# Gait Analysis to Track Parkinson's Disease Evolution Using Reproducible Research Practices

***Abstract*—A research work is reproducible when all research artifacts such as as text, data, figure and code are available for independent researchers reproduce the results. In this paper, we present a reproducible gait analysis to track Parkinson's Disease evolution by monitoring walking abnormalities. We applied Principal Component Analysis into gait data to detect user's abnormalities that may indicate the progression of Parkinson's Disease. We validated our approach with a public database of foot sensor data, which includes vertical ground reaction force records of subjects with healthy gait and Parkinson's Disease patients. We used the euclidean distance as data classifier. We reached a classification accuracy of 81.00% with leave-one-out cross-validation, which demonstrates the feasibility of our approach for tracking PD's symptoms based on user gait. All relevant data to reproduce our results are available n a public web page.**

## 1. Introduction

Gait analysis is the systematic study of human locomotion, including qualitative and quantitative assessment. Physicians and physiotherapists apply gait analysis subjectively in clinical evaluation, which sometimes is followed by a survey regarding gait quality [1]. This research area has attracted the interest of multidisciplinary researchers from medicine, physiotherapy, electronic engineering and computer science. The first gait analysis systems used video camera and were important to characterize the human locomotion and correlated studies. Nevertheless, with the advances and reduced costs of wearable sensors, gait analysis research has evolved in the last years. Nowadays, an effective approach is using foot sensors to collect gait data from the forces between the foot and the ground defined as Vertical Ground Reaction Force (VGRF) [2].

Gait analysis has been strongly applied to evaluate the evolution of neurological diseases such as Parkinson's Disease (PD) [1], which affects about 2% of the world population [3]. It is promising for PD research, because it enables the quantified progress tracking of PD patients using sensor-based gait analysis [4].

Yang et al. [5] present an accelerometer-based gait analysis system. They provide a graphical interface to display and analyze the gait acceleration data recorded by an accelerometer attached to the lower back of subjects. This system calculates a total of four gait features: spatio-temporal, frequency domain, regularity and symmetry. However, they do not calculate walking velocity; it is required that the users input this information.

Dillman et al. [6] used Principal Component Analysis (PCA) to analyze movements of the upper and lower extremities during treadmill walking with healthy subjects and two groups of PD's patients. This study considered 35 control subjects and 36 PD patients. They aimed to analyze the value of PCA to characterize dynamic changes of gait, focusing on the gait velocity and the severity of the disease.

Mazilu et al. [7] investigated the acceptance of a wearable gait assistance for PD's patients at their home. They developed a gait training system that detected Freezing of Gait (FoG) events. With the FoG detection, they used a rhythmic auditory signal that serves as gait stimulation to support the user to alleviate the FoG symptom and resume the gait movement.

Rodrigo et al. [8] used Kohonen map to identify subjects with PD based on their gait patterns. The extracted features were based on mean coefficient of variation and mean sum of VGRF during consecutive stance phases to describe the inter-subject variability. From the analysis of these results the Kohonen map was able to recognize: True Positive (*TPRate*) of 72.09% and True Negative (TN) of 70.21%.

In this paper, we propose a gait analysis approach to track PD evolution by monitoring walking abnormalities. We apply PCA into gait data to detect abnormalities that may indicate the progression of PD. We validated our approach with a public database of foot sensor data, which includes vertical ground reaction force records of healthy subjects and PD patients. We applied the PCA into the data and used the euclidean distance as data classifier reaching a classification accuracy of 81.00% with leave-one-out cross-validation [9].

This paper is organized as follows: in Section 2, we introduce the motivations and practices regarding reproducible research; in Section 3, we introduce Gait analysis; PCA is described in Section 4; in Section 5, we present our approach; in Section 6, the materials and methods of the research are presented; in Section 7, we show the results; and finally, in Section 8, we discuss conclusions and future works.

## 2. Reproducible Research

Computational sciences such as computer science, statistics, many fields of engineering, and signal processing are theoretical and experimental. These sciences are result

of a combination of theorems proving and development of computer codes to validate the research. According to Vanderwalle [10], to reproducible research increases the publication impact in these sciences. A research work is called reproducible if all information relevant to the work, including, text, data and code, are available for the independent researchers can reproduce the results. Donoho [11] encourages to publish research articles using reproducible research practices to widespread the knowledge. The suggested format for publishing reproducible research is a Web page containing the following items [10]:

1) title of the publication;
2) authors;
3) abstract;
4) full reference of the paper;
5) all the code to reproduce the results, images and tables;
6) all the data to reproduce the results;
7) list of configurations on which the code was tested (software version, platform)
8) a possibility to give comments and remarks (and to report bugs).

Therefore, to share our research and receive feedback from independent researchers, we created a web page containg all the relevant information to reproduce our results [1]. The source code is under GPL licence version 3.0 [2], which means that researchers can use it to developr their own researchs.

## 3. Gait Analysis

Gait analysis is a systematic study of human locomotion. It involves measuring, describing and evaluating gait characteristics to diagnose, rehabilitate or track the progress of patients.

The human gait is a periodic movement of the limbs during locomotion over a solid substrate. Each gait cycle starts when a foot initiates contact (i.e., heel strike) with the ground and restarts when it touches the ground again, as shown in Figure 1. Thus, each cycle begins at stance phase, initiated when the reference foot's heel strikes the ground and undergoes a swing phase, which is initiated when the reference foot's toe is off the ground and finalized when its heel strikes the ground (i.e., beginning of new stance phase). In healthy subjects, the stance phase represents, approximately, $60\%$ of the cycle, and the swing phase, $40\%$.

Gait analysis studies the forces and moments of the movement of body segments in a human gait, including the measurement of VGRF. The patients use adapted force sensors under the feet and attached to the shoes to measure the VGRF [4]. The result of this acquired data is the movement signal.

Toledo *et. al* [12] compared PD's patients with healthy subjects [12], [13] and identified the PD's patient gait characteristics is having shorter steps, reduction in the joints amplitude and an extension of the stance phase reaching almost $80\%$ for each gait cycle [14].

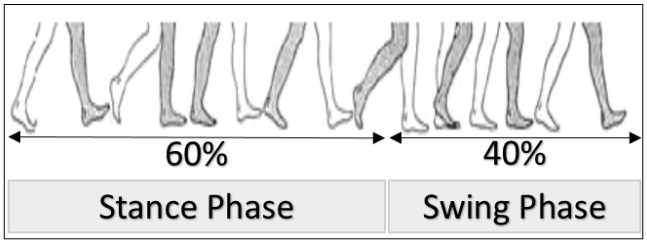


Figure 1. Gait Cycle Phases.

Given that the gait kinetics' studies provide clinical information regarding patient's clinical condition [4], we analyzed the signal processing of the sensor raw data of the Parkinson Disease Database [15] to evaluate the PD progress based on the gait characteristics.

## 4. Principal Component Analysis

Principal Component Analysis (PCA) is a statistic procedure to reduce data and eliminate redundancies. It identifies the data variance and applies linear data transformation to detect the most relevant data components on the first dimension, called the main axis. The second remaining variance is the secondary axis and so on [16].

PCA consists of the following steps [17]:

1) Scale the measurement data into an *m* x *n* matrix, where *m* is the number of measurement types and *n* is the number of samples;
2) Subtract the mean for each measurement type;
3) Calculate the *eigenvectors* and *eigenvalues* of the covariance matrix.
4) The Calculated *eigenvectors* and *eigenvalues* can be used to project the data into a new space called *eigenspace*.

In Figure 2, we show the difference of the mean gait between healthy and PD subjects. We used PCA to identify the difference of the mean gait between PD patients and healthy subjects.

## 5. Our Approach

The biomechanical analysis of human gait is part of the diagnosis and treatment process for PD. During this analysis, the patients are required to walk. The doctor analyses the gait by looking at the swing and stance phase and the gait posture. To automate the identification of each gait phase in the VGRF data, it is necessary to use signal processing techniques. In this work, we focus on the VGRF of each foot and identifying when the foot initiates contact (i.e., start of stance phase) with the ground and when it is off the ground (i.e., start of swing phase). For this purpose, we used the peaks and valleys technique as shown in details in Section 4.1.

1. Website URL: http://gaitparkinson.wordpress.com
2. http://www.gnu.org/licenses/gpl-3.0.en.html

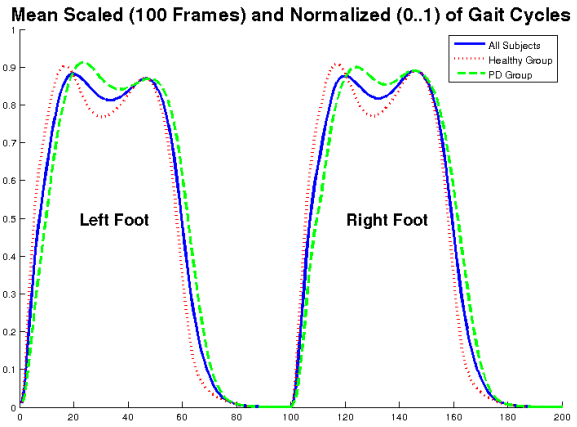


Figure 2. Mean Vector of the Gait Cycles.

After identifying the cycle, we extract a window length with the cycle movement and the gait characteristics and transform the VGRF data into PCA values. We project the values into the *eigenspace* providing a visualization tool to doctor. With this information, the doctor evaluates the patient's gait.

In Figure 3, we present an overview of our approach applied to a health monitoring system. Its architecture is composed of the gait system acquisition and signal processing component. The gait acquisition component is responsible for collecting user data and storing them in a database. The signal processing component is responsible for processing the data and making the results available to the healthcare professional.

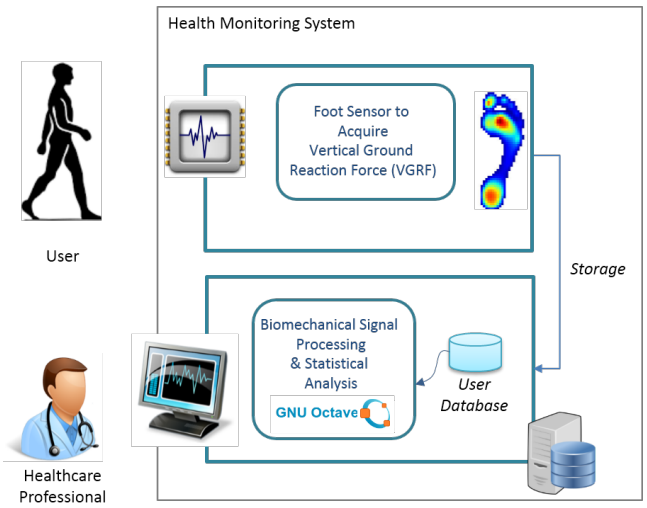


Figure 3. System Overview

## 5.1. Signal Processing Technique

In this work, we applied peaks and valleys detection technique to identify the beginning and end of each gait cycle. This technique identifies the peaks and discard very low values when they are considered noise. The peak is the highest point between the two lower points, which are considered the cycle valleys. The technique is applied to the signal using low pass filter to remove the signal noise. Thus, our biomechanical signal processing uses raw data, filters noise, identifies motion cycles and extracts feature vector characterizing the motion. Furthermore, we applied machine learning techniques for data classification. The complete process of gait analysis based on PCA is shown in Figure 4. It is divided into two phases: biomechanical signal processing and statistical analysis.

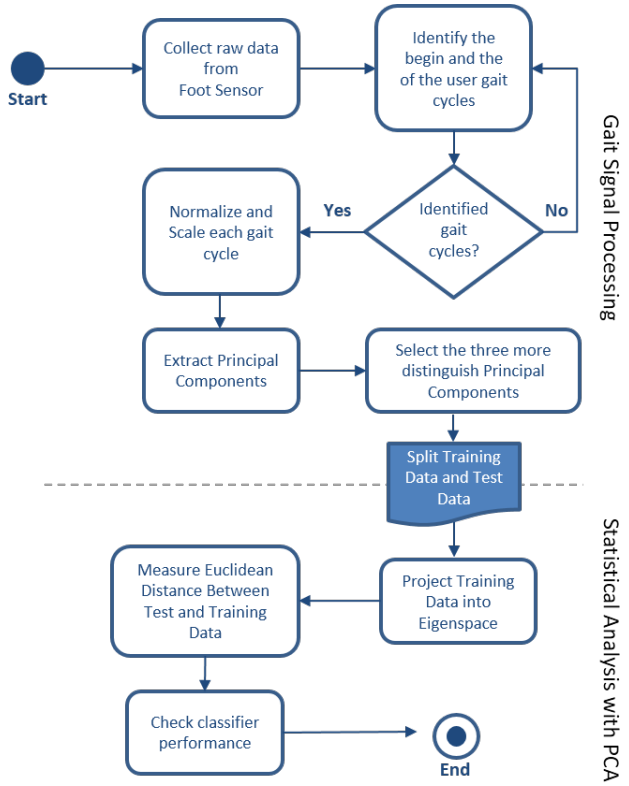


Figure 4. Biomechanical signal processing.

Considering the periodic movement of the human gait, we applied the peak and valley technique to identify each gait cycle of the raw data as shown in Figure 5. After identifying the gait cycle, we normalized the values from 0 to 1. Calculating the principal components requires a quadratic matrix [17] ($m$ x $n$). Therefore, we scaled each gait cycle into 100 frames to generate the principal components and consequently extract the *eigenvalues* and *eigenvectors*. We used a total of 120 gait cycles for each subject.

# 6. Material and Methods

In this work, we used the data available at physionet[3], which contains the VGRF records of subjects as they walked

3. http://physionet.org/pn3/gaitpdb/

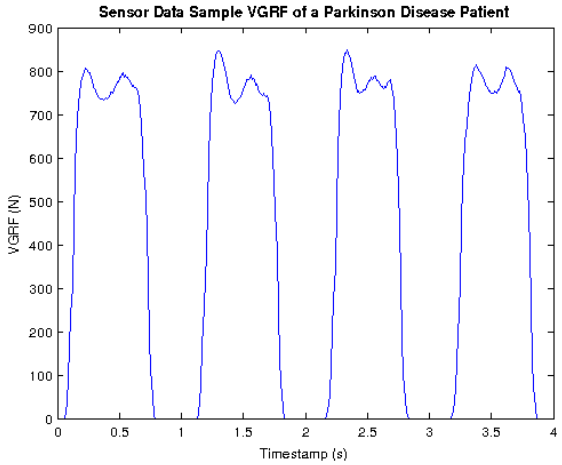


Figure 5. Sample of Acquired Signal of Foot Sensor.

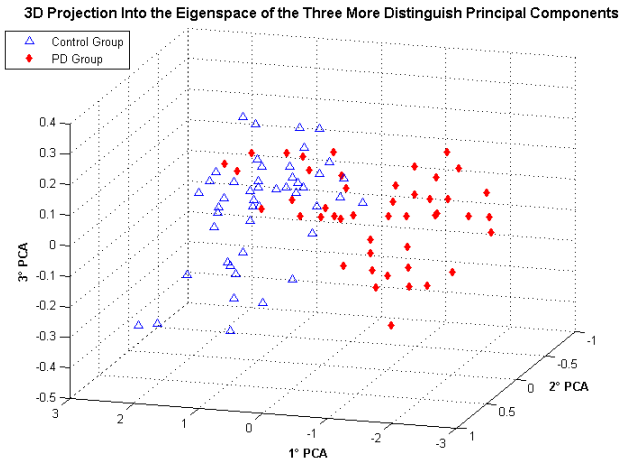


Figure 6. Projection of the Three More Distinguish Principal Components

at their usual pace for approximately 2 minutes on level ground. To populate this database, data was collected using 8 sensors (Ultraflex Computer Dyno Graphy, Infotronic Inc.[4]) attached underneath each subject's foot. The sensors measure force (in Newtons) as a function of time. We used two signals, which represent the sum of the 8 sensor outputs for each foot. Even though we used a public database, this work can be adapted to other foot sensors capable of capturing VGRF.

Our source code is under GPL License Version 3.0 and was developed using an Open Source Application (Octave Version 3.8.1 [5]) for numerical computations. The Octave language is similar to Matlab [6] and most programs are easily portable. All data necessary to reproduce our results are available online, as presented in Section 2;

We used an open database under the ODC Public Domain Dedication and License (PDDL) v1.0, which allows to freely share, modify, and use this work for any purpose and without any restrictions. We used The Parkinson's Disease database available at [15] to develop a gait analysis system under GPL license to track PD symptoms. Given data quality, we selected data from fifty PD patients and fifty healthy control subjects [7]. Considering both groups, we had mean age of 66.3 years and 63% men. We applied signal processing technique for the VGRF signal data acquired by foot sensors. We used a supervised learning approach to train our system to identify the subject group (PD or Control) and with the PCA projection into the *eigenspace* we evaluated the PD's progress through the gait characteristics.

By applying the signal processing techniques, we identified gait cycles and calculated the PCA as our feature vectors. The first three PCA were used in our gait analysis PD's tracking system through the projection of the PCA into the *eigenspace* [17]. Thus, we identified two clusters of the group classes and used the class identification to separate them (i.e. control group and PD patients) as shown in the Figure 7.

## 7. Validation

The goal of using PCA for gait analysis is to classify the data and identify the presence of PD. The theory of statistical learning provides a set of techniques for data analysis that allows the acquisition of knowledge by supervised learning methods. PCA [17] is mathematically defined as an orthogonal linear transformation that transforms data into a new coordinate system in which the greatest variance by any projection of data lies along the first coordinate (called the first component), the second greatest variance lies along the second coordinate, and so on. Thus, we used the euclidean distance of the test data to the training data to classify each subject as PD's Group or Control Group.

In Figure 7, we show the projection of the data from both groups, in which the red dots represent PD patients and blue dots, healthy subjects. By analyzing, it is clear that there are two clusters. To verify if the PCA values could be used to track the gait characteristics of a subject, we project the PCA values and draw a trajectory with different measurements of the same person over time, seeing how the symptoms increases or the patient's health improves.

In our analyses the changes in the patient's gait pattern will result in a moving of this point. For instance, if the point moves towards the direction of the subjects with PRGD, so the system detects an increase of PRGD symptoms with the patient.

For each subject, we defined an ID in the database to design and identify them in the *eigenspace* as shown in Figure 7. Using the PCA technique [17], we identified the two distinct groups of individuals: PD patients and healthy subjects. Changes in one of Parkinson's gait pattern can result in a change of location in the *eigenspace*. If the change

4. www.infotronic.nl/items/ExamplesEdwin.pdf
5. http://physionet.org/pn3/gaitpdb/
6. http://www.mathworks.com
7. The list of used subjects are described in our Reproducible Research website

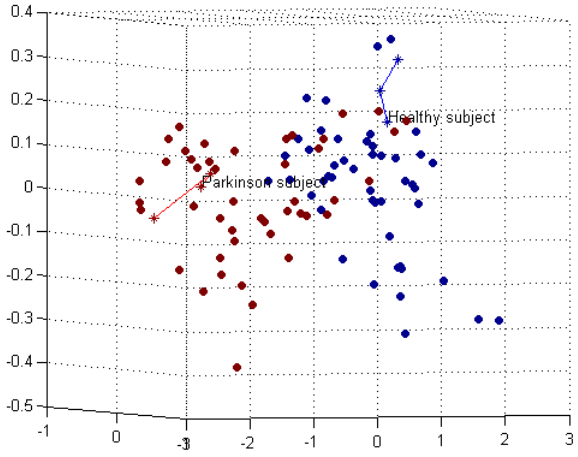


Figure 7. Projection of PRGD (RED) and HEALTHY (BLUE dots) Subjects at Eigengaits Space. RED and BLUE lines indicates the track of the same subject over time

is towards data of individuals who do not have the disease, we assume that there was an improvement in the individual state. If it is towards individuals diagnosed with PD, we assume that the model indicates worsening of symptoms.

The calculated distance is the similarity measure and can be expressed by the distance $D$ colon. However, the Euclidean distance, as estimated from the original variables, is influenced by: the scale of measurement, number of variables and the correlation between them. Therefore, standardization of data is necessary to maintain the same variance in the data and be able to identify the similarities [6]. In this research, the results were evaluated by means of the Euclidean distance between the vector of a test subject designed in the *eigenspace*, compared to all individuals in the training group. So, if the test subject is closer to an individual diagnosed with PD, he/she is classified as parkinsonian. On the other hand, if the test subject is closer to an individual without the diagnosis of PD, he/she is classified as a control subject.

### 7.1. Cross-Validation Results

To verify the accuracy of the classification of healthy versus PD, we performed Cross-validation [9]. Cross-validation is a technique used to assess the generalization of a prediction model in a data set. At first, it holds the partition data set into mutually exclusive subsets and then this subset is used to verify the accuracy of the search model.

To prevent over-fitting, accuracy is measured using a leave-one-out cross-validation. We hold a subject as the test data and the remainder group we used as training. The choice of the training group includes all classes in the database [9]. This technique is widely used to mitigate the occurrence of bias in research, because the training process is repeated several times with different samples including different cases in each class. Given a single sample, a method of prediction error rate in machine learning is to use a leave-one-out cross-validation [9].

To test the accuracy of the PCA method with cross-validation, we applied it to classify the data in the database, resulting in 81.00% of accuracy.

The classification performance obtained in this study is presented in the confusion matrix for two classes which consists of a matrix 2 x 2 with (TP, FP, TN and FN) presented in Table 1. A true positive (TP) indicates correctly classified with abnormal movement, true negative (TN) indicates correctly classified with normal movement.

Accordingly, the false positives (FP) indicate the normal movement classified as abnormal ones and the false negatives (FN) indicate the real abnormal movement not correctly detected. Table 2 shows the classification performance, where the *TpRate* is TP divided by the total number of positives, which is TP + FN; the *FpRate* is FP divided by the total number of negatives, which is FP + TN. The *Accuracy* is the number of correct classifications divided by the total number of samples [9].

TABLE 1. CONFUSION MATRIX OF PCA CLASSIFICATION LEAVE-ONE-OUT CROSS VALIDATION

| | ***Predicted Class*** | |
|---|---|---|
| | **Parkinson** | **Control Group** |
| **Parkinson** | 43 | 7 |
| **Control Group** | 12 | 38 |

TABLE 2. PERFORMANCE OF PD'S AND CONTROL GROUP CLASSIFICATION

| **Classifier Metrics** | |
|---|---|
| **TpRate** | 86.00% |
| **FpRate** | 24.00% |
| **Precision** | 78.18% |
| **Accuracy** | 81.00% |
| **F-Score** | 81.90% |

## 8. Conclusion

In this work, a gait analysis approach to track Parkinson's Disease evolution by monitoring walking abnormalities. This approach was evaluated by assessing the biomechanical data measurements from a public database [15]. We monitored PD gait symptoms extracting each gait cycle and calculating their respective Principal Components.

To evaluate our approach, we performed an experimental study with 100 research subjects divided into PD and Control groups. We used PCA and the euclidean distance to identify PD's gait characteristics and had an accuracy of 81.00% and 86% of *TpRate*.

Our study's limitation is that we do not consider the variation of gait movement between different people. Therefore, it is not a conclusive test to diagnose PD [4], [18]. Despite this limitation, it is accurate to track PD's symptoms based on the user gait.

For future works, we plan to use another public database hosted on Physionet [15] including different diseases as:

Huntington's Disease, Amyotrophic Lateral Sclerosis and Stroke patients [8] to perform multi-class classification.

## References


[1] F. Casamassima, A. Ferrari, B. Milosevic, P. Ginis, E. Farella, and L. Rocchi, "A wearable system for gait training in subjects with parkinson's disease," *Sensors*, vol. 14, no. 4, 2014. [Online]. Available: http://www.mdpi.com/1424-8220/14/4/6229

[2] D. A. Neumann, *Kinesiology of the musculoskeletal system: foundations for physical rehabilitation*, 2nd ed. St. Louis: Mosby, 2009.

[3] WHO, *Neurological Disorders: Public Health Challenges*, ser. Non-serial Publication. World Health Organization, 2006.

[4] W. Tao, T. Liu, R. Zheng, and H. Feng, "Gait analysis using wearable sensors," *Sensors*, vol. 12, no. 2, 2012.

[5] M. Yang, H. Zheng, H. Wang, S. McClean, and D. Newell, "igait: An interactive accelerometer based gait analysis system," *Computer Methods and Programs in Biomedicine*, vol. 108, no. 2, pp. 715 – 723, 2012.

[6] U. Dillmann, C. Holzhoffer, Y. Johann, S. Bechtel, S. Graber, C. Massing, J. Spiegel, S. Behnke, J. Brmann, and A. K. Louis, "Principal component analysis of gait in parkinson's disease: Relevance of gait velocity," *Gait & Posture*, vol. 39, 2014.

[7] S. Mazilu, U. Blanke, M. Dorfman, E. Gazit, A. Mirelman, J. M. Hausdorff, and G. Tröster, "A wearable assistant for gait training for parkinson's disease with freezing of gait in out-of-the-lab environments," *ACM Transactions on Interactive Intelligent Systems*, vol. 5, 2015.

[8] S. E. Rodrigo, C. N. A. Lescano, and R. H. Rodrigo, "Application of kohonen maps to kinetic analysis of human gait," *Revista Brasileira de Engenharia Biomedica*, vol. 28, pp. 217 – 226, 09 2012.

[9] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011.

[10] P. Vandewalle, J. Kovacevic, and M. Vetterli, "Reproducible research in signal processing," *Signal Processing Magazine, IEEE*, vol. 26, no. 3, pp. 37–47, May 2009.

[11] D. L. Donoho, "An invitation to reproducible computational research," *Biostatistics*, vol. 11, no. 3, pp. 385–388, 2010. [Online]. Available: http://biostatistics.oxfordjournals.org/content/11/3/385.short

[12] S. Frenkel-Toledo, N. Giladi, C. Peretz, T. Herman, L. Gruendlinger, and J. M. Hausdorff, "Effect of gait speed on gait rhythmicity in parkinson's disease: variability of stride time and swing time respond differently." *J Neuroengineering Rehabil*, vol. 2, Jul. 2005.

[13] C.-W. Cho, W.-H. Chao, S.-H. Lin, and Y.-Y. Chen, "A vision-based analysis system for gait recognition in patients with parkinson's disease," *Expert Systems with Applications*, vol. 36, pp. 7033–7039, 2009.

[14] D. Zwartjes, T. Heida, J. van Vugt, J. Geelen, and P. Veltink, "Ambulatory monitoring of activities and motor symptoms in parkinson's disease," *Biomedical Engineering, IEEE Transactions on*, vol. 57, no. 11, 2010.

[15] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, 2000.

[16] M. Mustafa, G. Georgoulas, and G. Nikolakopoulos, "Principal component analysis anomaly detector for rotor broken bars," in *IEEE Annual Conference of the Industrial Electronics Society*, 2014.


8. http://www.physionet.org/physiobank/database/gaitndd/


[17] J. Shlens, "A tutorial on principal component analysis," in *Systems Neurobiology Laboratory, Salk Institute for Biological Studies*, 2005.

[18] N. I. for Health and C. E. G. Britain, *Parkinson's Disease: Diagnosis and Management in Primary and Secondary Care*, ser. NICE clinical guideline. National Institute for Health and Clinical Excellence, 2006.