Further details on access are provided in the Code availability section.

Plans

This work aims to strengthen diagnostic procedures and clinical decision-making related to fall risk evaluation. The proposed methodology and accompanying online tool are designed for broad applicability across diverse wearable camera systems mounted at various body sites, thereby complementing IMU-derived mobility metrics. Nevertheless, such adaptations introduce distinct technical difficulties. A notable example is head-mounted cameras, where natural head rotations can move the intended terrain out of view. Planned extensions may combine accelerometer signals to quantify head orientation relative to the desired downward-facing angle, ensuring that video capture remains focused on foot-terrain interaction zones. Preliminary strategies for addressing terrain occlusion from the expected central viewpoint have already been explored using deep learning-based segmentation techniques, as reported earlier [25]. This strategy isolates the relevant floor area in each frame via a binary mask, permitting targeted classification independent of the floor's location within the image.

Future efforts to improve fall risk assessment will draw on complementary methods reported in related literature. One relevant example involves motion encoding through stacked spectrograms computed across successive time windows [30]. That framework applied chained CNNs for motion feature learning, combined with an LSTM network to handle sequential video streams, yielding robust classification of multiple daily activities (such as walking, running, and stair climbing) from head-mounted first-person perspective cameras. Stair ascent detection is especially pertinent, since earlier investigations have established stair negotiation as a key

extrinsic contributor to elevated fall risk, notably in people with Parkinson's disease [31, 32]. Waist-mounted downward-facing cameras also hold promise for documenting everyday gait and ambulation patterns in real-world free-living conditions. Such data could enable automated, markerless measurement of clinically relevant gait parameters in PD [12]. Recent work in this area introduced two complementary CNN models for ground-facing, waist-worn cameras [33]. The initial model (FootRegionProposer) identifies candidate foot regions through bounding boxes to facilitate trajectory and contact-point analysis [33]. A subsequent model (LocomoNet) then refines these proposals by suppressing false positives, which commonly arise when multiple feet appear simultaneously [33]. Advances of this nature are progressively enabling purely vision-based quantification of free-living gait without IMU dependency, although they continue to benefit from robust automated contextual understanding. The next phase of the current research will prioritize integrating object detection algorithms to recognize potential hazards and obstacles along a participant's route. Longer-term objectives will increasingly converge with ongoing efforts to develop AI systems for precise, automated extraction of detailed gait characteristics.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Allen NE, Schwarzel AK, Canning CG. Recurrent falls in Parkinson's disease: a systematic review. *Parkinson's Dis.* 2013;2013.
2. Mirelman A, Bonato P, Camicioli R, Ellis TD, Giladi N, Hamilton JL, et al. Gait impairments in Parkinson's disease. *Lancet Neurol.* 2019;18(7):697-708.
3. Dasgupta P, VanSwearingen J, Godfrey A, Redfern M, Montero-Odasso M, Sejdić E. Acceleration gait measures as proxies for motor skill of walking: a narrative review. *IEEE Trans Neural Syst Rehabil Eng.* 2020;29:249-61.

4. Jarchi D, Pope J, Lee TK, Tamjidi L, Mirzaei A, Sanei S. A review on accelerometry-based gait analysis and emerging clinical applications. *IEEE Rev Biomed Eng.* 2018;11:177-94.
5. Godfrey A, Conway R, Meagher D, ÓLaighin G. Direct measurement of human movement by accelerometry. *Med Eng Phys.* 2008;30(10):1364-86.
6. Hutabarat Y, Owaki D, Hayashibe M. Quantitative gait assessment with feature-rich diversity using two IMU sensors. *IEEE Trans Med Robotics Bionics.* 2020;2(4):639-48.
7. Del Din S, Kirk C, Yarnall AJ, Rochester L, Hausdorff JM. Body-worn sensors for remote monitoring of Parkinson's disease motor symptoms: vision, state of the art, and challenges ahead. *J Parkinsons Dis.* 2021;11(s1):S35-47.
8. Micó-Amigo ME, Bonci T, Paraschiv-Ionescu A, Ullrich M, Kirk C, Soltani A, et al. Assessing real-world gait with digital technology? Validation, insights and recommendations from the Mobilise-D consortium. *J NeuroEng Rehabil.* 2023;20(1):1-26.
9. Mazzà C, Di Marco R, Paraschiv-Ionescu A, Caruso M, Caulfield B, Cereatti A, et al. Technical validation of real-world monitoring of gait: a multicentric observational study. *BMJ Open.* 2021;11(12):e050785.
10. Lord S, Galna B, Rochester L. Moving forward on gait measurement: toward a more refined approach. *Mov Disord.* 2013;28(11):1534-43.
11. Del Din S, Godfrey A, Mazzà C, Lord S, Rochester L. Free-living monitoring of Parkinson's disease: lessons from the field. *Mov Disord.* 2016;31(9):1293-313.
12. Del Din S, Godfrey A, Galna B, Lord S, Rochester L. Free-living gait characteristics in ageing and Parkinson's disease: impact of environment and ambulatory bout length. *J NeuroEng Rehabil.* 2016;13(1):1-12.
13. Van Schooten KS, Pijnappels M, Rispens SM, Elders PJ, Lips P, van Dieën JH. Ambulatory fall-risk assessment: amount and quality of daily-life gait predict falls in older adults. *J Gerontol A Biol Sci Med Sci.* 2015;70(5):608-15.
14. Weiss A, Brozgol M, Dorfman M, Herman T, Shema S, Giladi N, et al. Does the evaluation of gait quality during daily life provide insight into fall risk? A novel approach using 3-day accelerometer recordings. *Neurorehabil Neural Repair.* 2013;27(8):742-52.
15. Nouredanesh M, Godfrey A, Howcroft J, Lemaire ED, Tung J. Fall risk assessment in the wild: a critical examination of wearable sensor use in free-living conditions. *Gait Posture.* 2021;85:178-90.
16. Hickey A, Stuart S, O'Donovan K, Godfrey A. Walk on the wild side: the complexity of free-living mobility assessment. *J Epidemiol Community Health.* 2017.
17. Hashmi MZUH, Riaz Q, Hussain M, Shahzad M. What lies beneath one's feet? Terrain classification using inertial data of human walk. *Appl Sci.* 2019;9(15):3099.
18. Klous G, Smit LAM, Borlée F, Kromhout H, Heederik D, Huss A. Mobility assessment of a rural population in The Netherlands using GPS measurements. *Int J Health Geogr.* 2017;16(1):1-13.
19. Hausdorff JM, Cudkowicz ME, Firtion R, Wei JY, Goldberger AL. Gait variability and basal ganglia disorders: stride-to-stride variations of gait cycle timing in Parkinson's disease and Huntington's disease. *Mov Disord.* 1998;13(3):428-37.
20. Hickey A, Del Din S, Rochester L, Godfrey A. Detecting free-living steps and walking bouts: validating an algorithm for macro gait analysis. *Physiol Meas.* 2016;38(1):N1.
21. Banerjee T, Yefimova M, Keller JM, Skubic M, Woods DL, Rantz M. Exploratory analysis of older adults' sedentary behavior in the primary living area using Kinect depth data. *J Ambient Intell Smart Environ.* 2017;9(2):163-79.
22. Godfrey A. Wearables for independent living in older adults: gait and falls. *Maturitas.* 2017;100:16-26.
23. Nouredanesh M, Godfrey A, Powell D, Tung J. Egocentric vision-based detection of surfaces: towards context-aware free-living digital biomarkers for gait and fall risk assessment. *J NeuroEng Rehabil.* 2022;19(1):79.
24. Moore J, Stuart S, McMeekin P, Walker R, Celik Y, Godfrey A, et al. Enhancing free-living fall risk assessment: contextualizing mobility based IMU data. *Sensors.* 2023;23(2):891.
25. Moore J, Stuart S, Walker R, McMeekin P, Young F, Godfrey A. Deep learning semantic segmentation for indoor terrain extraction: toward better informing free-living wearable gait assessment. In: 2022 IEEE-EMBS International Conference on Wearable and

- Implantable Body Sensor Networks (BSN); 2022; Ioannina, Greece. Piscataway (NJ): IEEE; 2022. p. 13-6.
26. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016; p. 2818-26.
 27. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-First AAAI Conference on Artificial Intelligence; 2017.
 28. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556; 2014.
 29. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, et al. Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861; 2017.
 30. Abebe G, Cavallaro A. A long short-term memory convolutional neural network for first-person vision activity recognition. In: Proceedings of the IEEE International Conference on Computer Vision Workshops; 2017; p. 1339-46.
 31. Nilsson MH, Hariz GM, Iwarsson S, Hagell P. Walking ability is a major contributor to fear of falling in people with Parkinson's disease: implications for rehabilitation. *Parkinson's Dis.* 2012;2012.
 32. Mactier K, Lord S, Godfrey A, Burn D, Rochester L. The relationship between real world ambulatory activity and falls in incident Parkinson's disease: influence of classification scheme. *Parkinsonism Relat Disord.* 2015;21(3):236-42.
 33. Nouredanesh M, Li AW, Godfrey A, Hoey J, Tung J. Chasing feet in the wild: a proposed egocentric motion-aware gait assessment tool. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops; 2018.