Gait & Posture 109 (2024) 259–270

Contents lists available at ScienceDirect

Gait & Posture

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

Review

Systematic review of automatic post-stroke gait classification systems

Yiran Jiao, Rylea Hart, Stacey Reading, Yanxin Zhang [*]

Department of Exercise Sciences, Faculty of Science, University of Auckland, Auckland 1023, New Zealand

ARTICLE INFO

Keywords:
Gait analysis
Gait assessment
Hemiplegic gait
Machine learning

ABSTRACT

Background: Gait classification is a clinically helpful task performed after a stroke in order to guide rehabilitation therapy. Gait disorders are commonly identified using observational gait analysis in clinical settings, but this approach is limited due to low reliability and accuracy. Data-driven gait classification can quantify gait deviations and categorise gait patterns automatically possibly improving reliability and accuracy; however, the development and clinical utility of current data driven systems has not been reviewed previously.
Research question: The purpose of this systematic review is to evaluate the literature surrounding the methodology used to develop automatic gait classification systems, and their potential effectiveness in the clinical management of stroke-affected gait.
Method: The database search included PubMed, IEEE Xplore, and Scopus. Twenty-one studies were identified through inclusion and exclusion criteria from 407 available studies published between 2015 and 2022. Development methodology, classification performance, and clinical utility information were extracted for review.
Results and significance: Most of gait classification systems reported a classification accuracy between 80%–100%. However, collated studies presented methodological errors in machine learning (ML) model development. Further, many studies neglected model components such as clinical utility (e.g., predictions don't assist clinicians or therapists in making decisions, interpretability, and generalisability). We provided recommendations to guide development of future post-stroke automatic gait classification systems to better assist clinicians and therapists. Future automatic gait classification systems should emphasise the clinical significance and adopt a standardised development methodology of ML model.

Introduction

Stroke is one of the primary causes of mortality and morbidity with 6.2 million fatal cases of stroke reported globally in 2017 [1]. Stroke survivors are typically left with impairment of functional capacity due to upper motor neuron (UMN) syndrome. As many as 80% of stroke survivors are affected by motor impairment that disrupts normal gait [2]. Stroke survivors have an increased risk of falls, reduced functional independence, and reduced quality of life. Thus, re-gaining lost gait function is a primary goal for many patients following stroke.

In the past few decades, instrumented gait analysis systems have been widely used to quantify gait deviations based on quantitative motion data (e.g., kinematic, EMG data). Furthermore, incorporating quantitative gait data and artificial intelligence allows for the development of automatic gait analysis systems (i.e., technology-driven tools designed to quantitatively assess gait performance). Early automatic gait analysis systems utilised rule-based expert systems to identify abnormal gait patterns using quantitative data input and prior knowledge. For example, one such system, QUAWDS, identified abnormal gait when lower extremity rotation deviated by more than 10° compared to normal gait patterns [3,4]. However, automatic gait analysis systems have not been widely used in clinical settings for post-stroke gait assessment for several reasons. First, rule-based systems require complex and extensive domain-specific knowledge. Second, identifying complex movement patterns requires complex rules which can be hard to extract and articulate [5]. Finally, few systems have been designed for patients following stroke [6] and attempting to apply generic systems to all neurological diseases is overly ambitious.

Recently, data-driven learning techniques such as machine learning (ML) models have been applied to explore and characterise gait patterns (e.g., gait classification tasks) to aid the clinical assessment of gait [7–9]. Gait classification requires analysis and categorisation of patients' gait patterns into clinically significant groups , which is a clinically meaningful task for the treatment of neurological diseases (e.g., stroke,

[*] Correspondence to: Building 907 - Bldg 907, Level 2, Room 235, 368 Khyber Pass, Newmarket, Auckland 1023, New Zealand.
E-mail address: yanxin.zhang@auckland.ac.nz (Y. Zhang).

https://doi.org/10.1016/j.gaitpost.2024.02.011
Received 19 May 2023; Received in revised form 11 January 2024; Accepted 12 February 2024
Available online 15 February 2024

cerebral palsy). For example, identifying different gait patterns (e.g., equinus gait) can help clinicians determine mechanisms behind walking impairment (e.g., muscle imbalance between the posterior tibialis muscle and the peroneus muscle) and generate specific surgical and rehabilitation strategies [10,11]. When compared to traditional expert systems, data-driven techniques could both explore potential unseen rules based on data mining and improve efficiency since they usually allow parallel processing [5]. Therefore, the development of automatic gait classification systems (i.e., automatic systems performing gait classification tasks specifically) may be an inevitable trend within clinical settings.

Although data-driven models have been applied in developing post-stroke automatic gait classification systems [12,13], it is not certain whether these models are well-developed and provide clinically meaningful information. A clinically applicable automatic gait classification system requires both clinical utility and correct learning algorithm development. Correct procedures around learning algorithm development can promote classification accuracy while considering clinical utility promotes the interpretability and generalisability of the developed system. However, previous reviews in this area only focused on learning algorithm/technique development [12–19] (e.g., selected algorithms and application of inertial sensors, biofeedback devices, and assistive robots) and have not critically reviewed the developed models in terms of both ML model development and clinical utility.

Thus, this paper will review and evaluate the literature surrounding the development of automatic gait classification systems and their effectiveness/utility in the clinical management of stroke-affected gait. Common issues would be identified regarding ML model development and clinical effectiveness. Based on this, our review will provide recommendations for the development of automatic gait classification systems for patients following stroke.

Methods

Systematic search strategy

A systematic search was conducted within three databases, PubMed, IEEE Xplore, and Scopus, using the PICO process in March 2022. All authors reviewed search terms. The final search terms used for this review are shown in Table 1. Relevant truncation symbols were used to retrieve all possible suffix variations of a root word. Due to the different database user interfaces, studies involving these keywords in the field of "Text Word" (in PubMed), "ALL Metadata" (in IEEE Xplore), and “Article title Abstract and Keywords” (in Scopus) are identified. A targeted search, including cross-referencing, was performed to complement the electronic search.

Table 1
Search terms.

Subject	Search terms
Patient	(stroke OR poststroke OR "post stroke" OR hemiplegic) AND
Method ("Intervention" was regarded as "Method" here)	("principal component*" OR "feature selection" OR "regression model*" OR "machine learning" OR "decision support system*" OR "automatic data processing" OR "artificial intelligence" OR "neural network*" OR "deep learning" OR "data mining" OR "Intelligent data analys*" OR clustering OR "cluster analysis" OR "support vector" OR "random forest*" OR "naive Bayes" OR k-nearest OR "k means" OR "decision tree*" OR "reinforcement learning" OR Bayesian) AND
Control	not applicable
Outcome	AND (gait OR walking) AND (diagnosis OR classif* OR assessment OR recognition OR "gait analysis" OR quantif* OR evaluat* OR detect* OR "gait pattern*" OR categor*)

Literature screen

The screening and eligibility assessment is described in detail in Fig. 1. The two study reviewers (YJ and RH) are PhD students with expertise in clinical biomechanics and machine learning techniques. The third reviewer (YZ) is an expert in the application of data-driven techniques in clinical gait analysis. The initial search was based on three electronic databases, led by the primary reviewer (YJ), revealing a total of 407 identified articles after removing duplicates. All studies identified from the databases were uploaded to Endnote 20 (Clarivate PLC, US) for screening and review. Then two reviewers (YJ and YZ) independently screened records based on the titles and abstracts using the inclusion criteria and exclusion criteria below. The remaining records from two reviewers were imported into Endnote 20 and the full texts were downloaded. YJ and YZ read the full text of the remaining studies independently and then had a face-to-face discussion to decide on the studies included. Any discrepancies between them were addressed with the involvement of the third reviewer (RH), fostering resolution through discussion among the three reviewers.

Inclusion criteria

Peer-reviewed journal articles published in English from 2015 onwards, satisfying the following conditions: (1) studies with data from patients following stroke, (2) studies mainly on gait classification, namely allocating gait patterns into homogeneous groups, categories, or clusters based on gait-related variables [20]; (3) studies utilising automatic methods or systems (e.g., machine learning) for gait classification instead of manual classification by therapists or experts; (4) studies using numerical data from intelligent facilities (e.g., kinematic, kinetic, spatiotemporal, bioelectric signal, video data) as input parameter.

Exclusion criteria

Excluded studies encompassed the following: (1) studies solely focused on healthy subjects simulating stroke gait patterns, groups with other neurological diseases, or mixed groups involving stroke and other diseases; (2) studies targeting posture recognition, activity recognition, gait event distinction, and discrimination between paretic and non-paretic limbs; (3) non-automatic studies relying on manual classification by clinicians; (4) studies with non-gait related input parameters (e.g., blood tests), only results from gait functional tests (e.g., total time of timed-up-and-go test) or non-gait functional activities (such as [21]); (5) studies with only quantitative measurements of gait deviation such as indices but lacking classification outcomes; (6) studies involving obstacle walking, prosthetic legs, lower-limb exoskeleton robots, and data from robotic rollators [22].

Data extraction

Information about ML model development and clinical utility was extracted to answer two research aims: (1) evaluate whether existing automatic gait classification systems have been developed using appropriate data-driven learning techniques, and (2) whether the classification results (i.e., groups/clusters) have clinical utility. The general characteristics extracted consisted of authors, year of publication, sample information about participants number, age, gender, and stroke type.

Considering the context of post-stroke gait analysis and utilising recommendations from previous reviews on machine learning in human movement biomechanics [23,24], we identified several characteristics to answer the first research aim. This included sample information, feature engineering (feature type, feature extraction method), model learning (feature selection, hyperparameter tuning, best-performed classification algorithm), and performance evaluation (validation procedure and classification performance). Due to the variety of possible

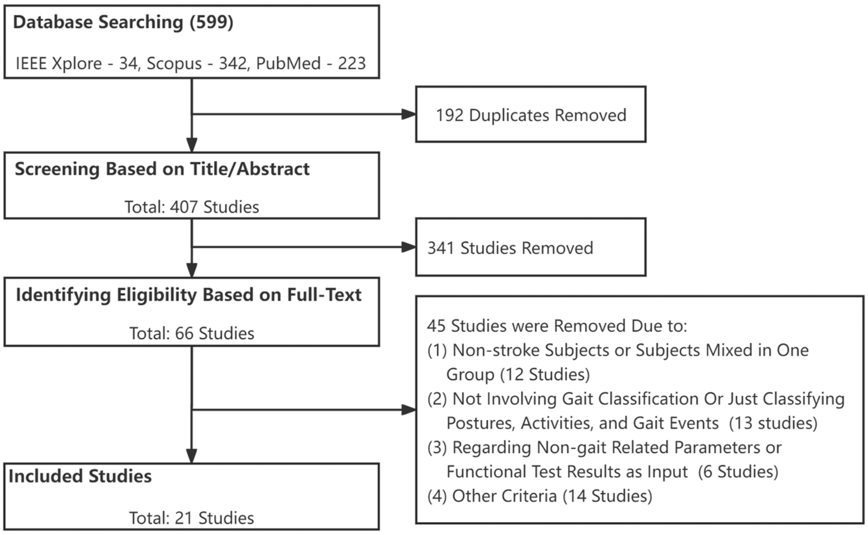
Database Searching (599)
IEEE Xplore - 34, Scopus - 342, PubMed - 223
192 Duplicates Removed
Screening Based on Title/Abstract
Total: 407 Studies
341 Studies Removed
Identifying Eligibility Based on Full-Text
Total: 66 Studies
45 Studies were Removed Due to:
(1) Non-stroke Subjects or Subjects Mixed in One Group (12 Studies)
(2) Not Involving Gait Classification Or Just Classifying Postures, Activities, and Gait Events (13 studies)
(3) Regarding Non-gait Related Parameters or Functional Test Results as Input (6 Studies)
(4) Other Criteria (14 Studies)
Included Studies
Total: 21 Studies

Fig. 1. Flow diagram of review.

ML techniques, the reviewed studies were categorised into either supervised ML or unsupervised ML based on their level of supervision [23].

Since clinical utility is not discussed in previous reviews, the authors of this study (including one clinical exercise expert SR) discussed and developed several characteristics to answer research aim 2. These characteristics are data collection methods (device and sampling), classification type (i.e., multi-class/binary-class [25], multi-label/single-label [26], global/individual classifier [24]), and output groups. The definition of classification types was based on previous studies [24–26], and the definition of sampling in this study is described in Section 2.6. Data collection methods, classification type, and output groups could reflect the classification output utility and feasibility. The sample information extracted previously, and sampling could also affect the generalisability of ML models, which were included in the discussion of ML model clinical utility.

Based on the consensus of authors, the most pivotal criterion for clinical utility is based on the clinical indication of predictions or outputs provided by the ML models. Reviewed studies were then grouped into three categories based on the outputs they provided: differentiating stroke and normal gait (category 1), differentiating gait related to stroke from gait related to other diseases (category 2), and identifying homogeneous gait subgroups within datasets of patients with stroke (category 3). Finally, two independent reviewers participated in the extraction of the data from the included paper in this review.

Quality assessment

Considering the absence of the validated quality measure for machine learning studies in this field, we set minimal requirements for methodological aspects crucial for result reproducibility [27]. These were based on the context of gait analysis and recommendations from previous studies regarding machine learning in human movement [23,24,27,28]. Studies with more than four adequate aspects were regarded as having an overall adequate quality. Two independent reviewers critically appraised the methodological quality of the included studies. In the event of a disagreement, a discussion between the two reviewers took place to reach an agreement.

1. Number of Events: The adequate number of events used in ML model development is >200 (100 events and 100 non-events [27,28].).
2. Sampling: Utilising multiple data sequences from a single participant within ML model training is considered inappropriate due to the high repeatability of gait patterns [24,29].
3. Feature selection: Feature selection is an important step for improving model computational efficiency with reduced risks of overfitting data [23,30]. Studies including the feature selection step were considered as adequate feature selection.
4. Algorithm Interpretability: Interpretability is important for medicine and health care [31]. The interpretability of algorithms was determined based on a previous study [32].
5. Validation: A study with an external validation procedure using a separate dataset is recognised as an adequately validated study [24].
6. Performance: A study with at least three performance evaluation metrics is recognised as a study with adequate performance evaluation metrics [24].

For data analysis of clinical utility, there are no standardised assessment criteria to justify whether the automatic gait classification systems are suitable in clinics. Therefore, the detailed characteristics of clinical utility mentioned before will be discussed in the Discussion.

Results

Literature search results

A total of 407 studies (excluding duplicates) were identified through database searching (see Fig. 1). Two independent reviewers disagreed on whether the two studies met the inclusion criteria (2) based on full-text screening. After discussion and involvement of the third reviewer, we deleted these two and a final tally of 21 studies was retained for eligible reports. Based on our previous definitions, all reviewers had agreement with ten studies fitting in category 1, three studies in category 2, and eight studies in category 3. Sixteen of the included studies used supervised ML, while five used unsupervised ML techniques, namely cluster analysis (CA). Information about general characteristics, clinical aspects, and characteristics of ML model development are presented in Table 2 and Table 3, respectively. Table 4 presents the quality assessment results of collated studies.

Machine learning model development

Sample information

The number of participants recruited in the majority of the reviewed studies was small with half including 20 or fewer participants [7,11,33–40]. The remaining studies recruited 30 or more patients with stroke [8,41–48]. Eight of the included studies (n = 21) had sample sizes of at least 200 events (100 events and 100 non-events), as suggested [8,11,35,

Table 2
Summary of experimental protocols and participant characteristics.

Study	Participant: Number (age-Mean±SD years, sex, stroke type)	Device (position) and Sampling	Classification Type and Groups
Category 1: Healthy people vs Stroke people			
(Iosa et al.,2021) [33]	SG: 16 participants (age: 54.6 ± 13.7, sex: NS, stroke type: NS) HG: 17 participants (age: 45.7 ± 13.4, sex: NS, stroke type: NS)	1 IMU on waist	Multi-class (HG vs SG vs SR), single-label, global classifier
(Li et al.,2019) [34]	SG: 15 participants (mean age 46, sex: 9M6F, stroke type: NS) HG:15 (mean age 29, sex: 9M6F, stroke type: NS)	EFS system; Multiple data sequences collected from one subject	Binary-class (HG vs SG), single-label, global classifier
(Hussain and Park, 2021) [41]	SG: 48 participants (age: 72.2±5.6, sex: 62%M, stroke type: ischemic) HG: 75 participants (age: 77, sex: 31% M, stroke type: ischemic)	EMG sensors (legs); Multiple data sequences collected from one subject	Binary-class (HG vs SG), single-label, global classifier
(Cui, 2018) [35]	SG: 21 participants (age: 47.88 ± 12.32, sex: 5F16M, stroke type: NS) HG: 21 participants (age: 47.38 ± 11.09, sex: 5F16M, stroke type: NS)	EMG sensors (legs), Qualysis system, force plate; Multiple data sequences collected from one subject	Binary-class (HG vs SG), single-label, global classifier
(Lee et al., 2018) [36]	SG: 20 participants (age: 63.2 ± 8.9, sex: 7F13M, stroke type: NS) HG: 15 participants (age: 59.7 ± 11.9, sex: 8F7M, stroke type: NS)	1 IMU (waist); Multiple data sequences collected from one subject	Binary-class (HG vs SG), single-label, global classifier
(Hsu et al., 2021) [37]	SG: 12 participants (age: 63.4 ± 6.9, sex: NS, stroke type: NS)HG: 7 participants (age: 23.5 ± 1.9, sex: NS, stroke type: NS), 4 (70.5 ± 5.3, sex: NS, stroke type: NS)	1 IMU (waist), Qualysis system, force plate Multiple data sequences collected from one subject	Binary-class (HG vs SG), single-label, global classifier
(Mathur and Bhatia, 2022) [42]	SG: 40 participants (same age group, sex: NS, stroke type: NS) HG: 40 participants (same age group, sex: NS, stroke type: NS)	17 IMUs (upper and lower limb)	Binary-class (HG vs SG), single-label, global classifier
(Altilio et al., 2021) [49]	SG: 25 participants (same age group, sex: NS, stroke type: NS) HG: 35 participants (same age group, sex: NS, stroke type: NS)	A smartphone device (shanks)	Binary-class (HG vs SG), single-label, global classifier
(Choi et al., 2021) [43]	SG: 61 participants (age: ≥65, sex: NS, stroke type: NS) HG: 61 participants (age:≥65, sex: NS, stroke type: NS)	EEG sensors (head); Multiple data sequences collected from one subject	Binary-class (HG vs SG), single-label, global classifier
(Hussain and Park, 2021) [44]	SG: 48 participants (age: 72.2 ± 5.6, sex: 30M18F, stroke type: ischemic) HG: 75 participants (age: 77, sex: 23M52F, stroke type: ischemic)	EEG sensors (head)	Binary-class (HG vs SG), single-label, global classifier
Category 2: Stroke people vs People with other diseases			
(Hsu et al., 2018) [38]	SG: 11 participants (age: 65.2 ± 13.7 yrs, sex: NS, stroke type: NS) OG: 9 participants (age: 66.4 ± 9.16 yrs, sex: NS, stroke type: NS)	7 IMUs (waist and lower limb)	Binary-class (OG vs SG), single-label, global classifier
(Wang et al., 2020) [7]	SG:13 participants (age: 61±15 yrs, sex: 9M4F, stroke type: NS) PD: 15 participants (age: 76±7 yrs, sex: 9M6F, stroke type: NS) PN: 8 participants (age: 40±8, sex: 3M5F, stroke type: NS) HG: 13 participants (age: 49±20, sex: 7 M6F, stroke type: NS)	2 IMUs (shank)	Multi-class (HG vs PD vs PN vs SG), single-label, global classifier
(Mannini et al., 2016) [39]	SG: 15 participants (age: 61.3±13, sex: 5F10M, stroke type: NS) HD: 17 participants (age: 54.3±12.2, sex: 7F10M, stroke type: NS) HG: 10 participants (age: 69.7±5.8, sex: 6F4M, stroke type: NS)	3 IMUs (shanks and waist), force mat	Multi-class (SG vs HD vs HG), single-label, global classifier
Category 3: Identifying homogeneous gait subgroups within stroke patient datasets			
(Wang et al.,2021) [11]	SG: 8 participants (age: 58.75±9, sex: 6M2F, stroke type: NS) HG:7 participants (age: 24.43±1, sex: 7 M, stroke type: NS)	2 IMUs (shanks); Multiple data sequences collected from one subject	Binary-class (SG vs HG), multi-class (subgroups of stroke: the drop-foot gait, circumduction gait, hip hiking gait, and back knee gait.), multi-label, global classifier
(Lee et al., 2021) [8]	SG: 206 participants (age: 63.24 ± 14.36, sex: 108M98F,	Smartphone camera	Binary-class (subgroups of stroke: IA vs DA), single-label, global classifier

(*continued on next page*)

Table 2 (*continued*)

Study	Participant: Number (age-Mean±SD years, sex, stroke type)	Device (position) and Sampling	Classification Type and Groups
	Stroke type was recorded)		
(Punt et al., 2017) [45]	SG: 40 participants (non-fallers age: 58.4±14.3, fallers age: 64.6±8.5, sex: 24M16F, stroke type: NS)	Vicon system, force platforms	Binary-class (subgroups of stroke: non-fallers vs fallers), single-label, global classifier
(Sekiguchi et al., 2021) [46]	SG: 72 participants (cluster 1 age: 52.6 ±9.5, cluster 2 age: 55.7±12.2, cluster 3 age: 58.5 ±11.4, sex: 53M19F, stroke type: NS)	Motion system, force plates	Multi-class (subgroups of stroke: 3 clusters), single-label, global classifier
(Tan et al., 2019) [9]	SG: 30 participants (age: NS, sex: NS, stroke type: NS)	Force sensitive resistors (toe and heel), 2 IMUs (shanks and feet), EMG system	Multi-class (subgroups of stroke: 3 clusters), single-label, global classifier
(Dolatabadi et al., 2017) [47]	SG: 68 participants (age: 61.5 ± 13.5, sex: 42F26M, stroke type: NS) HG: 20 participants (age: 8.8± 7.1, sex: 10F10M, stroke type: NS)	Force mat	Multi-class (subgroups of stroke: 3 clusters), single-label, global classifier
(PAUK et al., 2016) [40]	SG: 18 participants (age: 44.4±17.1, sex: 66.7%M, stroke type: NS)	Motion system, force platforms; Multiple data sequences collected from one subject	Multi-class (subgroups of stroke: 3 biclusters), multi-label, global classifier
(PAUK et al., 2016) [48]	SG: 41 participants (age: 48.6±19.6, sex: 48.8%M, stroke type: NS)	Motion system, force platforms; Multiple data sequences collected from one subject	Multi-class (subgroups of stroke: 3 biclusters), multi-label, global classifier

BBS: berg balance scale; DA: dependent ambulation; EFS system: electrostatic field sensing; HD: Huntington's disease; HG: healthy group; Multi-label classification: one instance can be classified into more than one category; FAC: functional ambulation category; MMSE: Mini-Mental State Examination; IA: independent ambulation; M/ F: male and female respectively; NS: not specified; OG: group with other diseases; PD: Parkinson's disease; PN: peripheral neuropathy; SG: stroke group; SNR: stroke people that have not returned to work; SR: stroke group that returns to work.

41,43,44,46,47].

Several studies in categories 1 and 2 did not mention participants' stroke type (e.g., ischemic or haemorrhagic) (10/13) [7,33,35–39,42,43,49], participant age (4/13) [34,42,43,49], or recruit age-matched control groups into their study (6/13) [7,34,37–39,49]. Two studies controlled gender ratios between groups [34,35]. Seven studies in category 3 similarly did not record stroke type [9,11,40,45–48].

Feature engineering

Feature types utilised by studies in this review included spatio-temporal data [7–9,33,38,42,45,47,49]), kinematic data [7,9,11,33,35,45], kinetic data [35,40,46,48], bioelectric signals (i.e., EEG, EMG) [35,41,43,44], and others (i.e., time-frequency-power features, video data) [8,34,36–39,49]. Seven studies employed multi-modality feature sets [8,9,33,35,38,39,49] while four studies compared multi-modality feature sets and the single-modality feature set gait classification. Two studies performed automatic feature reduction (i.e., Principal component analysis) through automatic method [35,45], while other studies extracted features based on prior knowledge.

Model learning

Feature selection was only performed by six studies from categories 1 (n = 10) and 2 (n = 3) [34,36–38,41,44]. Five studies used statistics-based feature selection methods [34,37,38,41,44], while eight studies compared classification performance between different feature sets [8,9,35,38,39,43,45,49]. Eight of reviewed papers utilised some forms of hyperparameter tuning, including three supervised ML papers [36,37,39] and all CA papers [9,40,46,48]. However, standard hyper-parameter tuning methods [23] (e.g., grid search or Bayesian optimisation) were not specified in four of them [36,37,40,48].

Studies within this review focussed on either supervised classification (16/21) [7,8,11,33–39,41–45,49] or unsupervised cluster analysis (CA) (5/21) [9,40,46–48]. Commonly adopted supervised classification algorithms with the best performance in reviewed papers included Support Vector Machines (SVM) [7,35,37,39], k-nearest neighbours (kNN) [34], Random Forests (RF) [36,43], Neural Networks (NN) [33,41,49], deep learning [8,11], and so on. Thirteen of supervised ML studies were in categories 1 and 2, while three of supervised ML studies were in category 3. CA methods used in collated studies included hierarchical cluster analysis [9,46], Gaussian mixture clustering (GMC) [47], or Bicluster analysis approaches [40,48]. All studies adopted CA techniques aimed to categorise subgroups of stroke population (i.e., category 3).

Performance evaluation

Seven studies that employed supervised ML assessed the developed classification performance on the training set and not a held-out testing set [7,33,34,39,45,49]. Ten studies that used supervised ML split their datasets into at least two groups (e.g., training set and held-out set) and evaluated the classification performance based on the held-out dataset [8,11,35–38,41–44]. Three studies clearly ensured that data from the same individual was confined to either the training or testing sets [11,37,42]. Fourteen studies that used supervised ML adopted k-fold cross-validation (CV) [8,11,34–38,41,43–45,49] or leave-one-subject-out (LOSO) CV [7,39]. Three studies using unsupervised ML (3/5) used ANOVA to find significant differences between clusters in some clinically meaningful parameters [9,40,46].

All supervised ML studies adopted accuracy as a performance evaluation metric. Six of them only reported accuracy [11,34,35,39,42,49] while others used additional metrics (e.g., sensitivity, specificity). Eleven of supervised ML studies reported accuracy within 90%−100%, ten of which were from categories 1 and 2 (10/13) [7,33–37,39,42,43,49] while the remaining study [11] was from category 3.

Clinical utility

Data collection method

Nine of the reviewed studies used IMUs in data collection [7,9,11,33,36–39,42]. Marker-based optical motion capture systems were used in six studies [35,37,40,45,46,48]. Four studies used EEG [43,44] or surface EMG sensors [35,41], and two studies used smartphones [8,49]. Eight of included studies clearly stated that they collected multiple trials or data sequences from one subject [11,35–37,40,41,43,48].

Classification type

Ten of studies in categories 1 and 2 used binary classification [34–38,41–44,49], while other three studies used multi-class classification [7,33,39]. Five of studies in category 3 were multi-class classification [9,40,46–48], while other three studies adopted binary classification [8,11,45]. Only three of included studies used multi-label classification [11,40,48], and they were in category 3. The remaining studies used single-label classification. All included studies adopted global classification instead of individual classification.

Table 3
Summary of model types, validation, and performance.

	Features Type (number)	Model Tuning	Algorithm (best performed one)	Validation Method	Best Classification Performance
Category 1: Differentiating **normal gait from stroke pathological gait**					
(Iosa et al.,2021) [33]	Temporal parameters and kinematics parameters (e.g., angle) (total 17)	Two algorithms were compared.	Supervised ML (ANN with FFNN)	NS	ACC/SEN/SPE: SG vs HG: 90.9% /93.8%/ 88.2% SR vs SNR: 93.8%/90.0%/ 100%
(Li et al.,2019) [34]	Electrostatic signal: DTW distance, SampEn, and Stability Index	Feature selected (Mann–Whitney test); four algorithms were compared.	Supervised ML (kNN)	10-fold CV	ACC: 94%
(Hussain and Park, 2021) [41]	EMG parameters (total 20, final 11 after selection)	Feature selected (Pearson chi-squared test); six algorithms were compared.	Supervised ML (NN)	Training set (70%) and testing set (30%); 10-fold CV	ACC/ SEN/SPE/PPV/AUC/ Gini: 65%/57%/74%/72%/69%/ 38% (testing set)
(Cui et al., 2018) [35]	Kinematics, GRF, and EMG parameters PCA for feature extraction	Single modal features and multimodal features were compared; seven algorithms were compared.	Supervised ML (SVM with multimodal features)	Training set (3/9), validation set (2/9), and testing set (4/9); 3-fold CV; split*	ACC: 98.21% (testing set)
(Lee et al., 2018) [36]	Time-frequency parameters from IMU (total 165, final 2–4 after selection)	Feature selected (sequential forward search algorithm); hyperparameters tunning (the number of trees)	Supervised ML (RF with 50 trees)	Training set (75%), testing set (25%); 4-fold CV	ACC/SEN/SPE/PPV: 100%/100%/100%/100% (testing set)
(Hsu et al., 2021) [37]	Time-frequency domain parameters from IMU (total 42, final 4 after selection)	Feature selected (post hoc test results and signal-to-noise ratio); hyperparameters tunning (kernels of SVM)	Supervised ML (SVM with Quadratics)	Training set (18/23), testing set (5/23);10-fold CV; not split*	ACC/ SEN/SPE: 96.55% /94.44% /100.00% (testing set)
(Mathur and Bhatia, 2022) [42]	Spatiotemporal parameters (total 7)	Four algorithms were compared.	Supervised ML (XGBoost)	Training set (60/84), testing set (24/84); not split*	PPV/SEN/F1-score: 96%/96%/96% (testing set)
(Altilio et al., 2021) [49]	Spatiotemporal parameters and frequency domain parameters from the smartphone sensor (total 5, final 2)	Five feature sets were compared; 9 algorithms were compared.	Supervised ML (PNN)	10-fold stratified validation	ACC: 91.13%
(Choi et al., 2021) [43]	EEG features (total 66)	Eight different feature sets were compared; ten algorithms were compared.	Supervised ML (RF with total of 66 features)	Training set (67% or 80%) and testing set (33% or 20%), 5-fold CV, 10-fold CV, 20-fold CV; split*	ACC/F-1 score/SEN/PPV: 92.37%/92.4%/92.4%/ 92.6% (testing dataset)
(Hussain and Park, 2021) [44]	EEG features (total 244, final 177 after selection)	Featured selected (Pearson's chi-square test); 4 algorithms were compared	Supervised ML (C5.0)	Training set (70%), testing set (33%); 10-fold CV	ACC/SEN/SPE/PPV, NPV, AUC: 89%/94%/84%/88%/92&/ 90% (testing set)
Category 2: Differentiating stroke-related gait from other pathological gaits					
(Hsu et al., 2018) [38]	Temporal parameters (total 17) and time domain parameters from IMUs (total 192)	Feature selected (based on ANOVA F-value); two different modal feature sets were compared	Supervised ML (MLP with multi-modal features sets)	Training set (60%), testing set (40%); 5-fold CV	ACC/PPV/SEN: 84.78%/88%/85% (testing set)
(Wang et al., 2020) [7]	Spatiotemporal parameters and kinematic parameters (total 8)	NS	Supervised ML (SVM)	Leave-one-subject-out CV	ACC/SEN: Four subgroups: 93.9%/ 92.3%
(Mannini et al., 2016) [39]	HMM-based parameters and time-frequency domain parameters (total 90)	Two different modal feature sets were compared; hyperparameters tuning (grid search: SVM kernel parameters), two algorithms were compared.	Supervised ML (SVM with multi-modal features)	Leave-one-subject-out CV	ACC/SEN: 90.5%/ 86.7% (not impaired side), 100% (impaired side)
Category 3: Identifying homogeneous gait subgroups within stroke patient datasets – Supervised ML					
(Wang et al.,2021) [11]	Kinematic parameters (total 2)	NS	Supervised ML (Deep Neural Network)	Training set (76.86%) and testing set (23.14%); 4-fold CV; not split*	ACC/ F1-score: SG vs HG: 99.34%/99.39% (testing set) Four subgroups: 97.31%/ 0.966 (training set)
(Lee et al., 2021) [8]	Video data and temporal gait parameters (swing time asymmetry)	Two different feature sets were compared	Supervised ML (3D-CNN with multi-modal feature sets)	Training set (80%), testing set (20%); 5-fold CV	ACC, PPV, SEN, F1-score: 88.7%/ 89.1%/ 95.7%/ 0.922 (testing set)
(Punt et al., 2017) [45]	Spatiotemporal, stability, symmetry, smoothness, and variability parameters (total 25, final 10 after features reduction); PCA for feature extraction	Two feature sets were compared.	Supervised ML (LR)	10-fold CV	ACC/SEN/SPE/AUC: 72%/85%/65%/73%
Category 3: Identifying homogeneous gait subgroups within stroke patient datasets —CA					
(Sekiguchi et al., 2021) [46]	QJS (total 2)	Using visual inspection of the dendrogram to decide the number of clusters	CA (hierarchical cluster analysis)	ANOVA	There is significant difference in SIAS, use of AFOs, kinematic and kinetic parameters among clusters.

(*continued on next page*)

Table 3 (*continued*)

	Features Type (number)	Model Tuning	Algorithm (best performed one)	Validation Method	Best Classification Performance
(Tan et al., 2019) [9]	Spatiotemporal and kinematic parameters (total 3), and related parameters (total 1)	Two different feature sets were compared); using agglomeration coefficient to decide the number of clusters	CA (hierarchical cluster analysis)	ANOVA	The clusters have significant differences in spatial-temporal parameters and ankle angles.
(Dolatabadi et al., 2017) [47]	Spatiotemporal parameters (total 7)	Fourteen models were compared; using Bayesian Information Criterion to decide the number of clusters	CA (a model-based Gaussian mixture clustering approach)	NS	The clusters' mean and SD values of 7 spatiotemporal parameters were described.
(PAUK et al., 2016) [40]	Kinetic parameters (total 3)	Hyperparameters tunning	CA (bicluster analysis)	ANOVA	The biclusters have significant differences in spatial-temporal parameters.
(PAUK et al., 2016) [48]	Kinetic parameters (joint moments) (total 3)	Hyperparameters tunning; two algorithms were compared	CA (new bicluster analysis: KMB)	NS	The biclusters' spatiotemporal parameters were described.

ACC: accuracy; ANN: artificial neural network; CA: cluster analysis; CV: cross-validation; DTW distance: dynamic time warping distance; GRF: ground reaction force; FFNN: feedforward neural network in ANN; HMM: Hidden Markov model; LR: logistic regression; kNN: k-nearest-neighbour; MLP: multilayer perceptron neural networks; ML: machine learning; NPV: negative predictive value; NS: not specified; NN: neural network; NS: not specified; PCA: principal component analysis; PNN: probabilistic neural network; PPV: positive predictive value (which is also called precision); QJS: ankle quasi-joint stiffness; RF: random forest; SampEn: sample entropy; SEN: sensitivity; SPE: specificity; SIAS: stroke impairment assessment; SVM: support vector machine; "Split" or "not split" refers to whether data from one subject were split into two different sets (e.g., training dataset, validation dataset and test dataset) or not; XGBoost: extreme gradient boosting; 3D-CNN: three-dimensional convolutional neural network;.

Table 4
Quality assessment results.

	Number of events	Sampling	Feature selection	Validation	Performance	Algorithm interpretability
(Iosa et al.,2021) [33]	0	0	0	0	1	0
(Li et al.,2019) [34]	0	0	1	0	0	0
(Hussain and Park, 2021) [41]	1	0	1	1	1	0
(Cui, 2018) [35]	1	0	0	1	0	0
(Lee et al., 2018) [36]	0	0	1	1	1	0
(Hsu et al., 2021) [37]	0	0	1	1	1	0
(Mathur and Bhatia, 2022) [42]	0	0	0	1	1	1
(Altilio et al., 2021) [49]	0	0	0	0	0	0
(Choi et al., 2021) [43]	1	0	0	1	1	0
(Hussain and Park, 2021) [44]	1	0	1	1	1	1
(Hsu et al., 2018) [38]	0	1	1	1	1	0
(Wang et al., 2020) [7]	0	0	0	0	0	0
(Mannini et al., 2016) [39]	0	0	0	0	0	0
(Wang et al.,2021) [11]	1	0	0	1	0	0
(Lee et al., 2021) [8]	1	0	0	1	1	0
(Punt et al., 2017) [45]	0	0	0	0	1	1
(Sekiguchi et al., 2021) [46]	1	0	0	0	0	0
(Tan et al., 2019) [9]	0	0	0	0	0	0
(Dolatabadi et al., 2017) [47]	1	0	0	0	0	0
(PAUK et al., 2016) [40]	0	0	0	0	0	0
(PAUK et al., 2016) [48]	0	0	0	0	0	0

0 represents "not adequate" or "not specified", while 1 suggests that the step was completed "adequately". Regarding algorithm interpretability, 1 represents "interpretable" while 0 refers to "not interpretable".

Quality assessment

Table 4 presents the quality assessment results for the reviewed papers. No study had adequate quality in all aspects, but one study had adequate quality in five aspects [44]. In terms of each aspect with adequate quality, there were eight studies for the number of events [8,11,35,41,43,44,46,47], one study for sampling method [38], six studies for feature selection [34,36–38,41,44], ten studies for validation [8,11,35–38,41–44], ten studies for performance evaluation metrics [8,33,36–38,41–45], three studies for algorithm interpretability [42,44,45].

Discussion

This systematic review aimed to evaluate previous studies examining the use of data-driven models for stroke gait classification in terms of ML model development and clinical utility. Supervised ML models were used predominantly in categories 1 and 2 to differentiate stroke gait from normal gait (category 1) or other pathological gait types (category 2). Most of them demonstrated high levels of accuracy (>90%). Studies in category 3 identified subgroups in the examined stroke populations through supervised ML or CA models. Methodological errors were found within the reviewed learning models when compared to established recommendations [23,24]. Furthermore, additional considerations of clinical significance and efficacy of the developed models were often neglected. The following sections will address the aims of this review by discussing methodological considerations (both technical and practical) for ML model development in this area and discuss the importance of clinical significance and efficacy. Finally, recommendations to overcome these limitations will be provided to guide direction of research and to aid development of improved learning models.

Machine learning model development

Sample definition

Many of reviewed studies recruited fewer than 30 patients with stroke for ML model development [8,41–48]. While some studies collected several trials from each participant to enlarge the dataset, most studies had fewer observations than the suggested number of events (100 events and 100 non-events [28]. The small number of patients with stroke may negatively affect ML model development and generalisation to new data [23]. Compensation through the use of multiple trials from the same subject for model training may overestimate the classification performance due to the decreased variation within their training and testing set [24].

Most studies provided limited descriptions regarding participant information and their recruitment, which might affect models' accuracy and limit generalisability. Participant demographics (age, gender, etc.) [50,51] and clinical characteristics (e.g., stroke type, etc.) [52,53] are factors contributing to variation in gait performance. Not controlling for these factors can lead to increased inter-group variability, negatively affecting model classification [54]. Class imbalance may also lead to model bias towards the majority class [23]. Future studies should provide rigorous inclusion criteria or information to provide readers with a clear idea of the developed models' utility.

Feature engineering

Collated studies that employed multi-modality feature sets (e.g., EMG and kinematics data) [8,35,38,39] demonstrated better classification performance compared to studies that adopted just one modality (e.g., kinematic data from motion capture) [35]. An explanation for this may be that single-modality feature sets contain less information than feature sets created from multiple modalities. Furthermore, incorporating clinically interpretable features (e.g., gait spatiotemporal parameters) with uninterpretable features (e.g., frequency domain data from IMU) may improve both the model interpretability and performance. On their own, spatiotemporal features provide understandable interpretation but lack information on joint movement strategies and muscle contributions [38]. Time-frequency-power features extracted from bioelectric sensors (e.g., EMG, EEG) [9,35,41,43,44] and IMUs [36–39,49] offer more information on neuromuscular and movement strategies especially about potential causes of gait deviations but lack clinical insight. Similarly, kinetic and kinematic data could provide information from other perspectives. Combining multiple modality feature sets may provide complementary perspectives for gait classification, and lead to interpretable and high-performing models [19,35,55]. However, time-frequency-power feature sets are typically much larger than the number of observations. This should be noted during ML model development to avoid overfitting the developed model to a given dataset [24,41].

Automatic methods, such as Principal Component Analysis (i.e., PCA) [35,45] could improve the model training efficiency by reducing feature set sizes and identifying trends within collected data. However, automated methods are often criticised due to their extracted features lacking interpretability [23]. In general, it is recommended that both automatic and manual feature extraction should be used to enrich the extracted input feature set and increase overall interpretability [24]. Regardless of what method is used, developers should ensure that the number of input features does not outweigh the number of observations used for ML model development. Failure to do so may cause model overfitting and subsequent poor classification performance on new data [23].

Model tuning

Model-tuning represents ML model developmental procedures such as feature selection and hyperparameter tuning. Such procedures should be incorporated due to their widely accepted benefits on model performance [23,24]. Only less than half included studies performed feature selection. Most of these studies utilised statistics-based feature selection [34,37,38,41,44], which are examples of filter methods. Such methods identify feature importance based on intrinsic properties within the dataset. Filter methods are efficient, but they neglect feature influence on classification performance, which may lead to important features being removed. One study utilised sequential forward selection [36] which is an example of wrapper feature selection. Such methods are advantageous compared to filter methods as they assess feature importance based on their influence on classification accuracy [56]. Furthermore, the included studies lacked standardised hyperparameter tuning procedures for both ML and CA, as recommended by Halilaj et al. [23]. Subsequently, the developed models may not have been optimally fitted to the collected datasets, resulting in sub-optimal model performance and utility [23]. Finally, model tuning methodology should be clearly reported so readers can understand the entire ML model development process.

Learning algorithm selection

Studies collated in this review adopted supervised ML algorithms or unsupervised algorithms (i.e., CA). Selecting the appropriate algorithm for a given classification task can be difficult. Therefore, it is recommended to trial several different learning algorithms for your given task, as some may perform better on a given dataset [23].

Interestingly, supervised ML was mostly used in categories 1 and 2 (i.e., categorising pathological gaits with normal gaits), while studies rarely used supervised algorithms for category 3 (i.e., classifying subgroups of stroke gait). One possible reason is that labelling many different post-stroke gait patterns observed in clinical assessment requires significant time and expertise. Conversely, labelling groups based on the presence or absence of stroke gait patterns from healthy controls or other clinical conditions provides a straightforward task. Unfortunately, such analysis provides little clinically utility. This point will be further discussed within the clinical utility section.

CA techniques were used for category 3. Traditional CA techniques (e.g., hierarchical CA) [9,46] can struggle to identify correct cluster numbers, select optimal clustering algorithms, and deal with outliers. In some scenarios, these factors can result in none of the developed ML models being established as the best choice [47]. Conversely, GMC can select the best mixture algorithm by determining appropriate cluster numbers using an Approximate Bayes Factor, and provide a soft clustering membership based on an index for each data point [47]. Bicluster analysis performs more data mining compared to traditional CA by considering both rows and columns of input features simultaneously [57]. It may also overcome the main limitation of standard clustering approaches, namely the excessive sensitivity towards variations in the gait trials, through considering both patients and samples information simultaneously in two dimensions [48].

Model evaluation

Almost half of the supervised ML papers assessed classification performance on the training dataset [7,33,34,39,45,49], neglecting any meaningful interpretation of the model's generalisability to new data [23]. While most of the supervised ML models evaluated performance using a held-out testing set, few ensured sequences from the same individual were organised in the training or testing set [11,37,42]. Incorporating movement trials from the same individual in the training and testing sets can replicate similar issues as validating classifier on the training set due to high levels of homogeneity in within-subject movement trials [23,58]. Therefore, future developers should consider this in their own developmental procedures.

Most supervised ML studies adopted k-fold CV or LOSO CV, which is recommended when performing ML model development and validation on small datasets [23,30]. Furthermore, if model tuning steps are performed then they should be evaluated before model evaluation on the testing set. This would involve the dataset being split into three sets: the training set, a model tuning validation set and a testing set. K-fold or

LOSO CV do not account for this process, therefore, it is unclear how studies in this review using this approach guided model-tuning decisions. Instead, a nested k-fold CV design would need to be adopted. Nested k-fold CV follows a similar structure to conventional CV where an outer loop iteratively trains and evaluates a ML model on different folds. Simultaneously, an inner loop performs feature selection and hyperparameter tuning, providing an unbiased estimation [59].

Accuracy was the most commonly used evaluation metric for reporting classification performance; however, the lack of additional evaluation metrics (e.g., sensitivity, specificity) in many studies [11,34,35,39,42,49] limited the reader's understanding of the ML model's performance. These estimates involve properties important to clinicians in different clinical contexts. For example, if gait classification is performed in context of fall risk screening, a high sensitivity may be necessary [21]. Comprehensive evaluation metrics are recommended to report their model's classification performance [23,58]. For studies that adopted CA, traditional statistical testing was used by several studies to evaluate whether identified clusters were significantly different from one another [9,40,46]. But remaining studies did not evaluate their identified clusters providing little insight into the efficacy of the identified clusters. Future research should provide some degree of cluster assessment so readers can adequately evaluate the developed ML models.

Clinical utility

Classification output utility

Classification outputs in this review often lacked clinical relevance. Most studies fall into categories 1 or 2, which focus on differentiating normal or other pathological gait from stroke gait. However, these ML models offer limited clinical utility as gait-based stroke diagnosis is not as reliable or valid as established stroke diagnostic techniques like MRI or CT scans [43,60]. Instead, the aim of clinical gait analysis should be to identify the different deviations and their severity within observed gait performance [61,62]. Models prioritising clinically meaningful outputs, such as identifying specific deviations or severity levels among post-stroke gaits may better contribute to guiding exercise prescriptions and tracking rehabilitation progress [11,63]. Studies in category 3 followed this notion by identifying specific gait patterns among post-stroke gaits. However, subgroups identified in category 3 studies are limited and may not cover all possible post-stroke gait deviations. Additionally, the developed CA models can lack clinical significance since they may discover artificial clusters which have little or no clinical relevance [9,40,46–48]. Supervised and semi-supervised learning may be suitable for this analysis in terms of interpretability [23,64]. They can classify new data into pre-existing categories which benefit output interpretation and allow for straightforward model validation. Interpretable algorithms and algorithms combining the strength of expert knowledge and data-driven techniques may further improve clinical utility [65–67].

Furthermore, the classification types adopted by many of the reviewed studies may lack clinical utility. Most studies (e.g., studies in category 1 and 2) used binary classification to categorise performances into one of groups using a single label [7,33–39,41–44,49]. This classification is clinically insufficient since it cannot identify various gait subgroups. Multi-class approaches can improve feedback insight overcoming this limitation. However, the single label nature of binary and multi-class classification may be ineffective since multiple deviations can coexist within one gait cycle, such as knee hyperextension and drop foot [11,20]. Multi-label classification can overcome this limitation by identifying multiple gait patterns simultaneously by allocating multiple labels to each patient (e.g., drop-foot gait and circumduction gait simultaneously [11],[26]). This approach was adopted by few studies [11,40,48] allowing for clinically meaningful insight. Future studies should continue to explore multi-label methods to improve clinical utility.

The inability to identify specific joint deviations is a large limitation in the developed automatic gait classification systems. Most studies within category 3 mentioned functional or overall differences in spatial-temporal characteristics (e.g., velocities) or clinical functional test results among different subgroups [8,9,40,45–48]. However, specific joint motions were not identified clearly. Only Wang and Chen [11] combined the observed typical joint deviation with classification outcomes, namely successfully classifying subgroups (e.g., drop-foot gait, hip hiking gait), which indicated specific movement characteristics. From a clinical perspective, the functional impairments are normally represented by multiple joint deviations [13,68], and then joint deviations could be associated with specific causes within the neuromuscular system [68]. Therefore, specific joint deviation identification (e.g., excess flexion or extension) is beneficial to finding precise causes that can inform clinical decision-making [69]. This scarcity of detailed movement strategies of output groups may lead to little insight for guiding rehabilitation therapy. Furthermore, both specific joint deviations and the coordination of multiple joints should be considered in the model. Such an approach would closely mimic experienced therapists since they pay more attention to the overall performance by assessing the coordination of multiple joints [70].

The stage and degree of abnormality (i.e., level of difference between pathological and normal gait patterns) can provide insight into rehabilitation progression (e.g., monitoring changes in gait patterns throughout a rehabilitation program) and help guide therapist decision-making [12,63]. Unfortunately, most studies in this review used a black-or-white approach to classification, which cannot detect subtle changes between gait patterns [71]. The studies assessing joint motions [9,11,40,46,48] also did not provide insight into the severity of gait deviations and only showed the kinematic characteristics of groups in tables. A possible solution is to integrate multi-class or advanced classification that provides scores or indexes (e.g., probability) [35] within a multi-label ML model. For example, post-stroke gait data could be first organised into subgroups based on the presence or absence of each potential gait abnormality (e.g., drop foot gait, knee hyperextension). Once organised, the severity of each abnormality can be identified using a multi-class classifier (e.g., small, medium or large deviation), which is similar to the severity stage classification for Parkinson's gaits [63,72]. Also, the data from different rehabilitation phases could be included in the training dataset to develop a gait classification system that is sensitive to rehabilitation progression.

Some studies in category 3 utilised CA to quantitatively identify stroke gait subgroups, where the output is not interpretable. Lack in interpretability of the outputs damages clinical applicability [31]. This is because therapists in clinics tend to develop rehabilitative strategies based on previously known gait styles/groups (e.g., drop foot). However, the unsupervised nature of CA can result in the ML model discovering clinically irrelevant "artificial" groups [20,23]. Although statistical analyses are normally performed after clustering to explain differences between these groups, it is still hard to categorise new data based on these clusters. Consequently, caution should be applied when interpreting and/or utilising the output from CA for stroke gait analysis.

To further improve the interpretability of the outputs, the future system should aim to analyse and interpret quantitative gait data. For example, providing clinical interpretation (e.g., the reasoning for abnormal gait) to clinicians should also be considered by future systems to contextualise the observed deviations. Having insight into the joint deviation and likely causative factors can help guide the clinician or therapist when deciding on the most suitable therapy. Prior structured knowledge such as Rancho Los Amigos gait analysis framework [69,73], or unstructured data extracted and processed by natural language processing techniques could be considered in performing reasoning process based on knowledge engineering techniques (e.g., knowledge graph) to output possible causes [74,75].

Generalisability and feasibility

All classifiers were developed using observations from multiple individuals, enabling them to classify new, previously unseen individuals (i.e., global classifiers) [24]. But many of the developed global classifiers utilised multiple gait cycles from each participant to increase the dataset size [11,34–37,40,41,43,48]. As mentioned in 4.1.1, the decreased variability within the collected dataset may negatively affect the ML model's generalisability, decreasing its utility for clinical gait analysis [23]. Additionally, these classifiers end up classifying gait data sequences instead of patients, which may result in poor classification performance for patients, since gait cycles from one patient may be assigned into different categories [7,39]. Therefore, global classifiers should ideally be developed using individual gait cycles from many individuals [24]. However, recruiting many patients is difficult in a practical context, which is the main limitation in developing these ML models. In this regard, novel methods such as transfer learning [76] may be used in the future to combine datasets by training a ML model on one dataset and using the learned features to extract relevant features from the other datasets. This may allow for the collation of data collected using different equipment from various laboratories, and even for the transfer of knowledge from another population to the field of post-stroke gait analysis [77].

As mentioned in Section 4.1.2, the utilised data collection method could decide the portability and implementation ability of a given system. Several studies in this review used accurate but less portable data collection approaches such as marker-based motion capture [37,40,45,46,48] which are too complicated in clinical settings [78]. Some portable data collection devices (e.g., markerless motion capture techniques [79]) offer a more feasible and simplistic approach to data collection used in the future to allow for implementable systems.

Limitations

Firstly, non-English studies were excluded, potentially leading to the omission of relevant research. Secondly, relying on keyword searches rather than full-text analysis may have resulted in missing certain studies. Lastly, studies using functional gait assessment scores and data from walking trolleys, robotic rollators, or lower-limb exoskeleton robots as input features were excluded. However, considering that patients with stroke often use lower-limb exoskeleton robots during gait training or daily life, utilising data from these devices could be an alternative option for gait classification.

Future work

As mentioned in Section 4, the current gait classification systems collated in this review cannot satisfy practical requirements in real clinical settings. The main problems limiting the application of previously developed models are their low clinical significance (e.g., distinguishing hemiplegic people from normal people) and methodological flaws in the development process (e.g., small sample size). Current systems have not routinely followed standardised machine learning methodology in human movement biomechanics [23]. Based on the discussion, we provide several recommendations for future clinical and ML model development (Table 5).

Conclusion

Computerised data-driven models (supervised ML and CA) can assist in gait classification for patients with stroke. Most studies collated in this review demonstrated acceptable classification accuracy of 80%–100%. However, we note that the classification performance of some papers might be overestimated due to small sample size and unsuitable assessment methods. Furthermore, key problems in development methodology, clinical utility and significance have been identified for previous studies, particularly in models focused on the diagnosis of stroke (i.e., distinguishing stroke population from other populations) compared to those that distinguished subgroups of stroke. Even though

Table 5
Recommendations for future models.

Aspects	Recommendations
Clinical utility	• Clinical meaningful outputs: Developed models should distinguish stroke subgroups (i.e., different post-stroke gait patterns or severity levels) instead of differentiating between healthy, stroke, and other disease groups. • Both the deviations of each joint and the coordination between different joints should be considered [70]. • Models should aim to identify multiple gait deviations and their severity levels simultaneously (e.g., multi-class and multi-label approaches) [11,72]. • The classification model should provide reasoning for specific gait patterns by integrating the most likely causes for the identified deviations. Reasoning could come from prior structured knowledge such as Rancho Los Amigos gait analysis framework [69], and unstructured data extracted through natural language processing techniques, and be integrated based on knowledge engineering techniques [74,75]. • A more simplified and stroke-targeted data-driven model may contribute to clinical application. • Data collection techniques should be accessible to real-world settings (e.g., portable marker-less validated tools such as OpenCap [79,80]).
Model development Accuracy (data preprocessing, model training, tuning and validation)	• Collected data should be preprocessed and organised using conventional filtering, segmentation, normalisation, and class balancing methods [23]. • Feature sets incorporating data from multiple modalities can enrich the feature space and improve model accuracy [19,24,31]. • Manual (e.g., parameters with clinical significance) and, automatic feature extraction/selection techniques (e.g., PCA, wrapper or embedded feature selection) and hyperparameter tuning should be performed [24]. • The collected dataset should be split into a training dataset for training, a validation dataset for model tuning, and a testing dataset for evaluation to avoid overestimating prediction performance [23]. Nested-cross-validation designs could be used to achieve this on small datasets [59]. • Comprehensive model evaluation should be conducted to provide sufficient insight into model performance [23].
Model interpretability	• Adopting interpretable features (e.g., joint angles), algorithms (i.e., supervised learning or semi-supervised learning), and outputs can (e.g., insightful analysis) contribute to greater model insight [23]. If non-interpretable feature sets (e.g., frequency domain data) are utilised, then developers could consider incorporating interpretable features to balance model accuracy and interpretability. Unsupervised learning (i.e., cluster analysis) may offer low clinical interpretability and may better suit research settings. Interpretable algorithms and algorithms integrating prior knowledge and data-driven knowledge may benefit interpretability [66,67]
Model generalisability	• Models should be developed and validated using a sufficiently large sample size that represents the target population and considers the class balance (e.g., age, sex ratios, condition severity) [23,54]. Transfer learning may potentially help by combining small

(*continued on next page*)

Table 5 (*continued*)

Aspects	Recommendations
	datasets from different laboratories, boosting dataset sizes [77]. • Global classifiers with one data sequence from one patient should be developed so the model can be generalised to new users[24].

some models attempted to split patients with stroke into different classes, the low interpretability or utility of the models may not facilitate or support the rehabilitation process. Recommendations provided in this review can help improve classification performance, interpretability, and generalisability to satisfy clinical requirements.

CRediT authorship contribution statement

Zhang Yanxin: Writing – review & editing, Supervision, Methodology, Investigation, Formal analysis. **Reading Stacey:** Writing – review & editing, Supervision. **Hart Rylea:** Writing – review & editing, Methodology, Investigation, Formal analysis. **Jiao Yiran:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation.

Declaration of Competing Interest

The authors do not have any conflict of interest which could have influenced the results of this work.

Acknowledgement

The authors would like to acknowledge Qi Zhang (i.e., a therapist) for her time as consultant in the writing of the script and Chinese Scholarship Council for living expenses support.

References

[1] R.V. Krishnamurthi, T. Ikeda, V.L. Feigin, Global, regional and country-specific burden of ischaemic stroke, intracerebral haemorrhage and subarachnoid haemorrhage: a systematic analysis of the global burden of disease study 2017, Neuroepidemiology vol. 54 (2) (2020) 171–179, https://doi.org/10.1159/000506396.

[2] P. Langhorne, F. Coupar, A. Pollock, Motor recovery after stroke: a systematic review, Lancet Neurol. vol. 8 (8) (2009) 741–754, https://doi.org/10.1016/S1474-4422(09)70150-4.

[3] M.A. Weintraub, T. Bylander, S.R. Simon, quawds: a composite diagnostic system for gait analysis, 1990/05/01/, Comput. Methods Prog. Biomed. vol. 32 (1) (1990) 91–106, https://doi.org/10.1016/0169-2607(90)90089-R.

[4] T. Bylander, M. Weintraub, S.R. Simon, A study of an expert system for interpreting human walking disorders, March 1994, Proc. Tenth Conf. Artif. Intell. Appl. 1-4 (1994) 178–184, https://doi.org/10.1109/CAIA.1994.323676.

[5] A.C. Lapham, R.M. Bartlett, The use of artificial intelligence in the analysis of sports performance: a review of applications in human gait analysis and future directions for sports biomechanics, 1995/06/01, J. Sports Sci. vol. 13 (3) (1995) 229–237, https://doi.org/10.1080/02640419508732232.

[6] J.M. Dzierzanowski, J.R. Bourne, R. Shiavi, H.S.H. Sandell, D. Guy, Gaitspert: an expert system for the evaluation of abnormal human locomotion arising from stroke, IEEE Trans. Biomed. Eng. 32 (11) (1985) 935–942, https://doi.org/10.1109/TBME.1985.325626.

[7] L. Wang, Y. Sun, Q. Li, T. Liu, J. Yi, Two shank-mounted IMUs-Based gait analysis and classification for neurological disease patients, IEEE Robot. Autom. Lett. vol. 5 (2) (2020) 1970–1976, https://doi.org/10.1109/LRA.2020.2970656.

[8] J.T. Lee, E. Park, T.D. Jung, Machine learning-based classification of dependence in ambulation in stroke patients using smartphone video data, Art no. 1080, J. Pers. Med. Artic. vol. 11 (11) (2021), https://doi.org/10.3390/jpm11111080.

[9] M.G. Tan, J.H. Ho, H.T. Goh, H.K. Ng, L. Abdul Latif, M. Mazlan, A new fractal-based kinetic index to characterize gait deficits with application in stroke survivor functional mobility assessment (Article), Biomed. Signal Process. Control vol. 52 (2019) 403–413, https://doi.org/10.1016/j.bspc.2018.09.014.

[10] J.M. Rodda, H.K. Graham, L. Carson, M.P. Galea, R. Wolfe, Sagittal gait patterns in spastic diplegia, J. Bone Jt. Surg. Br. vol. 86 (2) (2004) 251–258, https://doi.org/10.1302/0301-620x.86b2.13878.

[11] F.C. Wang, et al., Detection and classification of stroke gaits by deep neural networks employing inertial measurement units, Art no. 1864, Sens. Artic. vol. 21 (5) (2021) 1–18, https://doi.org/10.3390/s21051864.

[12] J. Wikstrom, G. Georgoulas, T. Moutsopoulos, A. Seferiadis, Intelligent data analysis of instrumented gait data in stroke patients-a systematic review, Comput. Biol. Med. vol. 51 (2014) 61–72, https://doi.org/10.1016/j.compbiomed.2014.04.004.

[13] D.M. Mohan, A.H. Khandoker, S.A. Wasti, S. Ismail Ibrahim Ismail Alali, H. F. Jelinek, K. Khalaf, Assessment methods of post-stroke gait: a scoping review of technology-driven approaches to gait characterization and analysis, Front. Neurol. vol. 12 (2021) 650024, https://doi.org/10.3389/fneur.2021.650024.

[14] G.J. Luvizutto, et al., Use of artificial intelligence as an instrument of evaluation after stroke: a scoping review based on international classification of functioning, disability and health concept, Top. Stroke Rehabil. (2021) 1–16, https://doi.org/10.1080/10749357.2021.1926149.

[15] A. Vienne, R.P. Barrois, S. Buffat, D. Ricard, P.P. Vidal, Inertial sensors to assess gait quality in patients with neurological disorders: a systematic review of technical and analytical challenges, Front. Psychol. vol. 8 (2017) 817, https://doi.org/10.3389/fpsyg.2017.00817.

[16] W. Deng, I. Papavasileiou, Z. Qiao, W. Zhang, K.Y. Lam, S. Han, Advances in automation technologies for lower extremity neurorehabilitation: a review and future challenges, IEEE Rev. Biomed. Eng. vol. 11 (2018) 289–305, https://doi.org/10.1109/RBME.2018.2830805.

[17] B.O. Bat-Erdene, J.L. Saver, Automatic acute stroke symptom detection and emergency medical systems alerting by mobile health technologies: a review, (in eng), J. Stroke Cereb. Dis. vol. 30 (7) (2021) 105826, https://doi.org/10.1016/j.jstrokecerebrovasdis.2021.105826.

[18] A. Viswakumar, V. Rajagopalan, T. Ray, P. Gottipati, C. Parimi, Development of a robust, simple, and affordable human gait analysis system using bottom-up pose estimation with a smartphone camera, (in eng), Front. Physiol. vol. 12 (2021) 784865, https://doi.org/10.3389/fphys.2021.784865.

[19] I. Boukhennoufa, X. Zhai, V. Utti, J. Jackson, K.D. McDonald-Maier, Wearable sensors and machine learning in post-stroke rehabilitation assessment: A systematic review, Art no. 103197, Biomed. Signal Process. Control Review vol. 71 (2022), https://doi.org/10.1016/j.bspc.2021.103197.

[20] F. Dobson, M.E. Morris, R. Baker, H.K. Graham, Gait classification in children with cerebral palsy: a systematic review, Gait Posture vol. 25 (1) (2007) 140–152, https://doi.org/10.1016/j.gaitpost.2006.01.003.

[21] Y.C. Hsu, et al., A novel approach for fall risk prediction using the inertial sensor data from the timed-up-and-go test in a community setting, IEEE Sens. J. vol. 20 (16) (2020) 9339–9350, https://doi.org/10.1109/JSEN.2020.2987623.

[22] J. Ballesteros, C. Urdiales, A.B. Martinez, M. Tirado, Automatic assessment of a rollator-user's condition during rehabilitation using the i-walker platform, Art no. 7911313, IEEE Trans. Neural Syst. Rehabil. Eng. Artic. vol. 25 (11) (2017) 2009–2017, https://doi.org/10.1109/TNSRE.2017.2698005.

[23] E. Halilaj, A. Rajagopal, M. Fiterau, J.L. Hicks, T.J. Hastie, S.L. Delp, Machine learning in human movement biomechanics: best practices, common pitfalls, and new opportunities, J. Biomech. vol. 81 (2018) 1–11, https://doi.org/10.1016/j.jbiomech.2018.09.009.

[24] R. Hart, H. Smith, Y. Zhang, Systematic review of automatic assessment systems for resistance-training movement performance: a data science perspective, Comput. Biol. Med. vol. 137 (2021) 104779, https://doi.org/10.1016/j.compbiomed.2021.104779.

[25] M. Grandini, E. Bagli, G. Visani, Metrics for multi-class classification: an overview, arXiv Prepr. arXiv (2020).

[26] M.-L. Zhang, Z.-H. Zhou, A review on multi-label learning algorithms, IEEE Trans. Knowl. Data Eng. vol. 26 (8) (2014) 1819–1837, https://doi.org/10.1109/tkde.2013.39.

[27] M. Sajjadian, et al., Machine learning in the prediction of depression treatment outcomes: a systematic review and meta-analysis, Psychol. Med. vol. 51 (16) (2021) 2742–2751, https://doi.org/10.1017/S0033291721003871.

[28] F.E. Harrell, F.E.H. jrl, Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis, Springer, 2001.

[29] O. Faude, L. Donath, R. Roth, L. Fricker, L. Zahner, Reliability of gait parameters during treadmill walking in community-dwelling healthy seniors, Gait Posture vol. 36 (3) (2012) 444–448, https://doi.org/10.1016/j.gaitpost.2012.04.003.

[30] S.B. Kotsiantis, I. Zaharakis, P. Pintelas, Supervised machine learning: a review of classification techniques, Emerg. Artif. Intell. Appl. Comput. Eng. vol. 160 (1) (2007) 3–24.

[31] A. Vellido, The importance of interpretability and visualization in machine learning for applications in medicine and health care, Neural Comput. Appl. vol. 32 (24) (2019) 18069–18083, https://doi.org/10.1007/s00521-019-04051-w.

[32] X. Wu, et al., Top 10 algorithms in data mining, 2008/01/01, Knowl. Inf. Syst. vol. 14 (1) (2008) 1–37, https://doi.org/10.1007/s10115-007-0114-2.

[33] M. Iosa, et al., Artificial neural network analyzing wearable device gait data for identifying patients with stroke unable to return to work, Art no. 650542, Front. Neurol. Artic. vol. 12 (2021), https://doi.org/10.3389/fneur.2021.650542.

[34] M. Li, S. Tian, L. Sun, X. Chen, Gait analysis for post-stroke hemiparetic patient by multi-features fusion method, Art no. 1737, Sens. Artic. vol. 19 (7) (2019), https://doi.org/10.3390/s19071737.

[35] C. Cui, et al., Simultaneous recognition and assessment of post-stroke hemiparetic gait by fusing kinematic, kinetic, and electrophysiological data, IEEE Trans. Neural Syst. Rehabil. Eng. vol. 26 (4) (2018) 856–864, https://doi.org/10.1109/TNSRE.2018.2811415.

[36] J. Lee, S. Park, H. Shin, Detection of hemiplegic walking using a wearable inertia sensing device, Art no. 1736, Sens. Artic. vol. 18 (6) (2018), https://doi.org/10.3390/s18061736.

[37] W.C. Hsu, et al., Can trunk acceleration differentiate stroke patient gait patterns using time-and frequency-domain features?, Art no. 1541, Appl. Sci. vol. 11 (4) (2021) 1–14, https://doi.org/10.3390/app11041541.

[38] W.C. Hsu, et al., Multiple-wearable-sensor-based gait classification and analysis in patients with neurological disorders, Art no. 3397, Sensors Article vol. 18 (10) (2018), https://doi.org/10.3390/s18103397.

[39] A. Mannini, D. Trojaniello, A. Cereatti, A.M. Sabatini, A machine learning framework for gait classification using inertial sensors: application to elderly, post-stroke and huntington's disease patients, Sensors Article vol. 16 (1) (2016), https://doi.org/10.3390/s16010134.

[40] J. Pauk, K. Minta-Bielecka, A new classification of hemiplegia gait patterns based on bicluster analysis of joint moments, (in eng), Acta Bioeng. Biomech. vol. 18 (4) (2016) 33–40.

[41] I. Hussain, S.J. Park, Prediction of myoelectric biomarkers in post-stroke gait, Art no. 5334, Sensors Article vol. 21 (16) (2021), https://doi.org/10.3390/s21165334.

[42] D. Mathur, D. Bhatia, Gait classification of stroke survivors - An analytical study, *J.* Interdiscip. Math. Artic. vol. 25 (1) (2022) 163–181, https://doi.org/10.1080/09720502.2021.2006332.

[43] Y.A. Choi, et al., Machine-learning-based elderly stroke monitoring system using electroencephalography vital signals, Art no. 1761, *Appl. Sci. A*rtic. vol. 11 (4) (2021) 1–18, https://doi.org/10.3390/app11041761.

[44] I. Hussain, S.J. Park, Quantitative evaluation of task-induced neurological outcome after stroke, Art no. 900, Brain Sci. Artic. vol. 11 (7) (2021), https://doi.org/10.3390/brainsci11070900.

[45] M. Punt, S.M. Bruijn, H. Wittink, I.G. Van De Port, J.H. Van Dieën, Do clinical assessments, steady-state or daily-life gait characteristics predict falls in ambulatory chronic stroke survivors? J. Rehabil. Med. Artic. vol. 49 (5) (2017) 402–409, https://doi.org/10.2340/16501977-2234.

[46] Y. Sekiguchi, K. Honda, D. Owaki, S.I. Izumi, Classification of ankle joint stiffness during walking to determine the use of ankle foot orthosis after stroke, Art no. 1512, Brain Sci. Artic. vol. 11 (11) (2021), https://doi.org/10.3390/brainsci11111512.

[47] E. Dolatabadi, A. Mansfield, K.K. Patterson, B. Taati, A. Mihailidis, Mixture-model clustering of pathological gait patterns, IEEE J. Biomed. Health Inform. vol. 21 (5) (2017) 1297–1305, https://doi.org/10.1109/JBHI.2016.2633000.

[48] J. Pauk, K. Minta-Bielecka, Gait patterns classification based on cluster and bicluster analysis, Biocybern. Biomed. Eng. Artic. vol. 36 (2) (2016) 391–396, https://doi.org/10.1016/j.bbe.2016.03.002.

[49] R. Altilio, A. Rossetti, Q. Fang, X. Gu, M. Panella, A comparison of machine learning classifiers for smartphone-based gait analysis, Med. Biol. Eng. Comput. Article vol. 59 (3) (2021) 535–546, https://doi.org/10.1007/s11517-020-02295-6.

[50] A. Gabell, U.S.L. Nayak, The effect of age on variability in gait1, J. Gerontol. vol. 39 (6) (1984) 662–666, https://doi.org/10.1093/geronj/39.6.662.

[51] S.H. Cho, J.M. Park, O.Y. Kwon, Gender differences in three dimensional gait analysis data from 98 healthy Korean adults, Clin. Biomech. vol. 19 (2) (2004) 145–152, https://doi.org/10.1016/j.clinbiomech.2003.10.003.

[52] R. Adedoyin, A. Obembe, Differences in gait between haemorrhagic and ischaemic stroke survivors, J. Med. Med. Sci. vol. 3 (2012) 556–561.

[53] K. Kaczmarczyk, A. Wit, M. Krawczyk, J. Zaborski, J. Gajewski, Associations between gait patterns, brain lesion factors and functional recovery in stroke patients, Gait Posture, Artic. vol. 35 (2) (2012) 214–217, https://doi.org/10.1016/j.gaitpost.2011.09.009.

[54] A. Kline, et al., Multimodal machine learning in precision health: a scoping review, NPJ Digit Med. vol. 5 (1) (2022) 171, https://doi.org/10.1038/s41746-022-00712-8.

[55] Y. Celik, S. Stuart, W.L. Woo, E. Sejdic, A. Godfrey, Multi-modal gait: a wearable, algorithm and data fusion approach for clinical and free-living assessment, Inf. Fusion vol. 78 (2022) 57–70, https://doi.org/10.1016/j.inffus.2021.09.016.

[56] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. vol. 3 (2003) 1157–1182.

[57] S.C. Madeira, A.L. Oliveira, Biclustering algorithms for biological data analysis: a survey, IEEE/ACM Trans. Comput. Biol. Bioinform. vol. 1 (1) (2004) 24–45, https://doi.org/10.1109/TCBB.2004.2.

[58] M. O'Reilly, B. Caulfield, T. Ward, W. Johnston, C. Doherty, Wearable inertial sensor systems for lower limb exercise detection and evaluation: a systematic review, Sports Med. vol. 48 (5) (2018) 1221–1246, https://doi.org/10.1007/s40279-018-0878-4.

[59] A. Vabalas, E. Gowen, E. Poliakoff, A.J. Casson, Machine learning algorithm validation with a limited sample size, e0224365, PLoS One vol. 14 (11) (2019), https://doi.org/10.1371/journal.pone.0224365.

[60] D.S. Liebeskind, A.V. Alexandrov, Advanced multimodal CT/MRI approaches to hyperacute stroke diagnosis, treatment, and monitoring, https://doi.org/10.1111/j.1749-6632.2012.06719.x, Ann. N. Y. Acad. Sci. vol. 1268 (1) (2012) 1–7, https://doi.org/10.1111/j.1749-6632.2012.06719.x.

[61] R. Baker, Gait analysis methods in rehabilitation, 2006/03/02, J. Neuroeng. Rehabil. vol. 3 (1) (2006) 4, https://doi.org/10.1186/1743-0003-3-4.

[62] S. Armand, G. Decoulon, A. Bonnefoy-Mazure, Gait analysis in children with cerebral palsy, (in eng), EFORT Open Rev. vol. 1 (12) (2016) 448–460, https://doi.org/10.1302/2058-5241.1.000052.

[63] B. E, B. D, B. R, Supervised machine learning based gait classification system for early detection and stage classification of Parkinson's disease, Appl. Soft Comput. vol. 94 (2020), https://doi.org/10.1016/j.asoc.2020.106494.

[64] J.E. van Engelen, H.H. Hoos, A survey on semi-supervised learning, Mach. Learn. vol. 109 (2) (2019) 373–440, https://doi.org/10.1007/s10994-019-05855-6.

[65] M.A. Ahmad, C. Eckert, and A. Teredesai, Interpretable Machine Learning in Healthcare, presented at the Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, Washington, DC, USA, 2018. [Online]. Available: https://doi.org/10.1145/3233547.3233667.

[66] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat. Mach. Intell. vol. 1 (5) (2019) 206–215, https://doi.org/10.1038/s42256-019-0048-x.

[67] J. Sun, et al., Combining knowledge and data driven insights for identifying risk factors using electronic health records, (in eng), AMIA Annu Symp. Proc. vol. 2012 (2012) 901–910.

[68] B. Balaban, F. Tok, Gait disturbances in patients with stroke, PM& R. vol. 6 (7) (2014) 635–642, https://doi.org/10.1016/j.pmrj.2013.12.017.

[69] R.L.A.N.R. Center, Observational Gait Analysis. Los Amigos Research and Education Institute, Rancho Los Amigos National Rehabilitation Center, 2001.

[70] A. Shumway-Cook and M.H. Woollacott, Motor Control: Translating Research Into Clinical Practice (.). Lippincott Williams & Wilkins, 2007.

[71] Y. Liao, A. Vakanski, M. Xian, D. Paul, R. Baker, A review of computational approaches for evaluation of rehabilitation exercises, Comput. Biol. Med. vol. 119 (2020) 103687, https://doi.org/10.1016/j.compbiomed.2020.103687.

[72] B. Vidya, S. P, Gait based Parkinson's disease diagnosis and severity rating using multi-class support vector machine, Appl. Soft Comput. vol. 113 (2021), https://doi.org/10.1016/j.asoc.2021.107939.

[73] M.G. Abi Hayla, Automation of the Interpretation of Clinical Gait Data: the Development of a Novel Computerised Technique, Ph.D., University of Surrey (United Kingdom), Ann Arbor, 28126953, 2012. [Online]. Available: http://ezproxy.auckland.ac.nz/login?url=https://www.proquest.com/dissertations-theses/automation-interpretation-clinical-gait-data/docview/2430743193/se-2?accountid=8424. https://openurl.auckland.ac.nz/resolve?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&genre=dissertations+%26+theses&sid=ProQ:ProQuest+Dissertations+%26+Theses+Global&atitle=&title=Automation+of+the+Interpretation+of+Clinical+Gait+Data+%3A+the+Development+of+a+Novel+Computerised+Technique&issn=&date=2012-01-01&volume=&issue=&spage=&au=Abi+Hayla%2C+Myriam+G&isbn=&jtitle=&btitle=&rft_id=info:eric/&rft_id=info:doi/http://epubs.surrey.ac.uk/855012/.

[74] X. Chen, S. Jia, Y. Xiang, A review: knowledge reasoning over knowledge graph, Expert Syst. Appl. vol. 141 (2020), https://doi.org/10.1016/j.eswa.2019.112948.

[75] G. Zhou, et al., Clinical decision support system for hypertension medication based on knowledge graph, Comput. Methods Prog. Biomed. vol. 227 (2022) 107220, https://doi.org/10.1016/j.cmpb.2022.107220.

[76] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. vol. 22 (10) (2010) 1345–1359, https://doi.org/10.1109/tkde.2009.191.

[77] T.T. Verlekar, P. Lobato Correia, and L.D. Soares, Using Transfer Learning for Classification of Gait Pathologies, Presented at the 2018 Ieee International Conference on Bioinformatics and Biomedicine (BIBM), 2018.

[78] S.R. Simon, Quantification of human motion: gait analysis—benefits and limitations to its application to clinical problems, 2004/12/01/, J. Biomech. vol. 37 (12) (2004) 1869–1880, https://doi.org/10.1016/j.jbiomech.2004.02.047.

[79] B. Horsak, et al., Concurrent validity of smartphone-based markerless motion capturing to quantify lower-limb joint kinematics in healthy and pathological gait, J. Biomech. vol. 159 (2023) 111801, https://doi.org/10.1016/j.jbiomech.2023.111801.