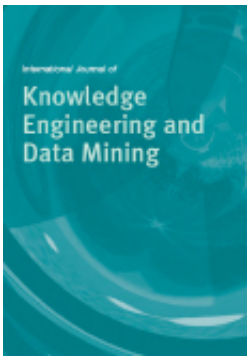


International Journal of Knowledge Engineering and Data Mining

ISSN online: 1755-2095 - ISSN print: 1755-2087
https://www.inderscience.com/ijkedm

Adaptive density-peaks clustering for gait analysis

Sinan Onal, Somaiya Khan Islam

DOI: 10.1504/IJKEDM.2022.10043817

Article History:

Received:	16 August 2021
Accepted:	24 October 2021
Published online:	10 October 2022

Adaptive density-peaks clustering for gait analysis

Sinan Onal* and Somaiya Khan Islam

Department of Industrial Engineering,
Southern Illinois University, Edwardsville
1 Hairpin Dr. Campus Box 1805,
62026, Edwardsville, IL, USA
Email: sonal@siue.edu
Email: sokhan@siue.edu
*Corresponding author

Abstract: Gait analysis compares the gait characteristics of people with health issues to those of a control group in order to detect gait abnormalities. This comparison is carried out by evaluating a number of gait parameters with discrete values. Gait data, on the other hand, is time-series data and must be assessed using a different approach. The purpose of this study was to develop a quantitative measure that takes into account time-series data for comparing the gait characteristics of two groups of individuals using clustering. The gait data were collected using an optical motion capture system. An adaptive density-peaks clustering technique with a shape-based similarity measure was employed to compare gait characteristics. The results demonstrate that the proposed adaptive density-peaks clustering technique, which employs dynamic derivative time wrapping distance measurement, outperforms three state-of-the-art clustering algorithms for comparing the gait characteristics using time-series gait data.

Keywords: density-peaks clustering; time-series analysis; gait analysis; motion capture system; biomechanics.

Reference to this paper should be made as follows: Onal, S. and Islam, S.K. (2022) 'Adaptive density-peaks clustering for gait analysis', *Int. J. Knowledge Engineering and Data Mining*, Vol. 7, Nos. 3/4, pp.145–162.

Biographical notes: Sinan Onal is an Associate Professor in the Department of Industrial Engineering and the Director of the NSF-funded Motion Capture and Analysis Laboratory. His research interests lie in the broad areas of computational biomechanics, machine learning and data mining, medical image processing and its applications, and engineering education. His research has been sponsored by the National Science Foundation, and SIUE Seed Grants for Transitional and Exploratory Projects. He was recognised by the Graduate School as the recipient of the FY2020-2022 Hoppe Research Professor, a prestigious award that annually distinguishes and supports the scholarly endeavors of a faculty member.

Somaiya Khan Islam earned a Masters degree in Industrial Engineering from Southern Illinois University, Edwardsville. Throughout her graduate studies, she worked at the Motion Capture and Analysis Laboratory (MOCAL) on a variety of projects involving human motion and data analysis. She received her Bachelorette degree in Industrial Engineering from Ahsanullah University of Science and Technology in Bangladesh. She is currently employed in St. Louis, Missouri, as a Supply Chain Engineer

1 Introduction

Gait analysis plays a crucial role in diagnosing and rehabilitating patients with movement disorders (Papi et al., 2018). The terms 'kinematics' and 'kinetics' are used to describe the gait patterns. Kinematic gait parameters describe the linear and angular displacement, velocity, and acceleration of motion captured using multiple optical cameras, whereas kinetic parameters collected from force platforms describe the forces that cause motion and joint moments. Both parameters comprise valuable quantitative information in gait pattern analysis.

Traditional clinical gait analysis is used to compare gait characteristics and identify gait pattern differences in spatiotemporal, kinematic, and kinetic gait parameters of people with particular health conditions (e.g., post-stroke, cerebral palsy, autism spectrum disorder, etc.) using multiple gait cycles. This comparison analysis is usually done using descriptive statistics and independent t-tests. Pogemiller et al. (2020) investigated if differences exist in gait patterns between Charcot-Marie-Tooth (CMT) disease type I and type II in childhood to young adults using a two-tailed Student t-test (Pogemiller et al., 2020). Starbuck at al. (2021) used the Shapiro-Wilks test to determine whether or not walking patterns varied between individuals with late-onset Pompe disease. Discrete outcome measures such as maximum hip adduction angles during stance and maximum sagittal and frontal hip, knee, and ankle angles during stance were obtained from the kinematic and kinetic data. Fujita et al. (2020) assessed women's gait characteristics with distal radius fracture (DRF) using a Student t-test for continuous variables and the Chi-square tests for categorical variables. Krauss et al. (2012) compared gait variables for knee joint kinematics in subjects with knee osteoarthritis and healthy controls using paired t-test (Krauss et al., 2012). Wu et al. (2021) investigated the effects of levodopa on mild parkinsonian symptoms patients' gait performance (MPS). They conducted an acute levodopa challenge test to determine the effect of levodopa on the gait performance of (MPS) using the Kolmogorov–Smirnov test. Khan et al. (2020) examined if repeated movements used in sports training had an effect on children's gait. They examined the gait patterns of physically active children to those of age-matched controls. To evaluate their walking patterns, a motion capture and analysis system and a student t-test were utilised. However, traditional gait analysis uses discrete values such as mean, maximum, and range of gait parameters extracted from a sequence of gait data and neglects essential gait information embedded in the dynamic sequences. Gait data, on the other hand, is a time series that is an ordered sequence of values of a gait variable at equally spaced time intervals. Thus, the analysis of time-series gait data requires a more appropriate approach to investigate gait alterations. Hence, researchers have used other methods to analyse time-series gait data such as attractor attributes, nonlinear dynamics analysis (Iqbal et al., 2015), multifractal analysis, cross-correlations (Muñoz-Diosdado, 2005), maximal Lyapunov exponent (Vieten et al., 2013), and clustering. However, the examination of pattern similarity for time-series data is a challenging task.

Among these methods, clustering is one of the well-known unsupervised methods in data mining that does not require a prior knowledge of group allocation and has been applied to analyse and quantify various time-series data such as the electrocardiogram (ECG) (Hautamaki et al., 2008) and gait data (Kuntze et al., 2018; Zgolli et al., 2018). Syczewska et al. (2021). performed a clustering analysis to categorise the patients based on the gait characteristics and indices analysed. The clustering was accomplished through the use of connectivity-based clustering and the weighted group approach using medians

(averaged linkage clustering). Yeh et al. (2010) proposed a simple fuzzy c-means based method for clustering the heartbeats from ECG signals. Tseng et al. (2020) analysed the possible relationship between common aging diseases such as diabetes, obesity, and hypertension using participants' ECGs. They implemented K-means clustering which is one of the most common clustering algorithms used for time-series data. The gait pattern of children with cerebral palsy was clustered using sparse K-means by Abbasi et al. (2020). Their clustering approach presented that it was capable of weighting and ranking the influential variables inside the clustering. However, this conventional method uses the statistical means as the centres of the clusters resulting in clusters that are not interpretable and fluctuate due to random initialisation. To overcome this problem, Arthur et al. (2007) proposed the k-means++ algorithm that initialises the centroids far away from each other to avoid some poor clustering that the original k-means causes. As an alternative to these two algorithms, k-medoids was introduced. Huy et al. (2016) presented an efficient implementation of k-medoids clustering for time-series data with dynamic time wrapping (DTW) distance measurement. The idea behind this method was to use the actual members of the dataset as clusters' centres (Hautamaki et al., 2008; Niennattrakul and Ratanamahatana, 2007). K-medoids is simple, effective, and robust to outliers, however, it is unsuitable for clustering arbitrary shaped groups of objects. Although k-means, k-menas++, and k-medoids are simple to run, they are not highly accurate with noisy data. Density-based spatial clustering of applications with noise (DBSCAN) Ester et al. (1996) was presented to overcome some of the drawbacks mentioned above. DBSCAN does not require the number of clusters specified by users and typically has higher accuracy than k-means. Wang et al. (2021) proposed a model for QRS detection in ECG signals based on U-Net and DBSCAN that comprise of three steps: preprocessing, building a U-Net model, and spatial clustering of applications with noise using density-based clustering (DBSCAN). Unlike conventional techniques, the U-net model can extract features automatically with little preprocessing and user adjustments. Following U-Net prediction, DBSCAN was capable of achieving clusters to find the R-peaks in the absence of knowledge of the number and duration of the QRS complexes. To accomplish non-invasive blood glucose monitoring and prediabetes/diabetes screening using ECG, Li et al. (2021) developed a method of combining DBSCAN and CNN (DBSCAN-CNN). The results indicated that the percentages of correct categorisation were increased. However, its drawbacks are ineffective distance measure, high computation cost, and inability to cluster data with different densities. Finally, the density-peaks clustering algorithm proposed by Rodriguez and Laio (2014) in 2014 has recently become popular due to its insensitivity to the 'density parameter.' Regardless of the shape or dimensionality of the data, the density-peaks technique groups it into clusters in a single step. Cluster centres have a higher density than surrounding areas and are separated from sites with higher densities by a relatively considerable distance (Jiang et al., 2019). However, several challenges remain and need to be addressed. The first and most important one is that the arbitrarily defined cut-off distance might affect the local density of data points, resulting in misclassified data points.

The main motivation of this paper is the presentation of a study that aims to develop a quantitative method that incorporates time-series gait data analysis, in contrast to the traditional gait analysis approach, in order to investigate gait pattern alterations associated with a specific target group using an adaptive density-peaks clustering approach. Specifically, this study investigated which gait parameters significantly

differed on the target group and contributed to separate gait patterns for individuals in this group. The proposed method was first validated using synthetically generated time series control charts and then applied to the gait data collected from target and control groups; 5–12 years old children active in sports, and an age-matched control group, respectively. Gait data were collected in a motion capture and analysis laboratory using ten optical cameras and two force platforms.

2 Methods

2.1 Participants

A total of 24 children aged 5–12 years old participated in this study. Of these, 14 children (eight girls, six boys) with a mean age of 10.1 ± 2.3 years, a mean height of 152.1 ± 14.9 cm, and a mean weight of 46.3 ± 18.4 kg were considered active in sports, particularly swimming with regular daily training. The control group comprised an additional ten children (five girls, five boys), with a mean age of 9.7 ± 1.14 years, a mean height of 146.2 ± 8.9 cm, a mean weight of 38.6.0 ± 7.6 kg, and no regular training in any sports at the time of data collection. All participants were able to walk freely without the use of a cane or mechanical aid and injury-free at the time of data collection.

2.2 Data collection

The current study was approved by the University's institutional review board. Participants and their parents were given verbal explanations about the purpose and methods of the study before signing written informed consent forms. The data was gathered in the motion capture and analysis laboratory. The three-dimensional trajectories of reflective markers placed on the participant's skin were tracked using a ten-camera Vicon optical motion capture system (Oxford Metrics; Oxford, UK) sampling at 100 Hz. To collect kinetic data, two 60 cm wide, 90 cm long, and 15 cm tall force plates (Bertec; Worthington, OH, USA) with a sampling frequency of 1000 Hz were used.

The lower body plug-in model Vicon Curzon-Jones and Hollands (2018) was used as a guide for marker placement. 16 reusable, reflective markers have been applied to the body with double-sided adhesive tape. Table 1 contains a collection of marker names, descriptions, and locations. The body weight was measured with a weight measurement, and the height was measured with the wall-mounted stadiometer before starting the data collection. A measuring tape and a vernier caliper were used to measure leg length, knee width, and ankle width. Once body measurements were collected, the anthropometrical data was entered as input for the subject information in the Vicon Nexus 2.6 software. Several practical trials were conducted to acclimate participants to the laboratory setting. The data collection rate was set to 100 samples per second, and the children were urged to complete at least eight trials. Five of the participant's best trials were chosen for analysis, with best defined as a trial in which the person stepped on each force plate once without glancing down at their feet and walked at their normal pace. The trials took place on an unmarked 16-foot corridor within the motion capture and analysis facility. Each participant was assigned a starting position. The start line was placed eight feet from the centre of two force plates, allowing participants to reach steady-state velocity before

reaching the force plate region. They were instructed to walk at a comfortable pace, looking straight ahead, not down, and to foot just once on each force plate.

Table 1 Vicon plug-in gait lower body marker configuration adapted from Vicon documentation

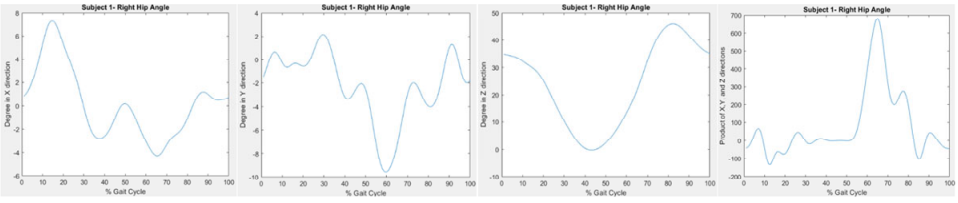
Marker label	*Definition*	*Position on body*
LASI	Left ASIS	Left anterior superior iliac spine (ASIS)
LPSI	Left PSIS	Left Posterior superior iliac spine (PSIS)
RASI	Right ASIS	Right anterior superior iliac spine (ASIS)
RPSI	Right PSIS	Right Posterior superior iliac spine (PSIS)
LTHI	Left thigh	Over the lower lateral 1/3 surface of the left thigh
LKNE	Left knee	On the flexion-extension axis of the left knee
LTIB	Left tibia	Over the lower 1/3 surface of the left shank
LANK	Left ankle	On the lateral malleolus along an imaginary line that passes through the transmalleolar axis
LHEE	Left heel	On the calcaneous at the same height above the plantar surface of the foot as the toe marker
LTOE	Left toe	Over the second metatarsal head, on the mid-foot side of the equinus break between fore-foot and mid-foot
RTHI	Right thigh	Over the upper lateral 1/3 surface of the right thigh
RKNE	Right knee	On the flexion-extension axis of the right knee
RTIB	Right tibia	Over the upper 1/3 surface of the right shank
RANK	Right ankle	On the lateral malleolus along an imaginary line that passes through the transmalleolar axis
RHEE	Right heel	On the calcaneous at the same height above the plantar surface of the foot as the toe marker
RTOE	Right toe	Over the second metatarsal head, on the mid-foot side of the equinus break between fore-foot and mid-foot

2.3 Data analysis

After collecting data with the motion capture system, gait events were determined using force plates that detected foot-strike and foot-off events. The system identified markers based on the lower body marker configuration, and spatiotemporal parameters were computed using the Vicon Nexus 2.6 software. To confirm that the markers were properly labeled, the procedure begins with calculations using the static plug-in gait pipeline, followed by the plug-in gait dynamic pipeline. The start and endpoints of dynamic trials were modified to give sufficient continuous data, and any gaps in the dynamic trials were filled to get an appropriate model. After processing the data with the Nexus 2.6 program, the data was transferred to visual 3-D software, where a standard gait model was constructed to capture all the gait kinematic and kinetic parameters. To offer a more complete picture of the gait cycle, data from both the right and left sides were incorporated. Each individual received five successful trials. While being retrieved, all the spatiotemporal, kinematic, and kinetic data were z-normalised, and time normalised in Visual 3-D. To remove the high-frequency components in the signal, the kinetic data

were filtered using a fourth-order, zero-lag, low-pass Butterworth filter with a cut-off rate of 6.0 Hz. Figure 1 depicts a representative sample of a subject's normalised gait kinematic.

Figure 1 The processed gait data of a subject in x, y, and z directions and the product of them respectively (see online version for colours)



Note: right hip angle.

A correlation filter technique presented by Curzon-Jones and Hollands (2018) was employed to exclude significantly related features from our cluster analysis. The filter approach detects and filters strongly linked characteristics automatically, depending on a user-defined threshold value, using the pairwise Pearson correlation function. 34 spatiotemporal, kinematic, and kinetic gait characteristics remained after this analysis to be employed in clustering. (The parameters are labeled in Figures 2 and 3).

2.4 Analysis of time series gait data

We present a clustering-based method to compare the gait kinematics and kinetics of two groups of subjects. Clustering is an unsupervised method that analyses data and groups them into clusters. The clustering, given a dataset of n data points $T = \{t_1, t_2,\ldots t_n\}$, the process of unsupervised partitioning of T into $C = \{C_1, C_2,\ldots C_k\}$. In such a way that homogeneous data is grouped together based on a specific similarity metric. Then, C_i is called a cluster,

$$\text{Where } T = U_{i=1}^{k} C_i \text{ and } C_i \cap C_j = \varnothing \text{ for } i \neq j \tag{1}$$

As a first step of the proposed method, a shape-based distance metric, derivative dynamic time warping (DDTW) Keogh and Pazzani (2001) is used to calculate similarities between two time-series gait data before grouping them into clusters. The traditional dynamic time warping (DTW) has been successfully used in many domains, but it might also provide pathological results. The main drawback observed in DTW is trying to explain the variability in the y-axis by warping the x-axis, resulting in misalignments where a single point on one time series maps onto a large subsection of another time series. However, this can be prevented by not considering the y-values of data points in DDTW. In the DTW, time normalised distance between two time series data A and B is:

$$D(A,B) = \left(\frac{\sum_{i=1}^{n} d(q_i, c_i).w(q_i, c_i)}{\sum_{i=1}^{n} w(q_i, c_i)} \right) \tag{2}$$

where n represents the number of observations in each time series, d (q_i, c_i) is the distance between q_i and c_i and $w(q_i, c_i)$ is a weight between q_i and c_i. The best alignment path between A and B is:

$$P_0 = \arg\min\left({}_P D(A,B)\right) \tag{3}$$

The distance measured (q_i, c_i) obtained using the DDTW is the square of the difference between the estimated derivatives of q_i and c_i. The following estimate method is used because it is more resistant to outliers.

$$D_x[q] = \frac{(q_i - q_{i+1}) + \left((q_{i+1} - q_{i-1})/2\right)}{2} \qquad 1 < i < m \tag{4}$$

where m is the number of data points in the squence. This estimate is just the average of the slopes of the lines connecting the point and its left neighbour, as well as the slopes of the lines connecting the point and its left neighbour and right neighbour.

DTW and DDTW are used to transform time series data to the upper triangular matrices (D_m) that show the DTW and DDTW distances between each subject and trial, T_n =(1, 2 ,…, n) for each gait parameter for the clustering analysis.

Density-peaks clustering Li et al. (2021) seeks to discover cluster centres having a larger density than surrounding regions in a single step, independent of the form of the data. As a result, they have a significant distance to places with a larger density. Densities are determined using a cut-off kernel, with neighbourhood specified by a predefined cutoff distance (d_c). This specifies a hyperball with a d_c) radius in D-dimensional space. The algorithm then determines the number of data points contained within this ball. Three essential parameters must be considered; ρ_i is the local density of data point i; δ_i is the minimum distance between data point i and other data points with higher density; and γ_i = ρ_i x δ_i is the product of the other two (Rodriguez and Laio, 2014). The original algorithm selects k points as the cluster centres based on ρ and δ. This is because cluster centres are expected to have a high value for both of them. However, it was not defined how exactly the selection should be made. In general, the problem is how to threshold the selected feature δ. We are proposing an adaptive cut-off distance to find this value. Cluster centres are assigned based on density and minimum distance in the original algorithm. This is due to the fact that cluster centres are anticipated to have a high value for both. However, it was not specified how the decision should be done precisely. In general, the issue is how to determine the threshold for the selected minimum distance. Thus, we propose using an adaptive cut-off distance to determine this value.

Let M ={x_1, x_2,…, x_n} a dataset with n data points. Each x_i has M attributes. Therefore, x_{ij} is the jth attribute of a data point x_i. The Euclidean distance between the data points x_i and x_j can be expressed as follows:

$$d_{ij} = d\left(x_i, x_j\right) = \left\| x_i - x_j \right\| \tag{5}$$

The local density ρ_i of the data point x_i is defined as

$$\rho_i = \sum\nolimits_{j \neq i} \omega\left(d\left(x_i, x_j\right) - d_c\right) \tag{6}$$

with

$$\omega(x) = \begin{cases} 1, x < 0 \\ 0, x > 0 \end{cases} \tag{7}$$

where d_c is the cut-off distance. In fact, ρ_i is the number of data points adjacent to the data point x_i. The minimal distance δ_i between data point x_i and any other data points x_i' with a higher density ρ_i' is given by

$$\delta_i = \begin{cases} \min_{j:\rho_i^{'}>\rho_i}(d_{ij}), & if\ \exists_i^j, s.t.\rho_i^{'} > \rho i \\ \max_{i'}(d_{ii'}), & otherwise \end{cases} \tag{8}$$

The local density is computed using the following equation:

$$\rho_i = \sum_j exp\left(-\frac{d_{ij}^2}{d_c^2}\right) \tag{9}$$

An adaptive cut-off distance is formulated as exponential moving average and used to eliminate the influence from the cut-off distance, as follows:

$$d_c^i = \left[Vt\left(\frac{s}{1+n}\right)\right] + d_c^{i-1}\left[1-\left(\frac{s}{1+n}\right)\right] \tag{10}$$

where d_c^i is the cut-off distance at data point *i*

V_t the value of data point *i*

d_c^{i-1} the cut-off distance at data point i-1

s smoothing factor which follows the formula: [2/(*selected #of points to be averaged* + 1)] and *n* is the number of points to be averaged.

2.5 Validation of the proposed method

The accuracy of the proposed method was first tested externally using a synthetic dataset. The proposed method's results were then evaluated internally based on the collected gait data. The two most practical internal cluster validation methods are used to evaluate the results, Dunn Index (DI) (Dunn, 1973) and Silhouette Index (SI) (Rousseeuw, 1987).

The DI aims to identify sets of clusters with a slight variance between the cluster's data points and well separated, where the centres of different clusters are far apart from each other. A higher (DI) value indicates a better clustering.

The DI for c number of clusters is defined as:

$$DI = \min_{1\le i\le k}\left(\min_{i+1\le j\le k}\left(\frac{dist(c_i, c_j)}{\max_{1\le l\le k} diam(c_l)}\right)\right) \tag{11}$$

where

$dist(c_i, c_j)$ is the distance between cluster c_i and c_j where

$dist(c_i, c_j) = \min_{xi\in ci,\, xj\in cj} d(x_i, x_j)$, where

$d(x_i, x_j)$ the distance between data points $x_i \in ci, x_j \in cj$

$diam(c_l)$ the diameter of the cluster c_l, where

$$diam(c_1) = \max_{xl_1, xl_2 \in cl} d(xl_1), xl_2)$$

The SI for each point, on the other hand, indicates how similar it is to members of its own cluster when compared to data points from other clusters. The silhouette value SI_i for the i^{th} point is defined as

$$SI_i = (b_i - a_i) / \max(a_i, b_i) \tag{12}$$

where a_i is the average distance between the i^{th} point to the other points in the same cluster as i, and b_i is the minimum average distance from the i^{th} point to points in a different cluster minimised over clusters. The silhouette value ranging from –1 to 1 shows that i is well suited to its own cluster but not well correspondent to others. If it approaches one, the data point is assigned to a suitable cluster. If it is close to –1, it indicates that data has been misassigned. Finally, if it is close to zero, the data point may be assigned to additional clusters.

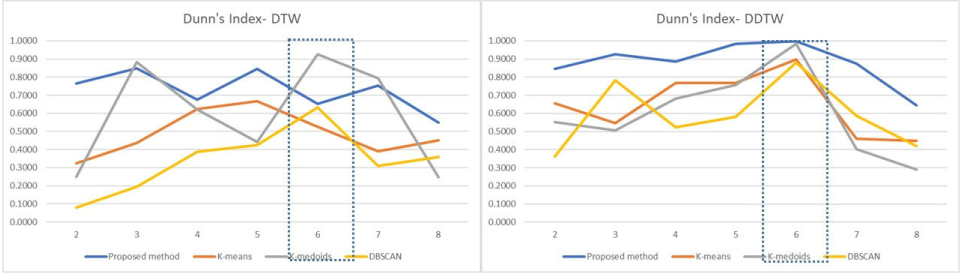
3 Results

The statistical analyses were conducted using MATLAB 2020a custom code. The proposed method's accuracy was first tested externally using a dataset that contains 600 examples of control charts (Pham and Chan, 1998) with six different classes. For the time series pattern, z-normalisation is an essential pre-processing step that allows the warping techniques or the clustering techniques to be applied to mainly focus on the structural similarities or dissimilarities rather than on the amplitude-driven ones. Z-score normalisation is a process of normalising the data that mainly to avoid the outlier issue. It transforms the data by converting the data to a common scale where an average number equals zero, and a standard deviation is one. The formula used for Z-normalisation is given by, $z = \frac{value - \mu}{\sigma}$, where μ is the mean value of the feature and σ is the standard deviation of the feature z-normalisation makes the features of the data less sensitive to the outliers in contrast to the min-max scaling.

After data normalisation, the proposed model and three other state-of-the-art clustering algorithms, k-means, k-medoids, and DBSCAN, were executed to measure the inter-cluster distances and intra-cluster distances. Figure 2 presents a graph of DI values (y-axis) as a function of the number of clusters (x-axis) for different clustering algorithms. However, the main interest was to measure DI values for six clusters, as we know there are six different data groups in the control charts dataset. Because of this, all the clustering techniques performed weakly for the other numbers of clusters. From graph (a), it is clear that the proposed algorithm reached out to the highest DI values in DDTW (DI = 0.9986) but the second-highest DI value in DTW (DI = 0.6324). K-medoids also equally performed with the proposed model in DDTW (DI = 0.9845) but better in DTW (DI = 0.9245). However, as seen from the graphs, DI values are much higher in DDTW

than DTW in general. These results indicate that DDTW might perform better over DTW using our gait data.

Figure 2 Comparison of Dunn's indexes of the proposed method along with three other state-of-the-art clustering methods using (a) DTW and (b) DDTW on control charts with six classes (see online version for colours)



(a) (b)

Note: y-axis-di and x-axis=#of clusters.

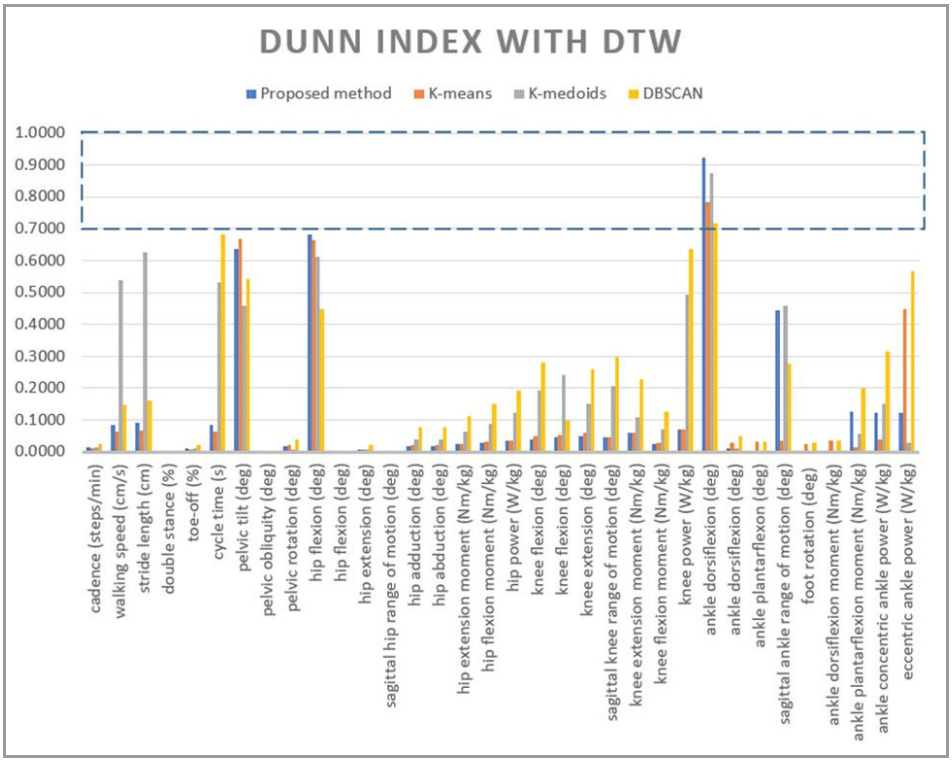
After validating the proposed method using the synthetic dataset, the method was applied to the study of participants' gait data. Before this process, a one-way analysis of variance (ANOVA) test was run to assess if participants in the study have similarities in height, weight, and age. No significant differences were found for age ($F= 1.22$, $P= 0.23$), height ($F = 0.67$, $P = 0.10$), or weight ($F = 1.8$, $P = 0.13$). Therefore, the two groups' characteristics were considered matched because all the P-values were greater than 0.05.

A total of 120 gait trials were collected from all the participants. Thus, the size of the distance matrices for each gait parameter was 120 by 120. The number of clusters was set to two because the aim was to identify differences between the two subject groups. After building DDTW based distance matrices, the proposed model with traditional clustering algorithms, k-means, k-medoids, and DBSCAN, were applied to group these calculated distances into two clusters representing the target and control groups. Each gait parameter was clustered independently. If the data points are cumulated together and very close to each other, the target group's gait pattern in that gait parameter is similar to the control group. Otherwise, a significant difference occurs. We also compared the performance of the proposed model with traditional clustering algorithms using DTW and DDTW distance measurements. The only input for the proposed model was the pair distance metrics between datapoints. A maximum of 200 iterations was run to execute the k-means algorithm. Random two points representing the two clusters were also initialised while executing k-means. To execute the k-medoids algorithm, the partitioning around medoids algorithm (PAM) was used [reference]. In contrast to the k-means algorithm, the PAM selects data points as centres. The PAM algorithm is based on searching all dataset elements for k numbers of medoids. For DBSCAN, three parameters were specified as the number of minimum points: 5, epsilon: 0.5, and the distance function: euclidian.

The cluster results were measured with the DI using DTW and DDTW, as shown in Figures 3 and 4, respectively. The value of both indexes is from 0 to 1, and as the value is higher, the greater the result of the cluster. The primary aim of the DI measurement is to maximise the distance between clusters while minimising the distances within the cluster. If a threshold for this fit is specified, we can check if the results of the cluster are applicable by comparing their respective threshold value and the threshold value can

influence the outcome of the assessment. We set DI levels to 0.7 in order to keep a more precise clustering result (Huang et al., 2014).

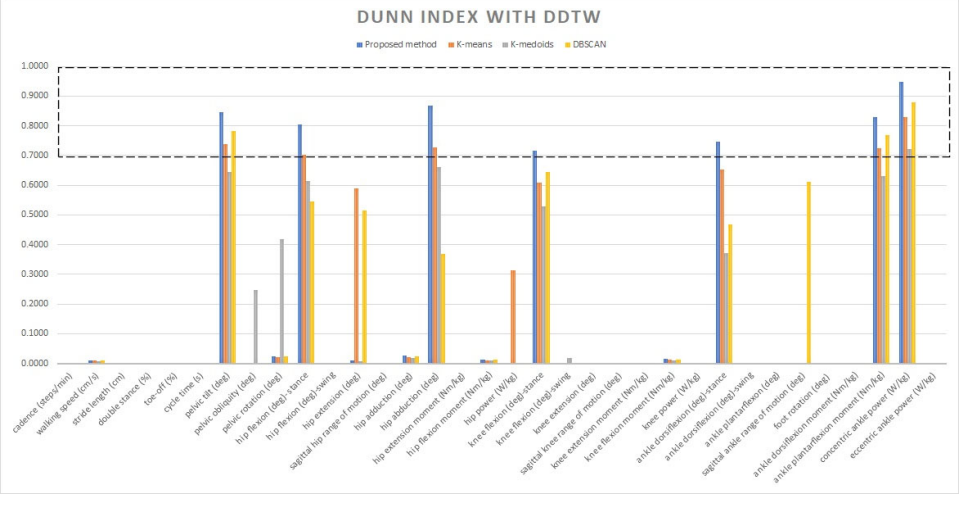
Figure 3 Dunn index for internal clustering validation using DTW measurement (see online version for colours)



It can be observed in Figure 3 that only a single parameter, ankle dorsiflexion was able to reach a higher DI that indicates the distances between data points in the target and control groups are maximised by all four algorithms. Although some of the DI values of the clustering methods were close to or higher than the 0.7 threshold value, none of the clustering methods could group gait parameters into two clusters, except ankle dorsiflexion. For instance, DI of the proposed method, k-means, k-medoids, and DBSCAN were higher than the threshold, 0.9240, 0.7843, 0.8756, and 0.7182 on ankle dorsiflexion, respectively. DI value gathered for cycle time was very close to the threshold value for the DBSCAN method, 0.6812. Similarly, DI for knee power and eccentric ankle power for DBSCAN were 0.6353 and 0.5677, respectively. The chart also shows that DIs for pelvic tilt given by the proposed model (DI=0.6357) and k-means (DI = 0.6678) were promising. This was also true for hip flexion obtained by the proposed model (DI = 0.6834), k-means (DI = 0.6634), and k-medoids (DI = 0.6315).

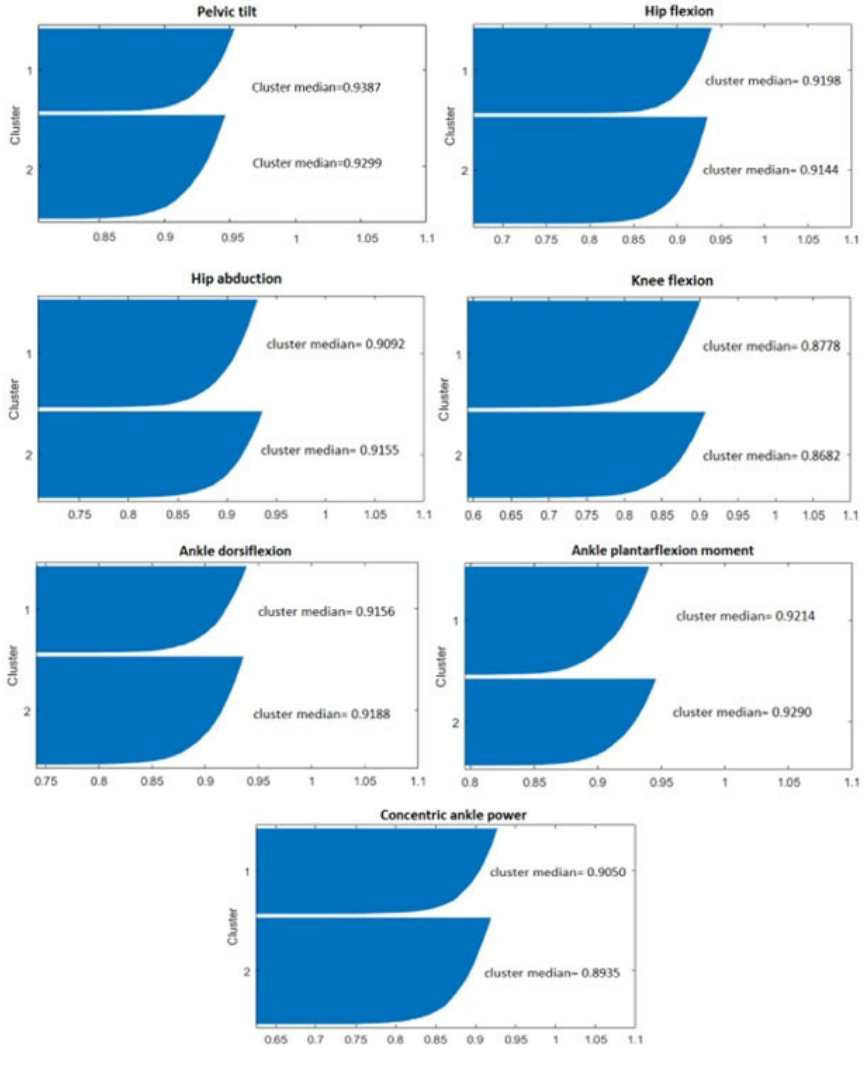
The same experiment was run using the DDTW similarity measurement. As seen in Figure 4, seven gait parameters were grouped into two clusters with higher DI values by the proposed model. These gait parameters are pelvic tilt-degree (DI = 0.8450), hip flexion-degree (DI = 0.8055), hip abduction-degree (DI = +0.8678), knee flexion-degree (DI = 0.7156), ankle dorsiflexion-degree (DI = 0.7465), ankle plantarflexion moment-Nm/kg (DI = 0.8289), and concentric ankle power-W/kg (DI = 0.9480) with have a higher index that indicates the distances between data points in the target and control groups are maximised. It can also be observed that there is a consistency with these gait parameters among the other clustering methods. DIs for the same gait parameters obtained by the other clustering methods were either higher than the threshold or very close to the threshold value. For instance, DIs for the concentric ankle power for all the clustering methods were higher than the threshold, proposed model (DI = 0.9480), k-means (DI = 0.8280), k-medoids (DI = 0.7215), and DBSCAN (DI = 0.8775).

Figure 4 Dunn index for internal clustering validation using DDTW measurement (see online version for colours)



The SI also supports DI results. Figure 5 shows only the significant gait parameters with higher DI values between the two groups with their cluster medians. As shown in the figure, SI values are for pelvic tilt-degree (sport trainees median = 0.9387 and control group median = 0.9299), hip flexion-degree (sport trainees median = 0.9198 and control group median = 0.9144), hip abduction-degree (sport trainees median = 0.9092 and control group median = 0.9155), knee flexion-degree (sport trainees median = 0.8778 and control group median = 0.8682), ankle dorsiflexion-degree (sport trainees median = 0.9156 and control group median = 0.9188), ankle plantarflexion moment-Nm/kg (sport trainees median = 0.9214 and control group median= 0.9290), and concentric ankle power-W/kg (sport trainees median = 0.9050 and control group median = 0.8935). These high SIs indicate that a data point is well matched to its own cluster and poorly matched to another cluster. Thus, the clustering solution is appropriate.

Figure 5 Clustering evaluations in terms of silhouette index (SI) (see online version for colours)



4 Discussion

This study presented a density-peaks clustering algorithm using an adaptive cut-off distance based on the exponential moving average to compare gait parameters of two groups of children. This algorithm was also compared with the other three state-of-the-art clustering algorithms using different similarity measurements, DTW and DDTW. The main contribution of this work is that it incorporated a clustering algorithm to analyse gait data, unlike other research works that have used descriptive statistics and independent t-tests that neglect essential gait information. The time-series gait data should and can be analysed by a more appropriate approach, such as the proposed adaptive density-peaks clustering algorithm using DDTW distance measurement.

There are several advantages to using the proposed adaptive density-peaks clustering algorithm over other well-known clustering techniques. Unlike DBSCAN, it identifies density-peaks as points surrounded by enough other points with lower density by detecting outliers of a density-distance plot. As a result, it seeks a greater distance between two density-peaks, and this distance must be greater than the distance between another arbitrary point. Many clustering algorithms require the user to set many parameters. For instance, k-means and k-medoids require assigning randomly selected initial points and iterations to execute the algorithms, but these parameters are intuitive and particularly sensitive to user choice. Randomly selecting initial points for each centroid might result in different clusters. In contrast, the density-peaks algorithm requires only one input, cut-off distance. However, this cut-off distance can affect the local density of data points, influencing the clustering results. We addressed this problem by developing an adaptive cut-off distance using an exponential moving average in our work. Thus, we eliminated defining intuitive cut-off distance by users. Compared with other clustering approaches, the proposed density-peaks clustering algorithm achieved the highest DI and SI values that mean data points in each cluster were well matched to its own cluster.

Traditional DTW has been used successfully in a variety of domains for time-series data clustering, but it might also provide pathological results. The main drawback observed in DTW is trying to explain the variability in Y-axis by warping X-axis. This can result in counterintuitive alignments. However, this problem can be prevented by not considering the Y-values of data points in DDTW. We evaluated this by running the proposed algorithm using DTW and DDTW as the similarity measurements. To do that, we used a publicly available dataset that contains 600 examples of control charts with six different classes. Among the four clustering algorithms that we tested, only k-medoids and density-peaks clustering using DDTW were able to group data points into six clusters with very high accuracy; DI values were 1 for both methods. The number of clusters that were correctly identified using DTW was just two. As a result of our findings, we can conclude that DTW is inaccurate and should not be considered as a subroutine in DBSCAN, density-peaks clustering, or k-means. However, it performed well with k-medoids.

To the best of our knowledge, no work has examined gait differences between two groups of children using clustering. Our main findings revealed that five key kinematic and two kinetic parameters, including pelvic tilt (degree), hip flexion (degree), hip abduction (degree), knee flexion (degree), ankle dorsiflexion (degree), ankle plantarflexion moment (Nm/kg), and concentric ankle power (W/kg), differ between children participating in sports and age-matched control groups, with no difference in spatiotemporal data. The clinical research backs up the findings of this clustering analysis. Below, we discuss possible explanations for these findings.

Because of repetitive adduction, athletes place a great deal of strain on their hip adductor muscles during training and competition (Keskinen et al., 1980; Grote et al., 2004; Pollard and Fernandez, 2004). As a result, a number of musculoskeletal injuries associated with this repetitive motion have been documented (Kennedy et al., 1978; McMaster and Troup, 1993; Grote et al., 2004). Because of the additional stress placed on those muscles by sports trainees, the repetitive adduction activated by hip adduction muscles may also affect gait patterns. During walking, the ankle usually moves in the sagittal plane. Each gait cycle includes two stages of plantar flexion and dorsiflexion. Sports trainees have greater ankle flexibility than other novice children, which explains

why the ankle plantarflexion moment, concentric ankle power, and ankle dorsiflexion differ between these two groups in our findings (Johnson et al., 1987). In each gait cycle, the knee also goes through two phases of flexion and extension. The results revealed that the target group's knee joint kinematics patterns differed from those of the control groups. Sports trainees' knees have excessive thigh abduction due to hip and knee flexes, according to Stulberg et al. (1980).

In each gait cycle, only one arc of hip extension and flexion occurs. Our findings also revealed that hip flexion during the stance phase differed between the target and controls. Even though the upper limb produces the majority of the force in many sports, lower limb joints are used in tandem with the upper limb to produce maximum force (Stulberg et al., 1980), affecting sports trainees' hip movement and stability. Sport trainees are also subjected to extreme stress on the hip abductor muscle as a result of repetitive adduction during their activities, which may explain differences in their gait patterns. The pelvis moves in all three planes while walking, and the magnitudes of pelvic motion are affected by walking speed (Lewis et al., 2017). The sagittal plane pelvic tilt effect suggests that both groups may use different control strategies while walking. Significant changes in pelvic tilt may result in differences in hip kinematics, as the pelvis is important in balancing the centre of the body's mass during progression. These findings may indicate that the control group shifts the body mass centre further than the target group leading to greater anterior pelvic inclination and bending of the hip in the two phases (Kindregan et al., 2015). For a variety of reasons, sports trainees may be less forward, including tight hip flexors/extender muscles. Hip extensor weakness, hip flexor contracture, or the spasticity of the hip flexor and balance and distal deformity can all cause anterior pelvic tilts (Brunner and Rutz, 2013). As the spatiotemporal parameters for our participants were not different for both sides, the difference in the structure of the participant's body could hardly be determined and played a role.

Although the accuracy achieved appears to be impressive, the proposed algorithm can be improved further by incorporating some other methods for calculating the densities of the data points. Optimisation techniques can also be used to improve the effectiveness of the proposed clustering method. Users must also decide on the number of clusters, but a manual selection of cluster centres can influence the clustering result. This issue must be addressed in the future.

5 Conclusions

The study made two significant contributions:

1 developing an effective quantitative method for analysing time-series gait data in order to investigate the gait patterns of children participating in sports

2 improving the density peak clustering algorithm by incorporating an adaptive cut-off distance measure. In this article, an extensive comparison of four clustering algorithms on DTW and DDTW distance measurements has been made for gait datasets collected from our motion capture laboratory.

This study demonstrates that the proposed density-peaks clustering-based method using DDTW distance measurement outperforms the other three state-of-the-art clustering algorithms and is a viable means to compare gait parameters of individuals from two

groups. Our clustering method is density-peaks-based, but cut-off distance is adaptive and initial input to execute the method is unnecessary. Out of all the accuracies obtained, pelvic tilt, hip flexion, hip abduction, knee flexion, ankle dorsiflexion, ankle plantarflexion moment, and concentric ankle power differ between sports trainees and age-matched control groups. In the future, we plan to improve the proposed method's effectiveness by employing optimisation techniques.

Acknowledgements

The authors would like to thank Dr. Xin Chen from the Department of Industrial Engineering at Southern Illinois University, Edwardsville, for his valuable review and feedback.

References

Abbasi, L., Rojhani-Shirazi, Z., Razeghi, M. and Raeisi-Shahraki, H. (2021) 'Kinematic cluster analysis of the crouch gait pattern in children with spastic diplegic cerebral palsy using sparse K-means method', *Clinical Biomechanics*, 1 January 2021, Vol. 81, No. 3, p.105248.

Arthur, D. and Vassilvitskii, S. (2007) 'k-means++: the advantages of careful seeding', *Presented at the Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, New Orleans, Louisiana.

Brunner, R. and Rutz, E. (2013) 'Biomechanics and muscle function during gait', *Journal of Children's Orthopaedics*, Vol. 7, No. 5, pp.367–371.

Curzon-Jones, B.T. and Hollands, M.A. (2018) 'Route previewing results in altered gaze behaviour, increased self-confidence and improved stepping safety in both young and older adults during adaptive locomotion', *Experimental Brain Research*, Vol. 236, No. 4, pp.1077–1089.

Dunn, J.C. (1973) 'A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters', *Journal of Cybernetics*, 1 January 1973, Vol. 3, No. 3, pp.32–57.

Ester, M., Kriegel, H-P., Sander, J. and Xu, X. (1996) 'A density-based algorithm for discovering clusters in large spatial databases with noise', presented at the *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland, Oregon.

Fujita, K., Iijima, H., Eguchi, R., Kuroiwa, T., Sasaki, T., Yokoyama, Y. et al. (2020) 'Gait analysis of patients with distal radius fracture by using a novel laser Timed Up-and-Go system', *Gait and Posture*, 1 July 2020, Vol. 80, pp. 223–227.

Grote, K., Lincoln, T.L. and Gamble, J.G. (2004) 'Hip adductor injury in competitive swimmers', *Am. J. Sports Med.*, January-February, Vol. 32, No. 1, pp.104-108.

Hautamaki, V., Nykanen, P. and Franti, P. (2008) 'Time-series clustering by approximate prototypes,' in *2008 19th International Conference on Pattern Recognition*, pp.1–4.

Huang, S., Cheng, Y., Lang, D., Chi, R. and Liu, G. (2014) 'A formal algorithm for verifying the validity of clustering results based on model checking', *PloS One*, Vol. 9, No. 3, p.e90109.

Huy, V.T. and Anh, D.T. (2016) 'An efficient implementation of anytime k-medoids clustering for time series under dynamic time warping', presented at the *Proceedings of the Seventh Symposium on Information and Communication Technology*, Ho Chi Minh City, Vietnam.

Iqbal, S., Zang, X., Zhu, Y., Saad, H.M.A.A. and Zhao, J. (2015) 'Nonlinear time-series analysis of different human walking gaits', in *2015 IEEE International Conference on Electro/Information Technology* (*EIT*), pp.25–30.

Jiang, Z., Liu, X. and Sun, M. (2019) 'A density peak clustering algorithm based on the K-nearest shannon entropy and tissue-like P system', *Mathematical Problems in Engineering*, 31 July 2019, Vol. 2019, p.1713801.

Johnson, J.E., Sim, F. and Scott, S. (1987) 'Musculoskeletal injuries in competitive swimmers', *Mayo Clinic Proceedings*, Vol. 62 4, pp.289–304.

Kennedy, J.C., Hawkins, R. and Krissoff, W.B. (1978) 'Orthopaedic manifestations of swimming', Am. J. Sports Med., November-December, Vol. 6, No. 6, pp.309–22.

Keogh, E.J. and Pazzani, M. (2001) 'Derivative Dynamic Time Warping,' in *SDM*.

Keskinen, K., Eriksson, E. and Komi, P. (1980) 'Breaststroke swimmer's knee. A biomechanical and arthroscopic study', *Am. J. Sports Med.*, July-August, Vol. 8, No. 4, pp.228–231.

Khan, S.I., Onal, S., Cho, S. and Smith, B. (2020) 'Gait comparison of physically active and inactive children', *IIE Annual Conference, Proceedings*, pp.1–6.

Kindregan, D., Gallagher, L. and Gormley, J. (2015) 'Gait deviations in children with autism spectrum disorders: a review', *Autism Research and Treatment*, Vol. 2015, p.741480.

Krauss, I., List, R., Janssen, P., Grau, S., Horstmann, T. and Stacoff, A. (2012.) 'Comparison of distinctive gait variables using two different biomechanical models for knee joint kinematics in subjects with knee osteoarthritis and healthy controls', *Clin Biomech* (Bristol, Avon), March, Vol. 27, pp.281–286.

Kuntze, G., Nettel-Aguirre, A., Ursulak, G., Robu, I., Bowal, N., Goldstein, S. et al. (2018) 'Multi-joint gait clustering for children and youth with diplegic cerebral palsy,' PloS one, Vol. 13, pp.e0205174-e0205174.

Lewis, C.L., Laudicina, N.M., Khuu, A. and Loverro, K.L. (2017) 'The human pelvis: variation in structure and function during gait', *The Anatomical Record*, 4 January 2017, Vol. 300, No. 4, pp.633–642.

Li, J., Tobore, I., Liu, Y., Kandwal, A., Wang, L. and Nie, Z. (2021) 'Non-invasive monitoring of three glucose ranges based on ECG by using DBSCAN-CNN', *IEEE Journal of Biomedical and Health Informatics*, Vol. 25, No. 9, pp.3340–3350.

McMaster, W.C. and Troup, J. (1993) 'A survey of interfering shoulder pain in United States competitive swimmers', *Am. J. Sports Med*., January-February, Vol. 21, No. 1, pp.67–70.

Muñoz-Diosdado, A. (2005) 'A non linear analysis of human gait time series based on multifractal analysis and cross correlations', *Journal of Physics: Conference Series*, 1 January 2005, Vol. 23, pp.87–95.

Niennattrakul, V. and Ratanamahatana, C. (2007) *On Clustering Multimedia Time Series Data Using K-Means and Dynamic Time Warping*.

Papi, E., Bo, Y.N. and McGregor, A.H. (2018) 'A flexible wearable sensor for knee flexion assessment during gait', *Gait and Posture*, Vol. 62, pp.480–483.

Pham, D.T. and Chan, A.B. (1998) 'Control chart pattern recognition using a new type of self-organizing neural network', *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, 3 January 1998, Vol. 212, No. 2, pp.115–127.

Pogemiller, K., Garibay, E., Pierz, K., Acsadi, G. and Õunpuu, S. (2020) 'Comparison of gait patterns and functional measures between Charcot-Marie-Tooth disease type I and II in children to young adults', *Gait and Posture*, 3 January 2020, Vol. 77, pp.236–242.

Pollard, H. and Fernandez, M. (2004) 'Spinal musculoskeletal injuries associated with swimming: a discussion of technique', *Australasian Chiropractic and Osteopathy: Journal of the Chiropractic and Osteopathic College of Australasia*, Vol. 12, No. 2, pp.72–80.

Rodriguez, A. and Laio, A. (2014) 'Clustering by fast search and find of density peaks', *Science*, Vol. 344, No. 6191, pp.1492–1496.

Rousseeuw, P.J. (1987) 'Silhouettes: a graphical aid to the interpretation and validation of cluster analysis', *Journal of Computational and Applied Mathematics*, January 11 1987, Vol. 20, pp.53–65.

Starbuck, C., Reay, J., Silk, E., Roberts, M., Hendriksz, C. and Jones, R. (2021) 'Are there common walking gait characteristics in patients diagnosed with late-onset Pompe disease?', *Human Movement Science*, 1 June 2021, Vol. 77, p.102777.

Stulberg, S.D., Shulman, K., Stuart, S. and Culp, P. (1980) 'Breaststroker's knee: pathology, etiology, and treatment', *Am. J. Sports Med.*, May-June, Vol. 8, No. 3, pp.164–71.

Syczewska, M., Święcicka, A., Szczerbik, E., Kalinowska, M., Dunin-Wąsowicz, D. and Łukowicz, M. (2021) 'Types of gait deviations in children and adolescents with Guillain-Barre syndrome identified using cluster analysis', *Biomedical Signal Processing and Control*, January 4 2021, Vol. 66, p.102496.

Tseng, K.K., Li, J., Tang, Y.J., Yang, C.W., Lin, F.Y. and Zhao, Z. (2020) 'Clustering analysis of aging diseases and chronic habits with multivariate time series electrocardiogram and medical records', *Front Aging Neurosci.*, 5 May, Vol. 12, p.95, doi: 10.3389/fnagi.2020.00095, PMID: 32477093; PMCID: PMC7232580.

Vieten, M.M., Sehle, A. and Jensen, R.L. (2013) 'A novel approach to quantify time series differences of gait data using attractor attributes', *PLOS ONE*, Vol. 8, No. 8, p.e71824.

Wang, H., He, S., Liu, T., Pang, Y., Lin, J., Liu, Q. et al. (2021) 'QRS detection of ECG signal using U-Net and DBSCAN', *Multimedia Tools and Applications*, 5 December 2021.

Wu, Z., Xu, H., Zhu, S., Gu, R., Zhong, M., Jiang, X. et al. (2021) 'Gait analysis of old individuals with mild parkinsonian signs and those individuals' gait performance benefits little from Levodopa', *Risk Management And Healthcare Policy*, Vol. 14, pp.1109–1118.

Yeh, Y-C., Wang, W-J. and Chiou, C.W. (2010) 'A novel fuzzy c-means method for classifying heartbeat cases from ECG signals', *Measurement,* 12 Janury 2010, Vol. 43, pp.1542–1555.

Zgolli, F., Henni, K., Haddad, R., Mitiche, A., Ouakrim, Y., Hagemeister, N. et al. (2018) 'Kinematic data clustering for healthy knee gait characterization', in *2018 IEEE Life Sciences Conference* (*LSC*), pp.239–242.