doi:10.1093/comjnl/bxy022

Optimized Clustering Techniques for Gait Profiling in Children with Cerebral Palsy for Rehabilitation

Chandra Prakash[1,2]*, Rajesh Kumar[3] and Namita Mittal[1]

[1] *Department of Computer Science and Engineering, Malaviya National Institute of Technology, Jaipur 302017, India*
[2] *Indira Gandhi Delhi Technical University for Women, Delhi 110006, India*
[3] *Department of Electrical Engineering, Malaviya National Institute of Technology, Jaipur 302017, India*
**Corresponding author: cse.cprakash@gmail.com*

Cerebral palsy (CP) is a neuro-development disease in children. It is quite an intricate task to categorize gait pattern into normal and CP based pathology. In this study, nature-inspired metaheuristic algorithms are explored on a publicly available gait dataset of 156 subjects for automatic gait profiling of children with cerebral palsy. Five cases are considered to explore the feature selection criteria before applying clustering technique. Finding the optimal number of clusters is a challenging task in the unsupervised learning area. In this study, an optimal number of gait profiles in the datasets is identified based on voting from mean square error, silhouette coefficient and Dunn index. The study demonstrates that optimized based gait profile clusters could assist quantitatively in clinical rehabilitation evaluation for the children affected by CP.

Keywords: gait profiling, clustering; gait analysis; rehabilitation; gait applications; k-means; particle swarm optimization; genetic algorithm

Received 24 February 2017; revised 7 February 2018; editorial decision 10 February 2018
Handling editor: Dimitrios Tzovaras

1. INTRODUCTION

Cerebral palsy (CP) is a neuro-development disease, common in children. It is associated with the floppy or rigid limbs, exaggerated reflexes and involuntary motion, poor speech and learning ability which, is considered as a non-progressive disease [1, 2]. The major reason for this motor disability is abnormal brain development, which often occurs before the birth of a child. On behalf of motor impairment of the limb, CP is classified into three major categories; spastic, ataxic and athetoid [3]. Different studies, around the globe highlight the severity of this disease and projects that worldwide CP cases vary from 1.5 to 4 per 1000 children. The condition is more severe in the developing countries [1, 4]. Thus a huge amount of money is involved in the rehabilitation and intervention policies related to the diagnosis of gait pathology related to cerebral palsy [5, 6].

The diversity in the gait pattern from children suffering from CP makes it a challenging task for researchers in the physical therapy and surgical community. This will enable health-care professionals to differentiate gait pattern into clinically significant categorize that assist in diagnosis, assessment, evaluation of the treatment outcomes and clinical decision-making [7–11]. All pathological pattern cannot be captured through just visual observation. There is a need to take the aid of statistical (frequency domain characteristics, spectral components, harmonic content, the coefficient of variation, etc.) and machine learning techniques [12]. Clinical gait analysis is a method to identify the hidden pattern in gait such as stride length, step length, cadence, stance and swing phase, etc. Clinicians can utilize gait classification concepts in their routine clinical practice to evaluate a patient's status, treatment and rehabilitation for CP disorders using the spatio-temporal and kinematics parameters [7, 13, 14].

There are several attempts made by researchers to characterize the gait pattern and categorize them into normal and gait pathology using computational techniques [15–19]. Supervised learning is a popular choice for the research community to classify the gait pattern of CP pathology for automated analysis. Neural network is most widely used for normal gait analysis, robotic rehabilitation, sports monitoring, and tactics, geriatric care, surveillance, activity recognition [7, 20–23]. Supervised learning technique is a dominant methodology when labeled training data points are available. Zheng *et al.* [24] implemented the trained neural network classifier to identify the abnormality in older person gait. Gait pathology can be identified using ground reaction force with learning vector quantization. Multilayer perceptron, Linear discriminant analysis classifiers, kernel Fisher discriminant and Bayesian classifier have been used to classify gait pattern in [25, 26]. Support vector machine (SVM) is used for the study of normal and age-related differences, normal and abnormal gait pathology. The authors in [27, 28] showed that it exhibits good results in CP pathology. SVM is a powerful classifier suitable from small to the medium dataset. The limitation of these supervised approaches is that their accuracy depends on the training sample. The larger the training sample, the more accurate the system is. Another limitation is that these techniques cannot be applied to the data where no prior information of the data is given.

To overcome these limitations, researchers have explored unsupervised, or clustering approaches [29, 13, 30]. Cluster models are not only popular in signal and image processing, but they also have wide applications in web mining and pattern recognition, sensor network, robotics, seismology, medical science in disease clustering, etc [31, 32].

In gait analysis, clustering techniques are most widely used for categorizing gait data into groups of disorders based on common hidden features in the subject dataset. K-means, fuzzy c-means, hierarchical clustering, self-organizing map (SOM) are some of the examples of clustering techniques that have been explored in diagnosing CP related gait abnormalities [7].

Finding the optimal number of clusters in a dataset is a challenging task in the unsupervised learning area [31].

Selection of number of clusters is considered as obscure as the interpretations not only depend on the shape and distribution of the data points but also on the desired clustering resolution of the user. In this study, authors have proposed a hypothesis that nature-based clustering approaches can help in finding the optimal number of gait profiles in the datasets. The resulting analysis suggests that the proposed approach is better and closer to ground truth than the clustering techniques (k-means, fuzzy clustering mean (FCM), genetic algorithm (GA) and particle swarm optimization (PSO)) used in this study. Results indicate that it helps in diagnosis, assessment and evaluation of treatment outcomes. It acts as an assistive tool to the doctors working in this area.

This paper is organized as follows. In Section 2, problem formulation is presented. Section 3 discusses the experimental methodology. Sections 4 and 5 present the results and conclusion with a future scope.

2. PROBLEM FORMULATION

2.1. Gait dataset description

The gait dataset used in this research is taken from the publicly accessible data from O'Malley *et al.* [13]. These gait data consist of two groups, first consisting of 88 children with the spastic diplegia, CP and another is the collection of 68 neurologically intact children and have no history of any motor disease. Data are collected using six cameras of Vicon System and processed with Vicon Clinical Manager at Motion Analysis Laboratory, University of Virginia [13].

The spatio-temporal gait parameters considered in the study are stride length and cadence. Age and limb length are considered as anthropometric parameters and are used to normalize the stride length and cadence. O'Malley *et al.* used fuzzy based clustering techniques for the classification of control and CP children based on stance and cadence parameters. For study and comparison, in this study, the authors have used the same gait parameters as in [13].

Authors considered these parameters after averaging of data measured through at least three trials of both legs. The mean age of children with CP is 9.89 years while it is 7.09 years for the control group. Table 1 presents the overview of the CP gait dataset considered in this study.

2.2. Traditional clustering approaches

Partitional, overlapping and hierarchical are the main three types of clustering approaches. Automated clustering techniques play an important role in image classification, intrusion detection, document clustering and medical imaging [33].

TABLE 1. Anthropometric and the spatio-temporal gait parameters of the participating subjects [13].

	88 Children with spastic-diplegia, CP				68 neurologically intact children			
	Min	Max	Mean	Std	Min	Max	Mean	Std
Age (years)	2	20	9.89	4.34	2	13	7.09	2.89
Leg length (m)	0.41	0.94	0.67	0.14	0.34	0.86	0.57	0.12
Stride Length (m)	0.31	1.25	0.74	0.21	0.67	1.44	1.03	0.19
Cadence (step/min)	10.46	210.24	120.00	33.56	104.88	174.24	136.84	15.81

The objective of clustering in this gait profiling problem is to group the considered CP gait dataset $X = (X_1, X_2, \ldots, X_N)$ of N objects into K groups such that $K \leq N$. Consider F features for each object, then dataset X can be represented as a matrix of size $N \times F$, where N is the number of rows and F is number of columns.

The purpose of clustering is to find K clusters each having a centroid point C_k in such a way that the similarity between data objects within the cluster to the centroid of the same cluster is minimum.

2.2.1. *K-means*

The K-means algorithm is the simplest and most popular partition clustering approach. It is also known as hard clustering approach. In K-means, the overall distance between the cluster members and the cluster centroids (intra-cluster distance) needs to be optimized [34]. Mean square error (MSE) is considered as the objective function illustrated in Equation (1).

$$\frac{1}{N}\sum_{k=1}^{K}\sum_{x_i \epsilon c_k} || \, x_i - C_k \, ||^2 \qquad (1)$$

where K is the total number of pre-defined clusters, k represents the cluster index and C_k is the centroid of the cluster k. This expression is needed to be minimized. The algorithm of K-means is presented in Algorithm 1.

Data points assigned to a cluster are based on their degree of closeness, measured by the Euclidean distance from the points to the cluster's center. K-means approach is easy to implement but the time complexity increases with larger datasets and very sensitive to the initially provided centroids.

2.2.2. *Fuzzy c-means*

Fuzzy logic is introduced by Zadeh during 1960s for handling the uncertain and imprecise knowledge in real-world applications [35]. The fuzzy c-means algorithm uses the reciprocal of distances between data in instances to decide the cluster centers. It is considered as the extension of the K-means and also known as soft clustering technique. The centroid of a cluster in a fuzzy c-means method is calculated as the mean of all points value, weighted by their degree of belongingness to the cluster. When the nature of cluster is overlapping then fuzzy clustering is preferred [35].

Algorithm 1 K-means Algorithm

```
Initialize number of cluster as (K) and $max\_iter$
Select initial centroids randomly
for $iter = 0$: $max\_iter$ or other termination criteria do
    Allocate each data-point to a cluster based on the similarity measure
    Update the centroids by taking means of all points in the particular cluster
end for
```

In FCM, to determine the best value of partition matrix U the following objective function $J_k(U, Cj)$ is minimized

$$J_k(U, Cj) = \sum_{i=1}^{k}\sum_{j=1}^{c_j}\mu_{ij}^{\;m}(\|X_i - C_j\|)^2$$
$$1 < m < \infty \qquad (2)$$

where m is any real number greater than one, μ_{ij} is the degree of membership of X_i in the cluster kth. Degree of membership μ_{ij} and cluster center C_j is given by Equations (3) and (4), respectively:

$$\mu_{ij} = \frac{1}{\sum_{p=1}^{k}\left(\frac{\| X_i - C_j \|}{\| X_i - C_p \|}\right)^{\frac{2}{m-1}}} \qquad (3)$$

$$C_j = \frac{\sum_{i=1}^{N}\mu_{ij}^{\;m} X_i}{\sum_{i=1}^{N}\mu_{ij}^{\;m}}. \qquad (4)$$

In 1997, O'Malley *et al.* [13] examined FCM on cerebral palsy children based on 168 subjects, spatio-temporal parameters (stride length and cadence). Carriero *et al.* [36] demonstrate the possibility of principal component analysis in quantitative classification of CP gait pattern. FCM analysis is used to cluster the data of 40 subjects (20 healthy and 20 soastic diplegic CP patients) with 27 parameters each.

2.2.3. *Other approaches*

In 1983, Wong *et al.* [29] explore k-nearest neighbor to classify the gait pathologies using walking speed, hip, and ankle movement of 62 CP patients, considering k as 5. Self-organizing feature map (SOM), Hierarchical clustering and K-means are used to differentiate normal and pathological gait pattern based on stride length and cadence [30]. Hierarchical clustering is used to cluster healthy group from pathological patients [37]. In 2007, Taro *et al.* apply hierarchical clustering on the sagittal kinematic gait data of 67 subjects (56 CP, 11 healthy). Thirteen gait clusters are formulated using sagittal plane hip, knee, and ankle kinematics [38]. The issue with the clustering techniques is that number of the cluster is required to be known in advance. Researchers have also explored the nature-inspired optimization algorithm for the clustering purpose [31].

Algorithm 2 Fuzzy c-means algorithm

```
Initialize number of cluster as k
Initialize $\mu_{ij}$ for all i,j from (4)
for Until termination criteria is met do
    Calculate cluster centers $V_j$
    Update $\mu_{ij}$ for all i,j
end for
```

3. EXPERIMENTAL METHODOLOGY

In this section, the methodology used and nature-inspired optimized based clustering approaches are discussed followed by the five different cases, clustering evaluation indices and parameter setting. Clustering techniques are employed on the CP dataset and based on the best possible number of cluster identified in this study, the condition after the post-surgery is examined to validate the clustering result. Figure 1 illustrates the work flow of the proposed methodology.

In K-means clustering approach, the result depends on the choice of the randomly selected cluster centroids. If they are close, takes the time to converge, while if the distance is on two extreme sides, then better cluster formation takes place. Thus the selection of the first cluster head points can be optimally selected using meta-heuristic approaches. In this study, nature-inspired optimization algorithms are used for the clustering purpose [31]. GA and PSO are the most common in evolutionary algorithms. In this study GA, PSO and a hybrid version of the both are proposed as clustering techniques with the following objective function [39].

The objective function used in this study is the combination of intra and inter-cluster distance. Intra-cluster distance ($D_{\text{inter}}\, c_n$) is the overall distance between the cluster members and its centroid. While inter-cluster distance $D_{\text{intra}}\, c_n$ is the overall distance between the centroids.

Mathematically, it can be modeled as a multi-objective optimization problem to find a cluster centroid C_k that has maximum similarity. The fitness function $f\,()$ can be represented as Equation (5):

$$
\begin{aligned}
f\left(X_{N\times F},\, C_k\right) &= \sum_{n=1}^{K} \left\{ w_1 * D_{\text{inter}}\left(c_n\right) - w_2 * D_{\text{intra}}\left(c_n\right) \right\} \\
&\quad \text{where} \\
D_{\text{inter}}\, c_n &= \max \,||\; C_i - C_j \;||, \\
&\quad \text{Here } i, j \;\epsilon\, [1, 2, 3 \ldots k],\; i \neq j, \\
&\forall\, k = 1, 2 \ldots K \\
D_{\text{intra}}\, c_n &= \sum_{i=1}^{N} \text{Min} \left\{ \left\| X_{i\times F} - C_{k\times F} \right\|^{2} \right\} \\
\text{Max} \quad & C_k^{\text{Optmize}} \qquad (5)
\end{aligned}
$$

where i is the data point of the k cluster, K is the total number of pre-defined clusters, k represents the cluster index, C_k is the centroid of the cluster k. w_1 and w_2 are weighted parameters. In this study, we have considered both w_1 and w_2 as 0.5 for the study purpose [39]. This expression needed to be maximized.

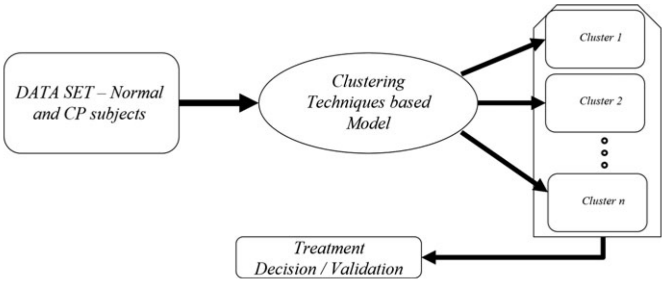
DATA SET – Normal and CP subjects
Clustering Techniques based Model
Cluster 1
Cluster 2
Cluster n
Treatment Decision / Validation

FIGURE 1. Clustering-based methodology used for gait profiling of CP children.

3.1. GA and proposed variant

Evolutionary techniques are a good approach to the optimization problem. Based on Darwin's principle of natural selection, GA is first meta-heuristic optimization algorithm proposed by Goldberg and Holland in 1988 [40].

In 1994, Bezdek *et al.* [41] explore basic GA to be used as a clustering approach. Similar to basic GA, individual features are characterized in the form of strings called chromosomes. Based on the number of clusters (K) and the population size (N), the chromosomes are considered as the cluster centroids candidate. The algorithm initiates by creating the random set of chromosomes in the search space followed by an iterative process of selection (maximum number of iterations maxiter), crossover probability (C_{prob}) and mutation probability (M_{prob}) to find the optimal K solution (cluster center (centroids)) in the search space.

3.1.1. Hybrid GA variant

In 1999, Krishna and Murty [42] proposed a novel hybrid Genetic K-means algorithm. K-means help in the search operation during crossover.

In this study, GA is hybridized with K-means algorithm to get the optimal cluster centers and effective fitness values as shown in Algorithm 3 [43].

Hybrid GA optimizes the location of starting centroid (cluster centers) from the population with the minimum fitness function. Use these cluster centers as the initial centroids in the k-mean and iterate, the steps until no significant changes in consecutive cluster centers.

3.2. PSO and proposed variant

PSO technique mimics swarm (birds, fish, etc.) social behavior for food searching and is developed by Eberhart and Kennedy [44]. Because it has a fast convergence rate and easy to implement, it is a popular choice among the researchers. Omran *et al.* [45], proposed cluster analysis using PSO for image processing. One year later, Merwe explored PSO for data clustering on different datasets [46].

In PSO-based clustering algorithm, swarm are defined in search space according to the constraints. The pseudo-code of PSO is presented. Population, initial weights, iteration is defined initially. Each particle position is defined as

$$x_i = [\mu_{i1}, \mu_{i1}, \ldots, \mu_{iK}] \tag{6}$$

where K cluster centroid vectors. The cluster C_{ik} has μ_{ik} as the cluster center point (centroid).

Algorithm 3 Pseudo-code for the hybrid-GA-based clustering

```
1:  Initialize $K$, $N$, $maxiter$, $C_{prob}$ and $M_{prob}$
2:  Generate initial population randomly where each
    chromosomes act as set of cluster centers
3:  for $iter = 0$: $maxiter$ do
4:      Apply selection in the population
5:      Apply crossover in the population
6:      Apply mutation in the population
7:  end for
8:  Output is the $K$ optimal cluster center with minimum
    MSE
9:  Apply K-means by initializing $max\_iter\_k$
10: Select initial centroids as output in step 8
11: for $iter = 0$: $max\_iter\_k$ or other termination criteria
    do
12:     Allocate each data-point to a cluster based on the
        similarity measure
13:     Update the centroids using eq. (5)
14: end for
```

Algorithm 4 Hybrid-PSO

```
Initialize number of clusters as k
Initialize population size ($N$), maximum number of
iterations $maxiter$ and inertia weight $w$
Generate initial population randomly with centroids as
its parameters
for $iter = 0 : maxiter$ do
    Find fitness function value for all particles
    Update velocity of particles by eq. (9)
    Update position of particles by eq. (10)
    Update $g_b$ and $p_{b_i}$
end for
Apply K-means Algorithm to return the centroids
```

Fitness function is considered as the Equation (5). Based on this, previous position is considered as particle best position P_b as

$$P_b = [P_{b1}, P_{b2}, \ldots, P_{bK}]. \tag{7}$$

In the beginning, P_b is consider as X_i as $X = (X_1, X_2, \ldots, X_N)$. Global solution G_b is the best position of swarm in the next iteration (t) and represented as follows.

$$G_b = [G_{b1}, G_{b2}, \ldots, G_{bt}]. \tag{8}$$

Cluster centroids positions are updated with updation of velocity (v_{ik}) and positions (x_i) of the particles as shown below.

$$\begin{aligned} v_{ik}(t+1) = w \times v_{ik}(t) + c_1 \times r_1 \times P_{b_k}(t) - x_{ik}(t)) \\ + c_2 \times r_2 \times (G_b(t) - x_{ik}(t)) \end{aligned} \tag{9}$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \tag{10}$$

where t is the number of iteration, r_1 and r_2 are the random numbers between [0,1], w is the inertia weight, c_1 and c_2 are the acceleration constant. In this study, inertia weight is taken as a function of time which varies from w is 0.9–0.5. This iteration continues till it reaches pre-defined number of iteration or there is no further change in the centroid of the clusters.

3.2.1. Hybrid PSO variant for clustering approach

Van der Merwe and Engelhrecht [46] developed a hybrid algorithm based on K-means and PSO in 2003. In this study, combination of PSO and k-means is used as expressed in Algorithm 4.

3.3. Case studies considered

In this study, five different cases are considered as shown in Table 2, to explore the best possible features selection.

Case 1: Considering all the four gait parameters (stride length, cadence, leg length and age) without normalization and scaling.

Case 2: The first two gait parameters (stride length and cadence) in their original form are considered. In [13, 15, 36], authors suggest that stride length and cadences are clinically

TABLE 2. Cases considered in this study.

CASE	Gait parameter/s considered	Data size	Normalization	Scaling
Case 1	Stride length, cadence, leg length and age	156 × 4	×	×
Case 2	Stride length and cadence	156 × 2	×	×
Case 3	Stride length and cadence	156 × 2	✓ Polynomial	✓
Case 4	Stride length and cadence and different cluster size $2 \le K \le 7$	156 × 2	✓ Polynomial	✓
Case 5	Validate test case using stride length and cadence and cluster size from case 4.	156 × 2	✓ Polynomial	✓

more significant for CP analysis than another kinematics parameter when focused on single joints, in both classification and clustering study. Stride length and cadences could be affected by age and leg length. Thus before applying any machine learning techniques, normalization model is required to remove trends that/if exist, in the CP and normal dataset with respect to age and leg length.

Case 3: Considering stride length and cadence, after Polynomial normalization with leg length and age respectively. O'Malley *et al.* [13] suggest that detendring normalization technique is better than offset and decorrelation based normalization methods. They suggested first and second order polynomial model to normalize stride length with respect to leg length and cadence concerning age, respectively for each subject. In our study, we consider stride length and cadence after normalization and scaling procedure suggested by [13]. They are statistically independent and significant in CP analysis. Table 3 presents an overview of the case considered in the study.

TABLE 3. Testing gait data of four subjects A, B, C and D.

Subject	State/pathology	Stride length (m)	Cadence (m)	Leg length (m)	Age (year)	Treatment condition
A	Neurologically Intact	1.29	112.8	0.78	13	Normal
B	Neurologically Intact	1.29	122.6	0.79	19	Normal
C1	Spastic-diplegia, CP	0.59	134.0	0.66	8	Prior to surgery
C2	Spastic-diplegia, CP	0.89	110.0	0.67	9	1 year post-surgery
C3	Spastic-diplegia, CP	1.04	119.0	0.71	11	3 year post-surgery
D1	Spastic-diplegia, CP	0.20	49.5	0.45	3	Prior to surgery
D2	Spastic-diplegia, CP	0.51	74.0	0.47	4	1 year post-surgery
D3	Spastic-diplegia, CP	0.76	131.0	0.52	5	2 year post-surgery

TABLE 4. Clustering result for case 1 and 2; when four and first two gait parameters (stride length, cadence, leg length and age) are considered respectively without normalization and scaling for $k = 2$.

Case	Algorithm		Cluster purity index	Intra-cluster	Inter-cluster	MSE	Silhouette Coefficient	Dunn Index
Case 1	K-means	Mean	0.647435897	727.414639437	2249.398883024	1454.829278874	0.534981035	0.021440854
	FCM	Mean	0.660256410	747.354298589	2218.037446776	1464.736588728	0.516406269	0.004728326
	GA	Mean	0.636538462	720.143215298	2397.168829263	1440.286430596	0.543759910	0.046664080
	H-GA	Mean	0.639743590	722.851700278	2453.155978148	144 5.703400555	0.553784374	0.045306570
	PSO	Mean	0.626923077	723.916326896	2254.787022798	1447.832653792	0.523427378	0.036291725
	H-PSO	Mean	0.639743590	722.851700278	2453.155978148	1445.703400555	0.553784374	0.045306570
Case 2	K-means	Mean	0.637179487	702.160532015	2190.654305937	1404.321064030	0.546380125	0.011359487
	FCM	Mean	0.666666667	722.581050726	2221.344900625	1451.894712461	0.532371425	0.005553906
	GA	Mean	0.647435897	689.652285216	2636.076130618	1379.304570433	0.589329956	0.026969577
	H-GA	Mean	0.647435897	689.520056309	2640.796659385	1379.040112617	0.589329956	0.026969577
	PSO	Mean	0.647435897	689.766697516	2642.220564311	1379.533395032	0.589329956	0.026969577
	H-PSO	Mean	0.647435897	689.520056309	2640.796659385	1379.040112617	0.589329956	0.026969577
Case 3	K-means	Mean	0.857051282	708.519282126	1969.690460032	354.259641063	0.600939744	0.018785763
		Stan Dev.	0.003096448	5.862232143	24.266868512	2.931116071	0.003002781	0.000000000
	FCM	Mean	0.857905983	717.704684628	870.750398306	358.852342314	0.592637055	0.010628756
		Stan Dev.	0.002616976	9.006015910	1036.169813098	4.503007955	0.008819135	0.007630181
	GA	Mean	0.880128205	649.313594100	1202.618384945	324.656797050	0.430536190	0.008211998
		Stan Dev.	0.022623352	2.793767682	113.016913587	1.396883841	0.000047791	0.004224293
	H-GA	Mean	0.858974359	720.253433800	1952.655520133	360.126716900	0.596759234	0.018785763
		Stan Dev.	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
	PSO	Mean	0.895512821	648.108629885	1192.770592444	324.054314943	0.430516403	0.006462975
		Stan Dev.	0.015716389	1.185202640	66.566954312	0.592601320	0.000031287	0.002765449
	H-PSO	Mean	0.858974359	720.253433800	1952.655520133	360.126716900	0.596759234	0.018785763
		Stan Dev.	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000

Case 4: Finding the optimal number of clusters in a dataset is an open research area. In the case of rehabilitation, the number of gait profiling of the patients can play a vital role. As *case 4*, an optimal number of gait profiles in the dataset are identified.

Case 5: Evaluation of the test subject is necessary to demonstrate the significance of the surgery. To validate the clustered gait profile, four subjects are taken as *case 5* as presented in Table 3 from [13]. A and B are normal subjects while C and D are examined before and after surgery.

3.4. Clustering evaluation indices

The clustering performance is evaluated using six clustering performances indices, including both internal (based on intrinsic characteristics, thus considered as unsupervised) measures and external (based on previous knowledge about data, thus can be considered as supervised) cluster validity indices. One external validity index (cluster purity index (CPI)) and five internal validity measures (distance Measures (intra-cluster distance and inter-cluster distance), MSE, silhouette coefficient (SC) and Dunn index (DI)) [31] are considered in this study. The external cluster validity index is considered to be close to user's semantic [32].

Affiliated probability index (API) is defined as the probability of belongingness of a given test sample Ts_i over the given set of cluster C and represented as

$$\mathrm{AP}(Ts_i, Cj) = \frac{\left(\frac{1}{D(Ts_i, Cj)}\right)}{\sum_{m=1}^{K}\left(\frac{1}{D(Ts_i, Cj)}\right)} \tag{11}$$

where $D(Ts_i, Cj)$ is the affiliation distance between a testing subject and a given cluster.

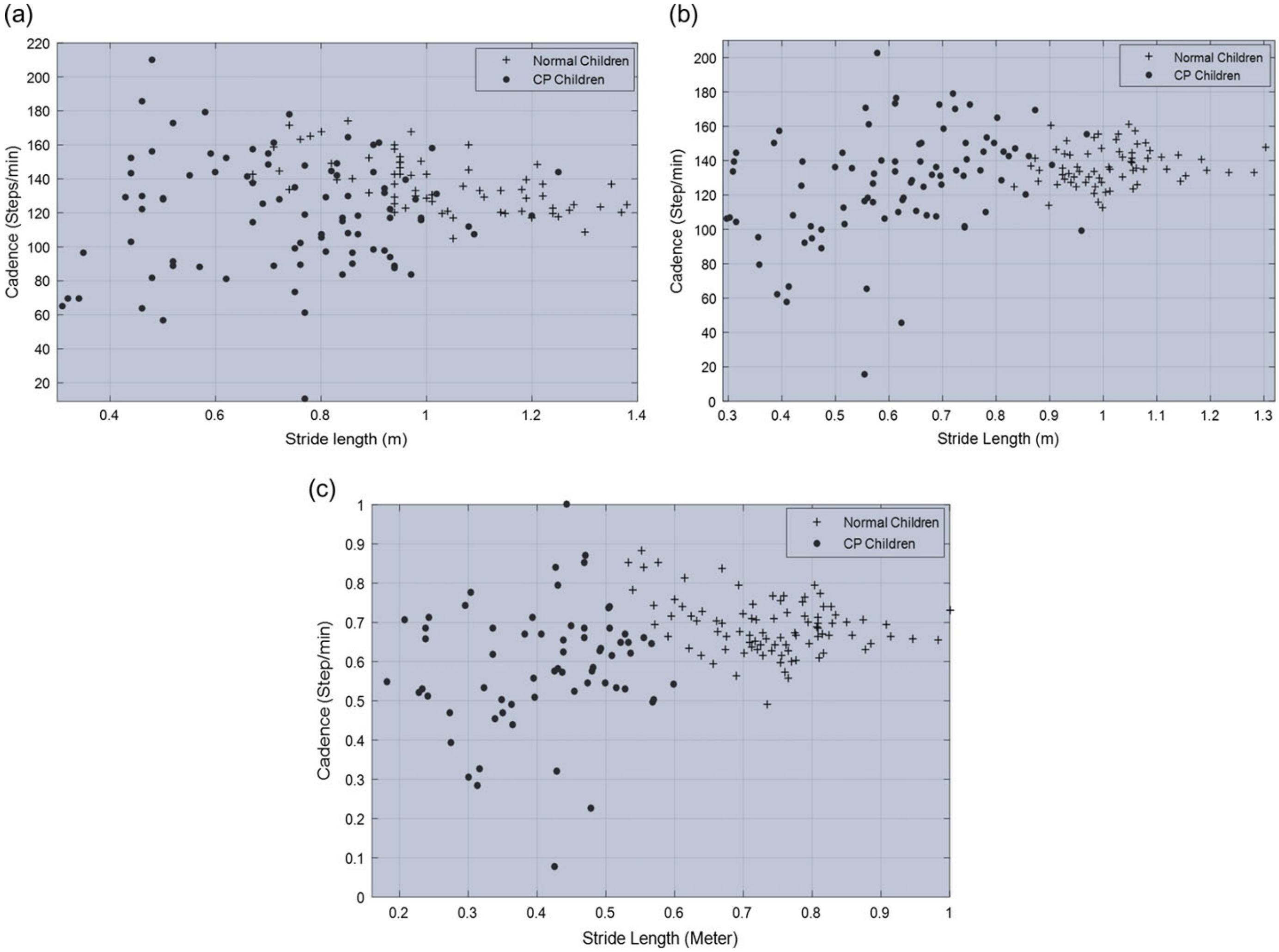

FIGURE 2. 2D plot of gait data for (**a**) CP and normal children natural data; (**b**) polynomial normalization and scale gait data with CP and normal children and (**c**) gait profiling after clustering approach.

3.5. Optimization parameter setting

The parameters affecting the performance of optimized clustering are taken as follows. In this study, population size is taken as 100 and 40 for GA and PSO, respectively. Crossover rate and mutation are chosen as 0.3 and 0.2. In this study, a time-varying inertia weight w varies from 0.9 to 0.5 is considered. w_1 and w_2 are 0.5 in objective function. All these parameters are tuned by using sensitivity analysis [47].

4. RESULT ANALYSIS

In this section, the analysis of the cases considered is discussed. On the basis of analysis of Table 4, it is found that the result of proposed variant (H-GA) and H-PSO is the same as the first cluster is optimally selected and same for both cases.

Case 1: Table 4 presents clustering result for case 1.

FCM yields the best external evaluation index; CPI when compared with the mean. Considering other internal evaluation measures, GA reports the best result in this case. The best one among all is highlighted in the Table 4. The value reported here is mean of 25 runs.

Case 2: H-GA-based optimized clustering algorithms performed best on four internal clustering performances indices. The best CPI is given by FCM as reported in Table 4 case 2. The overlapping nature of CP gait data with the normal children may be reasoned for this.

Considering other evaluation indices, H-GA-based clustering outperforms other traditional partitioning clustering techniques. Case 2 reported Minimum MSE than case 1.

Case 3: Fig. 2, illustrates the significance of the polynomial normalization for leg and age on stride length and cadence respectively. For this study, stride length can be considered a significant factor in discriminating the children with CP from the control group (neurologically intact children) when visualized by plotting a Figure. The red dot represents the children with CP, and blue is neurologically intact children. A clear separation in the dataset is observed after normalization, considered case 3 for $k = 2$.

Table 4 presents the result of K-means, GA, PSO, H-GA and PSO optimized clustering on gait data for $k = 2$.

Although classical performance metrics such as mean square error, silhouette coefficient and Dunn index are suitable methods of comparing the algorithms, they are not sufficient to find a difference in performance of the algorithms. To aggregate the performance comparison and statistical significance of computational intelligence algorithms, the popularity of parametric and nonparametric tests has increased in last few years. *T*-test (parametric) carried out for comparing different algorithms. The *t*-test assesses whether the mean of two groups of results is statistically different from each other or not. For testing, the two-tailed *t*-test is adopted with 5% significance level. The negative *t*-value with PSO as base algorithm along with low *P*-value and *h*-value of 1 w.r.t. all the other algorithms prove PSO to be significantly better than other algorithms including GA. The further comparison is made in Table 5 in the article.

Case 4: The result of different cluster sizing on three internal cluster validity indices (MSE, SC and DI) in case of $2 \leq k \leq 7$ is presented in Fig. 3. Figure 3a–c is the plot of K-means for MSE, SC and DI, respectively. Figure 3d–f is

TABLE 5. Result of *t*-test considering MSE on case 3.

Algorithms		K-means	FCM	GA	H-GA	PSO	H-PSO
K-means	*t*	0.0000	−2.2820	28.8308	−6.3298	31.9412	−6.3298
	h	0.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	P	1.0000	0.0349	0.0000	0.0000	0.0000	0.0000
FCM	*t*	2.2820	0.0000	24.6364	−1.7860	26.2677	−1.7860
	h	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000
	P	0.0349	1.0000	0.0000	0.0910	0.0000	0.0910
GA	*t*	−28.8308	−24.6364	0.0000	−80.2971	1.2556	−80.2971
	h	1.0000	1.0000	0.0000	1.0000	0.0000	1.0000
	P	0.0000	0.0000	1.0000	0.0000	0.2253	0.0000
H-GA	*t*	6.3298	1.7860	80.2971	0.0000	192.4919	0.0000
	h	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000
	P	0.0000	0.0910	0.0000	1.0000	0.0000	1.0000
PSO	*t*	−31.9412	−26.2677	−1.2556	−192.4919	0.0000	−192.4919
	h	1.0000	1.0000	0.0000	1.0000	0.0000	1.0000
	P	0.0000	0.0000	0.2253	0.0000	1.0000	0.0000
H-PSO	*t*	6.3298	1.7860	80.2971	0.0000	192.4919	0.0000
	h	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000
	P	0.0000	0.0910	0.0000	1.0000	0.0000	1.0000

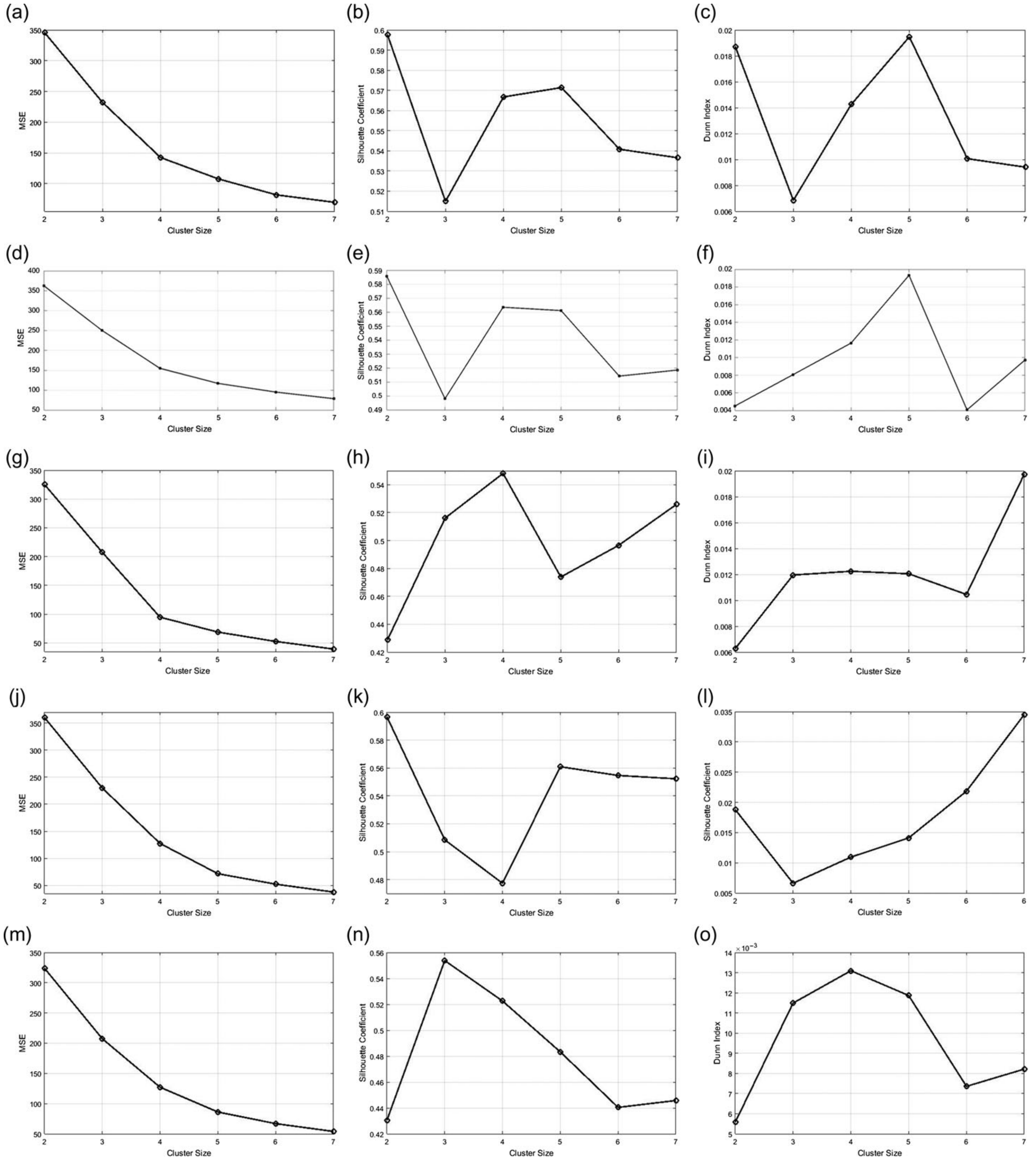

FIGURE 3. Composite error and indices plot for case 4 in terms of MSE, SC and DI for K-means, GA, H-GA and PSO clustering in case of $2 \leq k \leq 7$. (**a**) K-means and MSE, (**b**) K-means and SC, (**c**) K-means and DI, (**d**) FCM and MSE, (**e**) FCM and SC, (**f**) FCM and DI, (**g**) GA and MSE, (**h**) GA and SC, (**i**) GA and DI, (**j**) H-GA and MSE, (**k**) H-GA and SC, (**l**) H-GA and DI, (**m**) PSO and MSE, (**n**) PSO and SC and (**o**) PSO and DI.

the plot of FCM for MSE, SC and DI, respectively. Figure 3g–i plot is for GA-based clustering with MSE, SC and DI, respectively. Figure 3j–l is the plot of Hybrid GA clustering for MSE, SC and DI, respectively. Figure 3m–o is the plot of particle swam optimization based clustering for MSE, SC and DI, respectively. For each approach, cluster size is decided by majority voting. Lower the value of the MSE, better is the quality of clustering, and larger is the SC and DI value, the better is the cluster quality. Optimal cluster number (gait profiles) is chosen by majority voting among these validity indices [31].

K-means, FCM and H-GA-based clustering votes for five clusters as optimal choice, while GA and PSO-based clustering account four as the optimal clustering size. Figure 3 indicates that cluster size 5 exhibits a better quality than other given k setting. Thus conclusively, to validate test subject profile, assessment is carried out by considering k as 5 in the following analysis.

Case 5: Evaluation of the test sample is necessary to demonstrate the significance of the surgery as considered in case 5. This section discusses the analysis of generated CP gait profile, considering the optimal cluster sizing as $k = 5$ from Fig. 3, on four test subjects (brief description is in Table 1(b)) using affiliated probability index (API).

Figure 4 presents the affiliated probability distribution for two neurological intact subjects A, B and two patients C and D from pre and post-surgery in case of five clusters. Extensive analysis shows that cluster 5 in the Fig. 4 is for the subjects with the no gait pathology, the control group, followed by Cluster 3, Cluster 1, Cluster 4 and Cluster 2. Cluster 4 and 2 can be classified as gait pattern with CP cases.

The result confirms the gait pattern of subject A as normal. The study shows that subject B exhibits deviated gait pattern. After intensive analysis, it is found that the mean age of control group is 7.09 (2–13 age range), while the test subject considered is of 19-year-old. Thus this gait pattern is misclassified as cluster 4 instead of cluster 5; the control cluster.

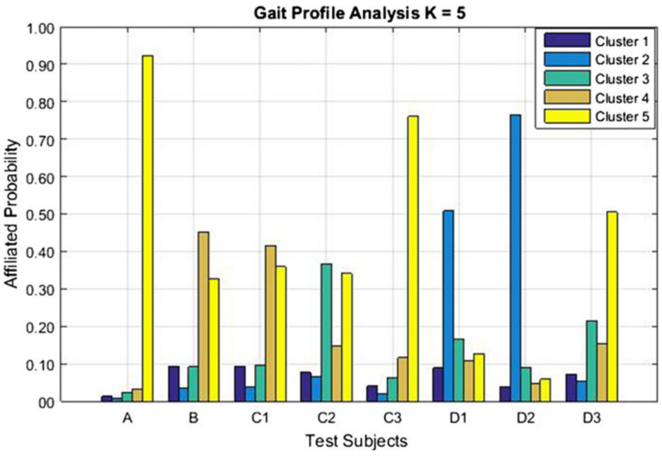

FIGURE 4. Gait profile plots for gait profile considering 5 clusters using affiliation probability.

Three observations are taken for both patients C and D. First observation of the third subject (C1) is performed before surgery, affiliated probability based gait profile also confirms the CP based pattern. In the second (C2) and third (C3) observations of the subject, C shifted from cluster 4 to cluster 3 and then to cluster 5 (the second and third observation are after post-surgery case). Gait profile based on affiliated probability index confirms that surgery helps the subject C to shift toward normal gait profile. Similar kind of observation is observed in patient D, where the first observation is before surgery, and D2 and D3 are post-surgery observations. The D1 and D2 in the plot exhibit that the problem is increased even after surgery. But after 2 years of surgery, observation D3 presents that the gait of the subject could be classified as normal gait profile.

Thus health-care professionals can take help of optimized clustering approaches to evaluate the recovery progress of patients after surgery. Doctors can monitor the recovery progress and if required change the strategy for treatment. But here it is important to note that the gait profile from automated optimization based clustering is an indicator of the gait state, thus aid doctors in making decisions regarding rehabilitation.

5. CONCLUSION AND FUTURE SCOPE

The current study shows that optimization based clustering approaches can be used for gait profiling for rehabilitation in children with Cerebral Palsy. Different cases regarding the data pre-processing are considered in this study. The dataset used is publicly available as CP gait dataset of 156 subjects having the age range from 2 to 20 for CP cases and 2 to 13 for the control group. K-means, FCM, GA, PSO and hybrid of both GA and PSO-based clustering approaches are used to find the gait profiles for the considered subjects. Best possible clustering sizing is selected based on voting based out of mean square error, silhouette coefficient, and Dunn index. The performance of the proposed clustering methods is evaluated using internal and external cluster performance index. This is validated with the unseen test samples, which undergoes surgery. The result indicates that optimized clustering technique outperforms the traditional and hybrid clustering approaches and can help in diagnosis, assessment and evaluation of treatment outcomes.

For study and comparison, the authors have used the same gait parameters as in [13] publication and there may be some missing factors (such as the number of volunteers to validate results may not be enough) and should be given prime consideration as future work. In diagnosis-related cases, there is no clear boundary between healthy and CP cases. The subject can lie in a different cluster at a time. As a future scope, one can analyze nature-inspired based optimized fuzzy clustering approach for treatment intervention for individuals with a disability. Selection of the optimal objective function can be further investigated for better CPI. The significance of the clustering approach can change with the dataset.

FUNDING

Department of Science and Technology, India (project under grant SR/S3/MERC/0101/2012).

REFERENCES

[1] Oskoui, M., Coutinho, F., Dykeman, J., Jetté, N. and Pringsheim, T. (2013) An update on the prevalence of cerebral palsy: a systematic review and meta-analysis. *Dev. Med. Child Neurol.*, **55**, 509–519.

[2] Rosenbaum, P. *et al* (2007) A report: the definition and classification of cerebral palsy april 2006. *Dev Med Child Neurol Suppl*, **109**, 8–14.

[3] Gage, J.R., Schwartz, M.H., Koop, S.E. and Novacheck, T.F. (2009) *The Identification and Treatment of Gait Problems in Cerebral Palsy*. John Wiley & Sons.

[4] Arneson, C.L., Durkin, M.S., Benedict, R.E., Kirby, R.S., Yeargin-Allsopp, M., Braun, K.V.N. and Doernberg, N.S. (2009) Prevalence of cerebral palsy: autism and developmental disabilities monitoring network, three sites, united states, 2004. *Disabil. Health J.*, **2**, 45–48.

[5] Vyas, A.G., Kori, V.K., Rajagopala, S. and Patel, K.S. (2013) Etiopathological study on cerebral palsy and its management by shashtika shali pinda sweda and samvardhana ghrita. *Ayu*, **34**, 56.

[6] Palisano, R., Rosenbaum, P., Walter, S., Russell, D., Wood, E. and Galuppi, B. (1997) Development and reliability of a system to classify gross motor function in children with cerebral palsy. *Dev. Med. Child Neurol.*, **39**, 214–223.

[7] Prakash, C., Kumar, R. and Mittal, N. (2018) Recent developments in human gait research: parameters, approaches, applications, machine learning techniques, datasets and challenges. *Artif. Intell. Rev.*, **49**, 1–40.

[8] Maenner, M.J., Blumberg, S.J., Kogan, M.D., Christensen, D., Yeargin-Allsopp, M. and Schieve, L.A. (2016) Prevalence of cerebral palsy and intellectual disability among children identified in two us national surveys, 2011–2013. *Ann. Epidemiol.*, **26**, 222–226.

[9] Gage, J.R. (1991) *Gait Analysis in Cerebral Palsy*. Mac Keith Press.

[10] Nicholson, K., Weaver, A., George, A., Hulbert, R., Church, C. and Lennon, N. (2017) Developing a clinical protocol for habitual physical activity monitoring in youth with cerebral palsy. *Pediatr. Phys. Ther.*, **29**, 2–7.

[11] Meyns, P., Pans, L., Plasmans, K., Heyrman, L., Desloovere, K. and Molenaers, G. (2017) The effect of additional virtual reality training on balance in children with cerebral palsy after lower limb surgery: A feasibility study. *Games Health J.*, **2017**, 39–48.

[12] Crouter, S.E., Kuffel, E., Haas, J.D., Frongillo, E.A. and Bassett, D.R., Jr (2010) A refined 2-regression model for the actigraph accelerometer. *Med. Sci. Sports Exerc.*, **42**, 1029.

[13] O'Malley, M.J., Abel, M.F., Damiano, D.L. and Vaughan, C.L. (1997) Fuzzy clustering of children with cerebral palsy based on temporal-distance gait parameters. *IEEE Trans. Rehabil. Eng.*, **5**, 300–309.

[14] Zhang, B.-l., Zhang, Y. and Begg, R.K. (2009) Gait classification in children with cerebral palsy by bayesian approach. *Pattern Recogn.*, **42**, 581–586.

[15] Dobson, F., Morris, M.E., Baker, R. and Graham, H.K. (2007) Gait classification in children with cerebral palsy: a systematic review. *Gait Posture*, **25**, 140–152.

[16] Cola, G., Avvenuti, M. and Vecchio, A. (2017) Real-time identification using gait pattern analysis on a standalone wearable accelerometer. *Comput. J.*, **60**, 1–14.

[17] Prakash, C., Gupta, K., Kumar, R. and Mittal, N. (2016) Fuzzy logic-based gait phase detection using passive markers. In *Proc. 5th Int. Conf. Soft Computing for Problem Solving*, pp. 561–572. Springer, Singapore.

[18] Prakash, C., Gupta, K., Mittal, A., Kumar, R. and Laxmi, V. (2015) Passive marker based optical system for gait kinematics for lower extremity. *Procedia Comput. Sci.*, **45**, 176–185.

[19] Cook, R.E., Schneider, I., Hazlewood, M.E., Hillman, S.J. and Robb, J.E. (2003) Gait analysis alters decision-making in cerebral palsy. *J Pediatr Orthop B*, **23**, 292–295.

[20] Barshan, B. and Yüksek, M.C. (2014) Recognizing daily and sports activities in two open source machine learning environments using body-worn sensor units. *Comput. J.*, **57**, 1649–1667.

[21] Novatchkov, H. and Baca, A. (2013) Artificial intelligence in sports on the example of weight training. *J. Sports Sci. Med.*, **12**, 27.

[22] Chai, Y., Ren, J., Zhao, H., Li, Y., Ren, J. and Murray, P. (2015) Hierarchical and multi-featured fusion for effective gait recognition under variable scenarios. *Pattern Anal. Appl.*, **2015**, 1–13.

[23] Prakash, C., Kumar, R. and Mittal, N. (2015) Vision based gait analysis techniques in elderly life -towards a better life. *CSI Commun.*, **2015**, 19–21.

[24] Zheng, H., Yang, M., Wang, H. and McClean, S. (2009) Machine learning and statistical approaches to support the discrimination of neuro-degenerative diseases based on gait analysis. In McClean, S. *et al* (eds.) *Intelligent Patient Management*, pp.57–70. Springer.

[25] Zhang, Z., Seah, H.S. and Quah, C.K. (2011) Particle swarm optimization for markerless full body motion capture. *Handbook of Swarm Intelligence*, pp. 201–220. Springer.

[26] Zhang, B.-l. and Zhang, Y. (2008) Classification of cerebral palsy gait by kernel fisher discriminant analysis. *Int. J. Hybrid Intell. Syst.*, **5**, 209–218.

[27] Nukala, B.T., Shibuya, N., Rodriguez, A., Tsay, J., Lopez, J., Nguyen, T., Zupancic, S. and Lie, D.Y.-C. (2015) An efficient and robust fall detection system using wireless gait analysis

sensor with artificial neural network (ann) and support vector machine (svm) algorithms. *Open J. Appl. Biosens.*, **3**, 29.

[28] Phinyomark, A., Hettinga, B.A., Osis, S.T. and Ferber, R. (2014) Gender and age-related differences in bilateral lower extremity mechanics during treadmill running. *PLoS One*, **9**, e105246.

[29] Wong, M.A., Simon, S. and Olshen, R.A. (1983) Statistical analysis of gait patterns of persons with cerebral palsy. *Stat. Med.*, **2**, 345–354.

[30] Xu, G., Zhang, Y. and Begg, R. (2006) Mining gait pattern for clinical locomotion diagnosis based on clustering techniques. In Li, X., Zaiane, O.R. and Li, Z. (eds.) *Advanced Data Mining and Applications*, pp.296–307. Springer.

[31] Nanda, S.J. and Panda, G. (2014) A survey on nature inspired metaheuristic algorithms for partitional clustering. *Swarm Evol. Comput.*, **16**, 1–18.

[32] Lai, H.P., Visani, M., Boucher, A. and Ogier, J.-M. (2012) An experimental comparison of clustering methods for content-based indexing of large image databases. *Pattern Anal. Appl.*, **15**, 345–366.

[33] MacQueen, J. *et al* (1967) Some methods for classification and analysis of multivariate observations. In *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*, pp. 281–297. Oakland, CA, USA.

[34] Rokach, L. and Maimon, O. (2005) *Clustering Methods.* Data Mining and Knowledge Discovery Handbook, pp.321–352. Springer.

[35] Zadeh, L.A. (1997) Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy Sets Syst.*, **90**, 111–127.

[36] Carriero, A., Zavatsky, A., Stebbins, J., Theologis, T. and Shefelbine, S.J. (2009) Determination of gait patterns in children with spastic diplegic cerebral palsy using principal components. *Gait Posture*, **29**, 71–75.

[37] Phinyomark, A., Osis, S., Hettinga, B.A. and Ferber, R. (2015) Kinematic gait patterns in healthy runners: A hierarchical cluster analysis. *J. Biomech.*, **48**, 3897–3904.

[38] Toro, B., Nester, C.J. and Farren, P.C. (2007) Cluster analysis for the extraction of sagittal gait patterns in children with cerebral palsy. *Gait Posture*, **25**, 157–165.

[39] Omran, M.G. (2005) Particle swarm optimization methods for pattern recognition and image processing. PhD Thesis, University of Pretoria.

[40] Goldberg, D.E. and Holland, J.H. (1988) Genetic algorithms and machine learning. *Mach. Learn.*, **3**, 95–99.

[41] Bezdek, J.C., Boggavarapu, S., Hall, L.O. and Bensaid, A. (1994) Genetic algorithm guided clustering. *Evolutionary Computation, 1994. IEEE World Congress on Computational Intelligence.,*In *Proc. First IEEE Conf.*, pp. 34–39. IEEE.

[42] Krishna, K. and Murty, M.N. (1999) Genetic k-means algorithm. *IEEE Trans. Syst. Man Cybern. B Cybern*, **29**, 433–439.

[43] Ozturk, C., Hancer, E. and Karaboga, D. (2015) Improved clustering criterion for image clustering with artificial bee colony algorithm. *Pattern Anal. Appl.*, **18**, 587–599.

[44] Eberhart, R.C., Kennedy, J. *et al* (1995) A new optimizer using particle swarm theory. In *Proc. 6th Int. Symp. Micro Machine and Human Science*, pp. 39–43. New York, NY.

[45] Omran, M., Salman, A. and Engelbrecht, A.P. (2002) Image classification using particle swarm optimization. In *Proc. 4th Asia-Pacific Conf. Simulated Evolution and Learning*, pp. 18–22. Singapore.

[46] Van der Merwe, D. and Engelbrecht, A.P. (2003) Data clustering using particle swarm optimization. *Evolutionary Computation, 2003. CEC'03. The 2003 Congress on*, pp. 215–220. IEEE.

[47] Eiben, A.E. and Smit, S.K. (2011) Parameter tuning for configuring and analyzing evolutionary algorithms. *Swarm Evol. Comput.*, **1**, 19–31.