Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Methods for Classifying Ambulatory Movements in people with Parkinson's: A Systematic Review and Meta-analysis

Seyedeh Mahboobeh Razaviasfali
sr23233@essex.ac.uk

The University of Essex
Fawei Deng
The University of Essex
Jackie Siaw Tze Wong
The University of Essex
Matthew JD Taylor
The University of Essex
Aleksandra V. Birn-Jeffery
The University of Essex

Article

Keywords: Parkinson's, gait disorders, Biomechanical parameters, classification metrics

Posted Date: May 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9448626/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Numerous gait-based methods have been developed to classify early-stage Parkinson's but, there remains no consensus on the most informative biomechanical parameters. Our review and meta-analysis aimed to synthesise biomechanical features and algorithms used to distinguish people with early-stage Parkinson's from people without Parkinson's, to identify which approaches perform best. PubMed, EBSCOhost, and Web of Science were searched from inception. Original studies using biomechanical parameters to classify gait between early-stage PD and controls were included. Of the 735 records screened, 31 studies met the inclusion criteria. For the meta-analysis we used a three-level random-effects model which provided pooled recall, accuracy, precision, specificity, and F1-score values of 0.92, 0.90, 0.89, 0.94, and 0.90, respectively. Overall, Ensemble machine learning, Support Vector Machines, and CNN-based deep learning performed the best. Conversely, K-nearest neighbours and classical statistical models consistently underperformed across all evaluated metrics. Substantial between-study heterogeneity was observed across all baseline models, confirming the necessity of a random-effects approach. This review highlights the need for Parkinson's classification methods based on stable, clinically interpretable biomechanical parameters. Although model performance was often high, heterogeneity likely reflected differences in gait features and protocols. Future work should develop clinically interpretable gait models centred on parameters with clear motor control priorities.

Introduction

Walking is perhaps the simplest form of locomotion which helps support activities of daily living and an independent lifestyle [1]. Therefore, any changes in walking can profoundly diminish quality-of-life [2]. Changes in movement can occur by alterations in coordination among multiple muscles and the neuro-motor system due to accident, aging, neurological diseases, degeneration of the musculoskeletal system [3–6]. Parkinson's is the second most common progressive neurodegenerative disorder, affecting millions worldwide and is expected to increase to 25 million by 2050 [7,8]. Parkinson's is caused by loss of dopaminergic neurons in the substantia nigra pars compacta and leads to motor symptoms such as tremor, rigidity, bradykinesia, postural instability, and abnormal gait [3,9]. These motor problems usually begin subtly and unilaterally in the early Hoehn and Yahr (H&Y) stages and then become more pronounced, bilateral, and disabling with progression to higher stages of the scale [9–11]. In general, walking in people with Parkinson's disease (PwP) is characterized by reduced stride length, shuffling steps, decreased walking speed [12,13], and increased gait variability (in stride and swing time), even among those at stage ≤ 2 H&Y [14,15]. Gait analysis is therefore important for early diagnosis, classification, supporting treatment in early stages (H&Y ≤ 2), and effective rehabilitation [16].

Traditionally, researchers have used human movement analysis as an assistive clinical tool to measure and characterize locomotion [17,18]. By comparing biomechanical outcomes such as spatiotemporal parameters, kinematics, and ground reaction forces, these methods have been used for diagnosis and classification purposes of many medical conditions (e.g. Parkinson's) [19,20]. However, even people without Parkinson's (PwNoP) across their lifespan show large differences and natural fluctuations in gait speed, step/stride length, cadence, and step-to-step variability [21,22]. Because these measures are both person-specific and noisy, they can only be estimated reliably when many steps are recorded, and walking speed is carefully controlled [22,23]. Parkinson's typically exacerbates these features, producing slower walking, shorter steps, altered cadence, longer double support, and greater variability, with experimental studies also reporting shorter and more variable first steps [24]. Therefore, because gait variability and methodological heterogeneity can overlap with Parkinson's-related changes, using spatiotemporal parameters alone for classification is prone to inaccuracy and misclassification [24].

Recent advances in technology have enabled prediction and comparison of gait in PwNoP and PwP using artificial intelligence (AI), advanced statistical methods, and other innovative large data volume techniques, with machine- and deep-learning (ML and DL) models increasingly applied to Parkinson's classification because they can automatically process large, heterogeneous datasets and detect subtle movement patterns [25–27]. While accuracy, sensitivity, specificity, precision/recall, F1-score, and receiver operating characteristic (ROC) curves (Table 1) are standard for assessing ML-, and DL-based models, they can overstate model reliability in small or imbalanced clinical datasets [28]. For instance, performance metrics can be misleading where the groups being compared (e.g., PwP vs. PwNoP) are very unequal in size, a high accuracy score often reflects the larger number of one group rather than a true ability to diagnose the disease [29]. High accuracy scores can be achieved where a model simply defaults to predicting the majority class, thereby masking the failure to identify the minority class [30,31]. These concerns are evident in current Parkinson's gait-classification studies, which often rely on limited samples, variable protocols, and datasets that lack detailed clinical information [32–34]. Moreover, without integrating key clinical context, such as disease stage, levodopa state (ON/OFF), comorbidities, and overall health, data-driven models risk learning patterns that are difficult to interpret clinically and may not generalize to real-world practice [20,35].

Table 1
Classification metrics' definition

Metric	Definition	Formula
Accuracy	The percentage of all predictions that were correct. It summarizes the model's overall performance.	$\varvec{T}\v$
Sensitivity (Recall)	The model's ability to correctly identify positive samples. It measures the proportion of actual positives that were correctly predicted.	$\varvec{T}\v$
Specificity	The model's ability to correctly identify negative samples. It measures the proportion of actual negatives that were correctly predicted.	$\varvec{T}\v$
Precision	Of all the samples the model predicted as positive, the proportion that were positive. It measures positive predictive value.	$\varvec{T}\v$
F1-Score (F-score)	The harmonic mean of precision and sensitivity. It provides a single score that balances both metrics and is especially useful for unbalanced datasets.	$2 \times \varv$

TP (True Positive): Positive samples correctly identified as positive. TN (True Negative): Negative samples correctly identified as negative. FP (False Positive): Negative samples incorrectly identified as positive. FN (False Negative): Positive samples incorrectly identified as negative.

Also, although these models have been used widely in clinical Parkinson's studies, there is limited information around the best biomechanical parameters to use within the models or which model (i.e. ML, DL) can best detect alterations in movement. Addressing these challenges is essential for advancing our understanding of movement in PwP and improving diagnostic and therapeutic strategies. Thus, our overarching aim was to synthesize the methods of ambulatory movement classification in PwP to identify key biomechanical parameters and model algorithms that were most effective in distinguishing between PwNoP and PwP-affected walking patterns at early H&Y stages.

Methods

Search strategy

This systematic review of the literature followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) approach [36]. The protocol was registered on PROSPERO (CRD 42025534168). No changes were made to the protocol.

Search terms

The following search terms were used across all three databases (PubMed, EBSCOhost, and Web of Science): (((gait classif* OR gait-based classif* OR gait predict*) AND ("neural network" OR Bayes OR Bayesian OR Markov Model OR "Markov chain Monte Carlo" OR "artificial intelligence" OR "machine learning" OR "deep learning")) AND (Parkinson*)) NOT ("non-human mammals" OR "animal*" OR "rat" OR "rodent*" OR "fish*" OR "bird*" OR "mice").

The final search was performed on 13/03/2026. We used PECO (Population (Adults ($\geq$ 18 years)), Exposure (People with Parkinson disease (H&Y ≤ 2), Comparator (PwNoP), Outcome (Health gait models built by different methodologies such as statistical methodology, mathematical model, machine learning, etc.) criteria to set the inclusion and exclusion criteria. All types of biomechanical data (e.g. kinematics, kinetics, and spatiotemporal) were included. All studies were available online as a full-text, and the search was carried out since inception. Studies were excluded if they were conducted on children or animals, biometric recognition, classification of gait events, were published as a book, editorial, meta-analysis, systematic review, or conference papers.

Studies were uploaded into CADIMA for duplicate removal, sifting, data extraction, and quality assessment. After removing duplicates, two reviewers (SR and FD) independently screened studies initially using titles and abstracts and then followed by full texts based on the PECO criteria. The selection criteria were applied within CADIMA in both phases. After each phase, disagreements were discussed and in case of doubt, a third independent reviewer (ABJ) was consulted, (19 inconsistent for full text and 23 for title and abstract screening).

Data collection process

Data extraction was performed by one investigator (SR), and the accuracy of data extraction was reviewed by a second investigator (ABJ) independently.

Data items

Data extracted included demographic data (e.g., lead author, publication year, dataset, sample size, age, mass, height, Hoehn-Yahr (H&Y) and The Unified Parkinson's Disease Rating Scale (UPDRS) score), the methodological data (task, instrumentation, data set, protocol, and biomechanical parameters) and model performance data (e.g., accuracy, recall (sensitivity), specificity, precision, and F1-Score). If confusion matrix and the sample size for each group and each stage of Parkinson's was reported but not any of the above mentioned metrics, we extracted them from the confusion matrix table and by using sample size [26,37–42]. For studies which had used one model across different sets of spatiotemporal parameters we took the average and reported [41,43,44]. Two studies had used one model but different augmentation approaches like (Jittering, Scaling, Rotation, Magnitude_warping, Permutation, Time_warping, and Random_sampling, etc) for which we reported the average of various augmentation methods for the reported metrics [45,46]. One study had used SMOT to

address the problem of imbalanced data sets, therefore, we took the average of two metrics reported in that study [47]. We reported the average of the metrics for another study which had used one unique model on the same biomechanical data with and without cropping data [48]. All the metrics were converted to percentage if reported as a proportion.

Risk of bias assessment

We used the check list provided by Downs and Black [49] that can be used to assess the methodological quality not only of randomized controlled trials but also non-randomized studies. This checklist consists of 27 items distributed between five sub-scales (Supplementary Information 2, Table S1). Each study was independently assessed by two reviewers (SR and FD), and discrepancies in scoring were resolved by the third reviewer (ABJ).

Synthesis methods

All statistical meta-analyses were conducted in R (RStudio 2025.09.1 + 401 "Cucumberleaf Sunflower" (RStudio Team, 2025)) using the *metafor* package [50] for multilevel modeling, with the *clubSandwich* and *wildmeta* packages subsequently employed to robustly account for statistically dependent effect sizes generated by studies evaluating multiple machine learning models on identical patient cohorts. To ensure mathematical stability and consider boundary constraints, raw classification metrics were first logit-transformed and assigned delta-method binomial sampling variances. The pooled performance was then estimated via a three-level random-effects working model, explicitly partitioning variance across the sampling level (Level 1), the observation level reflecting within-study model variations (Level 2), and the between-study level (Level 3). This model utilized the Bound Optimization by Quadratic Approximation (bobyqa) algorithm to achieve convergence despite the near-zero within-study variance (ceiling effects) typical of high-performing classification models. To rigorously correct for the nested within-study dependence and prevent artificially narrowed confidence intervals (pseudoreplication), we applied Cluster-Robust Variance Estimation (CRVE) with CR2 small-sample adjusted standard errors. True overall between-study heterogeneity was assessed using the intercept-only working model and is reported using three metrics: Cochran's Q statistic (to determine the statistical significance of the heterogeneity), the I^2 statistic (to report the percentage of total variation across studies attributable to heterogeneity rather than chance), and τ^2 (reflecting the absolute between-study variance). Finally, study-level leave-one-out sensitivity analyses were conducted to evaluate the robustness of the pooled estimates against highly influential publications, and forest and funnel plots were generated to visualize pooled effects, confidence intervals, and potential asymmetry.

Results

Study selection

A total of 735 unique papers were returned using database searches. Duplicates (n = 294) were removed manually by the first reviewer (SR), and three more publications were excluded since the abstract was not available online. After screening the abstract and title for the 438 remaining papers, 214 full-text papers were assessed for eligibility and 224 papers were excluded based on H&Y scores more than 2 (91%), only a PD population (0.005%), model did not include biomechanical parameters (0.005%), or they did not classify PwPD (0.015%) (Fig. 1). Through citation searching the 20 remaining publications, 883 studies were further identified. The same procedure was followed to screen the abstract and title and removed 863 out of these based on PECO criteria [51]. Twenty studies were included for full text review and a further 7 were excluded because they did not include H&Y stage 2. A further 13 studies from citation searching were added to the 20 publications after full text screening. In total 33 studies were included in our synthesis.

Characteristics of the included studies

The participants across the 33 studies varied in age with most studies ranging between 60–70 years (Table 2). Three papers did not report the age of both populations [52–54], and one lacked the age information for PwNoPD [48] (Table 2). The majority of the publications [26,37,38,40,42,45,47,52–63] (57.6%) had used a public dataset for Parkinson's provided by Physio Net which consists of three Parkinson's gait sub-data sets (Ga [64], Si [65], and Ju [66]). Only 24.24% of studies had reported the UPDRS (Unified Parkinson's Disease Rating Scale) scores for PwPD [41,43,54,67–71]. Almost 68% of the studies did not include mass and height information for their participants (Table 2). Only 9 publications [41,43,46,48,68,70–73] had used H&Y ≤ 2. For the rest of studies which had included PwPD of different H&Y levels, we extracted only the data for early stage of disease [26,37–40,42,44,45,47,52–63,67,69,74]. The biomechanical parameters used in the selected studies (Table 3) were ground reaction force (GRF) (42.42%) [26,37,38,46,52,55], kinematics (24.24%) [48] and spatiotemporal parameters of gait (39.39%) [39,39,56,68,71,72]. Only three studies used the both spatiotemporal and kinematic data [44,69,73] which two of them were not included in meta-analysis due to using non-walking tasks [69,73]. Walking was the most common task used, except in 2 studies [69,73] which used Timed-Up-and-Go (TUG) test (Table 3). TUG did not allow meaningful or robust pooled analysis, so it was not used in meta-analysis. 28 studies reported Recall [26,37–42,45,47,48,52–63,67,68,70,71,74,75], 20 studies accuracy [26,39–42,44,48,55,56,58–63,67,68,71,74,75], 18 studies precision [37,39–42,44,45,48,54–56,59,60,62,63,68,74,75], 16 studies F1-score [37,39–42,45,48,53–55,59,60,62,63,70,75], and 12 studies specificity [40,53–55,58–60,62,67,68,71,76].

Risk of bias in studies

Methodological quality for all the 33 included studies, assessed with the 27 item Downs & Black checklist, was uniformly poor (median = 10/27; range 9–11) [77]. Reporting items were generally met, but none of the studies undertook randomization, confounding adjustment, or participant blinding, and no sample size calculations were reported resulting in the poor risk of bias results.

Table 2
Demographic data of the included studies

Ref	Data Set	Sample Size (m:f)		Age (Years)		Mass (kg)		Height (m)		H&Y
		PwPD	*PwNoPD*	*PwPD*	*PwNoPD*	*PwPD*	*PwNoPD*	*PwPD*	*PwNoPD*	*PwPD*
37	Ga[a]	20:09	10:08	61.6 ± 8.8**	57.9 ± 6.7	73.1 ± 11.2	74.2 ± 12.7	nr	1.67 ± .07	
	Si[a]	22:13	18:11	67.2 ± 9.1	64.5 ± 6.8	70.3 ± 8.4	71.5 ± 11.0	nr	1.66 ± .07	2–3
	Ju[a]	16:13	12:14	66.8 ± 10.9	39.3 ± 18.5	75.1 ± 16.9	66.8 ± 11.	nr	1.87 ± .15	
42	Ga[a]	20:09	10:08	66.3*	66.3*	nr	nr	nr	nr	2–3
	Si[a]	22:13	18:11							
	Ju[a]	16:13	12:14							
52	Ga[a]	20:09	10:08	nr	nr	nr	nr	nr	nr	
	Si[a]	22:13	18:11	nr	nr	nr	nr	nr	nr	2–3
	Ju[a]	16:13	12:14	nr	nr	nr	nr	nr	nr	
26	Ga[a]	20:09	10:08	66.3*	66.3*	nr	nr	nr	nr	
	Si[a]	22:13	18:11			nr	nr	nr	nr	2–3
	Ju[a]	16:13	12:14			nr	nr	nr	nr	
58	Ga[a]	20:09	10:08	50–70, ≥ 70	50–70, ≥ 70	73.1 ± 11.3	74.2 ± 12.6	1.67 ± .07	1.68 ± .08	2–3
	Si[a]	22:13	18:11	50–70, ≥ 70	50–70, ≥ 70, <50	70.3 ± 8.35	71.5 ± 11.03	1.66 ± .07	1.69 ± .08	
	Ju[a]	16:13	12:14	50–70, ≥ 70, <50	50–70, ≥ 70	75.1 ± 16.90	66.8 ± 11.08	1.87 ± .15	1.83 ± .08	
45	Ga[a]	20:09	10:08	61.6 ± 8.8	57.9 ± 6.7	71.3 ± 12.7	74.2 ± 12.7	1.67 ± .07	1.68 ± .08	2–3
	Si[a]	22:13	18:11	67.2 ± 9.1	64.5 ± 6.8	70.3 ± 8.4	71.5 ± 11.0	1.66 ± .07	1.69 ± .08	
	Ju[a]	16:13	12:14	66.80 ± 10.85	39.31 ± 18.51	75.1 ± 16.89	66.8 ± 11.07	1.87 ± .15	1.83 ± .08	
59	Ga[a]	20:09	10:08	61.6 ± 8.8	57.9 ± 6.7	73.1 ± 11.2	74.2 ± 12.7	1.67 ± .07	1.68 ± .08	2–3
	Si[a]	22:13	18:11	67.2 ± 9.1	64.5 ± 6.8	70.3 ± 8.4	71.5 ± 11.0	1.66 ± .07	1.69 ± .08	
	Ju[a]	16:13	12:14	66.80 ± 10.85	39.31 ± 18.51	75.1 ± 16.89	66.8 ± 11.07	1.87 ± .15	1.83 ± .08	
40	Ga[a]	20:09	10:08	66.3*	66.3*	nr	nr	nr	nr	1–3
	Si[a]	22:13	18:11							
	Ju[a]	16:13	12:14							
55	Ga Si Ju[a]	58:35	40:33	65.4 ± 14.5	63.7 ± 40.3	69.5 ± 12.2	71.2 ± 22.6	1.68 ± .06	1.68 ± .08	2–3
38	Ga Si Ju[a,b]	93	72	66.3*	66.3*	nr	nr	nr	nr	Modified 1–5
57	Ga Si Ju[a,b]	58:35	40:33	66.3 ± 9.5	63.7 ± 8.7	nr	nr	nr	nr	2–3
56	Ga Si Ju[a]	58:35	40:33	65.4 ± 14.5	63.7 ± 40.3	69.5 ± 12.2	71.2 ± 22.6	1.68 ± .06	1.68 ± .08	1–3
60	Ga Si Ju[a]	58:35	40:33	66.3 ± 9.5	63.7 ± 8.64	72.4 ± 11.83	72.8 ± 12.26	1.67 ± .08	1.68 ± .08	2–3
61	Ga Si Ju[a]	63:37%	55:45%	66.3*	66.3*	nr	nr	nr	nr	2–3
78	Ga Si Ju[a]	58:35	40:33	66.3 ± 9.5	63.7 ± 8.64	72.4 ± 11.83	72.8 ± 12.26	1.67 ± .08	1.68 ± .08	2–3
62	Ga Si Ju[a]	58:35	40:33	66.3 ± 9.5	63.7 ± 8.64	72.4 ± 11.83	72.8 ± 12.26	1.67 ± .08	1.68 ± .08	2–3

Ref	Data Set	Sample Size (m:f)		Age (Years)		Mass (kg)		Height (m)		H&Y
		PwPD	*PwNoPD*	*PwPD*	*PwNoPD*	*PwPD*	*PwNoPD*	*PwPD*	*PwNoPD*	*PwPD*
63	Ga Si Ju[a]	58:35	40:33	65.43 ± 14.88	63.69 ± 40.23	69.64 ± 14.11	71.20 ± 24.28	1.68 ± 0.07	1.68 ± .08	1–3
53	Ga Si Ju[a]	91	71	nr	nr	nr	nr	nr	nr	2–3
54	Ga Si Ju[a]	93	72	nr	nr	nr	nr	nr	nr	0–3
69d	V-TIME[e]	65.6:34.4%	42:58%	68.0 ± 2.7	58.0 ± 3.5	nr	nr	nr	nr	
	BEAT-PD[e]	332	100							1–3
39	c	26:31	12:12	72.6 ± 8.1	70.7 ± 3.9	nr	nr	nr	nr	Modified 1–5
74	c	17:15	6:10	69.28 ± 8.56	67.19 ± 8.75	nr	nr	nr	nr	1–4
41	c	64:39	33:42	68.95 ± 5.95	68.91 ± 5.54	63.10 ± 9.74	63.94 ± 9.94	1.61 ± 8.57	1.60 ± 8.97	1.84 ± 0.37
43	c	37:41	21:29	67.51 ± 7.08	65.80 ± 5.51	63.96 ± 10.54	63.68 ± 10.04	1.61 ± 7.72	1.61 ± 8.38	1–2
67	c	27:36	35:30	66.90 ± 8.43	67.40 ± 7.17	67.83 ± 12.38	66.92 ± 11.37	1.65 ± 7.98	1.65 ± 7.96	1–3
70	c	15:22	9:10	64.29 ± 7.12	63.13 ± 5.53	69.26 ± 9.22	68.63 ± 4.83	1.66 ± 7.15	1.67 ± 0.05	1.37 ± 0.55
46	c	22:24	12:32	68.7 ± 9.8	73.9 ± 6.0	nr	nr	nr	nr	1–2
48	c	88b	94 b	68 (41–82)	nr	nr	nr	nr	nr	1–2
44	c	13:14	13:14	43–88	41–79	nr	nr	nr	nr	1–3
68	c	24:16	21:19	59.8 ± 10.6	59.8 ± 11.9	nr	nr	nr	nr	1.16 ± 0.24
71	c	20:10	20:10	66.4 ± 8.4	65.6 ± 10.5	65.5 ± 6.2	62.1 ± 9.6	nr	nr	1.8 ± 0.45
73d	c	32:53	40:45	62.9 ± 9.9	62.84 ± 5.8			1.65. ± .06	1.63 ± .05	1.66 ± 0.44
72	c	07:12	12:09	64.1 ± 7	63.4 ± 5.8	67.3 ± 9.7	63.6 ± 8.5	1.67 ± 0.08	1.62 ± 0.06	1.47 ± 0.45

*number in each age group; a Physionet; b no split by sex; c data collected by the researchers; d, these three studies are not included in meta-analysis, e; V-TIME (129) and BEAT-PD (303) two different dataset used.; EPD, Early-stage PD; f, female; PwP, People with Parkinson's; PwNoPD, People without Parkinson's ; m, male; MPD. Mild PD; nr; not reported; UPDRS, Unified Parkinson's Disease Rating Scale; * Group means ** Group means and standard deviations were combined using weighted averages and pooled variance - SD calculated from range

Table 3
Methodologies of the included studies

Ref	Task	Instrumentation	Data set	Model	Performance analysis	Biomechanical Parameters
57	Walking	Foot sensors	Ga Ju Si	Dual-Modal Convolutional Attention-enhanced Long Short-Term Memory (DCALSTM)	5-fold cross validation	VGRF
37	Walking	Foot sensors	Ga Ju Si	Hybrid ConvNet-Transformer	10-fold cross validation	VGRF
52	Walking	Foot sensors	Ga Ju Si	Convolutional neural network (CNN) + Long-Short Term Memory (LSTM)	10-fold cross validation	VGRF
26	Walking	Foot sensors	Ga Ju Si	One-Dimensional Convolutional Neural Network (1D-CNNs)	*nr*	VGRF
55	Walking	Foot sensors	Ga Ju Si	Deep 2D-CNNs	10-fold cross validation	VGRF
61	Walking	Foot sensors	Ga Ju Si	Convolutional neural networks (CNNs)	nr	VGRF
56	Walking	Foot sensors	Ga Ju Si	Long-Short Term Memory (LSTM) Network	10-fold cross validation	Spatiotemporal parameters
38	Walking	Foot sensors	Ga Ju Si	Temporal pyramid attention-based spatiotemporal (PAST) fusion mode	*nr*	VGRF
54	Walking	Foot sensors	Ga Ju Si	Convolutional long deep neural network (CLDNN)	ten-fold cross-validation	VGRF
63	Walking	Foot sensors	Ga Ju Si	Decision tree Random forest XGBoost LightGBM	5-fold cross-validation	VGRF
40	Walking	Foot sensors	Ga Ju Si	Single CNN Parallel CNN Quadruplet CNN LSTM	nr	VGRF
60	Walking	Foot sensors	Ga Ju Si	Information Set-based Decision Tree (IFS-DT)	ten-fold cross-validation	Spatiotemporal parameters
59	Walking	Foot sensors	Ga Ju	K-nearest neighbours (KNN) Naive Bayes (NB)	ten-fold cross-validation	Spatiotemporal parameters

Ga, Ju, Si, (Physionet with dataset);*†BEAT-PD and V-TIME: two different datasets; a, data collected by the researcher, GAITRite; a pressure sensitive walkway mat; GYENNO MATRIX, initial sensor wearable device; IMU, inertial measurement unit; RoM, Range of Motion; TUG, timed-up- and-go; VGRF, vertical ground reaction forces; *nr*; not reported;

Ref	Task	Instrumentation	Data set	Model	Performance analysis	Biomechanical Parameters
			Si	Ensemble classifier (EC) Support Vector Machine (SVM)		
78	Walking	Foot sensors	Ga Ju Si	Artificial Neural Network (ANN)	ten-fold cross-validation	Spatiotemporal parameters
62	Walking	Foot sensors	Ga Ju Si	Ensemble classifier	ten-fold cross-validation	Spatiotemporal parameters
45	Walking	Foot sensors	Ga Ju Si	One-dimensional-convolutional neural network (1D-ConvNet)	ten-fold cross-validation	VGRF
58	Walking	Foot sensors	Ga Ju Si	hybrid convolutional neural network-long short term memory (CNN–LSTM)	nr	VGRF
53	Walking	Foot sensors	Ga Ju Si	Ensemble K-nearest neighbor (EnKNN)	nr	VGRF
42	Walking	Foot sensors	Ga Ju Si	Deep convolutional neural network (DCNN) Parallel DCNN Quadruplet DCNN	nr	VGRF
46	Walking	IMU	a	Deep convolutional neural network (CNN)	5-fold cross validation	Kinematics
44	Walking	IMU	a	SVM with non-linear kernel (rbf) bases	5-fold cross validation	Spatiotemporal parameters and kinematics
74	Walking	APDM OPAL system with wearable IMUs	a	Convolutional Neural Network (CNN)	5-fold cross validation	Kinematics
43	Walking	Xsens DOT Wearable Sensor platform	a	Convolutional neural network (Residual Neural Network (ResNet)) Convolutional neural network (Dense Convolutional Network (DenseNet)) Convolutional neural network (SqueeezeNet))	fivefold cross-validation	Kinematics
67	Walking	ReadyGo system	a	gradient-boosted decision-tree	nr	Spatiotemporal parameters
41	Walking	IMU	a	Logistic regression KNN K-nearest neighbor NB Naïve Bayes LDA Linear discriminant analysis QDA Quadratic discriminant analysis SVM Support vector machine RF Random forest	five-fold cross-validation	Spatiotemporal parameters
70	Walking and turning	IMU	a	Convolutional neural network (CNN)	Cross-Entropy Loss function	Kinematics

Ga, Ju, Si, (Physionet with dataset);*†BEAT-PD and V-TIME: two different datasets; a, data collected by the researcher, GAITRite; a pressure sensitive walkway mat; GYENNO MATRIX, initial sensor wearable device; IMU, inertial measurement unit; RoM, Range of Motion; TUG, timed-up- and-go; VGRF, vertical ground reaction forces; *nr*; not reported;

Ref	Task	Instrumentation	Data set	Model	Performance analysis	Biomechanical Parameters
72	Walking and turning	IMU	a	Support vector machine (SVM) with linear kernel	*nr*	Spatiotemporal parameters
				SVM with non-linear kernel (rbf) bases		
48	Walking	Kinect®	a	Long-Short Term Memory (LSTM) model	5-fold cross validation	Kinematics
				One-dimensional convolution model (Conv1D)		
				Conv LSTM model		
				LSTM model		
				Conv1D		
				Conv LSTM model		
69	Walking* TUG†	body-fixed sensors	V-TIME*	Random Under-sampling Boosting (RUSBoost)	*nr*	Spatiotemporal parameters and kinematics
			BEAT-PD†			
73	TUG	GYENNO MATRIX device	a	Support vector machine (SVM)	*nr*	Spatiotemporal parameters and kinematics
71	Walking	IMU, Foot sensors		Decision tree (DT-RFE)	*nr*	Spatiotemporal parameters
			a			
39	Walking	IMU	a	Soft-margin kernel support vector machine (SVM)	10-fold cross validation	Spatiotemporal parameters
68	Walking	GAITRite	a	SVM (RBF) kernel	10-fold cross validation	Spatiotemporal parameters

Ga, Ju, Si, (Physionet with dataset);*†BEAT-PD and V-TIME: two different datasets; a, data collected by the researcher, GAITRite; a pressure sensitive walkway mat; GYENNO MATRIX, initial sensor wearable device; IMU, inertial measurement unit; RoM, Range of Motion; TUG, timed-up- and-go; VGRF, vertical ground reaction forces; *nr*; not reported;

Models' performance

Overall, our review of 31 independent studies, comprising of 44 model performance estimates for accuracy, 38 for precision, 54 for recall, 22 for specificity, and 37 for F1-score, revealed robust pooled performance metrics alongside significant moderator effects due to model choice. The pooled accuracy of 0.90 (95% CI 0.86–0.94) (Fig. 2), precision of 0.89 (95% CI 0.84–0.93) (Fig. 3), recall of 0.92 (95% CI 0.89–0.94) (see Supplementary Information 1, Fig. S1), specificity of 0.94 (95% CI 0.89–0.97) (see Supplementary Information 1, Fig. S2), and F1-score of 0.90 (95% CI 0.85–0.94) (see Supplementary Information 1, Fig. S3) indicate that, across these studies, the models achieve high classification performance. Even after allowing for the fact that multiple models came from the same studies, the overall metric performance remained high and statistically reliable for accuracy (t_{CR2}=10.1, 95% CI on the logit scale: 1.77 to 2.70, $p < 0.001$), recall (t_{CR2}=13.8, 95% CI on the logit scale: 2.04 to 2.76, $p < 0.001$), precision (t_{CR2}=15.8, 95% CI on the logit scale: 1.64 to 2.58, $p < 0.001$), f1-score (t_{CR2}=14.9, 95% CI on the logit scale: 1.73 to 2.72, $p < 0.001$), and specificity (t_{CR2}=8.67, 95% CI on the logit scale: 2.09 to 3.53, $p < 0.001$). However, the corresponding prediction interval (≈ 58% to 99%) for all the metrics indicates that the expected accuracy in an individual future study could vary widely. Significant heterogeneity was observed across all five metrics, indicating that the variability in reported model performance was greater than expected by chance alone. Heterogeneity was highest for F1-score (I^2 = 87.41%), followed by accuracy (I^2 = 68.12%), suggesting substantial inconsistency across studies for these outcomes. Precision showed moderate heterogeneity (I^2 = 43.24%), while recall (I^2 = 37.14%) and specificity (I^2 = 25.36%) showed comparatively lower heterogeneity, indicating more consistent performance across studies for these metrics. Leave-one-model-out sensitivity analyses indicated that the overall pattern was stable and not driven by any single model label ($p > 0.05$). Visual inspection of the funnel plots suggested some degree of asymmetry across the five metrics, indicating the possibility of small-study effects (see Supplementary Information 1, Fig. S4-8). However, these plots were interpreted cautiously because the analyses were based on logit-transformed performance proportions with approximated sampling variances and included dependent outcomes clustered within studies. Therefore, the funnel plots were treated as exploratory rather than confirmatory indicators of publication bias.

Based on forest plot, for accuracy (Fig. 2), the strongest models were SVM and CNN–LSTM, while the weakest were LSTM, KNN K-nearest neighbor, and LDA. For precision, EnKNN and SVM were highest, while LSTM and KNN/LDA variants were lowest (Fig. 3). For recall, CNN + LSTM and DCALSTM tied for the highest estimate, while Naïve Bayes, logistic regression, and LDA were lowest (see Supplementary Information 1, Fig. S1). For F1-score, Support Vector Machine, AdaBoost, and CLDNN were highest, while LSTM and KNN/LDA variants were lowest (see Supplementary Information 1, Fig. S2). For specificity, several models tied at the top around 0.99; besides the three shown, SVM-Kinematic also had the same estimate and logit value (see Supplementary Information 1, Fig. S3, Tabel 4).

Table 4
Best and worst performing models across accuracy, precision, recall and F1-score.

Metric	Rank	Model	Back-transformed estimate	Logit estimate
Accuracy*	1	SVM	0.9986	6.5935
Accuracy*	2	Support Vector Machine (SVM)	0.9950	5.3027
Accuracy*	3	CNN-long short-term memory (CNN–LSTM)	0.9876	4.3776
Accuracy**	1	Long-Short Term Memory (LSTM) model	0.5850	0.3433
Accuracy**	2	KNN K-nearest neighbor	0.6335	0.5473
Accuracy**	3	LDA Linear discriminant analysis	0.6470	0.6059
Recall*	1	CNN + LSTM	0.9990	6.9068
Recall*	2	DCALSTM	0.9990	6.9068
Recall*	3	SVM	0.9981	6.2583
Recall**	1	NB Naïve Bayes	0.5990	0.4013
Recall**	2	Logistic regression	0.6115	0.4536
Recall**	3	LDA Linear discriminant analysis	0.6120	0.4557
Precision*	1	Ensemble K-nearest neighbor (EnKNN)	0.9990	6.9068
Precision*	2	SVM	0.9970	5.8138
Precision*	3	Convolutional long deep neural network (CLDNN)	0.9862	4.2692
Precision**	1	Long-Short Term Memory (LSTM) model	0.5400	0.1603
Precision**	2	KNN K-nearest neighbor	0.6470	0.6059
Precision**	3	LDA Linear discriminant analysis	0.6530	0.6323
F1-score*	1	Support Vector Machine (SVM)	0.9820	3.9992
F1-score*	2	Ensemble learning based Adaboost (ELA) classifier	0.9808	3.9335
F1-score*	3	Convolutional long deep neural network (CLDNN)	0.9800	3.8918
F1-score**	1	Long-Short Term Memory (LSTM) model	0.5850	0.3433
F1-score**	2	KNN K-nearest neighbor	0.6230	0.5023
F1-score**	3	LDA Linear discriminant analysis	0.6435	0.5906
Specificity*	1	Decision tree (DT)	0.9990	6.9068
Specificity*	2	Ensemble K-nearest neighbor (EnKNN)	0.9990	6.9068
Specificity*	3	Support Vector Machine (SVM)	0.9990	6.9068
Specificity**	1	K-nearest neighbours (KNN)	0.6760	0.7354
Specificity**	2	decision tree (DT)-RFE	0.8300	1.5856
Specificity**	3	RBF-SVM classification models	0.8350	1.6215
* Best performance, **worse performance				

Discussion

The aim of this systematic review and meta-analysis was to identify the best models and biomechanical parameters for classifying early-stage Parkinson's using during walking. Our main finding showed that machine-learning models appeared more consistently strong across all metrics than deep-learning models, although some deep-learning hybrids were also excellent. ML and DL models using GRF consistently outperformed both models trained on spatiotemporal parameters and unprocessed kinematic time-series data (e.g. accelerations or joint angles). Taken together, these results indicate that, under current data and study-design conditions, ML and DL models trained on GRF data are the most robust and reliable approach for early-stage Parkinson's gait classification.

Across the studies included, the models demonstrated robust overall performance in distinguishing people with Parkinson's disease (PwPD) from healthy controls (PwNoPD), with the highest pooled performance observed for specificity and recall (Fig. 2–3, Supplementary Information 1, Fig. S1-3). Specifically, the pooled specificity was 95%, indicating that models were consistently effective at correctly identifying individuals without PD, and this metric showed the greatest stability across studies [79]. Pooled recall (sensitivity) was similarly high at 92%, suggesting that most approaches were also reliable in detecting PwPD. In contrast, pooled precision (positive predictive value) was lower at 89%, implying that positive PD predictions were comparatively less consistent

and more heterogeneous across datasets and modelling pipelines, likely reflecting a higher susceptibility to false positives and threshold/prevalence effects [79]. Collectively, these findings suggest that, at the meta-analytic level, the most reliable aspect of current ML/DL approaches is their ability to recognize PwNoPD (specificity of 0.94 (95% CI 0.89–0.97)) and capture PwPD cases (recall of 0.92 (95% CI 0.89–0.94)), whereas improving their precision (0.89 (95% CI 0.84–0.93)) to reduce false positive diagnoses remains a critical priority for enhancing clinical utility and reducing unnecessary follow-up among healthy individuals [79]. Across the model-specific outputs, top-performing labels repeatedly included SVM, decision tree / boosting / ensemble methods, AdaBoost, LightGBM, and EnKNN, [53,55,63] while among deep-learning models the strongest were mainly CNN-based hybrids such as CNN–LSTM, CLDNN, and DCALSTM [38,52,54,57,58] for classifying early-stage Parkinson's gait, frequently achieving scores of 97–100% (Fig. 2–3, Supplementary Fig. S1-3). A plausible explanation is that these models are better able to capture nonlinear patterns in biomechanical data, such as GRF and trim series data [80,81]. CNN-based hybrid models may have benefited from learning informative local and temporal representations directly from the biomechanical signal, while ensemble and kernel-based machine-learning models may have provided strong classification boundaries with good flexibility [82,83]. However, DL models like plain Long-Short Term Memory (LSTM) model and ML based models such as KNN K-nearest neighbor, LDA, and Logistic regression [41,43,48,70] consistently demonstrated the worst performance across different metrics (Fig. 2–3, Supplementary Information 1, Fig. S1-3). These weaker results may reflect sensitivity to noisy or high-dimensional feature spaces, stronger modelling assumptions, limited flexibility [82,84] and may also be related to their greater data requirements [85]. Nevertheless, all models in this review were ML- or DL-based and therefore share common limitations as data-driven black-box approaches [86]. They are vulnerable to bias and their internal decision processes are often difficult to interpret, meaning they classify PwPD vs PwNoP without clearly showing which gait features contributed to that decision or how [86]. There is a recognized trade-off between accuracy and interpretability [87] since more complex algorithms tend to be harder to explain [88], while clinicians need to understand causal relationships and potential treatment effects, which standard supervised ML does not provide [87]. In this review only one study had used explainable artificial intelligence (XAI) techniques including layerwise relevance propagation (LRP) to enhance the transparency of their deep learning models to interpret the intricate patterns in gait dynamics [40]. The paper provided preliminary support that LRP identified model-relevant regions, but the interpretability evidence is not strong enough to conclude that the explanations were fully reliable or clinically meaningful [40]. In general, although explainable ML/DL methods are emerging to support clinical decision-making [89–91], they may not capture complex interactions as effectively as high-capacity ML/DL models and can themselves be influenced by bias and subjective judgement, for example through biased training data or through human choices about which features and explanations to prioritize, which in turn limits their generalizability [92].

At the biomechanical level, the strongest pattern appeared to involve ground reaction force features combined with machine-learning or CNN-based deep-learning models [40,54,55,63], whereas weaker performance was more often observed for spatiotemporal features combined with conventional machine-learning classifiers [41,44,48]. These findings are broadly in line with previous work, where tree/ensemble methods (Random Forest, Extra Trees, gradient boosting variants such as XGBoost) and SVMs are repeatedly reported as leading performers under limited-data conditions [93,94] while LSTM and ConvLSTM models trained on short, noisy Kinect recordings with raw joint coordinates and minimal preprocessing show weaker performance [48]. Training on small, heterogeneous samples and relying on internal cross-validation alone also increases the risk of overfitting and unstable estimates [26,57]. A more direct comparison is possible in the subgroup of studies [26,37,38,40,42,45,47,52–63] that all used the same publicly available GRF database (73 PwNoPD, 93 PwPD) [95]. In this subset, the input data, pre-processing, and outcome definition are effectively held constant, so differences in performance can be attributed mainly to the model itself. Visual inspection of the forest plots (Fig. 2–3, Supplementary Information 1, Fig. S1-3) shows that, even under these matched conditions, ELA, DCALSTM, CNN+LSTM and SVM models still cluster at the top of the accuracy, precision, recall, specificity, and F1 distributions, whereas simpler 1D-ConvNet, LSTM, and deep models occupy the middle–lower range. This implies that model choice, rather than dataset differences, is the primary driver of performance in the current literature, while recognizing that these comparisons remain limited to a single source dataset. In addition, substantial residual heterogeneity remained for accuracy ($I^2 \approx 68.12\%$) and F1-score ($I^2 \approx 87.41\%$) which indicate that the disagreement between studies is not random but likely driven by study-level factors, notably the specific model architecture and the choice of biomechanical inputs. Funnel plot asymmetry for all four metrics (see Supplementary Information 1, Fig. S4-8) further suggests small-study effects or publication bias, which complicates interpretation of the pooled estimates. Moreover, age as a particularly important source of bias alone can reduce walking speed, shorten stride length, and increase gait variability [96], and many of these changes overlap with early Parkinson's gait [97]. In datasets where PwNoPD are younger than PwPD, a model may achieve high accuracy simply by distinguishing age-related gait patterns, effectively acting as an age detector rather than a Parkinson's classification model [96], which inflates performance and limits clinical application.

The choice of biomechanical parameters also significantly determines model performance [44,89], where most models achieved their best results when using GRF data, regardless of the model [26,37,38,52,55–57]. Alharthi et.al. [26] used deep learning-based models such as 1D-CNNs, deep 2D-CNNs, and LSTM network utilizing GRF data from PhysioNet dataset [95] and achieved a recall of 96–97%. However, the Physio-Net dataset [95] does not contain uniform biomechanical data, as it combines level-ground and treadmill walking under both single- and dual-task conditions, despite the substantial known differences in joint kinematics, temporal parameters, and dual-task effects on stride length, walking speed, and step-to-step variability in both PwNoPD and PwPD [98–100]. Therefore, the heterogeneous acquisition protocols (treadmill vs overground, single vs dual task) introduce confounding variability that can mask true Parkinson's gait patterns and limit model generalizability [26,37,38,52,55–57,72]. Djurić-Jovičić et al. [68] used an RBF-kernel SVM to compare models built on a standard spatiotemporal feature set with those using a reduced subset, and showed that step-time asymmetry and heel-to-heel base-of-support variability contributed more to classification accuracy than the traditionally used double-support time, which had little discriminative value and even reduced model performance when included. Balaji et. al. [56] used some spatiotemporal data (extracted from GRF data available on Physio-Net dataset [95]) to test their ML based models (Decision tree, SVM, Ensemble classifier, Bayes classifier). The accuracy, recall, and precision were achieved (97%) for most of their ML models [56]. However, it is worth noting that spatiotemporal gait parameters show measurable variability regardless of pathology across the lifespan, and reliability depends on the parameter and protocol [16,101], justifying careful selection of gait parameters and protocols when using spatiotemporal features to classify Parkinson's [12,14]. In contrast, some studies used kinematic time-series data, such as acceleration and segmental range of motion, with deep-learning

models and reported performance ranging from approximately 50% to 80%. However, there were two machine-learning models evaluated on these kinematic data, making it difficult to determine whether the weaker performance was attributable to the models themselves or to the biomechanical data used to distinguish early-stage Parkinson's disease from PwNoPD [44,74].

Limitations

This review faces several limitations that should be addressed in future research. First, variability in reported classification metrics across models was high, reflecting heterogeneity in sample size, class balance, data quality, and evaluation practice. Some of the Parkinson's datasets were small, unbalanced, or lacked key demographic information; only one study used equal sample sizes for people with and without Parkinson's, and most models were trained on the same publicly available GRF database. Although some authors applied k-fold cross-validation or data-balancing methods such as SMOTE, the widespread absence of independent external test sets means that overfitting is likely and reported performance metrics may be inflated and not fully generalizable to clinical populations. Second, the methodological quality of the primary studies was uniformly low since none of them had reported randomization, adjustment for confounders, blinding, or a priori sample-size calculations. Age, a critical confounder in gait analysis, was often inadequately controlled, with some studies failing to recruit age-matched controls and others not reporting age at all. In addition, we were not able evaluate diagnostic accuracy (recall/specificity pairs) by random-effects meta-analyses using bivariate models, since only a few studies had reported specificity. Moreover, because few studies reported confusion matrices, exact variance estimation was not possible for most outcomes. Therefore, sampling variances were approximated using the reported metric values and the most relevant available sample size for the corresponding metric. Finally, we did not explicitly account for medication status, even though dopaminergic treatment can attenuate gait deficits and thereby increase the risk of false negatives in classification.

Conclusion

In conclusion, under current data and study-design conditions, early-stage Parkinson's gait is best classified by regularized, feature based machine learning models (e.g., RBFSVM, tree/ensemble methods) trained on ground reaction forces and carefully selected spatiotemporal or variability features while deep learning on raw kinematics performs less reliably in small, heterogeneous cohorts. However, this review underscored the importance of developing a comprehensive method that relies on stable biomechanical parameters for Parkinson's classification and diagnosis. Although Artificial Intelligence-, Machine Learning-, and Deep Learning-based models showed acceptable recall and accuracy, their performance varied widely between studies, due to differing biomechanical parameters or merging heterogeneous datasets from different walking protocols (e.g., level vs. treadmill walking). This variation makes it difficult to determine which models and biomechanical parameters are most suitable, and current models have limited clinical interpretability because they do not explain gait deviations in Parkinson's or incorporate explicit neuromechanical principles. In total, the models reviewed here do not explicitly incorporate neuromechanical principles, such as muscle coordination, joint kinetics, or other physiological mechanisms underlying gait, via established biomechanical models (e.g. multisegmented or reduced order models) [102]. Instead, they are built purely on statistical patterns in the data without embedding prior biomechanical or clinical knowledge [20,37]. This lack of neuromechanical grounding constrains both interpretability and the depth of clinical insight that these models can offer [103]. Future work should develop clinically interpretable gait models centered on parameters under clear neuromechanical control.

Declarations

Acknowledgement and Funding Information

This work was supported by the School of Sport, Rehabilitation and Exercise Sciences (SRES) PhD Studentship, University of Essex (awarded to Seyedeh Mahboobeh Razaviasfali). The funder had no role in study design, data collection/analysis, or manuscript preparation.

Authors' contributions

Seyedeh Mahboobeh Razaviasfali: writing original draft preparation, Reviewer, Software, Visualization, Investigation. Fawei Deng: acted as the second reviewer and provided critical revisions. Aleksandra V. Birn-Jeffery: acted as the third reviewer. Jackie Siaw Tze Wong: Writing – review & editing. Matthew J. D. Taylor and Aleksandra V. Birn-Jeffery contributed equally in: Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Validation, Conceptualization. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Data availability

The search strategy and extracted data contributing to the meta-analysis is included in the main text; any additional data are available on request from the corresponding author.

References

1. Cappozzo, A. Gait analysis methodology. *Hum. Mov. Sci.* **3**, 27–50 (1984).

2. Mahlknecht, P. et al. Prevalence and Burden of Gait Disorders in Elderly Men and Women Aged 60–97 Years: A Population-Based Study. *PLoS ONE*. **8**, e69627 (2013).
3. Bloem, B. R., Grimbergen, Y. A. M., Cramer, M., Willemsen, M. & Zwinderman, A. H. Prospective assessment of falls in Parkinson's disease. *J. Neurol.* **248**, 950–958 (2001).
4. Aleixo, P., Vaz-Patto, J., Moreira, H. & Abrantes, J. Dynamic joint stiffness of the ankle in healthy and rheumatoid arthritis post-menopausal women. *Gait posture*. **60**, 225–234 (2018).
5. Brown, P. Abnormal oscillatory synchronisation in the motor system leads to impaired movement. *Curr. Opin. Neurobiol.* **17**, 656–664 (2007).
6. Alexander, N. B. Postural Control in Older Adults. *J. Am. Geriatr. Soc.* **42**, 93–108 (1994).
7. Baby, M. S., Saji, A. J. & Kumar, C. S. Parkinsons disease classification using wavelet transform based feature extraction of gait data. In2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT) 2017 Apr 20 (pp. 1–6). IEEE. in.
8. Geritz, J. et al. Cognitive parameters can predict change of walking performance in advanced Parkinson's disease – Chances and limits of early rehabilitation. *Front. Aging Neurosci.* **14**, 1070093 (2022).
9. Rabey, J. M. & Korczyn, A. D. The Hoehn and Yahr rating scale for Parkinson's disease. InInstrumental methods and scoring in extrapyramidal disorders 1995 (pp. 7–17). Berlin, Heidelberg: Springer Berlin Heidelberg. in 10.1007/978-3-642-78914-4_2
10. AlMahadin, G. et al. Parkinson's disease: current assessment methods and wearable devices for evaluation of movement disorder motor symptoms - a patient and healthcare professional perspective. *BMC Neurol.* **20**, 419 (2020).
11. Hoehn, M. M., Yahr, M. D. & Parkinsonism *Neurology* **17**, 427–427 (1967).
12. Hausdorff, J. M., Cudkowicz, M. E., Firtion, R., Wei, J. Y. & Goldberger, A. L. Gait variability and basal ganglia disorders: stride-to-stride variations of gait cycle timing in Parkinson's disease and Huntington's disease. *Mov. Disord*. **13**, 428–437 (1998).
13. Lee, S. H. & Lim, J. S. Parkinson's disease classification using gait characteristics and wavelet-based feature extraction. *Expert Syst. Appl.* **39**, 7338–7344 (2012).
14. Bryant, M. S. et al. Gait variability in Parkinson's disease: influence of walking speed and dopaminergic treatment. *Neurol. Res.* **33**, 959–964 (2011).
15. Frenkel-Toledo, S. et al. Effect of gait speed on gait rhythmicity in Parkinson's disease: variability of stride time and swing time respond differently. *J. Neuroeng. Rehabil*. **2**, 23 (2005).
16. Hausdorff, J. M. Gait dynamics in Parkinson's disease: Common and distinct behavior among stride length, gait variability, and fractal-like scaling. *Chaos: Interdisciplinary J. Nonlinear Sci.* **19**, 026113 (2009).
17. Buckley, C. et al. The Role of Movement Analysis in Diagnosing and Monitoring Neurodegenerative Conditions: Insights from Gait and Postural Control. *Brain Sci.* **9**, 34 (2019).
18. Winter, D. A. Biomechanics and Motor Control of Human Movement. *Biomech. Motor Control Hum. Mov.* **64** 10.1002/9780470549148 (2009).
19. Khusainov, R., Azzi, D., Achumba, I. E. & Bersch, S. D. Real-Time Human Ambulation, Activity, and Physiological Monitoring: Taxonomy of Issues, Techniques, Applications, Challenges and Limitations. *Sensors* **13**, 12852–12902 (2013).
20. Hermez, L. et al. Clinical Gait Analysis: Characterizing Normal Gait and Pathological Deviations Due to Neurological Diseases. *Sens. (Basel Switzerland)*. **23**, 6566 (2023).
21. Ayoubi, F., Launay, C. P., Annweiler, C. & Beauchet, O. Fear of Falling and Gait Variability in Older Adults: A Systematic Review and Meta-Analysis. *J. Am. Med. Dir. Assoc.* **16**, 14–19 (2015).
22. König, N., Singh, N. B., von Beckerath, J., Janke, L. & Taylor, W. R. Is gait variability reliable? An assessment of spatio-temporal parameters of gait variability during continuous overground walking. *Gait Posture*. **39**, 615–617 (2014).
23. Terrier, P., Turner, V. & Schutz, Y. GPS analysis of human locomotion: Further evidence for long-range correlations in stride-to-stride fluctuations of gait parameters. *Hum. Mov. Sci.* **24**, 97–115 (2005).
24. Roemmich, R. T. et al. Spatiotemporal variability during gait initiation in Parkinson's disease. *Gait Posture*. **36**, 340–343 (2012).
25. Zhu, W. et al. Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. InProceedings of the AAAI conference on artificial intelligence. Mar 5 (Vol. 30, No. 1). (2016). https://doi.org/10.1609/aaai.v30i1.10451 doi:10.1609/aaai.v30i1.10451.
26. Alharthi, A. & Ozanyan, K. & IEEE. Deep Learning for Ground Reaction Force Data Analysis: Application to Wide-Area Floor Sensing. *2019 IEEE 28TH INTERNATIONAL SYMPOSIUM ON INDUSTRIAL ELECTRONICS (ISIE)* 1401–1406 (2019).
27. Brand, Y. E. et al. Gait Detection from a Wrist-Worn Sensor Using Machine Learning Methods: A Daily Living Study in Older Adults and People with Parkinson's Disease. *Sensors (Basel)* **22**, (2022).
28. He, H. & Garcia, E. A. Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng.* **21**, 1263–1284 (2009).
29. Saito, T. & Rehmsmeier, M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*. **10**, e0118432 (2015).
30. Steyerberg, E. et al. Assessing the Performance of Prediction Models A Framework for Traditional and Novel Measures. *Epidemiol. (Cambridge Mass)*. **21**, 128–138 (2010).
31. Mitchell, T. M. Machine learning and data mining. *Commun. ACM*. **42** (11), 30–36 (1999).
32. Lai, D. T. H., Begg, R. K. & Palaniswami, M. Computational intelligence in gait research: a perspective on current applications and future challenges. *IEEE Trans. Inform. Technol. biomedicine: publication IEEE Eng. Med. Biology Soc.* **13**, 687–702 (2009).

33. Heikal, M. & Eldawlatly, S. An ensemble classification technique of neurodegenerative diseases from gait analysis. In 2020 15th International Conference on Computer Engineering and Systems (ICCES) (pp. 1–6). IEEE. in (2020)., December 10.1109/ICCES51560.2020.9334609
34. Aich, S. et al. A Supervised Machine Learning Approach to Detect the On/Off State in Parkinson's Disease Using Wearable Based Gait Signals. *Diagnostics* **10**, 421 (2020).
35. Tahir, N. M. & Manap, H. H. Parkinson disease gait classification based on machine learning approach. *J. Appl. Sciences(Faisalabad)*. **12** (2), 180–185 (2012).
36. Parums, D. V. & Editorial Review Articles, Systematic Reviews, Meta-Analysis, and the Updated Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 Guidelines. *Med Sci Monit* 27, e934475-1-e934475-3 (2021).
37. Naimi, S., Bouachir, W. & Bilodeau, G. 1D-convolutional transformer for Parkinson disease diagnosis from gait. *Neural Comput. Appl.* **36**, 1931–1946 (2024).
38. Pei, X., Fan, H. & Tang, Y. Temporal pyramid attention-based spatiotemporal fusion model for Parkinson's disease diagnosis from gait data. *IET Signal Proc.* **15**, 80–87 (2021).
39. Ogata, T. et al. Foot Trajectory Features in Gait of Parkinson's Disease Patients. *Front. Physiol.* **13**, 726677 (2022).
40. Alharthi, A. S. Interpretable machine learning comprehensive human gait deterioration analysis. *Front Neuroinform* **18**, (2024).
41. Hwang, J. et al. Machine learning for early detection and severity classification in people with Parkinson's disease. *Sci. Rep.* **15**, 234 (2025).
42. Alharthi, A. S., Casson, A. J. & Ozanyan, K. B. Gait Spatiotemporal Signal Analysis for Parkinson's Disease Detection and Severity Rating. *IEEE Sens. J.* **21**, 1838–1848 (2021).
43. Choi, H. et al. Convolutional neural network based detection of early stage Parkinson's disease using the six minute walk test. *Sci. Rep.* **14**, 22648 (2024).
44. Caramia, C. et al. IMU-Based Classification of Parkinson's Disease From Gait: A Sensitivity Analysis on Sensor Location and Feature Selection. *IEEE J. biomedical health Inf.* **22**, 1765–1774 (2018).
45. Huo, H., Jiao, S., Li, D., Ma, L. & Xu, N. Efficient quantification of Parkinson's disease severity using augmented time-series data. *PLOS ONE*. **20**, e0319826 (2025).
46. Uchitomi, H., Ming, X., Zhao, C., Ogata, T. & Miyake, Y. Classification of mild Parkinson's disease: data augmentation of time-series gait data obtained via inertial measurement units. *Sci. Rep.* **13**, 12638 (2023).
47. Veeraragavan, S., Gopalai, A. A., Gouwanda, D. & Ahmad, S. A. Parkinson's Disease Diagnosis and Severity Assessment Using Ground Reaction Forces and Neural Networks. *Front. Physiol.* **11**, 587057 (2020).
48. Reyes, J., Montealegre, J., Castano, Y., Urcuqui, C. & Navarro, A. LSTM and Convolution Networks exploration for Parkinson's Diagnosis. *2019 IEEE COLOMBIAN CONFERENCE ON COMMUNICATIONS AND COMPUTING (COLCOM 2019)* (2019).
49. Downs, S. H. & Black, N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *J. Epidemiol. Community Health*. **52**, 377–384 (1998).
50. Viechtbauer, W. Conducting Meta-Analyses in R with the metafor Package. *J. Stat. Softw.* **36**, 1–48 (2010).
51. Moher, D., Liberati, A., Tetzlaff, J. & Altman, D. G. The PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Med.* **6**, e1000097 (2009).
52. Zhao, A., Qi, L., Li, J., Dong, J. & Yu, H. A hybrid spatio-temporal model for detection and severity rating of Parkinson's disease from gait data. *Neurocomputing* **315**, 1–8 (2018).
53. Zhao, H. et al. Severity level diagnosis of Parkinson's disease by ensemble K-nearest neighbor under imbalanced data. *Expert Syst. Appl.* **189**, 116113 (2022).
54. Muñoz-Mata, B. G., Dorantes-Méndez, G. & Piña-Ramírez, O. Classification of Parkinson's disease severity using gait stance signals in a spatiotemporal deep learning classifier. *Med. Biol. Eng. Comput.* **62**, 3493–3506 (2024).
55. Wang, Q., Zeng, W. & Dai, X. Gait classification for early detection and severity rating of Parkinson's disease based on hybrid signal processing and machine learning methods. *Cogn. Neurodyn.* **18**, 109–132 (2024).
56. Balaji, E., Brindha, D. & Balakrishnan, R. Supervised machine learning based gait classification system for early detection and stage classification of Parkinson's disease. *APPLIED SOFT COMPUTING* **94**, (2020).
57. Xia, Y., Yao, Z., Ye, Q., Cheng, N. A. & Dual-Modal Attention-Enhanced Deep Learning Network for Quantification of Parkinson's Disease Characteristics. *IEEE Trans. neural Syst. rehabilitation engineering: publication IEEE Eng. Med. Biology Soc.* **28**, 42–51 (2020).
58. Vidya, B. & Sasikumar, P. Parkinson's disease diagnosis and stage prediction based on gait signal analysis using EMD and CNN–LSTM network. *Eng. Appl. Artif. Intell.* **114**, 105099 (2022).
59. E, B., Elumalai, D. B., K, U. & V. K. & Data-driven gait analysis for diagnosis and severity rating of Parkinson's disease. *Med. Eng. Phys.* **91**, 54 (2021).
60. Balakrishnan, A., Medikonda, J., Pramod, K. & Natarajan, M. Information Set-Based Decision Tree for Parkinson's Disease Severity Assessment Using Multidimensional Gait Dataset. *IEEE Access.* **12**, 129187–129201 (2024).
61. Setiawan, F. & Lin, C. W. Implementation of a Deep Learning Algorithm Based on Vertical Ground Reaction Force Time–Frequency Features for the Detection and Severity Classification of Parkinson's Disease. *Sensors* 21, (2021).
62. Balakrishnan, A., Medikonda, J., Namboothiri, P. K. & Natarajan, M. Mahalanobis Metric-based Oversampling Technique for Parkinson's Disease Severity Assessment using Spatiotemporal Gait Parameters. *Biomed. Signal Process. Control*. **86**, 105057 (2023).

63. Mittal, R. et al. Machine learning approach to gait analysis for Parkinson's disease detection and severity classification. *Front. Rob. AI*. **12**, 1623529 (2025).
64. Frenkel-Toledo, S. et al. Treadmill walking as an external pacemaker to improve gait rhythm and stability in Parkinson's disease. *Mov. Disord.* **20**, 1109–1114 (2005).
65. Yogev, G. et al. Dual tasking, gait rhythmicity, and Parkinson's disease: Which aspects of gait are attention demanding? *Eur. J. Neurosci.* **22**, 1248–1256 (2005).
66. Hausdorff, J. M. et al. Rhythmic auditory stimulation modulates gait variability in Parkinson's disease. *Eur. J. Neurosci.* **26**, 2369–2375 (2007).
67. Yin, W. et al. Gait analysis in the early stage of Parkinson's disease with a machine learning approach. *Front Neurol* **15**, (2024).
68. Djurić-Jovičić, M., Belić, M., Stanković, I., Radovanović, S. & Kostić, V. S. Selection of gait parameters for differential diagnostics of patients with *de novo* Parkinson's disease. *Neurol. Res.* **39**, 853–861 (2017).
69. Mirelman, A. et al. Detecting Sensitive Mobility Features for Parkinson's Disease Stages Via Machine Learning. *Mov. Disord*. **36**, 2144–2155 (2021).
70. Li, X. et al. Network Based Classification Method for Mild to Moderate Parkinson's Disease at Turns. in *12th Asian-Pacific Conference on Medical and Biological Engineering* (eds Wang, G. Springer Nature Switzerland, Cham, 371–378 (2024). 10.1007/978-3-031-51455-5_41
71. Tao, S., Wang, H., Kong, L., Lv, Z. & Wang, Z. Testing the Validity of a Spatiotemporal Gait Model Using Inertial Measurement Units in Early Parkinson's Patients. *IEEE ACCESS.* **11**, 80573–80587 (2023).
72. Meng, L. et al. Inertial-Based Gait Metrics During Turning Improve the Detection of Early-Stage Parkinson's Disease Patients. *IEEE Trans. Neural Syst. Rehabil. Eng.* **31**, 1472–1482 (2023).
73. Cai, G. et al. Specific Distribution of Digital Gait Biomarkers in Parkinson's Disease Using Body-Worn Sensors and Machine Learning. *J. Gerontol. Biol. Sci. Med. Sci.* **78**, 1348–1354 (2023).
74. Lin, C. H. et al. Early Detection of Parkinson's Disease by Neural Network Models. *IEEE Access.* **10**, 19033–19044 (2022).
75. Uchitomi, H., Ming, X., Zhao, C., Ogata, T. & Miyake, Y. Classification of mild Parkinson's disease: data augmentation of time-series gait data obtained via inertial measurement units. *Sci. Rep.* **13**, 12638 (2023).
76. Balaji, E., Brindha, D. & Balakrishnan, R. Supervised machine learning based gait classification system for early detection and stage classification of Parkinson's disease. *Appl. Soft Comput.* **94**, 106494 (2020).
77. Hooper, P., Jutai, J. W., Strong, G. & Russell-Minda, E. Age-related macular degeneration and low-vision rehabilitation: a systematic review. *Can. J. Ophthalmol.* **43**, 180–187 (2008).
78. Veeraragavan, S., Gopalai, A., Gouwanda, D. & Ahmad, S. Parkinson's Disease Diagnosis and Severity Assessment Using Ground Reaction Forces and Neural Networks. *FRONTIERS PHYSIOLOGY* **11**, (2020).
79. Altman, D. G. & Bland, J. M. Diagnostic tests 2: Predictive values. *BMJ* **309**, 102 (1994).
80. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
81. Chen, T., Guestrin, C. & XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794Association for Computing Machinery, New York, NY, USA, (2016). 10.1145/2939672.2939785
82. Alharthi, A., Yunas, S. U. & Ozanyan, K. Deep Learning for Monitoring of Human Gait: A Review. *Inst. Electr. Electron. Eng. Sens. J.* **19**, 9575–9591 (2019).
83. Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86, 2278–2324 (1998).
84. Aznar-Gimeno, R. et al. Deep Learning for Walking Behaviour Detection in Elderly People Using Smart Footwear. *Entropy (Basel Switzerland)* **23**, (2021).
85. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **9**, 1735–1780 (1997).
86. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
87. Abdullah, T. A. A., Zahid, M. S. M. & Ali, W. A. Review of Interpretable ML in Healthcare: Taxonomy, Applications, Challenges, and Future Directions. *Symmetry* **13**, 2439 (2021).
88. Gregor, S. et al. Interpretability of machine learning-based prediction models in healthcare. *WIREs Data Min. Knowl. Discov.* **10**, e1379 (2020).
89. Slijepcevic, D. et al. Explaining Machine Learning Models for Clinical Gait Analysis. *ACM Trans. Comput. Healthc.* **3** (1–14), 27 (2021).
90. Alharthi, A. S. Interpretable machine learning comprehensive human gait deterioration analysis. *Front. Neuroinformatics*. **18**, 1451529 (2024).
91. Rupprechter, S. et al. A Clinically Interpretable Computer-Vision Based Method for Quantifying Gait in Parkinson's Disease. *Sensors* **21**, 5437 (2021).
92. Sánchez-Fernández, L. P. Biomechanics of Parkinson's Disease with Systems Based on Expert Knowledge and Machine Learning: A Scoping Review. *Computation* **12**, 230 (2024).
93. di Biase, L. et al. Gait Analysis in Parkinson's Disease: An Overview of the Most Accurate Markers for Diagnosis and Symptoms Monitoring. *Sensors* **20**, 3529 (2020).
94. Saboor, A. et al. Latest Research Trends in Gait Analysis Using Wearable Sensors and Machine Learning: A Systematic Review. *IEEE Access.* **8**, 167830–167864 (2020).
95. Goldberger, A. L. et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation* **101** (23), e215–e220 (2000).
96. Gupta, U., Bansal, H. & Joshi, D. Gender specific and age dependent classification model for improved diagnosis in Parkinson's disease. arXiv preprint arXiv:1904.09651. 2019;10.

97. Sofuwa, O. et al. Quantitative Gait Analysis in Parkinson's Disease: Comparison With a Healthy Control Group. *Arch. Phys. Med. Rehabil.* **86**, 1007–1013 (2005).
98. Alton, F., Baldey, L., Caplan, S. & Morrissey, M. C. A kinematic comparison of overground and treadmill walking. *Clin. Biomech. Elsevier Ltd*. **13**, 434–440 (1998).
99. Semaan, M. B. et al. Is treadmill walking biomechanically comparable to overground walking? A systematic review. *Gait Posture*. **92**, 249–257 (2022).
100. Bello, O., Sánchez, J. A. & Vazquez-Santos, C. Fernandez-Del-Olmo, M. Spatiotemporal Parameters of Gait During Treadmill and Overground Walking in Parkinson's Disease. *J. Parkinson's Disease*. **4**, 33–36 (2014).
101. Hausdorff, J. M. et al. Impaired regulation of stride variability in Parkinson's disease subjects with freezing of gait. *Exp. Brain Res.* **149**, 187–194 (2003).
102. Song, S. et al. Deep reinforcement learning for modeling human locomotion control in neuromechanical simulation. *J. Neuroeng. Rehabil*. **18**, 126 (2021).
103. Wan, S., Guan, S. & Tang, Y. Advancing Bridge Structural Health Monitoring: Insights into Knowledge-Driven and Data-Driven Approaches. *J. Data Sci. Intell. Syst.* **2**, 129–140 (2024).

Figures

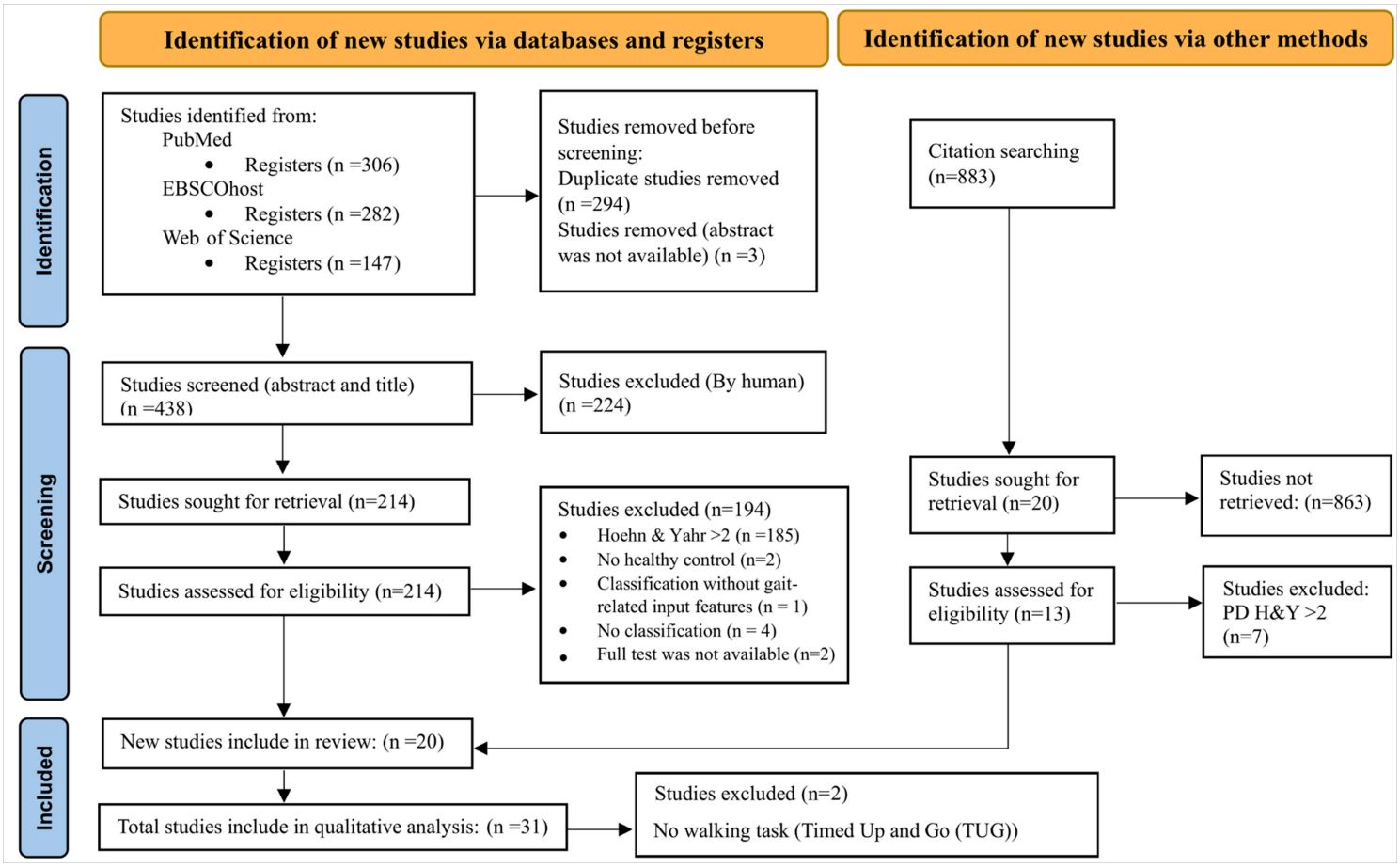


Figure 1

PRISMA flow diagram

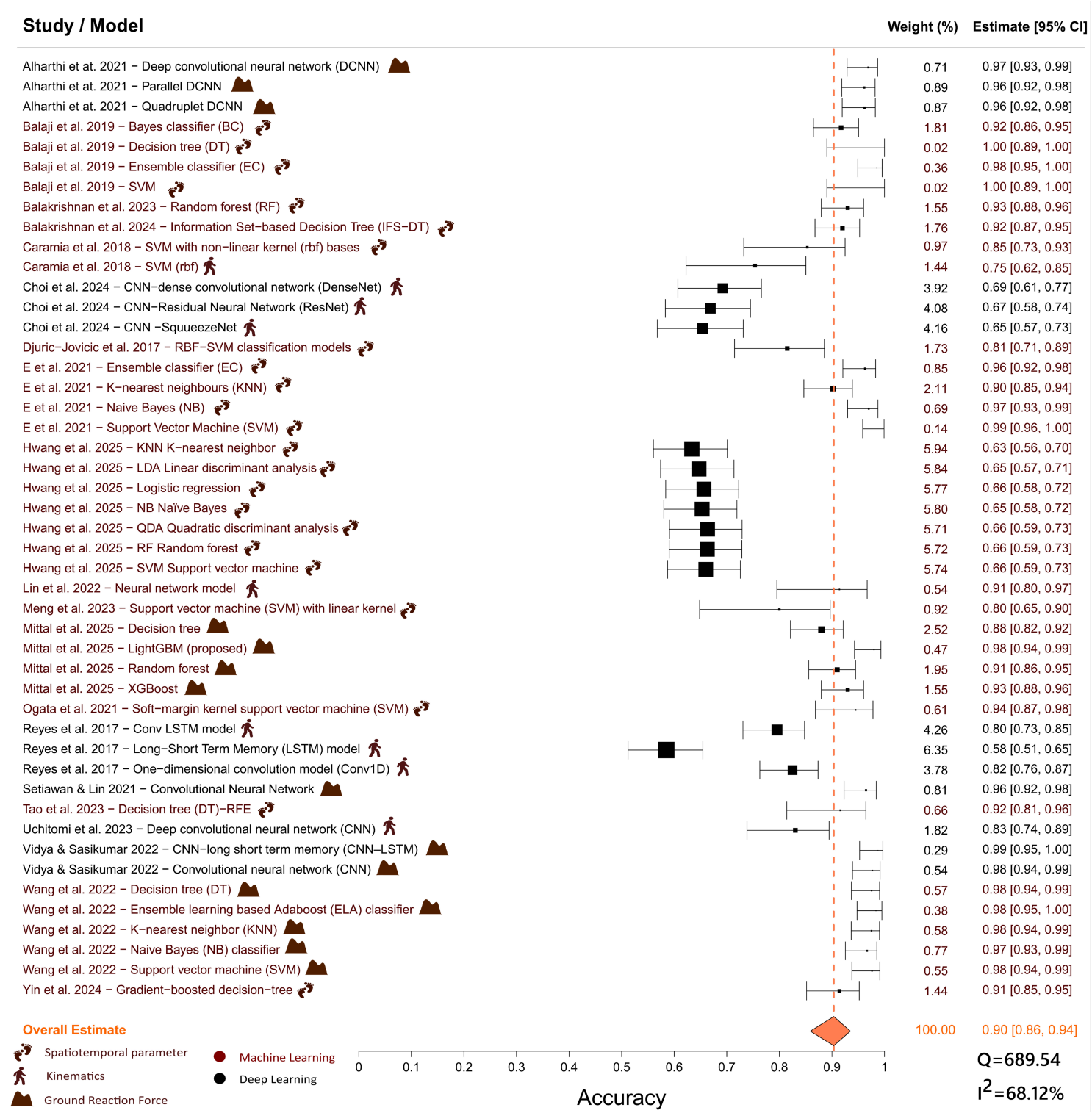


Figure 2

Forest plot for Accuracy

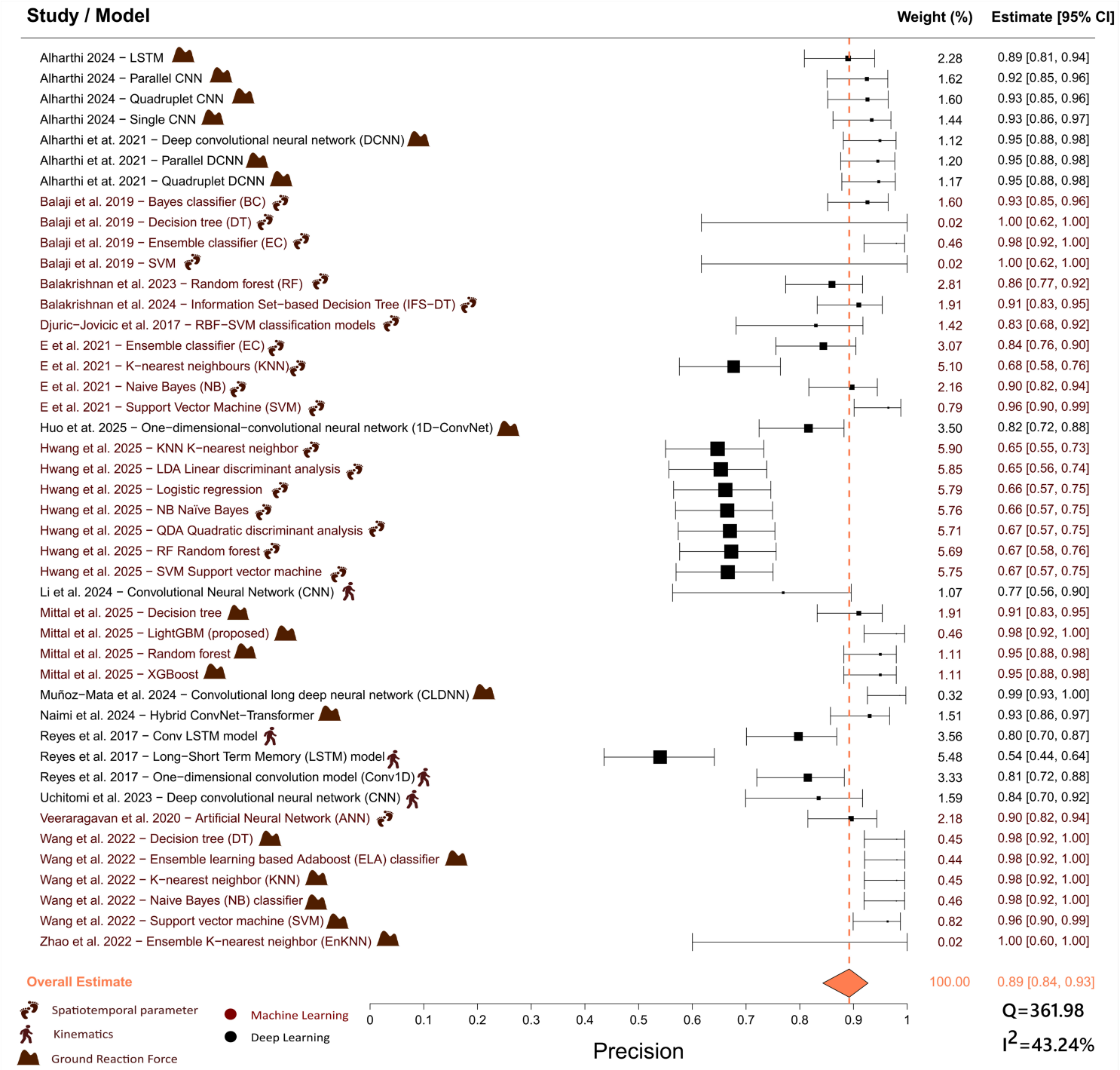


Figure 3

Forest plot for Precision

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation1.docx
- SupplementaryInformation2.xlsx