OPEN ACCESS Freely available online

International Journal of Physical Medicine & Rehabilitation

Research Article

Digital Twin-Based Controls in Gait Analysis: A Machine Learning Approach

Paul M Trusov[1], Dacia Martinez Diaz[2], Charles S Layne[2*]

[1]Montgomery Blair High School, Maryland, USA; [2]Department of Health and Human Performance, University of Houston, Texas, USA

ABSTRACT

Human gait refers to the way people walk, which can vary widely between individuals due to factors such as body structure, age and health conditions. Traditional gait analysis often compares the walking patterns of individuals with and without medical conditions, which may incorrectly attribute natural gait variations to these conditions, thus introducing biases in diagnosing and understanding gait abnormalities. This study proposes a novel approach using machine learning to create synthetic control subjects that emulate a "healthy twin" for affected individuals. This method allows for a more refined comparison by accounting for individual-specific gait characteristics, thereby isolating abnormalities more accurately attributable to the medical condition. The method utilizes a Long Short-Term Memory (LSTM) model to analyze gait waveform data. A gait waveform is a graphical representation of the cycles of movement an individual makes during walking. By focusing on segments of the waveform that are not influenced by the medical condition, the LSTM model generates synthetic gait trajectories that mirror what the individual's gait would potentially look like if unaffected. The results show that the proposed methodology produces more accurate predictions of individual gait trajectories compared to traditional methods, which rely on average data from control groups. Therefore, the proposed approach could provide a more precise benchmark for gait comparison, thereby enhancing the accuracy of diagnoses and the efficacy of subsequent treatments.

Keywords: Gait analysis; Machine learning; Long Short-Term Memory (LSTM); Synthetic control groups

INTRODUCTION

Gait analysis, the systematic study of human walking, plays a significant role in sports science, orthopedics and neurology. This analysis assesses the biomechanical and physiological aspects of walking, with a focus on movements, body mechanics and muscle activity. A thorough understanding of gait is important for optimizing athletic performance, informing orthopedic treatments and managing neuromuscular disorders. These disorders, which impact muscle control and coordination, can alter gait patterns and pose challenges to mobility. Accurate gait analysis is therefore critical for diagnosing gait abnormalities, developing effective rehabilitation protocols and improving prosthetic, orthotic, pharmacological and genetic interventions [1,2].

To gain a comprehensive understanding of gait and its complexities, gait analysis can be approached from both a descriptive and comparative perspective, each offering distinct insights into movement patterns and their variations. A descriptive approach in gait analysis focuses on examining the gait characteristics of individuals [1]. This method involves detailed observation and measurement of gait patterns within individuals allowing researchers to understand the specific nuances of their movement. By capturing the unique aspects of a person's gait, this approach provides baseline data that can be used to monitor changes over time or assess the impact of specific interventions. Comparative analysis involves comparing individuals across different groups, such as age, gender, and health status, to identify variations in gait patterns [3]. This approach is particularly valuable in assessing gait differences associated with various health conditions, allowing for the direct comparison of gait characteristics between affected individuals and healthy controls. Through such comparisons, researchers can identify distinct gait deviations, uncover underlying biomechanical inefficiencies and

Correspondence to: Charles S Layne, Department of Health and Human Performance, University of Houston, Texas, USA, E-mail: clayne2@uh.edu

Received: 20-Nov-2024, Manuscript No. JPMR-24-35294; **Editor assigned:** 22-Nov-2024, PreQC No. JPMR-24-35294 (PQ); **Reviewed:** 09-Dec-2024, QC No. JPMR-24-35294; **Revised:** 17-Dec-2024, Manuscript No. JPMR-24-35294 (R); **Published:** 25-Dec-2024, DOI: 10.35248/2329-9096.24.S27.003

Citation: Trusov PM, Diaz DM, Layne CS (2024). Digital Twin-Based Controls in Gait Analysis: A Machine Learning Approach. Int J Phys Med Rehabil. S27:003.

evaluate the effectiveness of interventions aimed at improving mobility and function.

One challenge in performing comparative gait analysis is identifying an appropriate control group. In gait studies, it is common to match individuals in control and affected groups based on characteristics such as age, gender, or physical activity level to ensure that observed differences in gait can be attributed to specific factors rather than unrelated variables. For example, Shah et al., used age-matched controls to compare gait measures in people with Multiple Sclerosis (MS) and Parkinson's Disease (PD) with controls in both lab settings and daily life [4]. Sofuwa et al., matched participants by age, weight, and height, focusing on spatiotemporal gait parameters in PD, observing reduced step length and velocity [5]. Ebersbach et al., matched participants by age, gender, and height, comparing gait between patients with Parkinson's, cerebellar ataxia, and subcortical arteriosclerotic encephalopathy, identifying differences in stride length variability and compensation strategies [6]. Phinyomark et al., used a matched case-control methodology to compare the gait kinematics of runners with Iliotibial Band Syndrome (ITBS) to healthy counterparts [7]. By pairing subjects based on similar characteristics except for the presence of ITBS they were able to directly assess the impact of the condition on gait patterns, effectively isolating the specific effects. This approach highlights the importance of matching individuals from control and patient groups in identifying and analyzing gait deviations related to various conditions.

Several statistical methods have been developed to refine the establishment of control groups in gait analysis studies. One broadly used method, Propensity Score Matching (PSM), distinguishes itself from direct matching techniques by using statistical probabilities to reduce selection bias and enhance the comparability of study groups [8]. Instead of matching individuals solely based on observable characteristics like age, gender, or physical condition, PSM calculates a propensity score for each individual-representing the likelihood of having a certain medical condition given a set of observed covariates (i.e., likelihood of belonging to either the treated or control group based on their characteristics). Individuals are then matched across groups based on similar propensity scores, ensuring that the groups are comparable in terms of these characteristics. This method facilitates more accurate comparisons of gait patterns between normal and disabled populations by balancing both known and potential confounders across groups, thus isolating the effects of specific factors on gait abnormalities more effectively than direct matching [9].

While PSM and matched case-control methodologies offer robust frameworks for establishing control groups in gait analysis, they present inherent limitations by focusing primarily on observable characteristics such as age, gender and physical condition. These approaches may inadvertently introduce biases and noise into the matching procedures due to factors that are not directly observable (confounds) [10,11]. For example, individual habits or peculiarities in gait trajectory that are not related to the medical condition but influence walking patterns, might be misattributed to the condition leading to incorrect conclusions about the condition's role in affecting gait or the impact of interventions and characteristics of the disorder. Consequently, while these traditional methods provide a structured approach to comparing gait patterns, they might still fall short of capturing the complete biomechanical landscape, highlighting the need for enhanced methodologies that can integrate both observable and latent factors to provide a more comprehensive analysis of gait dynamics.

To address this potential shortcoming, this paper introduces a machine learning approach to develop a statistically-based control condition that allows to account for intrinsic characteristics of the subject's gait patterns. This approach utilizes time-series kinematic data (waveform) such as joint angles commonly captured by standard gait analysis systems (e.g., Vicon® motion capture system) ensuring that the matching is rooted in the most direct and significant indicators of gait [12].

The intuition behind the proposed approach is as follows, at the core of the method is a time-series kinematic data collected from a sample of healthy individuals that is used to learn the relationships between various waveform fragments in individual gait (e.g., early *vs.* later stride trajectory of ankle movement) using machine learning techniques. Using the predictive ability of the trained model that captures the inferred relationships, one can reconstruct other fragments of the waveform given appropriate inputs are provided. For example, appropriate inputs labeled here as the anchor, can be the corresponding waveform fragments taken from an individual with a medical condition enabling the construction of an individual-specific waveform as if it were from a healthy individual. Essentially, this method creates a "healthy twin" or synthetic control subject that serves as a more precise benchmark for comparison with the actual gait of an individual affected by a specific medical condition.

By using this matching strategy, we aim to isolate the specific effects of the medical condition on gait, thereby providing a more accurate assessment of the deviations caused by the condition itself, rather than variations due to intrinsic characteristics. The proposed methodology can be easily integrated with traditional matching variables enhancing performance of the well-established techniques such as PSM (Figure 1).

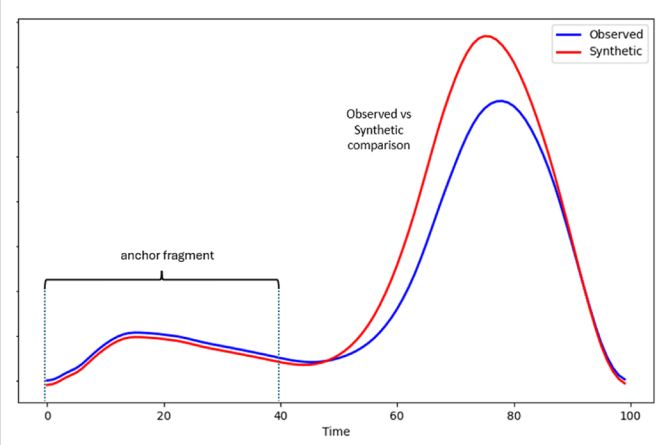

Figure 1: Conceptualization of the method.

The necessary condition for the proposed methodology to be effective is the ability to identify anchors or fragments of a waveform that, (a) reflect the intrinsic gait characteristics of the subject and (b) remain unaffected by the disease or disabling condition. We argue that this identification can be accomplished

either through expert knowledge or by utilizing empirical methods. For instance, applying data mining techniques to datasets containing waveforms from both healthy individuals and those with medical conditions enables the extraction of waveform fragments that are unique to each healthy subject, yet do not permit differentiation, in a classification sense, between healthy and affected groups.

The next section presents the methodological framework of our approach, detailing each step of the proposed methodology. The results based on synthetic data are then presented to demonstrate the preliminary efficacy of our data mining technique. Subsequent sections apply the methodology to actual individual data, providing a validation of our approach. The paper concludes with a discussion of the findings and key implications of our research.

MATERIALS AND METHODS

Methodological framework

Our methodological framework is presented in Figure 2 and includes the following six steps: Data collection, data preparation, identification of anchors, model training on healthy individuals, generating synthetic control subjects, and comparison analysis.

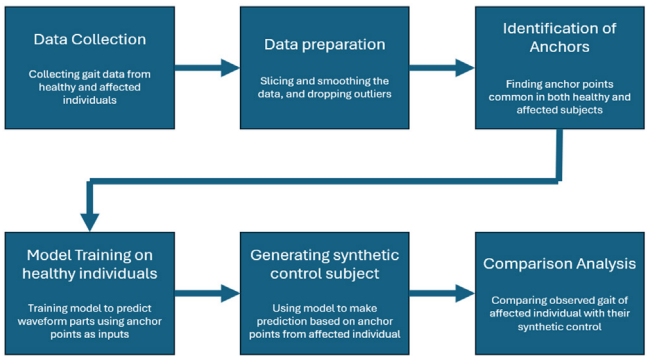

Figure 2: Methodological framework flowchart.

Data collection: The process begins with the collection of kinematic data from both healthy and affected individuals. Data is gathered using conventional motion capture systems, such as the Vicon® motion capture system. A larger dataset is recommended for healthy individuals, as machine learning models perform better with extensive data, allowing the model to capture the full range of normal gait variations accurately. Fewer recordings are necessary for affected individuals, as this data is used for comparative analysis rather than model training.

Data preparation: The collected waveform data is divided into individual strides. Smoothing techniques are applied to reduce noise. Outliers are identified and removed.

Identification of anchors: Individual strides data is inspected to identify parts of the waveform that vary across subjects while not showing systematic differences between healthy and affected individuals. The identification can be achieved either through data analysis techniques or expert input.

Model training: The stride data collected from healthy individuals is divided into two parts that is a fragment of the waveform serving as an anchor and additional fragments used for subsequent comparative analysis. For affected individuals, the corresponding (non-anchor) fragments are expected to display gait abnormalities. A machine learning model is trained using this data, with the anchor segment as the input and the remaining parts of the waveform as the output, requiring the model to effectively handle sequential data and capture temporal dependencies.

Generating synthetic controls: Anchor fragments from affected individuals are used to create a synthetic gait waveform through the trained model. This waveform represents what the gait would likely look like if the individual whose anchor fragments were used as input had normal biomechanical function, unaffected by the disorder.

Comparison analysis: A standard comparative analysis is conducted using the observed waveforms and the corresponding synthetic control generated by the model for the focal affected individuals. This analysis helps to identify and quantify deviations between the actual and synthetic gait patterns, highlighting the impact of the disorder on the individual's movement. The analysis can be performed in various ways, such as correlation or cross-correlation techniques or Statistical Parameter Mapping (SPM) [13,14].

Predictive model

The effectiveness of the proposed methodology depends on the model's ability to predict counterfactual segments of an affected individual's waveform, simulating how it would appear if the individual were healthy, while retaining the distinct characteristics of the subject's gait. Among various techniques suited for predicting time-series trajectory data, we have selected the Long Short-Term Memory (LSTM) model, a type of machine learning tool specifically developed for handling time-dependent data [15]. LSTMs are designed to model sequences where future values are influenced by past observations, making them well-suited for capturing patterns in kinematic data, such as knee angles during walking.

For instance, when provided with a fragment of a knee angle trajectory, an LSTM model can predict the continuation of the movement by learning from the temporal relationships in the previous data. Unlike models that treat each time point independently, the LSTM retains relevant information from previous time steps through its internal memory structure. This enables it to make more accurate predictions about the knee's future movement, whether that involves forecasting the next segment of the trajectory or reconstructing the entire waveform of knee movement.

Here the Figure 3 illustrates a possible configuration of the LSTM model, where the first 40 (T) timepoints of the waveform are used as input (anchors), and the remaining 60 (100-T) observations are treated as output (the part of the waveform to be predicted). Here, N refers to the total number of strides collected from multiple healthy individuals used for model training. The model consists of three layers: the first layer processes the input sequence, the second layer applies the LSTM transformations to capture temporal dependencies over the sequence, and the final dense layer combines the LSTM output to match the dimensions of the predicted waveform.

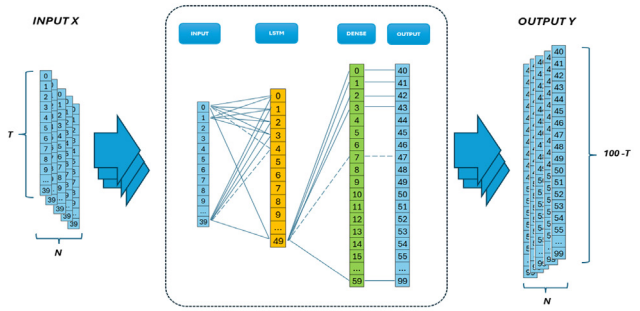

Figure 3: Modeling gait patterns using LSTM.

During training, the model is iteratively adjusted to minimize prediction errors (i.e., deviation of predicted Output Y from actual data) using the data from healthy individuals. The model progressively refines its parameters by processing diverse gait data, thereby enhancing its ability to generalize across different conditions (i.e., individual gait variations). One could think of the prediction process as matching the input data (anchor) from the individual with a medical condition to the corresponding waveform fragments taken from a sample of healthy individuals. The best-matching waveforms from the healthy set are then used as a synthetic control for the focal subject. While this description is an oversimplification of how the LSTM model is actually trained and makes predictions, it captures the core idea behind the approach (Figure 4).

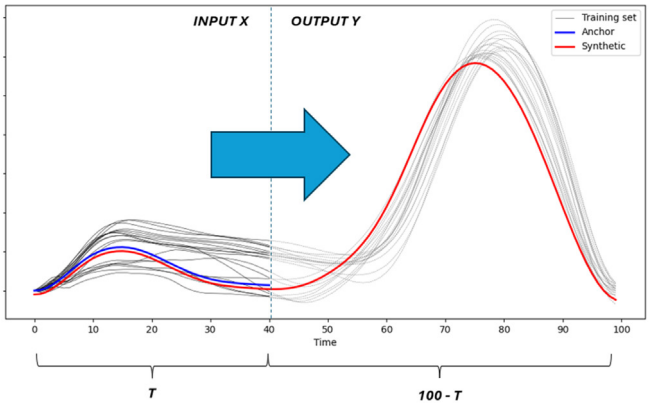

Figure 4: Predicting the last 60 timepoints using the first 40.

Sample size requirements

The model's performance relies on both the quality and quantity of training data, as adequate data is necessary for accurately capturing underlying patterns and temporal dependencies in waveforms. However, determining the optimal amount of data needed for reliable predictions can be challenging due to the risk of overfitting with small datasets or underfitting with insufficient training data. To address this concern, we first investigate the relationship between sample size and the prediction accuracy of our LSTM model using synthetic data. This approach allows us to explore various dimensions characterizing our training data without engaging in potentially time-consuming and costly data collection with actual subjects.

To generate synthetic data, we used a sample of 300 strides from a single healthy individual. We calculated the mean waveform and fitted a Fourier series model with 10 parameters to represent it. Random perturbations were then applied to each parameter to create a large sample of synthetic waveforms (strides). Through an iterative process of visually inspecting the generated curves and adjusting the perturbations, we produced a sample that closely resembles gait data typically collected from actual subjects.

With a virtually unlimited sample size, we investigated how the number of strides used in LSTM model training impacts its predictive performance. In all experiments, the first 50 data points were used as input to the model (i.e., anchor). As a benchmark, we used the sample average waveform, which aligns with the standard approach of using the average waveform from healthy individuals as a control condition for comparison with kinematic data from an affected subjects.

The results of our analysis depicted in Figure 5 reveal that the improvement in model accuracy (as measured by Mean Absolute Percentage Error (MAPE)) follows an exponential decay pattern initially, with significant reductions in predictive error observed as sample sizes increased to about 1000 strides. Beyond this point, the increments in accuracy become marginal. This plateau suggests that there exists a threshold in the sample size-identified here around 1000 data points beyond which the addition of more data does not substantially enhance the model's performance (Figure 5).

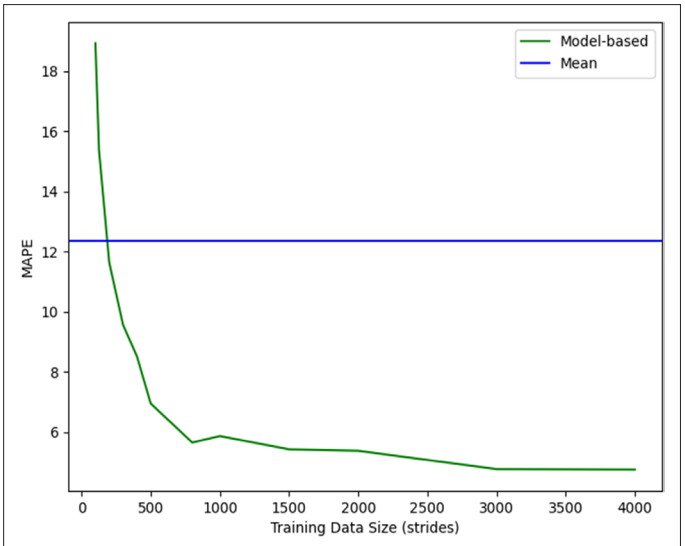

Figure 5: Mean Absolute Percentage Error (MAPE) as function of sample size.

It is important to note that these findings should not be interpreted as generalizable to our proposed model architecture, as the analysis is performed on well-behaved synthetic data. Actual data from field studies are typically subject to greater noise and significant irregularities across participants. Nevertheless, we find this analysis valuable, as it helps establish a lower bound for the sample size required to train the proposed model.

Method's validation with field data

Following the validation with synthetic data, the methodology

was applied to real-world field data to evaluate its performance. Gait waveforms from three subjects were used to replicate the synthetic data analysis, with the LSTM model generating synthetic control trajectories based on anchor segments from each subject's gait. We explored the predictive performance of several anchors of varying lengths, representing key stages in the gait cycle, from Foot Flat (5%-10%) to Terminal Swing (up to 87%). As a benchmark, the sample average the average waveform computed from a sample of healthy individuals was used for comparison. This application allowed us to assess how well the model performed in creating individualized predictions of healthy gait patterns using actual data, further validating its potential for identifying gait abnormalities in practical scenarios.

Field data collection and preparation

Three healthy young adults (two males) were recruited from the University of Houston campus. The participants' average age was 23.5 years (SD 2.0). Infrared reflective markers were bilaterally applied to the posterior and anterior superior iliac spine (hip), lateral knee, shank, lateral malleolus, 1st metatarsophalangeal joint, and heel. This setup allowed for the development of hip, knee, and ankle joint time-series waveforms using the Vicon® Plug-in-Gait model, which were then separated into individual strides with heel strike as the temporal demarcation event. Individual strides were time-normalized such that each stride was represented by 100 samples. After a five-minute acclimation period, the participants walked on a motorized treadmill at a self-selected comfortable speed for 15 minutes. This protocol resulted in approximately 540 strides being obtained for each participant.

RESULTS AND DISCUSSION

Following the approach used in our sample size exploration with synthetic data, the LSTM model was trained using the initial part of the gait as inputs (anchor) and the remaining part of the waveforms as the output. We used 80% of strides as a model training sample and the remaining 20% as the holdout. Table 1 presents our results and comparison with the naive (mean-based) model.

The results suggest two main findings: 1) Prediction accuracy improves as more waveform data is allocated to the anchor used for model training. For instance, the MAPE of the model trained on the "Foot Flat" stage of the gait is 553%, whereas for anchors starting from the "Pre-Swing" stage, the error drops below 75%; 2) More importantly, the proposed individualized approach to control condition formation consistently outperforms the naive (average-based) approach, supporting the key proposition of this paper (Table 1).

Table 1: Model predictive performance and comparison with average-based predictions for multiple anchors.

	%	LSTM model			Average		
		MSE	MAE	MAP	MSE	MAE	MAP
Foot flat	5	23.3	3.9	553.1	23.3	3.9	545.0
Midstance	10	16.4	3.0	433.1	24.3	4.1	570.2
Terminal stance	30	12.3	2.4	113.8	24.2	4.0	362.3
Pre-swing	50	12.4	2.4	74.8	22.2	3.6	105.1
Initial swing	60	5.9	1.8	34.1	24.8	3.9	47.5
Mid-swing	73	15.9	3.0	55.7	21.2	3.7	62.9
Terminal swing	87	1.8	1.0	22.8	22.8	3.9	123.6

Note: MSE: Mean Squared Error; MAPE: Mean Absolute Percentage Error; MAE: Mean Absolute Error

This study sought to enhance gait analysis by applying a machine learning-driven approach to overcome limitations in traditional methods. The core innovation involves using a predictive model to reconstruct subject-specific gait trajectories from waveform fragments that are unaffected by the medical condition. This allows for the creation of a more precise control group by reflecting the individual peculiarities of the focal subject. Consequently, this method could improve diagnostic accuracy by identifying subtle abnormalities that conventional methods might overlook.

The methodology presents significant potential in medical fields focused on biomechanics. By generating synthetic control subjects from the waveform data of healthy individuals, this approach provides an individualized benchmark for comparison, leading to more accurate diagnoses and personalized treatment plans. Furthermore, the proposed methodology could be expanded beyond gait studies to analyze other biomechanical waveforms, such as those related to cardiovascular or respiratory functions. By applying the same machine learning principles to these waveforms, researchers could develop tools for early detection of abnormalities in heart rhythms or breathing patterns, potentially leading to innovations in diagnosing and treating a wide range of health conditions [16].

CONCLUSION

As the immediate next steps in developing this methodology, we envision two key directions: 1) Given the critical role of the anchor, which must reflect the subject's intrinsic gait characteristics and remain unaffected by the medical condition, it is important to develop and test effective identification techniques; 2) To further support the ideas presented in this paper, it would be beneficial to benchmark the performance of the synthetic controls against alternative methods of comparative analysis, such as control group averages, PCA-based filtering, or propensity score matching. The primary challenge in such benchmarking lies in establishing the ground truth for the affected individuals' gait deviations from the healthy norm. We believe that a combination of expert knowledge and theoretical guidance could be helpful in defining these norms.

REFERENCES

1. Baker R. Gait analysis methods in rehabilitation. J Neuroeng Rehabil. 2006;3(1):4.
2. Tao W, Liu T, Zheng R, Feng H. Gait analysis using wearable sensors. Sensors. 2012;12(2):2255-2283.
3. Duhamel A, Bourriez JL, Devos P, Krystkowiak P, Destee A, Derambure P, et al. Statistical tools for clinical gait analysis. Gait Posture. 2004;20(2):204-212.
4. Shah VV, McNames J, Mancini M, Carlson-Kuhta P, Spain RI, Nutt JG, et al. Laboratory *versus* daily life gait characteristics in patients

with multiple sclerosis, Parkinson's disease, and matched controls. J Neuroeng Rehabil. 2020;17(1):159.

5. Sofuwa O, Nieuwboer A, Desloovere K, Willems AM, Chavret F, Jonkers I, et al. Quantitative gait analysis in Parkinson's disease: Comparison with a healthy control group. Arch Phys Med Rehabil. 2005;86(5):1007-1013.
6. Ebersbach G, Sojer M, Valldeoriola F, Wissel J, Muller J, Tolosa E, et al. Comparative analysis of gait in Parkinson's disease, cerebellar ataxia and subcortical arteriosclerotic encephalopathy. Brain. 1999;122(7):1349-1355.
7. Phinyomark A, Osis ST, Hettinga BA, Leigh R, Ferber R. Gender differences in gait kinematics in runners with iliotibial band syndrome. Scand J Med Sci Sports. 2015;25(6):744-753.
8. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41-55.
9. Biffi E, Beretta E, Storm FA, Corbetta C, Strazzer S, Pedrocchi A, et al. The effectiveness of robot *vs.* virtual reality-based gait rehabilitation: A propensity score matched cohort. Life. 2021;11(6):548.
10. Austin PC, Yu AYX, Vyas MV, Kapral MK. Applying propensity score methods in clinical research in neurology. Neurology. 2021;97(18):856-863.
11. Reiffel JA. Propensity score matching: The 'Devil is in the details' where more may be hidden than you know. Am J Med. 2020;133(2):178-181.
12. Merriaux P, Dupuis Y, Boutteau R, Vasseur P, Savatier X. A study of vicon system positioning performance. Sensors. 2017;17(7):1591.
13. Ogihara H, Tsushima E, Kamo T, Sato T, Matsushima A, Niioka Y, et al. Kinematic gait asymmetry assessment using joint angle data in patients with chronic stroke-a normalized cross-correlation approach. Gait Posture. 2020;80:168-173.
14. Alhossary A, Pataky T, Ang WT, Chua KSG, Kwong WH, Donnelly CJ. Versatile clinical movement analysis using statistical parametric mapping in MovementRx. Sci Rep. 2023;13(1):2414.
15. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735-1780.
16. Kukke SN, Curatalo LA, de Campos AC, Hallett M, Alter KE, Damiano DL. Coordination of reach-to-grasp kinematics in individuals with childhood-onset dystonia due to hemiplegic cerebral palsy. IEEE Trans Neural Syst Rehabil Eng. 2016;24(5):582-590.