Discriminating Between Indoor and Outdoor Environments During Daily Living Activities Using Local Magnetic Field Characteristics and Machine Learning Techniques

Original
Discriminating Between Indoor and Outdoor Environments During Daily Living Activities Using Local Magnetic Field Characteristics and Machine Learning Techniques / Marcianò, V., Cereatti, A., Bertuletti, S., Bonci, T., Alcock, L., Gazit, E., Ireson, N., Bevilacqua, A., Mazzà, C., Ciravegna, F., Del Din, S., Hausdorff, J.M., Ifrim, G., Caulfield, B., Caruso, M.. - In: IEEE SENSORS JOURNAL. - ISSN 1530-437X. - 25:1(2025), pp. 1507-1515. [10.1109/jsen.2024.3490253]

Availability:
This version is available at: 11583/2996406 since: 2025-01-09T05:42:22Z

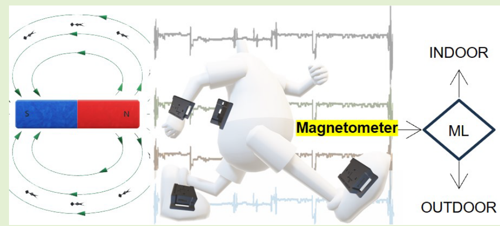
Publisher:
IEEE

Published
DOI:10.1109/jsen.2024.3490253

IEEE Sensors Council IEEE SENSORS JOURNAL, VOL. 25, NO. 1, 1 JANUARY 2025

Discriminating Between Indoor and Outdoor Environments During Daily Living Activities Using Local Magnetic Field Characteristics and Machine Learning Techniques

Vincenzo Marcianò, Andrea Cereatti, Stefano Bertuletti, Tecla Bonci, Lisa Alcock, Eran Gazit, Neil Ireson, Antonio Bevilacqua, Claudia Mazzà, Fabio Ciravegna, Silvia Del Din, *Member, IEEE*, Jeffrey M. Hausdorff, Georgiana Ifrim, Brian Caulfield, and Marco Caruso, *Member, IEEE*



***Abstract*—Wearable technology has rapidly advanced, opening new possibilities for context-aware applications in fields such as healthcare and gait analysis, where distinguishing between indoor and outdoor environments is essential. This is often accomplished through technologies like GPS, Wi-Fi, cellular, and Bluetooth which, however, come with privacy concerns, high power consumption, and dependency on external infrastructure. To address these challenges, recent studies have preliminary exploited the ambient magnetic field, though comprehensive validation with real-life data is lacking. This article seeks to validate machine learning techniques, i.e., random forest (RF), extreme gradient boosting, and stacked long short-term memory (LSTM) networks, for indoor–outdoor discrimination using exclusively magnetometer data from the daily activities of 20 participants in four cities across three countries. The study investigated the most effective magnetometer placement (feet, lower back, and nondominant wrist) and pre-processing techniques (e.g., features and window size). Reference data are obtained through a GPS-based algorithm coupled with a geographical database running on a smartphone. The extreme gradient boosting algorithm yielded the best results, with an accuracy of 0.91, an $F1$-score of 0.90, and an area under the ROC curve of 0.94. These findings confirm the feasibility of accurately estimating indoor/outdoor context information from a magnetometer at the same update frequency as a common GPS, but with important energy savings. The proposed model can be integrated into state-of-the-art gait analysis systems, being able to discriminate the location to avoid misinterpreting gait deviations in real-world settings, thus supporting continuous and ubiquitous gait monitoring. Datasets and algorithm implementations have been made publicly available.**

***Index Terms*—Gait analysis, indoor–outdoor detection, machine learning, magnetic field, magneto-inertial measurement units, magnetometer, time series classification, wearable sensors.**

Received 17 October 2024; accepted 30 October 2024. Date of publication 14 November 2024; date of current version 2 January 2025. This work was supported by the Mobilise-D project that has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No. 820820. This JU receives support from the European Union's Horizon 2020 research and innovation program and the European Federation of Pharmaceutical Industries and Associations (EFPIA). Silvia Del Din was also supported by the Innovative Medicines Initiative 2 Joint Undertaking (IMI2 JU) project IDEA-FAST - Grant Agreement 853981. Lisa Alcock and Silvia Del Din were also supported by the National Institute for Health Research (NIHR) Newcastle Biomedical Research Centre (BRC) based at The Newcastle upon Tyne Hospital National Health Service (NHS) Foundation Trust, Newcastle University and the Cumbria, Northumberland and Tyne and Wear (CNTW) NHS Foundation Trust. Lisa Alcock and Silvia Del Din were also supported by the NIHR/Wellcome Trust Clinical Research Facility (CRF) infrastructure at Newcastle upon Tyne Hospitals NHS Foundation Trust. This study was also supported by the National Institute for Health Research (NIHR) through the Sheffield Biomedical Research Centre (BRC, Grant Number IS-BRC-1215-20017). All opinions are those of the authors and not the funders. The content in this publication reflects the authors' view, and neither IMI nor the European Union, EFPIA, NHS, NIHR, DHSC, or any associated partners are responsible for any use that may be made of the information contained herein. The associate editor coordinating the review of this article and approving it for publication was Prof. Huang Chen Lee. *(Corresponding author: Marco Caruso.)*

Please see the Acknowledgment section of this paper for the author affiliations.

Digital Object Identifier 10.1109/JSEN.2024.3490253

I. Introduction

A WIDE range of applications, such as healthcare [1], [2], [3], [4], [5], fitness tracking [6], [7], [8], personalized advertising [9], pedestrian navigation [10], and environmental monitoring [11], rely on precise indoor–outdoor differentiation. For instance, in healthcare, the detection of a patient's transitions from indoor to outdoor allows clinicians to monitor their well-being, functional capacity, and safety (e.g., fall risk and falls) more effectively [1], [2], [3], [4], [5]. Outdoor environments may involve more variable terrains, such as uneven ground, inclines, or changes in surface type, all of which can alter gait patterns [12]. In contrast, indoor environments often feature smoother, more predictable surfaces. Without this contextual awareness, gait analysis models could misinterpret changes in gait parameters as potential pathological issues, when in fact they might simply result from environmental factors. In fitness tracking, knowing whether an individual is exercising and walking indoors or outdoors can influence the calculation of metrics such as distance, speed, and elevation, providing more accurate and personalized insights [6], [7], [8]. As a result, the ability to accurately detect whether a person is located indoors or outdoors has become increasingly significant with the rapid growth of mobile computing and location-based services [13], [14], [15], [16].

Existing approaches to indoor–outdoor detection have primarily relied on GPS [17], GNSS [18], Wi-Fi [19], cellular [20], light [21], or Bluetooth [22]. Often, they have relied on combining GPS data with a geographical database, such as an open street map, to improve the accuracy in distinguishing between indoor and outdoor environments [23]. The main advantage of the latter approach is that it can be implemented on a smartphone, it is versatile, and has applications in diverse domains. Conversely, other limitations exist such as jeopardizing privacy, high-power consumption, and dependency on external infrastructure or mapping surveys.

Recently, hall effect magnetometers, sensors capable of measuring the local magnetic field, have gained attention as a promising alternative for indoor–outdoor detection [24], [25]. Based on the idea that indoor and outdoor environments have a different “magnetic signature,” a magnetometer’s sensitivity to significant variations of the magnetic field patterns, ferromagnetic objects, electronic equipment, and building characteristics can be exploited for differentiating between indoor and outdoor environments and detecting transitions. Additional advantages include no privacy-related concerns, their small size, lightweight, low-energy consumption, and robustness against environmental disturbances such as weather conditions. Conveniently, many wearable inertial devices employed for clinical digital mobility already come equipped with a magnetometer.

A preliminary investigation into the use of ambient magnetic field data has already been conducted. Li et al. [25] developed an indoor–outdoor detector, based on the analysis of variance in magnetometer signals recorded by a smartphone. This method, based on an empirically determined threshold, was tested across 133 diverse environments, including outdoor areas (e.g., football fields, squares), semi-outdoor areas (e.g., pathways near buildings), and indoor spaces (e.g., offices, shopping malls). The accuracy amounted to approximately 78% in distinguishing indoor and semi-outdoor spaces from outdoor environments. However, no information on the number of participants was provided. As declared, the accuracy of the method is evidently dependent on the specific threshold values used, highlighting its sensitivity to these parameters. Ashraf et al. [24] recorded magnetic data with a smartphone in three different locations. An accuracy of around 85% was achieved using decision tree classifiers but no information about the number of subjects or the duration was provided, making a direct comparison difficult. In addition, experiments were conducted in limited, non-real-world conditions, which introduces reproducibility issues. Galván Tejada et al. [26] focuses on estimating specific indoor locations by mapping a building’s magnetic field “fingerprint.” This means that the performance of the classifiers directly depends on the specific environment.

Although these preliminary studies have demonstrated encouraging performance, validity was only partially investigated. The first goal of this study was to assess the influence of the device placement on the specific body segments. In fact, while it is known that magnetometers placed at different points in space may have different readings, depending on the local distortion of the Earth’s magnetic field [27], it is hypothesized that the performance of the indoor/outdoor detectors could be influenced by the specific sensor location but this aspect has not yet been explored. Second, this study assessed the generalization of the method performance across different geographical, urban environments and people’s habits, as there is a lack of information on these factors. Finally, this study compared the performance of simple machine learning techniques with more complex deep learning models, assessing how hyperparameter tuning can improve accuracy. Data were collected from four wearable devices attached to the left foot, right foot, lower back, and nondominant wrist of 20 participants across four sites in Europe and Asia. The reference data was supplied through a GPS-based algorithm operating on a smartphone carried by the participant. To promote transparency and enable future comparisons, we have shared both datasets and algorithm implementations. To the best of our knowledge, this is the first study to make machine and deep learning models in this field publicly available. At this stage, however, the absence of comparable models remains a significant limitation, preventing direct benchmarking against other studies.

II. Materials and Methods

A. Data Collection

A subset of the data acquired during the Mobilize-D technical validation study [28] was used, selected according to GPS data availability. This dataset comprised data from 20 participants, for a total of approximately 47 h of real-world data recording. It included both young and elderly healthy adults ($n = 3$ YHA and $n = 7$ HA, respectively) and participants with potentially altered mobility due to Parkinson’s disease ($n = 4$, PD), multiple sclerosis ($n = 4$, MS), and chronic obstructive pulmonary disease ($n = 2$, COPD). Participants were recruited from four sites: The Newcastle upon Tyne Hospitals NHS Foundation Trust, U.K., and Sheffield Teaching Hospitals NHS Foundation Trust, U.K. (ethics approval granted by the London – Bloomsbury Research Ethics committee, 19/LO/1507), Tel Aviv Sourasky Medical Center, Israel (ethics approval granted by the Helsinki Committee, Tel Aviv Sourasky Medical Center, Tel Aviv, Israel, 0551-19TLV), and University of Sassari, Sassari, Italy (19/LO/1507). Each participant provided written informed consent to participate in the study. Each participant was acquired for 2.5 h during their daily living activities in habitual environments (e.g., home, work, shops, etc.) without supervision. During this session, participants were asked to follow their usual routine while incorporating specific tasks to introduce variability and capture key data elements including outdoor walking, moving between rooms, walking up and down stairs, and walking on a ramp. Additional details about the experimental setup and protocol can be found in Mazzà et al. [28] and Scott et al. [29]. Each participant was equipped with four magneto-inertial measurement units (MIMUs) (INDIP system, University of Sassari,

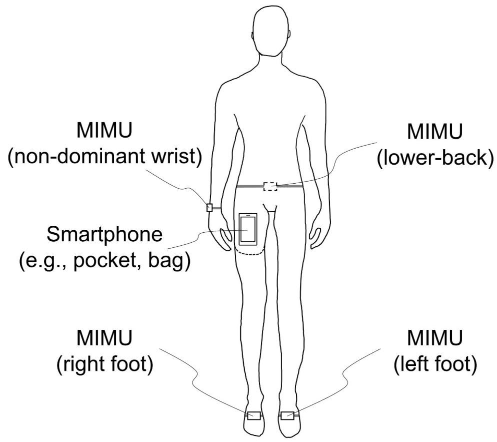


Fig. 1. Experimental setup.

Italy) [30], one on each foot, one on the lower back (LB), and one on the nondominant wrist (WR) (Fig. 1). Each MIMU integrated a triaxial accelerometer, a triaxial gyroscope, and a triaxial magnetometer (full scale: ±50 Gauss) capturing data at a sampling frequency of 100 Hz. In addition, each participant was equipped with a smartphone (model Samsung S9 – Fig. 1) to provide the GPS reference.

B. Magnetometer-Based Machine Learning Classification Models

To perform binary classification between indoor and outdoor environments, three supervised machine learning models were implemented and compared. Specifically, two statistical methods, XGBoost (XGB) and random forest (RF), were selected, alongside the more complex long short-term memory (LSTM) networks, which are based on deep recurrent neural networks.

1) *RF:* It is widely used for both classification and regression tasks [31], [32]. RF makes predictions by combining multiple decision trees during training, each based on a random subset of the data and features. For classification, each tree "votes" on the predicted class, and the majority vote determines the final prediction. This ensemble approach helps to avoid overfitting and to improve generalization, making RF more robust to noise and outliers in the data. Parameters: max depth = 4; max # of features = 50%; # of estimators = 200; bootstrap sampling = activated.
2) *XGB:* Chosen for its high performance and scalability and built-in features, XGB is an advanced implementation of gradient boosting [33]. XGB operates by sequentially adding weak learners, typically decision trees, where each new tree corrects the errors of its predecessors. XGB's novel tree-building algorithm is particularly efficient in handling sparse data, a common challenge in machine learning. Furthermore, it incorporates regularization techniques to avoid overfitting, which is a frequent issue in gradient-boosting models. Additional features, such as missing value handling and feature importance analysis, further enhance its utility for data analysis and prediction. Parameters: set of features' subsample = 10%; min child weight = 1; max depth = 2; inner learning objective = binary logistic classification.
3) Stacked LSTM: LSTMs are a type of recurrent neural network designed to capture long-term dependencies in sequential data [34], [35]. Unlike traditional neural networks, LSTMs have a memory cell that can maintain information for long periods, along with gates that regulate the flow of information. This makes them particularly suitable for time series data. In a stacked LSTM, multiple LSTM layers are placed on top of each other, allowing the model to learn hierarchical features at different time scales. Fig. 2 shows the architecture implemented to manage data recorded by the MIMUs. Parameters: # of RNN layers = 2; hidden RNN features = 256; linear layer features = 256; Dropout = 0.

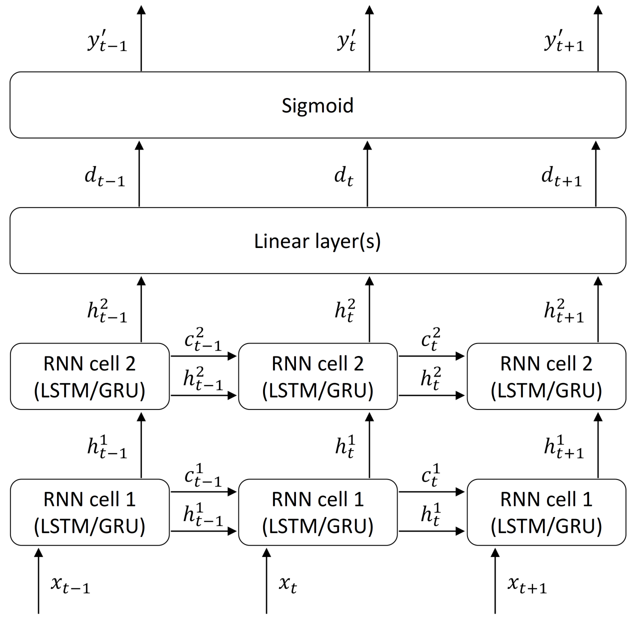


Fig. 2. Architecture of the stacked recurrent neural networks where x_i is the input signal, c_i is the cell state of the ith LSTM layer, h_i is the hidden state of the ith LSTM layer, d_i is the signal processed by the LSTM linear layer into 256 features, and y'_i is the final prediction.

C. Ground Truth Data Provided by GPS-Based Algorithm

The raw GPS data were acquired through the Aeqora App (Department of Computer Science, The University of Sheffield) which is capable of detecting GPS data at a frequency of 1 reading every 10 s or 10 m of distance. The app also captures other sensor readings, including step count, activity recognition, and more. The data were then aggregated into high-level ground truth labels characterizing the participants' activity context, such as trajectories and stay points. The stay points were identified as the centroids of spatial and temporal clusters where GPS data are located within a predefined radius of 20 m for over 2 min (or in the absence of data, considering that the app did not report the location within a radius of 10 m from the previous location). A multitier processing framework based on OpenStreetMap [36] allowed for the classification of

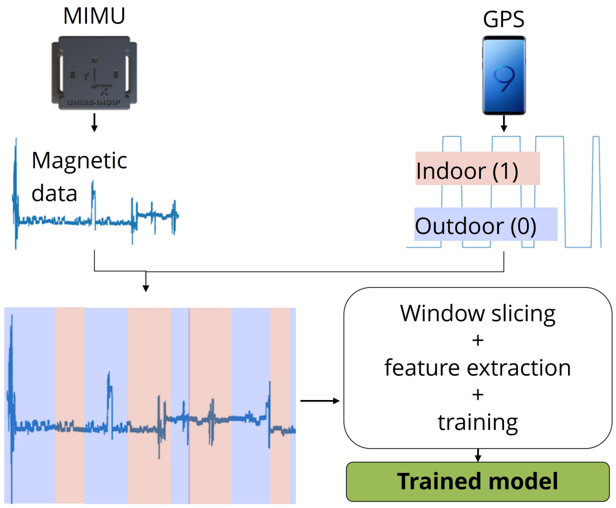


Fig. 3. Schematic view of the preprocessing pipeline.

the stay points as indoor or outdoor. Activities carried out within the predefined radius from indoors were also classified as indoor, to account for GPS imprecision. Activities carried out outside this range (and in areas not identified as indoors by OpenStreetMap) were considered outdoor activities. The result of aggregating GPS coordinates with trajectories and stay points is a set of binary labels produced at a frequency of 1 Hz, describing the context of the corresponding MIMU signal. Each label represents the continuous probability of that activity being performed indoors (1) or outdoors (0).

D. Data Pre-Processing

Data acquired by MIMUs were aggregated and synchronized on a per-second basis with the one provided by the AEQORA app (indoor/outdoor label) based on their common Unix epoch time previously set using a laptop.

The complete pre-processing pipeline is graphically described in Fig. 3.

1) Signal Handling: Triaxial local magnetic field data and their Euclidean norm from all MIMUs locations were obtained and concatenated resulting in a vector consisting of 16 features per each sampling interval. To provide context about the behavior of the features over time, a dedicated window is applied to the entire dataset. This produces a new dataset that can be viewed as a time series, with a dimensionality of N, F, T, where N represents the number of data points, F the number of features considered, and T corresponds to the window size (1 and 2 s). Specifically, when all four MIMUs are considered, F is set to 16, while when each single MIMU is considered separately F is equal to 4.

2) Feature Engineering: Due to the variability and complexity of the local magnetic field, additional time and frequency features with respect to those considered by Ashraf et al. [15] were considered for a more in-depth exploration (Table I).

3) Dataset Organization: Based on existing [37], a 60/40 ratio for the indoor-to-outdoor percentage was selected within the participants' pool as part of the binary classification methodology. This careful adjustment of the dataset's composition ensures sufficient representation of indoor and outdoor cases, thereby amplifying the model's capacity to accurately recognize and categorize patterns in various environmental settings. Fifteen participants were included in the training set (approximately 30 h of recorded data), while the remaining five were assigned to the testing set (approximately 17 h of recorded data). Table II provides details of the dataset organization.

TABLE I
LIST OF EXTRACTED FEATURES

State-of-the-art [24]	Additional
Mean, median, variance, mean absolute deviation, kurtosis, skewness, percentiles (1, 10, 25, 50, 75, 99), interquartile range, trimmed mean (10%).	Root mean square, sensor correlation, first and second dominant frequency, power at the first and second dominant frequency.

TABLE II
DETAILS FOR TRAINING (TR), VALIDATION (V), AND TESTING SETS (T)

Set	Cohort	#	Height (std) m	Weight (std) kg	Center
TR + V	YHA	1	1.83	67	Sassari
	HA	4	1.66 (0.11)	71 (11)	Newcastle
		2	1.66 (0.11)	71 (11)	Tel Aviv
	COPD	1	1.70 (0.08)	76 (5)	Newcastle
		1	1.70 (0.08)	76 (5)	Sheffield
	MS	2	1.68 (0.09)	80 (30)	Sheffield
		1	1.68 (0.09)	80 (30)	Tel Aviv
	PD	2	1.76 (0.09)	84 (17)	Newcastle
		1	1.76 (0.09)	84 (17)	Tel Aviv
T	YHA	2	1.84 (0.01)	76 (13)	Sassari
	HA	1	1.85	86	Newcastle
	MS	1	1.69	60	Sheffield
	PD	1	1.75	70	Tel Aviv

4) Dimensionality Reduction: Prior to using the feature space for training, validating, and testing the machine learning models, a PCA was employed to identify the most informative features thus reducing memory consumption and computational complexity. This approach aimed to enhance the performance of both RF and the XGB algorithms while reducing the high dimensionality brought by the time–frequency feature space. The explained variance ratio is used to select the most appropriate number of PCA dimensions. For this work, the heuristic value was set to 0.85, resulting in a total of 12 features. Preliminary tests explored the use of other dimensionality reduction techniques such as t-distributed stochastic neighbor embedding (t-SNE) and uniform manifold approximation and projection (UMAP). The PCA demonstrated superior performance in terms of stability and computational efficiency. Specifically, we found t-SNE to be computationally expensive, requiring several hours to compute the compressed space map for a single subject, making it less practical for our application. Conversely, while UMAP offered a more efficient computation time, it resulted in a 10%–15% reduction in accuracy across all the sensors. Overall, PCA was found to be the most suitable choice for this study.

5) Training and Testing: The proposed classification approach's performance was validated using the leave one subject out according to the scheme reported in Fig. 4. This approach allowed us to evaluate the methodology's inter-subject generalization capability. The correct

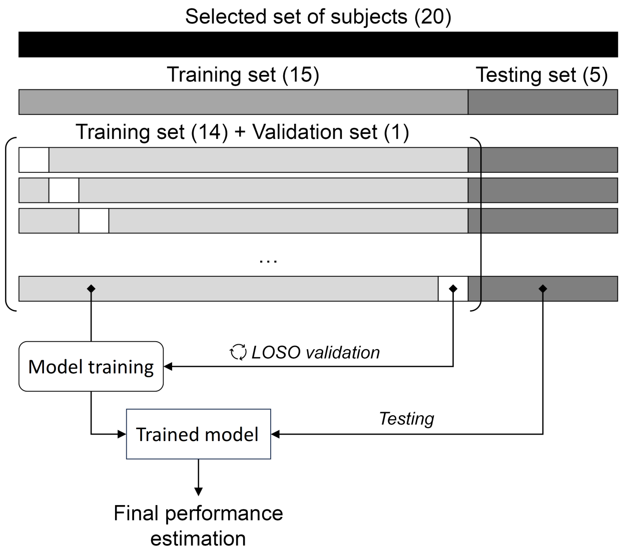


Fig. 4. Training, validation, and testing pipeline. LOSO = leave one subject out.

classification accuracy resulting from leaving one subject out of cross-validation was used to assess the hyperparameter tuning, evaluated using confusion matrices. The training process for a stacked LSTM involves optimizing a binary cross-entropy loss function using back-propagation through time to update the weights of the LSTM layers and the fully connected layers. The training, validation, and testing pipeline was validated offline on a laptop running Python, with the CUDA environment available for training the stacked LSTM. It was equipped with a 4 GB NVIDIA 1650 Ti for faster training inference.

6) Performance Metrics: To characterize the model's performance, accuracy, $F1$-score, and AUC ROC were computed. In particular, the $F1$-score was used to take account of both the false positive and false negative rates of the model. AUC ROC was calculated to provide a single score that summarizes the method's ability to distinguish between indoor/outdoor across all possible classification thresholds. The average and standard deviation values of each metric were calculated across participants for each model (i.e., RF, XGB, and stacked LSTM), device placement (i.e., LB, WR, and foot ensemble as the average value of left and right foot), and window size (i.e., 1 and 2 s). In addition, the same metrics while considering data from full configuration (feet + LB + WR) were also computed for each model. The performance was also evaluated in terms of inference time when using a processor Intel Core i7-9750H (6C/12T, 2.6/4.5 GHz, 12 MB).

III. Results

The distributions of the time series of the normalized magnetic norm are reported for each center, device placement, and conditions (indoor/outdoor) by means of a boxplot representation in 5. The percentage of time spent indoors/outdoors for each center amount to 65%, 73%, 73%, and 100%, respectively. Table III reports the model performance across different participants, including average and standard deviation values for accuracy, $F1$-score, and AUC ROC. These results are organized by device placement (foot ensemble, LB, WR, and full configuration) and window size (1 and 2 s). Results in terms of inference speed indicate that while both XGB and RF have an inference time of approximately 47 ms per one-second window, the LSTM model requires around 120 ms for the same duration. The bi-LSTM architecture composed roughly 930 000 parameters.

IV. Discussion

This article aimed to discriminate indoor from outdoor environments during activities of daily living by means of machine learning algorithms based only on local magnetic fields recorded by magnetometers. To overcome the limitations of previous datasets, in this study, data were collected from healthy and pathological subjects in different cities and environments while carrying out their daily life activities. This resulted in a heterogeneous dataset, as highlighted by the magnetic norm phenomena across the centers and device placement depicted in Fig. 5. It is possible to assess that these distributions greatly differed depending on the indoor/outdoor location, all other factors being equal.

Reported results showed that a single MIMU positioned on the foot can be proficiently used as input for a machine learning model (XGB – 1s window: accuracy = 0.91; $F1$-score = 0.90; AUC ROC = 0.94). Reduced performances were obtained when considering the LB (RF – 1s window: accuracy = 0.88; $F1$-score = 0.88; AUC ROC = 0.89) and full configurations (RF – 1s window: accuracy = 0.88; $F1$-score = 0.88; AUC ROC = 0.88; XGB – 1s window: accuracy = 0.87; $F1$-score = 0.88; AUC ROC = 0.89). The latter result indicates that redundancy of MIMUs does not necessarily result in better performance. The worst performances were obtained for the wrist placement (RF – 2s window: accuracy = 0.83; $F1$-score = 0.82; AUC ROC = 0.80). However, even in this scenario, the LSTM (2s window: accuracy = 0.67; $F1$-score = 0.66; AUC ROC = 0.61) was outperformed by RF and XGB (2s window: accuracy = 0.83; $F1$-score = 0.82; AUC ROC = 0.79). From Table III, it can be observed that a one-second window size yields slightly better results compared to a 2s window.

Discriminating between indoor and outdoor environments during gait can greatly benefit from ferromagnetic interference. However, the nondominant wrist placement may detect disturbances that are unrelated to the indoor–outdoor environments, such as those from a smartphone. In contrast, placing the sensor on the foot provides unique data due to its proximity to ferromagnetic materials, like underfloor pipes or electric cables, commonly incorporated into floors in residential buildings. This explains the poorer results obtained from the wrist configuration compared to the foot placement, which exhibited the best performance. It is important to note that the aim of this work is to contextualize gait analysis. Therefore, a brief misclassification during the transition between indoor and outdoor and vice versa (e.g., near building walls) does not pose a significant issue.

The use of PCA for XGB and RF models led to better performance with respect to those without PCA only when considering the full (both RF and XGB) and feet (only XGB)

TABLE III
OVERALL PERFORMANCES OF THE MODELS, ACROSS SUBJECTS, IN TERMS OF ACCURACY, $F1$-SCORE, AND AUROC, FOR EACH SENSOR PLACEMENT (FEET, LB, WR, AND FULL CONFIGURATION) AND WINDOW SIZE

Window Size	Device placement	Model	PCA	Accuracy (std)	F1-score (std)	AUC ROC (std)
1 second	Foot	RF	without PCA	0.90 (0.08)	0.88 (0.07)	0.94 (0.03)
			with PCA	0.88 (0.09)	0.87 (0.06)	0.93 (0.02)
		XGB	without PCA	0.89 (0.05)	0.89 (0.05)	0.92 (0.04)
			with PCA	**0.91 (0.05)**	**0.90 (0.04)**	**0.94 (0.05)**
		Stacked LSTM	-	0.81 (0.15)	0.80 (0.14)	0.83 (0.10)
	LB	RF	without PCA	0.88 (0.10)	0.87 (0.11)	0.91 (0.08)
			with PCA	0.81 (0.14)	0.80 (0.14)	0.87 (0.11)
		XGB	without PCA	0.86 (0.11)	0.87 (0.06)	0.90 (0.06)
			with PCA	0.66 (0.15)	0.65 (0.26)	0.71 (0.24)
		Stacked LSTM	-	0.26 (0.22)	0.25 (0.19)	0.30 (0.26)
	WR	RF	without PCA	0.80 (0.11)	0.79 (0.13)	0.83 (0.11)
			with PCA	0.73 (0.18)	0.72 (0.21)	0.77 (0.20)
		XGB	without PCA	0.78 (0.11)	0.77 (0.13)	0.81 (0.12)
			with PCA	0.81 (0.15)	0.80 (0.15)	0.84 (0.14)
		Stacked LSTM	-	0.70 (0.10)	0.69 (0.12)	0.73 (0.14)
	Full	RF	without PCA	0.78 (0.10)	0.77 (0.08)	0.82 (0.06)
		RF	with PCA	0.79 (0.05)	0.78 (0.07)	0.83 (0.04)
		XGB	without PCA	0.87 (0.06)	0.86 (0.08)	0.90 (0.07)
			with PCA	0.89 (0.07)	0.88 (0.08)	0.92 (0.05)
		Stacked LSTM	-	0.76 (0.10)	0.75 (0.16)	0.79 (0.20)
2 seconds	Foot	RF	without PCA	0.88 (0.07)	0.87 (0.06)	0.93 (0.03)
			with PCA	0.87 (0.08)	0.86 (0.06)	0.92 (0.02)
		XGB	without PCA	0.89 (0.05)	0.88 (0.05)	0.92 (0.04)
			with PCA	0.90 (0.05)	0.89 (0.04)	0.93 (0.05)
		Stacked LSTM	-	0.80 (0.15)	0.79 (0.14)	0.82 (0.10)
	LB	RF	without PCA	0.85 (0.11)	0.84 (0.14)	0.87 (0.08)
			with PCA	0.80 (0.17)	0.79 (0.17)	0.85 (0.11)
		XGB	without PCA	0.85 (0.06)	0.84 (0.05)	0.87 (0.06)
			with PCA	0.65 (0.26)	0.64 (0.24)	0.68 (0.25)
		Stacked LSTM	-	0.25 (0.26)	0.24 (0.23)	0.28 (0.27)
	WR	RF	without PCA	0.78 (0.12)	0.77 (0.11)	0.81 (0.10)
			with PCA	0.72 (0.11)	0.71 (0.10)	0.75 (0.12)
		XGB	without PCA	0.76 (0.08)	0.75 (0.08)	0.79 (0.09)
			with PCA	0.79 (0.14)	0.78 (0.14)	0.82 (0.14)
		Stacked LSTM	-	0.68 (0.10)	0.67 (0.10)	0.71 (0.14)
	Full	RF	without PCA	0.75 (0.08)	0.74 (0.07)	0.78 (0.06)
			with PCA	0.77 (0.07)	0.76 (0.07)	0.80 (0.05)
		XGB	without PCA	0.85 (0.05)	0.84 (0.05)	0.88 (0.05)
			with PCA	0.87 (0.05)	0.86 (0.05)	0.90 (0.05)
		Stacked LSTM	-	0.72 (0.19)	0.71 (0.20)	0.75 (0.20)

configurations. On the contrary, reduced performance was obtained for the LB and WR configurations.

The results reported in this work outperformed those from relevant similar studies based on magnetometer data only: Ashraf and colleagues [24] achieved an accuracy of 85.3%, while the approach proposed by Zhou et al. [25] was able to achieve an accuracy around 80%. This can be explained by considering that our study included a larger and more comprehensive dataset, with a total of 47 h of everyday activity data recorded in four cities with distinct characteristics (e.g., metropolitan versus industrial city) across different countries and cohorts of patients.

For all device placements, the obtained performance indicated that simpler machine learning models such as XGB and RF outperformed a multilayer and multichannel LSTM by up to 10% in terms of accuracy, highlighting that in this case simplicity can be more effective, despite the processing cost associated with handcrafting and manually extracting features for XGB and RF. These results imply that further investigation into more complex and deep learning methods (such as Transformers) may not be necessary, at least when a low-computational cost is a primary requirement.

The use of inertial signals to achieve discrimination between indoor and outdoor has been explored in the literature, although, to the best of our knowledge there are few studies using only the accelerometer and gyroscope (e.g., [38], [39]). Unfortunately, one important issue that emerges from these studies is the lack of shared machine/deep learning models which does not allow a mutual comparison with the results of this work. In [38], RF and AdaBoost algorithms were trained on six activities, recorded from healthy subjects. Although the model achieved over 97% accuracy, this may be due to the distinct and clearly distinguishable kinematic patterns of the activities, raising concerns about limited generalization to real-world scenarios. In fact, some activities, like walking, can produce similar inertial data indoors and outdoors, making it challenging to distinguish between environments using inertial data alone, such as long indoor corridors and straight outdoor roads. Additionally, the study relied on IMU orientation which, however, is not directly provided by all the

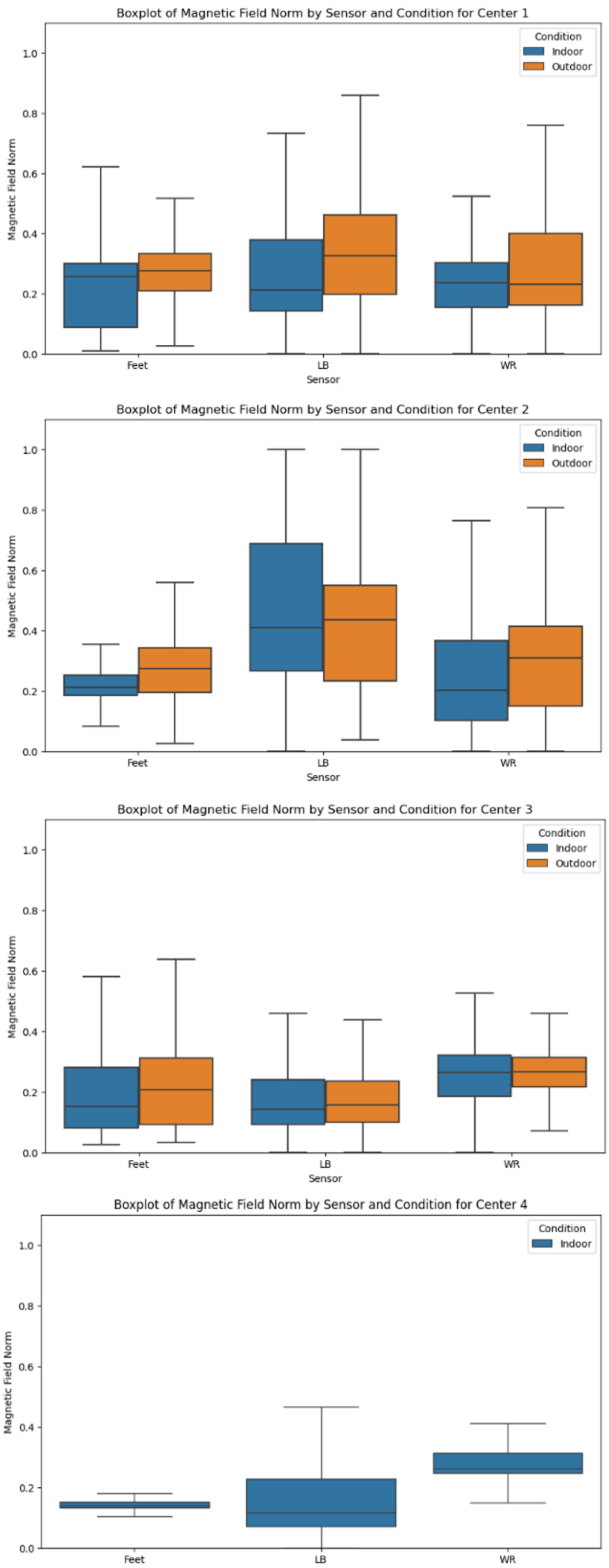


Fig. 5. Boxplots of the normalized magnetic norm distribution across the four geographic areas and three device placements in both indoor and outdoor conditions.

commercial vendors. Estimating the orientation, in addition to increasing the computational burden, requires the sensor fusion algorithms to be properly tuned [40] and special caution for long recordings [41]. Moreover, no investigation among the most promising placements was conducted. The second study [39] compared the performance of a deep learning-based indoor/outdoor detector across four datasets. An important finding was the decrease in accuracy and $F1$-score when magnetometer data were excluded from the dataset: the performance dropped from 92% to 82% when only accelerometer and gyroscope data were used. This can be explained by considering that inertial data are highly sensitive to subject-specific factors, such as walking dynamics and pathologies, which can significantly alter movement patterns even within the same cohort of subjects. These signals reflect not only environmental conditions but also walking characteristics, which are further influenced by varying terrains, such as uneven surfaces or inclines, distorting accelerometer, and gyroscope readings [12]. Conversely, magnetometer data respond to changes in the environment's magnetic field, offering a more reliable distinction between indoor and outdoor settings. Therefore, a magnetometer-based classifier is expected to be more robust, as it avoids the variability and complexity inherent in inertial data.

Finally, the inference time of the tested models suggests that the discrimination between indoor and outdoor environments can be performed every second which is comparable with the update frequency of the GPS, thus providing a low power (1.3 mW, as reported in the datasheet, versus between 4 and 1200 mW of the GPS [42] and between 600 and 2000 mW of the WiFi sensing on smartphones [43]), convenient, and privacy-safe alternative.

V. Limitations

Although the findings in the present study are very encouraging, there is a limitation concerning the technology used as a reference—GPS in conjunction with OpenStreetMap. This combination should be considered a silver standard rather than a gold standard although it is probably the most practical solution for tracking location during people's daily activities. It is expected that in urban settings, where factors such as buildings can cause GPS signal reflections, indoor–outdoor detection can be more complicated with the consequence of a reduction of the accuracy of the reference method.

VI. Conclusion

The model proposed in this article centers around the simplicity of using a single low-cost magnetometer, thus reducing the complexity of the experimental setup, data storage, and energy consumption while respecting the user's privacy. In summary, the best performances were obtained with the XGB model fed with foot data by applying PCA and considering a window size equal to 1 s.

The proposed model can be incorporated into state-of-the-art pipelines for gait analysis parameter extraction, particularly those related to real-world scenarios. The setup is highly compatible with both single-sensor configurations, such as a

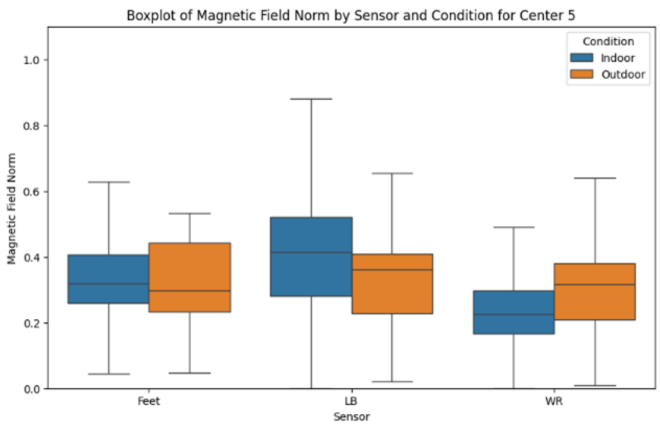


Fig. 6. Boxplot of the normalized magnetic norm distributions for the fifth center.

magnetometer embedded in a device placed on the pelvis [44], which is widely accepted by participants, wrist [45], or a multisensor system positioned on the feet and pelvis [30]. As widely recognized by recent literature, a clinically meaningful characterization of real-world mobility must consider contextual confounding factors [28] to avoid misinterpretation of gait deviations. Among the latter, the participant's location plays a crucial role. The flexibility of the proposed model can be very helpful in this sense being able to discriminate indoor and outdoor environments using only signals recorded by magnetometers, already integrated into most commercial MIMUs, enabling ubiquitous and continuous gait analysis monitoring.

Appendix

To further demonstrate the effectiveness of the proposed model, not only on unseen subjects but also on locations exhibiting magnetic field data that nominally differ from the data encountered during the training phase, two additional subjects were collected in a fifth center (Turin, Italy) for testing purposes only. The magnetic characteristics of this center are reported in Fig. 6. The classification was performed with the XGB fed with PCA-processed input signals from feet. This combination provided the best performance during the main analysis. The average accuracy, $F1$-score, and AUC-ROC amounted to 88%, 0.87%, and 0.90%, respectively, demonstrating the generalization capability of the proposed processing steps and statistical learning techniques.

Availability of Data and Material

The dataset presented in this article can be found on Zenodo online repository. Permanent link and DOI will be generated upon the paper's acceptance. Algorithm implementations can be found on https://github.com/SanBast/ml-indoor-outdoor-detection.

Conflicts of Interest

SilviaDel Din and Jeffrey M. Hausdorff report consultancy activity with Hoffmann-La Roche Ltd., outside of this article.

Acknowledgment

The authors acknowledge all the members of the Mobilize-D Work Package 2 for continuous discussion and critical input. They are particularly grateful to the participants in the study for their time and enthusiastic contribution, especially during the pandemic.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by The Newcastle upon Tyne Hospitals NHS Foundation Trust and Sheffield Teaching Hospitals NHS Foundation Trust: London—Bloomsbury Research Ethics Committee, 19/LO/1507; Tel Aviv Sourasky Medical Center: the Helsinki Committee, 0551-19TLV; Robert Bosch Foundation for Medical Research: Medical Faculty of the University of Tübingen, 647/2019BO2; University of Kiel: Medical Faculty of Kiel University, D540/19, under Application No. ISRCTN (12246987).

Vincenzo Marcianò, Andrea Cereatti, Stefano Bertuletti, and Marco Caruso are with the Department of Electronics and Telecommunications, Politecnico di Torino, 10129 Turin, Italy (e-mail: marco.caruso@polito.it).

Tecla Bonci is with the Department of Mechanical Engineering, The University of Sheffield, S1 3JD Sheffield, U.K.

Lisa Alcock and Silvia Del Din are with the Faculty of Medical Sciences, Translational and Clinical Research Institute, Newcastle University, NE1 7RU Newcastle upon Tyne, U.K.

Eran Gazit and Jeffrey M. Hausdorff are with the Center for the Study of Movement, Cognition and Mobility, Neurological Institute, Tel Aviv Sourasky Medical Center, Tel Aviv, 6423906, Israel.

Neil Ireson and Fabio Ciravegna are with the Department of Computer Science, The University of Sheffield, S1 4DP Sheffield, U.K.

Antonio Bevilacqua, Georgiana Ifrim, and Brian Caulfield are with the Insight Centre for Data Analytics, University College Dublin, Dublin 4, Ireland.

Claudia Mazzà is with the Department of Mechanical Engineering, The University of Sheffield, Sheffield S1 3JD, U.K. and also with the Department of Medical Science, Indivi , 4051 Basel, Switzerland.

References

[1] S. C. Mukhopadhyay, N. K. Suryadevara, and A. Nag, "Wearable sensors for healthcare: Fabrication to application," *Sensors*, vol. 22, no. 14, p. 5137, Jul. 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/14/5137

[2] T. Mustafa and A. Varol, "Review of the Internet of Things for healthcare monitoring," in *Proc. 8th Int. Symp. Digit. Forensics Secur. (ISDFS)*, Jun. 2020, pp. 1–6.

[3] T. Rault, A. Bouabdallah, Y. Challal, and F. Marin, "A survey of energy-efficient context recognition systems using wearable sensors for healthcare applications," *Pervas. Mobile Comput.*, vol. 37, pp. 23–44, Jun. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574119216301298

[4] F. Ali et al., "An intelligent healthcare monitoring framework using wearable sensors and social networking data," *Future Gener. Comput. Syst.*, vol. 114, pp. 23–43, Jan. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X1931605X

[5] L. Letts et al., "The physical environment as a fall risk factor in older adults: Systematic review and meta-analysis of cross-sectional and cohort studies," *Austral. Occupational Therapy J.*, vol. 57, no. 1, pp. 51–64, 2010.

[6] D. Jin, H. Halvari, N. Maehle, and A. H. Olafsen, "Self-tracking behaviour in physical activity: A systematic review of drivers and outcomes of fitness tracking," *Behav. Inf. Technol.*, vol. 41, no. 2, pp. 242–261, 2022, doi: 10.1080/0144929X.2020.1801840.

[7] S. Asimakopoulos, G. Asimakopoulos, and F. Spillers, "Motivation and user engagement in fitness tracking: Heuristics for mobile healthcare wearables," *Informatics*, vol. 4, no. 1, p. 5, Jan. 2017. [Online]. Available: https://www.mdpi.com/2227-9709/4/1/5

[8] X. Gui, Y. Chen, C. Caldeira, D. Xiao, and Y. Chen, "When fitness meets social networks: Investigating fitness tracking and social practices on werun," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, 2017, pp. 1647–1659, doi: 10.1145/3025453.3025654.

[9] Y. J. Park and M. Skoric, "Personalized ad in your Google glass? Wearable technology, hands-off data collection, and new policy imperative," *J. Bus. Ethics*, vol. 142, no. 1, pp. 71–82, Apr. 2017, doi: 10.1007/s10551-015-2766-2.

[10] M. S. Grewal, L. R. Weill, and A. P. Andrews, *Global Positioning Systems, Inertial Navigation, and Integration*. Hoboken, NJ, USA: Wiley, 2007.

[11] S. L. Ullo and G. R. Sinha, "Advances in smart environment monitoring systems using IoT and sensors," *Sensors*, vol. 20, no. 11, p. 3113, May 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/11/3113

[12] A. Torres-Pardo et al., "Legged locomotion over irregular terrains: State of the art of human and robot performance," *Bioinspiration Biomimetics*, vol. 17, no. 6, Oct. 2022, Art. no. 061002, doi: 10.1088/1748-3190/ac92b3.

[13] F. Feriol, D. Vivet, and Y. Watanabe, "A review of environmental context detection for navigation based on multiple sensors," *Sensors*, vol. 20, no. 16, p. 4532, Aug. 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/16/4532

[14] A. A. Abdallah, S. S. Saab, and Z. M. Kassas, "A machine learning approach for localization in cellular environments," in *Proc. IEEE/ION Position, Location Navigat. Symp. (PLANS)*, Apr. 2018, pp. 1223–1227.

[15] I. Ashraf, S. Hur, and Y. Park, "Smartphone sensor based indoor positioning: Current status, opportunities, and future challenges," *Electronics*, vol. 9, no. 6, p. 891, May 2020. [Online]. Available: https://www.mdpi.com/2079-9292/9/6/891

[16] V. F. Miramá, L. E. Díez, A. Bahillo, and V. Quintero, "A survey of machine learning in pedestrian localization systems: Applications, open issues and challenges," *IEEE Access*, vol. 9, pp. 120138–120157, 2021.

[17] V. Bui, N. T. Le, T. L. Vu, V. H. Nguyen, and Y. M. Jang, "GPS-based indoor/outdoor detection scheme using machine learning techniques," *Appl. Sci.*, vol. 10, no. 2, p. 500, Jan. 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/2/500

[18] Y. Zhu et al., "A fast indoor/outdoor transition detection algorithm based on machine learning," *Sensors*, vol. 19, no. 4, p. 786, Feb. 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/4/786

[19] V. Bellavista-Parent, J. Torres-Sospedra, and A. Perez-Navarro, "New trends in indoor positioning based on WiFi and machine learning: A systematic review," in *Proc. Int. Conf. Indoor Positioning Indoor Navigat. (IPIN)*, Nov. 2021, pp. 1–8.

[20] J. L. Bejarano-Luque, M. Toril, M. Fernández-Navarro, R. Acedo-Hernández, and S. Luna-Ramírez, "A data-driven algorithm for indoor/outdoor detection based on connection traces in a LTE network," *IEEE Access*, vol. 7, pp. 65877–65888, 2019.

[21] M. B. Rhudy, S. K. Dolan, C. Mello, and N. Greenauer, "Indoor and outdoor classification using light measurements and machine learning," *Appl. Artif. Intell.*, vol. 36, no. 1, Dec. 2022, Art. no. 2012001, doi: 10.1080/08839514.2021.2012001.

[22] P. Roy and C. Chowdhury, "A survey of machine learning techniques for indoor localization and navigation systems," *J. Intell. Robotic Syst.*, vol. 101, no. 3, p. 63, Mar. 2021, doi: 10.1007/s10846-021-01327-z.

[23] A. Bevilacqua et al., "Automated mobility context detection with inertial signals," *Signal Process. EESS*, 2022. [Online]. Available: https://arxiv.org/abs/2205.08409

[24] I. Ashraf, S. Hur, and Y. Park, "MagIO: Magnetic field strength based Indoor- outdoor detection with a commercial smartphone," *Micromachines*, vol. 9, no. 10, p. 534, Oct. 2018. [Online]. Available: https://www.mdpi.com/2072-666X/9/10/534

[25] M. Li, P. Zhou, Y. Zheng, Z. Li, and G. Shen, "IODetector: A generic service for indoor/outdoor detection," *CM Trans. Sensor Netw.*, vol. 11, no. 2, pp. 1–29, Dec. 2014, doi: 10.1145/2659466.

[26] C. E. Galván-Tejada et al., "Estimation of indoor location through magnetic field data: An approach based on convolutional neural networks," *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 4, p. 226, Apr. 2020. [Online]. Available: https://www.mdpi.com/2220-9964/9/4/226

[27] M. Caruso, A. M. Sabatini, M. Knaflitz, M. Gazzoni, U. D. Croce, and A. Cereatti, "Orientation estimation through magneto-inertial sensor fusion: A heuristic approach for suboptimal parameters tuning," *IEEE Sensors J.*, vol. 21, no. 3, pp. 3408–3419, Feb. 2021.

[28] C. Mazzà et al., "Technical validation of real-world monitoring of gait: A multicentric observational study," *BMJ Open*, vol. 11, no. 12, 2021, Art. no. e050785.

[29] K. Scott et al., "Design and validation of a multi-task, multi-context protocol for real-world gait simulation," *J. Neuroeng. Rehabil.*, vol. 19, no. 1, p. 141, 2022, doi: 10.1186/s12984-022-01116-1.

[30] F. Salis et al., "A multi-sensor wearable system for the assessment of diseased gait in real-world conditions," *Frontiers Bioeng. Biotechnol.*, vol. 11, Apr. 2023, Art. no. 1143248. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fbioe.2023.1143248

[31] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001. [Online]. Available: https://link.springer.com/article/10.1023/A:1010933404324

[32] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[33] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[34] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," *Phys. D, Nonlinear Phenomena*, vol. 404, Mar. 2020, Art. no. 132306. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167278919305974

[35] A. Bagnall, J. Lines, A. Bostrom, J. Large, and E. Keogh, "The great time series classification bake off: A review and experimental evaluation of recent algorithmic advances," *Data Mining Knowl. Discovery*, vol. 31, no. 3, pp. 606–660, May 2017, doi: 10.1007/s10618-016-0483-9.

[36] M. Haklay and P. Weber, "OpenStreetMap: User-generated street maps," *IEEE Pervasive Comput.*, vol. 7, no. 4, pp. 12–18, Oct. 2008.

[37] J. Heaton, "Ian goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning," *Genetic Program. Evolvable Mach.*, vol. 19, nos. 1–2, pp. 305–307, Jun. 2018.

[38] A. Esmaeili Kelishomi, A. H. S. Garmabaki, M. Bahaghighat, and J. Dong, "Mobile user indoor-outdoor detection through physical daily activities," *Sensors*, vol. 19, no. 3, p. 511, Jan. 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/3/511

[39] M. Bilal Akram Dastagir, O. Tariq, and D. Han, "TabCLR: Contrastive learning representation of tabular data classification for indoor-outdoor detection," *IEEE Access*, vol. 12, pp. 102505–102520, 2024.

[40] M. Caruso, A. M. Sabatini, M. Knaflitz, U. Della Croce, and A. Cereatti, "Extension of the rigid-constraint method for the heuristic suboptimal parameter tuning to ten sensor fusion algorithms using inertial and magnetic sensing," *Sensors*, vol. 21, no. 18, p. 6307, Sep. 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/18/6307

[41] R. Rossanigo, M. Caruso, F. Salis, S. Bertuletti, U. D. Croce, and A. Cereatti, "An optimal procedure for stride length estimation using foot-mounted magneto-inertial measurement units," in *Proc. IEEE Int. Symp. Med. Meas. Appl. (MeMeA)*, Jun. 2021, pp. 1–6.

[42] A. Grenier, E. S. Lohan, A. Ometov, and J. Nurmi, "A survey on low-power GNSS," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 3, pp. 1482–1509, 3rd Quart., 2023.

[43] K.-H. Kim, A. W. Min, D. Gupta, P. Mohapatra, and J. Pal Singh, "Improving energy efficiency of Wi-Fi sensing on smartphones," in *Proc. IEEE INFOCOM*, Apr. 2011, pp. 2930–2938.

[44] C. Kirk et al., "Mobilise-D insights to estimate real-world walking speed in multiple conditions with a wearable device," *Sci. Rep.*, vol. 14, no. 1, p. 1754, 2024.

[45] F. Kluge et al., "Real-world gait detection using a Wrist-Worn inertial sensor: Validation study," *JMIR Formative Res.*, vol. 8, May 2024, Art. no. e50035. [Online]. Available: https://formative.jmir.org/2024/1/e50035

Open Access funding provided by 'Politecnico di Torino' within the CRUI CARE Agreement