Artificial Intelligence for Clinical Gait Diagnostics of Knee Osteoarthritis: An Evidence-based Review and Analysis

Luca Parisi[1,2], Narrendar Ravichandran[2], Matteo Lanzillotta[2,3]

[1]*Faculty of Business and Law (Artificial Intelligence Specialism), Coventry University, Priory Street, Coventry, CV1 5FB, United Kingdom*
[2]*University of Auckland Rehabilitative Technologies Association (UARTA), University of Auckland, 11 Symonds Street, Auckland, 1010, New Zealand*
[3]*Department of Counselling Psychology, Institute of Systemic Psychotherapy "Centro Studi Eteropoiesi", Corso Francia 98, Turin, 10143, Italy.*

*** Corresponding author.**
E-mail address: luca.parisi@ieee.org (L. Parisi).
E-mail address: narrendar@ieee.org (N. RaviChandran).
E-mail address: amatt.do@gmail.com (M. Lanzillotta).

Abstract

Background
Knee osteoarthritis (OA) remains a leading aetiology of disability worldwide. With recent advances in gait analysis, clinical assessment of such a knee-related condition has been improved. Although motion capture (mocap) technology is deemed the gold standard for gait analysis, it heavily relies on adequate data processing to yield clinically significant results. Moreover, gait data is non-linear and high-dimensional. Due to missing data involved in a mocap session and typical statistical assumptions, conventional data processing methods are unable to reveal the intrinsic patterns to predict gait abnormalities.
Research question
Albeit studies have demonstrated the potential of Artificial Intelligence (AI) algorithms to address these limitations, these algorithms have not gained wide acceptance amongst biomechanists. The most common AI algorithms used in gait analysis are based on machine learning (ML) and artificial neural networks (ANN). By comparing the predictive capability of such algorithms from published studies, we assessed their potential to augment current clinical gait diagnostics when dealing with knee OA.
Methods
Thus, an evidence-based review and analysis were conducted. With over 188 studies identified, 8 studies met the inclusion criteria for a subsequent analysis, accounting for 78 participants overall.
Results
The classification performance of ML and ANN algorithms was quantitatively assessed. The test classification accuracy (ACC), sensitivity (SN), specificity (SP) and area under the curve (AUC) of the ML-based algorithms were clinically valuable, i.e., all higher than 85%, differently from those obtained via ANN.
Significance
This study demonstrates the potential of ML for clinical assessment of knee disorders in an accurate and reliable manner.

Keywords - ***Artificial Neural Networks; Machine Learning; Biomechanics; Gait; Knee; Osteoarthritis.***

1 Introduction

1.1 Gait analysis to aid diagnosis of knee-related disorders

Gait analysis involves the assessment of human motion. By revealing the underlying physiological walking, it is essential to recognise lower limb-related pathologies. Motion capture (mocap) technologies are deemed the gold standard for gait data acquisition. Mocap involves the use of reflective markers and infrared cameras to estimate the position of the human body in three dimensions (3D). Kinematic parameters (joint angles) describing the movement of the human body can be inferred from mocap data. Moreover, by integrating instrumented force platforms, kinetic data (forces and moments) can also be measured. Considering the impact of knee disorders, the disruption of the anterior cruciate ligament (ACL) is a widespread debilitating injury amongst athletes involved in sports activities, such as soccer, football and basketball [1]. ACL may lead to a long unwanted absence from sports, pain, disability, with the risk of developing knee osteoarthritis (OA) [1, 2]. Knee OA is the most common joint disorder worldwide [3, 4]. Such a knee-related condition typically requires surgical intervention and months of rehabilitation. Furthermore, the incidence of symptomatic knee OA is increasing due to the ageing population and obesity, particularly in developed countries [3]. Quantitative analysis of the gait can help clinicians diagnose knee-related conditions by recognising deviations from physiological gait, e.g., knee adduction moment (KAM) in patients with knee OA [5, 6]. However, analysis of mocap data heavily relies on adequate experimental setup, pre- and post-processing methods, e.g., gap filling between marker trajectories, adequate band-pass filter design and thresholds. Although statistical tools are widely used in gait analysis, given the nonlinearity and high dimensionality of gait data [7], they are only analytical and lack predictive capability to generalise to unseen data.

1.2 Artificial Intelligence to improve clinical gait diagnostics

Artificial Intelligence (AI)-based algorithms can help overcome the above-mentioned limitations. They can enable classification of gait data and generalisation in their predictive outcomes [8]. Furthermore, these techniques can be effective in both supervised [9, 10] and unsupervised scenarios [8]. Whilst unsupervised AI learning-based algorithms, such as the self-organising map (SOM) [11, 12], Random Forest (RF) [13, 14], can classify gait data used as inputs without preliminarily knowing their true classes, supervised classifiers instead, such as the multi-layer perceptron (MLP) [15], the radial basis function (RBF) networks [16] and the Support Vector Machine (SVM) [14] require that the true classes of the input data are preliminarily known. Amongst AI-based methods, Machine Learning (ML) [7, 17-19] and Artificial Neural Networks (ANN) have proven to be accurate when dealing with gait-related data on patients with knee OA [2, 5, 6, 20, 21].

Artificial Neural Networks (ANNs) are brain-inspired computational algorithms. They involve numerous inputs and outputs, with some intermediate layers defined as ‘hidden’ [22]. The adjustment of weights in these layers occurs during the ‘learning’ or ‘training’ phase, wherein their *moduli* determine the degree of relevance towards the required classification [12, 22]. During training, the weights across these layers are updated iteratively to minimise the error during training (e.g., the Mean Squared Error or MSE) between the actual and predicted outputs [22]. In the ‘testing’ phase, ANNs apply the ‘learnt’ relationship to classify unseen data [12]. Supervised and unsupervised ANNs are respectively feed-forward (e.g. multi-layer perceptron (MLP) and radial basis function (RBF)) and feedback-type of architectures (e.g. self-organising maps (SOM)) due to their corresponding direction (forward or backward) of conveying

information. Whilst the MLP deploys either the logistic/sigmoid or the hyperbolic tangent sigmoid transfer functions [23], RBF uses the Gaussian transfer function [16]. MLP is trained via the back-propagation algorithm, whereby the initially randomised weighted inputs are propelled forward, whilst the errors are iteratively propagated backwards until an optimal set of weights generates the least MSE [22, 24]. The SOM uses competitive learning and deploys the Kohonen neighborhood transfer function [11], whereby input data features are clustered based on distance metrics between data points [15].

Amongst Machine Learning (ML)-based methods for gait analysis, the SVM [25] is the most widely applied technique [19, 26]. SVMs map the training samples via kernel functions into a high dimensional space and apply a decision surface boundary as an optimal separating hyperplane (OSH) for classification [26].

Via AI, gait-related indices could be derived as subject-specific metrics to assess the impact of gait-driven rehabilitation in patients with knee OA. Due to the lack of evidence-based, quantitative analysis on the efficacy of these algorithms, the use of AI-based tools in clinical gait diagnostics is still limited [27]. Moreover, none of the studies published so far as a topical review in gait analysis [28] has been able to offer such a comprehensive and quantitative analysis. Let alone to select an algorithm for aiding either diagnosis or assessment of prognosis of knee OA, biomechanists currently do not have any objective ground truth whereby they could choose amongst several algorithms, as well as the time and expertise involved in understanding the tools in question (AI).

To the best of the authors' knowledge, this is the first study that has performed an evidence-based analysis on AI-related studies in clinical gait diagnostics tailored to patients with knee OA. Besides providing an evidence-based review, a star-rating system for quality assessment of relevant literature has been formulated. The scope of this rating system is not only limited to this study but can also be applied in any AI-related studies involving healthcare data.
By performing a methodological evaluation of the eligible articles, the validity of the inferences drawn in this study were further ascertained. This review is hoped to have a significant impact in the field of Clinical Biomechanics in promoting the clinical application of AI-based methods to aid diagnosis and/or assessment of prognosis of several lower limb pathologies.

2 Methods

The high-level objectives of this evidence-based review and analysis are the following:

1. To identify studies that have used AI for gait analysis on knee OA and conduct a methodological quality assessment for their inclusion to conduct an evidence-based analysis;
2. To assess the predictive capability of such studies by comparing supervised and unsupervised models via an evidence-based analysis approach, with respect to test classification accuracy (ACC) and further performance measures, such as sensitivity (SN), specificity (SP) and area under the curve (AUC);
3. To compare studies that have implemented ANN- and ML-based algorithms in clinical gait diagnostics for knee OA, with respect to the above-mentioned performance measures.

An evidence-based analysis was carried out via *Review Manager (RevMan) (Version 5.3. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration, 2014).* Further to performing the I^2 test to assess the heterogeneity amongst selected studies [29], a sensitivity analysis was carried out to discard studies that could have biased the results. Furthermore, a statistical analysis on the accuracy and reliability measures of the algorithms reviewed was performed via *IBM SPSS Statistics (IBM Corp. Released 2016. IBM SPSS Statistics, Version 24.0. Armonk, NY: IBM Corp.).* Although gait data include spatiotemporal and metabolic data more broadly, only kinetic and kinematic data were considered in this study, as they are gold standard variables to quantify human motion in clinical gait diagnostics, and, thus, were collectively referred to as "gait data".

2.1 Inclusion and exclusion criteria

Full-length papers and conference articles were initially screened, and their titles and abstracts were assessed for eligibility against the aims and high-level objectives of this study by all four authors independently.

All major electronic databases including PubMed, Web of Science, MEDLINE, ScienceDirect, Scopus, Google Scholar, IEEE Xplore, Springer, Wiley, O'Reilly, SAGE, Cochrane, Embase were parsed with relevant keywords of interest, e.g., knee osteoarthritis, machine learning and artificial neural networks. Subsequent subscription-based and open-access articles published from 1984 (since when optometric mocap systems were first introduced) until 29/04/2018 were searched for. However, no relevant articles were found in MEDLINE, Cochrane, O'Reilly, Embase and Sage databases, which were thus discarded.

The exact keywords used to guide the literature survey were the following: "machine learning", "gait", "motion capture", "neural network", "multi-layer perceptron", "random forests", "support vector machine", "self-organizing maps" (US spelling), "self-organising maps" (UK spelling), "k-nearest neighbor" (US spelling), "k-nearest neighbour" (UK spelling) and "radial basis function".

Selected articles must have reported gait-related data on human participants with knee OA**,** regardless of any other demographic factor except for the age **(**older than 12 but younger than 75**)**. Studies reporting gait-related data on human subjects with neurotrauma or with neurodegenerative disorders, such as Parkinson's Disease, Cerebral Palsy or Huntington's Disease, were discarded from this review.

Studies in which kinematic data collected only using optoelectronic systems that uses RGB-D sensors (e.g., Vicon, Qualysis and Asus Xtion) were considered for inclusion. Moreover, body segments must have been identified via passive markers and any other methods

for measuring or estimating such gait metrics were not considered. Selected articles must have reported data on walking gait (speed lower than that of a healthy human subject, approximately less than 5 km/h, due to the knee OA). Any studies involving walking on instrumented treadmills (with embedded force platforms) and fall detection systems were discarded.

An initial search was performed only based on the title. Subsequently, a second and final search was carried out, after which relevant key articles were selected for inclusion based on their abstract and full-text content. All bibliographies from the retrieved key articles were also searched for potential articles that might not have been considered previously. The "Preferred Reporting Items for Systematic reviews and Meta-Analysis" (PRISMA) guidelines [30] were followed throughout this study. 188 studies were identified following a further screening. By applying the above-mentioned inclusion and exclusion criteria, 180 articles were excluded and 8 were selected. A methodological quality assessment was performed on these studies for their inclusion in the meta-analysis, as outlined in 2.2.

In case of incongruencies in the selected articles that could have not been clarified amongst the four authors and reviewers, the corresponding author of the selected articles was contacted for clarification, thus ascertaining whether the articles in question were eligible for inclusion, instead of discarding them *a priori*.

2.2 Methodological quality assessment: The UARTA star-rating system.

Adapted from the MQAS scale [32], a quality assessment on selected articles was performed via a star-rating system developed by the authors L.P. and N.R. at the University of Auckland Rehabilitative Technologies Association (UARTA). The "UARTA Star-rating System for Assessing Clinical Significance of Artificial Intelligence-related Research" is deemed applicable to any research articles dealing with AI applied to healthcare-related data. Each of the following points corresponds to a star (★) attributed to selected papers for meeting the criterion described in the statement next to it. A maximum of fifteen stars was attributed to each of the selected articles. Articles carrying less than seven stars were not considered for the review and meta-analysis.

- ★ Selected articles must have outlined a clear purpose for the classification task, specifying inputs/outputs, reported any data pre-processing steps undertaken to ensure accuracy and consistency of the results presented, and to enable their reproducibility. If those were not applicable, the authors of the selected articles must have justified why they were not.
- ★ Selected articles must have reported any data post-processing steps undertaken to ensure accuracy and consistency of the results presented, and to enable their reproducibility. If those were not applicable, the authors of the selected articles must have justified why they were not.
- ★ Selected articles must have reported a measure (number and/or percentage of the whole dataset) quantifying the training set of the data used.
- ★ Selected articles must have reported a measure (number and/or percentage of the whole dataset) quantifying the cross-validation set of the data used.
- ★ Selected articles must have reported a measure (number and/or percentage of the whole dataset) quantifying the testing set of the data used.
- ★ Selected articles must have reported the name of the cross-validation algorithm (e.g., holdout validation, leave-one-out (LOO), nested or k-fold cross-validation, specifying the number k of partitions made where applicable) to avoid overfitting and ensure reproducibility of the results attained.
- ★ Selected articles must have reported any qualitative outputs showing the training- and cross-validation-related mean squared error (MSE) curves against the number of iterations or epochs to illustrate at which iteration/epoch overfitting occurs. This step is fundamental to stop the training accordingly.
- ★ Satisfying the above-mentioned criterion provides evidence on the avoidance of overfitting, thus ensuring that the algorithms tested were truly learning from the data which were trained on, rather than solely 'remembering' the input features.
- ★ Selected articles must have reported testing or out-of-sample classification accuracy.
- ★ Selected articles must have reported at least one measure of error, e.g., the mean squared error (MSE) or cross entropy, or it should be clearly inferable from the performance measures reported.
- ★ Selected articles must have reported the sensitivity (SN).
- ★ Selected articles must have reported the specificity (SP).
- ★ Selected articles must have reported the area under the receiver operating characteristic curve (AUC) or, at least, the Pearson's product-moment coefficient of determination (r^2).
- ★ Selected articles must have reported any of the above-mentioned performance measures with confidence intervals.
- ★ Selected articles must have reported any qualitative outputs on the receiver characteristic curve (ROC) under which the AUC was computed.

Based on the inclusion and exclusion criteria in 2.1 and the quality assessment performed via the UARTA star-rating scale, eight (N=8) key articles were retrieved, as per the selection procedure outlined in Fig. 1.

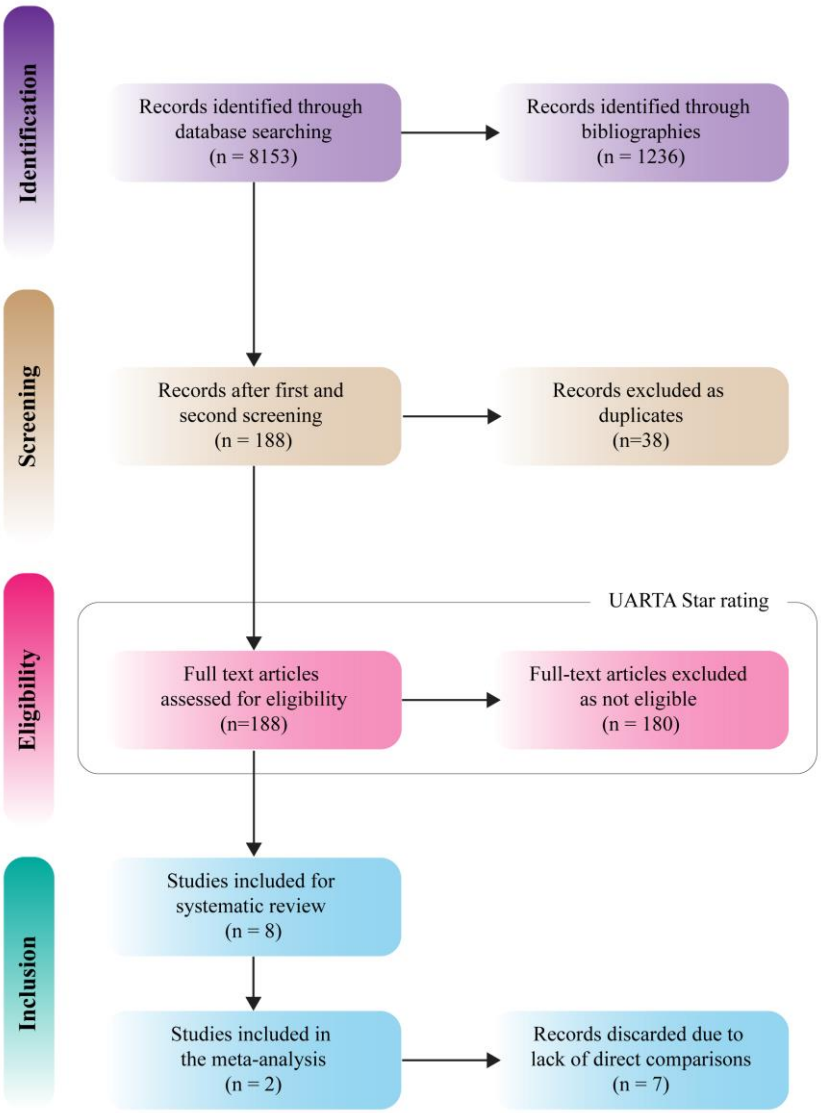

Figure 1. PRISMA flowchart showing the selection procedure adopted to recruit suitable articles for review.

Table 1 shows the stars attributed to each of the selected articles for meeting the UARTA star-rating scale-related criteria, as outlined above.

Table 1. Stars attributed to each of the selected articles for meeting the UARTA start-rating scale-related criteria.

Scoring attributes	[5]	[18]	[19]	[20]	[6]	[21]	[17]	[7]
Pre-processing	★	★		★		★	★	★
Post-processing	★				★	★	★	★
Training Dataset	★	★	★	★	★	★	★	★
Cross-validation dataset	★	★					★	★
Testing Dataset	★	★	★	★	★	★	★	★
Cross-validation Algorithm	★	★	★				★	★
Learning/training Measures	★		★				★	
Learning Performance	★							
Classification Accuracy	★	★	★	★	★	★	★	★
Measure of Error	★	★		★	★	★		
SN		★	★			★	★	
SP		★	★			★	★	
AUC			★		★		★	
CI								
ROC			★					
Total	10	10	9	5	7	8	11	7

Table 2 summarises the main elements derived from a comprehensive review of the selected studies, as per the UARTA star-rating quality assessment scale described above.

Table 2. Comprehensive review of included studies.

		[5]	[19]	[18]	[21]	[6]	[20]	[17]	[7]
Aim		To predict knee adduction moment using force plates and anthropometric measurements	To predict pathological gait patterns pre-operatively	To quantify pathological differences in gait phases and joint angles	To assess the viability of ensemble classifiers for recognising pathological gait patterns	To predict knee joint moments from the movement of body segments	To recognise pathological gait patterns using joint angles and spatio-temporal gait parameters	To classify pathological gait patterns pre-and post-operatively	To recognise gender-independent and gender-dependent differences in pathological gait.
Condition		Knee OA	Knee OA	Knee OA	Knee OA	Knee OA	Knee OA	Knee OA	Knee OA
Type of study		Regression	Classification	Classification	Classification	Regression	Classification	Classification	Classification
Subject classes (N)		Asymptomatic OA (N=28), Mild OA (N=28), Severe OA (N=28)	Healthy (N=12) Symptomatic OA (N=11)	Healthy (N=15) Symptomatic OA (N=32)	Healthy (N=91) OA (N=110)	Alkaptonuria with knee OA (N=31)	Healthy (N=91) OA (N=110)	Healthy (N=12) Symptomatic OA (N=11)	Healthy (N=43) Symptomatic OA (N=100)
Data	**Training set (N, %)**	-	51 trials (73.91%)	21 subjects (44.68%)	138 subjects (66.8%)	22 subjects (70%)	138 subjects (66.8%)	69 trials (100%)	112 subjects (80%)
	Cross-validation (CV) set	-	Leave-one-out (LOO)	5-fold	-	-	-	Leave-one-out (LOO)	10-fold
	Testing set (N, %)	-	18 (26.09%)	21 (44.68%)	63(31.3%)	9 (30%)	63(31.3%)	69 (100%)	14 (10%)
Mocap technology and sampling rate		120 Hz	-	Vicon (200 Hz)	-	Qualisys, Oqus and Vicon	-	Vicon (50Hz)	Vicon (120 Hz)
Force platform technology and sampling rate		-	Kistler force plate (400 Hz)	-	-	-	-	Kistler force plate (400 Hz) and GAITRite	Bertec instrumented treadmill
Gait data type		Kinetics and kinematics	Kinetics and Kinematics	Kinematics	Kinematics	Kinetics and kinematics	Kinematics	Spatio-temporal and symmetry indices	Kinematics
Data pre-processing		Low-pass filter	-	Normalisation per gait cycle	Removal of missing values	Normalisation per gait cycle	Removal of missing values	-	Normalisation per gait cycle
Gait-related input features (N)		Twelve (N=12) features: Ground reaction forces - x,y,z , relative velocities -x,y,z), relative displacements - x,y,z , stance duration, time point knee axis alignment.	Twelve (N=12) features: walking velocity, cadence, stride length, stride time, step time, step length, single support time, double support time.	Two (N=2) features: gait phase deviation and the joint function deviation reference-based indices (GCD-RBI and JFD-RBI)	Five (N=5) feature vectors: four for temporal changes of knee joint angle (KFlex, KMFlex, KMVal, KPTot), one for time-distance parameters	3,131 instances, consisting of 12 columns of angles (ankle - x, y, z, knee - x, y, z, hip - x, y, z, pelvis - x, y, z)	Five (N=5) feature vectors four for temporal changes of knee joint angle (KFlex, KMFlex, KMVal, KPTot), one for time-distance parameters	Four (N=4) spatio-temporal parameters: speed, cadence, stride length, stride time. Four (N=4) symmetry indices: step length, step time, single and double support time.	For each of the groups, two feature vectors were created based on the original discrete variables and a principal component analysis (PCA).
Attributes of the classifiers	**Architecture**	MLP	SVM	Quadratic SVM and KNN	MLP	RF, DT and MLP	MLP	SVM	SVM
	Type of learning/training	Backpropagation	SVMs were trained over the range C = {0.1, 1, 10, 100, 1000}	-	-	RF: ensemble of un-pruned decision tree. DT: C4.5 decision tree; MLP: -	-	SVM (kernel width s=0.5, penalty parameter C=10).	-
	Transfer function: Hidden Layer	Hyperbolic tangent sigmoid	Linear, polynomial and Gaussian kernels	SVM: quadratic kernel KNN: nearest-neighbor	Tangent sigmoid	-	Tangent sigmoid	Gaussian kernel	Linear kernel
	Output layer	Linear function	-	-	Tangent sigmoid	-	Tangent sigmoid	-	-

	Input vectors (N)	12	12	2	207*	3,131	207*	69	8
	Hidden layers (N)	1	-	-	2*	1	1*	-	-
	Hidden neurons in the first layer (N)	7	-	-	150*	10	50*	-	-
	Hidden neurons in the second layer (N)	-	-	-	40*	-	-	-	-
	Output vectors (N)	1	1	1	1*	1	1*	1	1
Training ...	**Type**	Levenberg-Marquardt	-	-	-	-	-	-	-
	Learning rate	-	-	-	-	-	-	-	-
	Momentum	-	-	-	-	-	-	-	-
	Epochs	-	-	-	-	-	-	-	-
Post-processing		-	-	-	-	-	-	-	-
Software deployed		MATLAB (Mathworks, MA)	-	-	MATLAB (Mathworks, MA)	-	MATLAB (Mathworks, MA)	-	-
Classifier performance	**Accuracy**	median R for group curves 99.8%	88.89%	SVM: 85%; KNN: 87%	90.48%	RF(r2):96.27% DT(r2):67.67% MLP(r2):86.16%	90.48%	94.2%	99%
	Sensitivity (SN)	median R: 74.2%	-	SVM: 91%; KNN:93%;	96.55%	-	96.55%	100%	-
	Specificity (SP)	77.78%	-	SVM: 75%; KNN: 75%	85.29%	-	85.29%	97%	-
	AUC	0.836	-	-	-	RF: 0.889 DT: 0.829 MLP: 0.874	-	-	-
Limitations/Lack of reporting on		SN; SP; ACC; ROC.	Pre-processing; post-processing; cross-validation dataset; learning performance; measure of error; CI.	Post processing; Learning/training measures; Learning performance; AUC; CI; ROC.	Cross-validation set; cross-validation algorithm; learning/training measures; learning performance; AUC; CI; ROC.	Pre-processing; cross-validation dataset; cross-validation algorithm; learning measures; learning performance; SN; SP; CI; ROC.	Post-processing; cross-validation dataset; cross-validation algorithm; learning/training measures; learning performance; SN; SP; AUC; CI; ROC.	Learning performance; measures of error; CI; ROC.	Learning measures; learning performance; measures of error; SN; SP; AUC; CI; ROC.

*data pertains to the MLP6 classifier, which was the most accurate individual MLP tested using the lowest number of hidden layers and hidden neurons amongst the multiple MLPs tested. This enables direct comparison of performance with the other studies.

3 Results

Table 3 and fig. 2.a illustrate the results obtained via a Forest plot indicating that the test classification accuracy obtained via ML-based algorithms was not statistically different with respect to that attained via ANN (mean difference (MD): 13.84%, 95% CI: -13.45, 41.12, p=0.32). Furthermore, the heterogeneity was considerably high (I^2=93%), which is due to the very low number of studies (N=2) that compared the performance of ML and ANN directly, i.e., in the same study.

Table 3. Forest plot showing the mean difference (MD) in test classification accuracy (mean±standard deviation) between machine learning (ML)-based algorithms and artificial neural networks (ANN) using gait-related kinetic and kinematic data.

Study	ML			ANN			Mean Difference	
	Mean	*SD*	*Total*	*Mean*	*SD*	*Total*	*Weight*	*IV, Random,95% CI*
[6]	81.10	14.06	31	52.41	37.56	31	46.80%	28.69 [14.57, 42.81]
[18]	80.05	5.76	47	79.26	6.12	47	53.20%	0.79 [-1.61, 3.19]
Total (95% CI)			78			78	100%	13.84 [-13.45, 41.12]
Heterogeneity: Tau^2 =362.48; Chi^2=14.58, df=1(p=0.00); I^2=93%								
Test for overall effect: Z=0.99 (P=0.32)								

Fig. 2.b shows a Funnel plot indicating that, since the studies (N=2) are plotted near the mid-line representing the average MD, there was no publication bias in the results obtained.

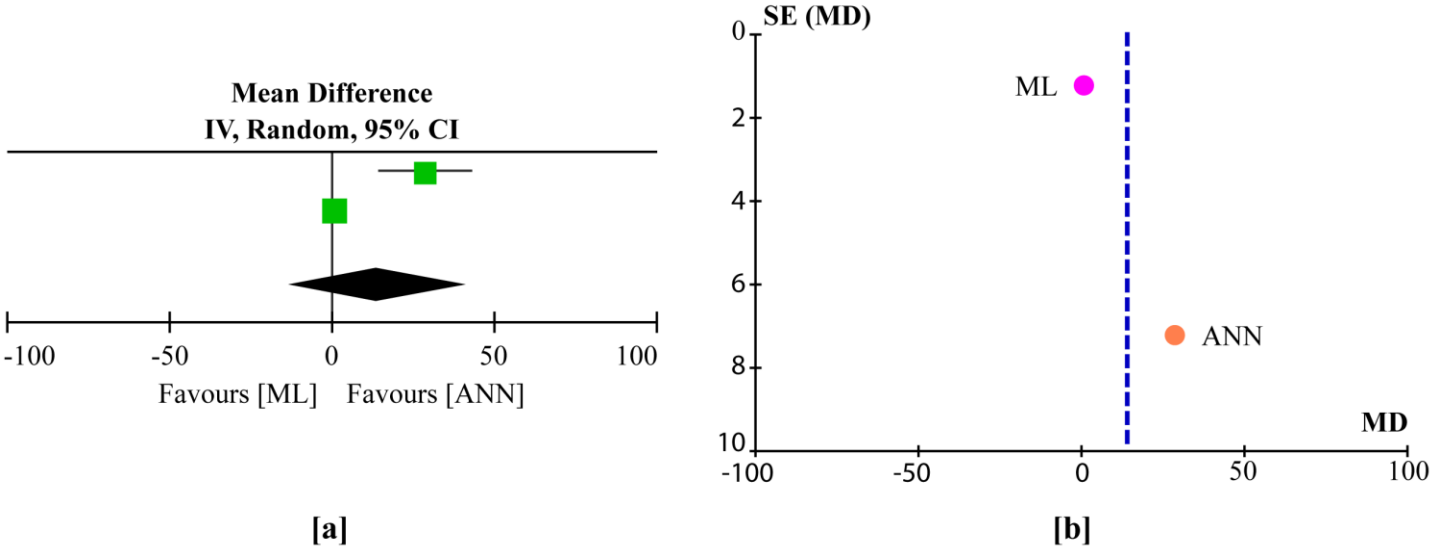

Figure 2. **a)** Forest plot comparing the test classification accuracy of ML- and ANN-based algorithms in clinical gait diagnostics from selected studies eligible for the evidence-based analysis as including both performance data in the same study; **b)** Funnel plot indicating no publication bias in the results obtained.

Table 4 summarises the results obtained from a statistical analysis on the classification outcomes reported in selected studies. Whilst the test classification accuracy was reported in eight (N=8) selected studies, the sensitivity and specificity were reported in four (N=4) studies, with the area under the curve being reported only in two (N=2) studies.

Table 4. Results from the descriptive statistics and correlation analysis.

Performance measure	Studies (N)	ML Mean (%)	ML Standard deviation (%)	ANN Mean (%)	ANN Standard deviation (%)	Correlation (r^2, p-value)
Accuracy (ACC)	8	89.81	6.88	90.69	5.42	1, p=0.01
Sensitivity (SN)	4	95.50	6.36	87.92	12.01	N/A
Specificity (SP)	4	86.00	15.56	79.36	5.32	N/A
Area under the curve (AUC)	2	0.86	0.00	0.85	0.03	N/A

4 Discussion

Previous reviews on the use of AI in clinical gait diagnostics were purely qualitative [28]. Instead, in this study not only a qualitative analysis of previous research findings was carried, but also a quantitative one was performed using a novel evidence-based analysis approach in AI for clinical gait diagnostics, which yielded two eligible studies that compared the classification performance of ML- and ANN-based methods directly, i.e., in the same study. Aljaaf et al. [6] were able to capture underlying gait patterns from the movement of body segments and quantified the knee adduction moment of the ankle, knee (KAM), hip and pelvis from their corresponding Euler angles during a single gait cycle. As compared to other ANN-based techniques, MLP was the most accurate algorithm tested (r^2=0.86, root mean squared error (RMSE)=0.07). Karg *et al*. [18] applied a quadratic SVM to quantify pathological discrepancies in gait phases and joint angles between patients with symptomatic OA-related gait abnormalities and healthy subjects deploying spatio-temporal parameters with a classification accuracy of 85%. Patients with knee OA tendentially were found to have decreased walking speed, cadence, stride and step lengths when considering both lower limbs, a reduced time taken in single support but a longer time when in double support.

To the best of the authors' knowledge, the UARTA-star rating quality assessment scale represents the first ever clinical gait diagnostics-equivalent (when dealing with AI-based clinical decision-making) of the MQAS scale [32] and the Newcastle-Ottawa quality assessment scale [31], the latter being used to assess findings from cohort studies (prospective or retrospective) published in the medical literature. Published studies lack consistency in reporting machine learning-based research findings as shown in Table 2 and as evident from the analysis in Table 4, where not all (N=8) selected studies had reported all the main performance measures of the AI-based algorithms tested on gait data. Remarkably, only two (N=2) studies reported the area under the curve (AUC) (Table 4), which is one of the most important performance measures for AI algorithms.

In selected papers that directly compared classification outcomes between machine learning (ML)- and artificial neural networks (ANN)-based architectures (N=2) [6, 18], ML-based algorithms were found to consistently have a higher accuracy and a lower standard deviation than those of ANN, thus being more stable in dealing with gait data (Table 3). The reduced performance in ANN-based algorithms may be partly explained by the re-sampling occurring within ANN-based architectures, such as the multi-layer perceptron (MLP), where weights and biases are adjusted iteratively, and so the data is continuously resampled until the mean squared error drops below a preset threshold that is deemed acceptable to stop training the ANN. However, such mean difference was not significant (p=0.32, Table 3 and Fig. 2.a) and the heterogeneity between studies was also considerably high (I^2=93%, Table 3). Instead, when considering outcomes on test classification accuracy of the AI-based algorithms from the eight studies (N=8) that reported such a performance measure, the ANN seems to have a slightly higher accuracy with a slightly lower standard deviation than the ML-based ones, being both highly correlated between one another (r-squared=1.00, p=0.01; Table 4). Nevertheless, as also shown in Table 4, all reliability-related performance measures of ML-based algorithms (SN=95.50%, SP=86.00, AUC=0.86) were consistently higher than those of ANN-based architectures (SN=87.92%, SP=79.36, AUC=0.85). These apparently contradictory results further support the development and use of the UARTA star-rating quality assessment scale for clinical gait diagnostics-related studies using machine learning for three main purposes:

1. qualitatively evaluating the technical rigour and quality of such studies;
2. quantitatively perform the first ever objective evidence-based analysis (in this study) on results reported in published studies;
3. providing guidelines to biomechanists and clinicians on which algorithm would be more accurate and reliable.

Fig. 2.b shows no publication bias, as the mean differences from the selected studies are close to the midline of the graph indicating the mean MD. Therefore, the lack of publication bias supports the reliability of the above-mentioned conclusions derived from analysing the results reported in Table 3. Moreover, Table 3 seems to suggest that ANN can be used when dealing with gait data collected on patients with knee OA [18], whilst ML seems to generalise to patients with any other knee-related conditions better [6, 18].

Whilst the test classification accuracy was reported in four studies, the sensitivity and specificity were mentioned in four studies, with the area under the curve being reported in two studies only. The limited size of the data available for the evidence-based analysis and statistical analysis is a major limitation of this study, which, indeed, highlights an even greater limitation in the reporting machine learning-related results in the literature. Nevertheless, the main limitation of this study remains the small sample of studies reviewed, which met the inclusion criteria for eligibility. Therefore, it is hard to draw definitive conclusions.

To summarise, with respect to applications in clinical gait diagnostics, whilst both ANN- and ML-based algorithms attempt to mimic the learning-related mechanisms occurring in the brain and can handle nonlinear and highly dimensional data, they require adequate data pre-processing (removal of outliers, at times normalization or standardisation of inputs) and do not directly yield physiologically interpretable results.

The development and validation of the UARTA star-rating quality assessment scale seeks to change such a *status quo* and obviate the lack of appropriate, consistent and thorough reporting on machine learning-related findings in the clinical gait diagnostics literature. The implementation of this set of standards for selection criteria is also intended to guide the development and testing of AI-based algorithms in clinical gait diagnostics, such that progress in AI research can be promptly translated in readily available and thoroughly validated tools that biomechanists and clinicians can easily use to aid diagnosis and/or assessment of prognosis of lower limb disorders and/or pathologies worldwide.

5 Conclusion

This study establishes clear design criteria for selecting and deploying Artificial Intelligence (AI)-based algorithms for diagnostic and/or prognostic purposes in clinical gait diagnostics, particularly when dealing with data on patients with knee-related conditions. A concise but comprehensive description of the main AI learning-based algorithms was provided (ANN and ML). A quantitative analysis enabled the definition of criteria for selecting the most accurate and reliable AI-based algorithm to apply in a clinical setting. Based on this analysis, the test classification accuracy (ACC), sensitivity (SN), specificity (SP) and area under the curve (AUC) of the ML-based algorithms analysed were found to be clinically valuable, i.e., all higher than 85% (ACC=89.81±6.88%; SN=95.50±6.36%; SP=86.00±15.56%; AUC=0.86±0.00), differently from those obtained via ANN (SP=79.36±5.32%).

Biomechanists have so far applied AI-based algorithms without having any standards for guiding adequate selection and implementation of such tools. Via the development and validation of the UARTA star-rating quality assessment scale for machine learning-based studies in clinical gait diagnostics, we attempted to define a set of initial standards, guidelines that can promote a thorough and prompt translational application of previous research findings and AI-based algorithms. It is hoped that the UARTA scale will be considered when international standards will be outlined on appropriate and consistent reporting of findings from clinical gait diagnostics-related studies in which machine learning was used.

AI can revolutionise and objectify best practices in clinical gait diagnostics, augmenting the capabilities of biomechanists to aid diagnosis and assessment of prognosis in patients with knee-related conditions.

ACKNOWLEDGMENTS

The authors would like to thank the University of Auckland Rehabilitative Technologies Association (UARTA) for giving them the chance of developing this collaborative research work. The authors would also like to thank Professor Tom Chau, pioneer and author of a qualitative review focused on Artificial Intelligence-based techniques in gait analysis in 2001, from the University of Toronto, Canada, for his assistance in reviewing this manuscript.

DECLARATION OF INTEREST

The authors declare no conflicts of interest.

CONTRIBUTORS

All authors directly participated in the planning, execution and analysis in the study. Each of the authors has read and concurs with the content in the final manuscript.

FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

[1] Moustakidis, S. P., Theocharis, J. B., & Giakas, G. (2010). A fuzzy decision tree-based SVM classifier for assessing osteoarthritis severity using ground reaction force measurements. *Medical Engineering and Physics*, 32(10), 1145-1160.

[2] Kusakunniran, W., Prachasri, N., Dirakbussarakom, N., & Yangchaem, D. (2017, February). Distinguishing ACL patients from healthy individuals using multilayer perceptron on motion patterns. In *Knowledge and Smart Technology (KST), 2017 9th International Conference* on (pp. 1-5). IEEE.

[3] Sowers, M. R., & Karvonen-Gutierrez, C. A. (2010). The evolving role of obesity in knee osteoarthritis. *Current opinion in rheumatology*, 22(5), 533.

[4] Neogi, T. (2013). The epidemiology and impact of pain in osteoarthritis. *Osteoarthritis and Cartilage*, 21(9), 1145-1153.

[5] Favre, J., Hayoz, M., Erhart-Hledik, J. C., & Andriacchi, T. P. (2012). A neural network model to predict knee adduction moment during walking based on ground reaction force and anthropometric measurements. *Journal of biomechanics*, 45(4), 692-698.

[6] Aljaaf, A. J., Hussain, A. J., Fergus, P., Przybyla, A., & Barton, G. J. (2016, July). Evaluation of machine learning methods to predict knee loading from the movement of body segments. In *Neural Networks (IJCNN), 2016 International Joint Conference* on (pp. 5168-5173). IEEE.

[7] Phinyomark, A., Osis, S. T., Hettinga, B. A., Kobsar, D., & Ferber, R. (2016). Gender differences in gait kinematics for patients with knee osteoarthritis. *BMC musculoskeletal disorders*, 17(1), 157.

[8] Parisi, L., Biggs, P. R., Whatling, G. M., & Holt, C. A. (2015). A Novel Comparison of Artificial Intelligence Methods for Diagnosing Knee Osteoarthritis. In *XXV Congress of the International Society of Biomechanics*, 1227-1229.

[9] Parisi, L. (2014a). Exploiting Kinetic and Kinematic Data to Plot Cyclograms for Managing the Rehabilitation Process of BKAs by Applying Neural Networks. *Int. J. Biomed. Biol. Eng.*, 8(10), 664–668.

[10] Parisi, L. (2014b). Neural Networks for Distinguishing the Performance of Two Hip Joint Implants on the Basis of Hip Implant Side and Ground Reaction Force. *Int. J. Medical, Heal. Pharm. Biomed. Eng.*, 8(10), 659–663.

[11] Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1), 59-69.

[12] Barton, J. G., & Lees, A. (1997). An application of neural networks for distinguishing gait patterns on the basis of hip-knee joint angle diagrams. *Gait & Posture*, 5(1), 28-33.

[13] Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.

[14] Criminisi, A., Robertson, D., Konukoglu, E., Shotton, J., Pathak, S., White, S., & Siddiqui, K. (2013). Regression forests for efficient anatomy detection and localization in computed tomography scans. *Medical image analysis*, 17(8), 1293-1303.

[15] Barton, G., Lisboa, P., Lees, A., & Attfield, S. (2007). Gait quality assessment using self-organising artificial neural networks. *Gait & posture*, 25(3), 374-379.

[16] Park, J., & Sandberg, I. W. (1991). Universal approximation using radial-basis-function networks. *Neural computation*, 3(2), 246-257.

[17] Levinger, P., Lai, D. T., Webster, K., Begg, R. K., & Feller, J. (2007, August). Support Vector Machines for detecting recovery from knee replacement surgery using quantitative gait measures. In Engineering in Medicine and Biology Society, 2007. *EMBS 2007. 29th Annual International Conference of the IEEE* (pp. 4875-4878). IEEE.

[18] Karg, M., Seiberl, W., Kreuzpointner, F., Haas, J. P., & Kulić, D. (2015). Clinical gait diagnostics: Comparing explicit state duration HMMs using a reference-based index. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 23(2), 319-331.

[19] Levinger, P., Lai, D. T., Begg, R. K., Webster, K. E., & Feller, J. A. (2009). The application of support vector machines for detecting recovery from knee replacement surgery using spatio-temporal gait parameters. *Gait & posture*, 29(1), 91-96.

[20] Koktas, N. S., Yalabik, N., & Yavuzer, G. (2006a, November). Combining neural networks for gait classification. In *Iberoamerican Congress on Pattern Recognition* (pp. 381-388). Springer, Berlin, Heidelberg.

[21] Koktas, N. S., Yalabik, N., & Yavuzer, G. (2006b, December). Ensemble classifiers for medical diagnosis of knee osteoarthritis using gait data. In Machine Learning and Applications, 2006. *ICMLA'06. 5th International Conference* on (pp. 225-230). IEEE.

[22] Michie, D., Spiegelhalter, D. J., Taylor, C. C. And Campbell, J. (1994). Machine learning, neural and statistical classification. Edited by D. Michie, D. J. Spiegelhalter, C. C. Taylor, and J. Campbell. Upper Saddle River, NJ, USA: Ellis Horwood.

[23] Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4), 303-314.

[24] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533.

[25] Vapnik, V. (1998). Statistical learning theory. 1998. *Wiley*, New York.

[26] Begg, R., & Kamruzzaman, J. (2005). A machine learning approach for automated recognition of movement patterns using basic, kinetic and kinematic gait data. *Journal of biomechanics*, 38(3), 401-408.

[27] Benedetti, M. G., Beghi, E., De Tanti, A., Cappozzo, A., Basaglia, N., Cutti, A. G., ... & Fantozzi, S. (2017). SIAMOC position paper on gait analysis in clinical practice: General requirements, methods and appropriateness. Results of an Italian consensus conference. *Gait & posture*, 58, 252-260.

[28] Chau, T. (2001). A review of analytical techniques for gait data. Part 2: neural network and wavelet methods. *Gait & posture*, 13(2), 102-120.

[29] Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. BMJ: *British Medical Journal*, 327(7414), 557.

[30] Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Prisma Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS medicine*, 6(7), e1000097.

[31] Stang, A. (2010). Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses. *European journal of epidemiology*, 25(9), 603-605.

[32] Parisi, L., RaviChandran, N., & Manaog, M. L. (2019). A novel hybrid algorithm for aiding prediction of prognosis in patients with hepatitis. *Neural Computing and Applications*, 1-14, https://doi.org/10.1007/s00521-019-04050-x.