

Machine Learning and Similarity Network Approaches to Support Automatic Classification of Parkinson's Diseases Using Accelerometer-based Gait Analysis

Elham Rastegari

University of Nebraska at Omaha

erastegari@unomaha.edu

Sasan Azizian

University of Nebraska at Omaha

sazizian@unomaha.edu

Hesham Ali

University of Nebraska at Omaha

hali@unomaha.edu

Abstract

Parkinson's Disease is a worldwide health problem, causing movement disorder and gait deficiencies. Automatic noninvasive techniques for Parkinson's disease diagnosis is appreciated by patients, clinicians and neuroscientists. Gait offers many advantages compared to other biometrics specifically when data is collected using wearable devices; data collection can be performed through inexpensive technologies, remotely, and continuously. In this study, a new set of gait features associated with Parkinson's Disease are introduced and extracted from accelerometer data. Then, we used a feature selection technique called maximum information gain minimum correlation (MIGMC). Using MIGMC, features are first reduced based on Information Gain method and then through Pearson correlation analysis and Tukey post-hoc multiple comparison test. The ability of several machine learning methods, including Support Vector Machine, Random Forest, AdaBoost, Bagging, and Naïve Bayes are investigated across different feature sets. Similarity Network analysis is also performed to validate our optimal feature set obtained using MIGMC technique. The effect of feature standardization is also investigated. Results indicates that standardization could improve all classifiers' performance. In addition, the feature set obtained using MIGMC provided the highest classification performance. It is shown that our results from Similarity Network analysis are consistent with our results from the classification task, emphasizing on the importance of choosing an optimal set of gait features to help objective assessment and automatic diagnosis of Parkinson's disease. Results illustrate that ensemble methods and specifically boosting classifiers had better performances than other classifiers. In summary, our preliminary results support the potential benefit of accelerometers as an objective tool for diagnostic purposes in PD.

Keywords

Gait Parameters, feature selection, machine learning, Parkinson's Disease, Geriatrics

1. Introduction

Parkinson's Disease (PD) is a progressive neurodegenerative disorder of the brain and central nervous system, affecting patients' motor system, causing reduced movement, tremor, postural instability and postural rigidity [1]. Fatigue, Small shuffling steps, freezing of gait, and bradykinesia are some of the symptoms of PD [2]. In spite of advances in medical care, gait disturbances have been shown to worsen as PD advances, which in turn leads to loss of independence and lower patients' quality of life [3], [4].

There is no reliable test that can discriminate between PD and other conditions with similar clinical symptoms. Clinical diagnosis is currently based on current test results and patients' history and there is still a need for an alternate diagnosis technique, which can offer simple, quick and non-invasive measurement. Comparing gait patterns in PD patients with their healthy counterparts through machine learning approaches can potentially help healthcare providers to diagnose PD, quantify the progression of disease, and therefore help millions of people suffering from PD around the world. Gait analysis by itself may not be a solution to PD. However, gait features could provide information that are complementary to other sources of information.

Walking is the most widely recognized form of human movement. Gait analysis involves estimation and evaluation of biomechanical features associated with walking. Gait analysis usually requires force platforms besides optical motion analysis systems consisting of body markers and video cameras. Although these traditional movement analysis systems

have been used for more than two decades [5] and provided accurate information about movement, they require expensive facilities and patients' frequent visit to doctors' office or locomotion laboratories. Thus, their use in clinical practice is limited by these factors. Moreover, patients wish to boost their quality of life while reduced number of clinic visits are desired. An alternative locomotion analysis approach, based on wearable monitoring devices such as inertial sensors has been shown to have the potential of being used as a quantitative method in clinical practice [6], [7]. Even though inertial sensors are not still routinely utilized for diagnosis of PD or treatment assessment purposes, they have already been used to investigate motor complications in PD [10], [11], [23], [24].

The main contribution of this study is to identify an optimal set of gait features extracted only from accelerometers along with the best classifiers that can help to diagnose PD in the early stages. Although some of the features are already introduced in the literature, we created a new feature set only using accelerometer data that can lead to high classification accuracy. We will identify patients with mild PD from healthy individuals and geriatrics who present similar gait deficiencies (e.g., asymmetric walking pattern) using machine learning classifiers and a Similarity Network model. Although gait disorders are common in elderly population, their prevalence increases with age. 85% of people have a normal gait at the age of 60 while this proportion drops to 18% at the age of 85 [8]. A geriatric, around or over the age of 80, whose gait deficiencies are due to aging should not be confused with a PD patient by any machine learning or Similarity Network model because of presenting similar gait patterns. No study has considered classification of these three population only using acceleration data to the extent of our knowledge. We will also contribute by introducing Maximum Information Gain Minimum Correlation (MCMIG) feature selection approach that will be proved to increase the classification performance. We wish to answer several questions in our research:

- Can accelerometer-based gait analysis provide the potential biomarkers of PD?
- Is it possible to use various machine learning algorithm and Similarity Network model together with accelerometer-based gait features to help in early detection of PD?
- How important is it to choose the optimal set of gait features in building diagnostic and predictive models for PD?

- Can MIGMC feature selection technique help in identifying optimal set of gait features?

The rest of this paper is structured as follow: First, we provide a background about automatic diagnosis of PD using machine learning techniques. Then, we explain our dataset and research method. Furthermore, results will be provided and discussed along with future improvements.

2. Background

Machine learning and data mining techniques have become an inevitable part of modern life. They are being widely used in the biomedical science and healthcare domain with the goal of early diagnosis, providing prognosis and understanding the classification of disease [9]–[11]. Machine learning techniques have been extensively used to automatically discriminate between healthy individuals and those with PD. This could be a step toward early diagnosis if patients are in the early stages of the disease. Based on a survey study done by Bind et al. [12], many studies used speech and voice signal [13]–[16], several studies have used magnetic resonance imaging [17], [18], some have considered genomics data [19], [20], and few of them have utilized gait (movement) patterns [21] for classifying individuals into healthy and PD groups.

Analysis of gait and posture is one of the components of the clinical assessment of PD. With the healthcare moving rapidly from the long-standing reactive treatment approach to the early detection and preventative era, using wearable devices together with machine learning approaches in the healthcare domain has rapidly increased.

Machine learning techniques together with wearable devices (e.g., accelerometers) have been used in many PD studies to recognize activities of daily living [22], identify diseases severity level by predicting UPDRS score [23], or predict FOG events, tremor, dyskinesia or bradykinesia [9], [24]–[26]. However, there are very few studies that have used wearable devices to discriminate between gait patterns of PD and healthy individuals [24], [27]–[30]. In the study by Patel et al. a system was used to analyze lower and upper extremities and no distinction was found between subjects from PD and control groups in this study [24]. However, they used eight accelerometers to rate symptoms associated with PD like tremor, bradykinesia, dyskinesia. Deepak Joshi et al. performed wavelet decomposition of spatiotemporal gait variables as an alternate technique to identify patients with PD from healthy adults.

Although they achieved the classification accuracy of 90.32%, PD patients and control subjects were not age-matched in their study which can significantly impact the results. It is not clear whether the high classification accuracy achieved by Deepak et al. is due to the gait alterations caused by ageing or the disease itself. Barth et al. has used pressure sensors along with inertial sensors [29], and Klucken et al. and Barth et al. used inertial sensors including accelerometers and gyroscopes for classifying individuals into healthy and PD groups [27], [28]. While both studies reached a high overall classification rate, developed models require data from both accelerometer and gyroscope [27], [28], and a combination of tasks, including a 10-meter walk, heel-toe tapping, and circling, is required to be performed by each individual [28].

Considering continuous monitoring of gait, which is a necessary step toward real-life applications of gait in the context of smart homes and healthcare domain [31]–[33], including more data in analysis (e.g., from accelerometer and gyroscope) leads to a higher dimensionality of feature space and consequently more computationally intensive process. If we can show that the same task can be completed only using accelerometers' data (without a need for data collected from gyroscopes), without compromising the accuracy, and with a smaller set of features, it would help saving memory and computational resources. Moreover, considering only individuals' walking patterns, instead of extracting and analyzing features associated with different tasks such as circling and tapping performed in the study by Kluchen et al. [28], will reduce the complexity of monitoring individuals seamlessly and continuously over time. This study aims at discriminating patients with mild PD from healthy individuals (same age range with PD patients and older adults in their 80s) only using accelerometers' data.

3. Material and method

Gait analysis in the healthcare domain with the aim of identifying groups with pathology in the early stages (early diagnosis of disease) can be structured in the following steps: data acquisition, data reprocessing, feature extraction, feature selection/reduction, and modelling. These steps can be seen in Figure 1.

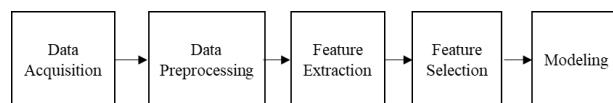


Figure 1. Gait analysis for early diagnosis

3.1. Signal acquisition and pre-processing

Gait analysis in the healthcare domain can be structured in the following steps: data acquisition, data reprocessing, feature engineering, feature selection/reduction, exploratory analysis, and modelling. These steps can be seen in Figure 1.

We used accelerometers-derived data from a publicly available data set collected by Barth and colleagues [34]. This data set includes three groups of people: healthy elderly, geriatrics, and people with mild PD. For the present study, we selected the data associated with a 40-meter walk protocol. In the 40-meter walk experiment, data was sampled at 102.4 HZ and collected from subjects' left and right ankles using SHIMMER while subjects walked 10 meters four times at their comfortable speed and in an obstacle-free environment. Subjects were instructed to turn in the transverse plane after each 10 meters. SHIMMER is a validated inertial sensing platform, including accelerometer and gyroscope [35]. X, Y, and Z axes of accelerometer represent signals in Anteroposterior (AP), Vertical, and Mediolateral (ML) directions, respectively. Figure 2 shows the directions of accelerometer axes according to the sensor placement. Table 1 shows Subjects' information.



Figure 2. The directions of accelerometer axes according to the sensor placement [34]

Table 1. Subjects' Characteristics

	Control	PD	Geriatrics
Number of subjects	10	10	10
Gender (M/F)	5:5	5:5	4:6
Age	64 ± 8.4	63.8 ± 9.3	81 ± 4.1
UPDRS III		12.7 ± 6.0	
H & Y		1.7 ± 0.9	

Prior to extracting gait parameters, we calibrated the data using the guideline provided by Barth et al [34]. Then, we used z-axis of gyroscope for stride segmentation. we removed each subject's first and last two strides to eliminate any irregularities associated

with the initiation and termination of gait [36]. We also removed strides associated with the beginning and end of each 10 meters.

3.2. Gait Features

The following features were extracted from the accelerometer signal. These features represent either the characteristics of strides (stride level features) or the characteristics of complete gait sequence (signal level features).

Average stride time. Mean of stride time has been used in evaluation of age-related gait differences [37] as well as disease-related gait differences [38]. Average stride time is considered as one of the features for the classification task.

RMS of Acceleration/Body Oscillation. Root mean square (RMS) of acceleration is a statistical measure of the magnitude of acceleration that has been frequently reported in gait research [39], [40]. RMS of acceleration has been used to quantify the attenuations of accelerations [41]. Normalized RMS of acceleration has been shown to provide results that are more consistent across gait analysis research and is referred to as body oscillation in either AP, ML or vertical direction. Normalized RMS of acceleration in the ML direction has been shown to be significantly higher in people with movement disorder than healthy subjects [40]. We calculated both RMS and normalized-RMS of acceleration in AP, ML, and vertical direction and added them to our original set of features. We calculated normalized RMS of acceleration using equation (1), in which *dir* represents the direction of the acceleration signal.

$$RMSR_{dir} = \frac{RMS_{dir}}{\sqrt{RMS_{AP}^2 + RMS_{ML}^2 + RMS_{Ver}^2}} \quad (1)$$

Maximum and Minimum Acceleration. Maximum and minimum acceleration of each stride for AP, ML, and vertical direction is averaged over all strides. It resulted in six gait features that are included in the original feature set. These features were calculated and included because they are shown to be significantly different between patients with PD, geriatrics, and healthy elderlies [28], [42]. The difference between the obtained maximum and minimum acceleration value in each direction was also included in our original feature set.

Variability of Signal per Stride. To have more details about strides' acceleration signals, we calculated variability of acceleration signal per stride for all three directions. Then, we took the average of all standard deviation values as seen in equation (2) in

which *s* is the number of strides, *d_s* is the number of data points per stride, *x_i* is the value of acceleration associated with *i*th data point, and *μ_s* is the average of acceleration value per stride. This feature was calculated for AP, ML and vertical acceleration signals, separately.

$$\frac{1}{s} \sum_{s=1}^{Ns} \sqrt{\frac{1}{d_s} \sum_{i=1}^{d_s} (x_i - \mu_s)^2} \quad (2)$$

Signal Vector Magnitude (VM). VM represents intensity or magnitude of acceleration in all directions. It is reported as one of the metrics for continuous gait monitoring using accelerometers [43]. Therefore, it was included in our original feature set.

Symmetry. Although several studies [44], [45], including our previous study [45] have reported that gait asymmetry is not significantly present in patients with PD regardless of the body site the accelerometers attached to, and of the disease severity level, there are other studies reported asymmetry as a significant gait parameter in patients with PD as well as elderlies [46]. To measure symmetry, we considered stride time, VM, and RMS of acceleration in all three directions and calculated the differences between the value of each gait parameter associated with the left ankle and the value of the same gait parameter associated with the right ankle as it is calculated in our earlier studies [42], [45].

Stride to stride variability. Gait variability, defined as stride-to-stride fluctuation, has been shown to be sensitive to aging and pathology [38], [47]. There are evidences for increased gait variability in patients with PD even in the early stages of the disease [48]. In our study, variability was calculated as the within-person coefficient variation (CV) of various features across strides considering both left and right sides. Stride-to-stride variability was calculated regarding stride time, VM, and RMS of acceleration in all three directions.

Velocity. Gait speed has been shown to be significantly slower in patients with PD than that of healthy older adults [39]. Although we did not have the exact speed information for the subjects, we estimated speed value for each subject indirectly using equation (3), in which 102.4 HZ is the frequency of data sampling, 40 is the distance all subjects walked, *N_i left* and *N_i right* are the number of data points from left and right ankles for the *i*th subject. Since we had strides' data from left and right ankle, and the number of data points collected for the left and right ankle might have been slightly different, we then took

the average of speed values associated with the left and right ankles.

$$Speed_i = \text{mean}\left(\frac{102.4*40}{N_{i_left}} + \frac{102.4*40}{N_{i_right}}\right) \quad (3)$$

Signal Smoothness. The smooth movement of a human body is achieved by balancing the forces of several muscles. Balance disorders are common in patients with PD [49] and in geriatrics [50]. If balance cannot be upheld, jerky motions are likely to happen during walking. To account for smoothness of walking signal, we used equation (4), in which d_{sig} represents the number of data points in the signal, and AC_i and AC_{i+1} are the acceleration value of two consecutive data points i and $i+1$. We calculated this feature for AP, ML and vertical acceleration signals, separately.

$$\frac{1}{d_{sig}} \sum_{i=1}^{d_{sig}} \text{abs}(AC_i - AC_{i+1}) \quad (4)$$

Standardization has been shown to provide great improvement in gait classification of patients with PD and healthy individuals [51]. Once we calculated all features, we standardized each gait feature using equation (5), in which xi is the value of feature x for the i th subject and $X_{\max}(\text{inter class})$ is the highest value of feature x amongst all three groups.

$$X_{\text{norm}} = \frac{xi}{X_{\max}(\text{inter class})} \quad (5)$$

3.3. MIGMC Feature Selection

Too many gait parameters have been used over the years. Selecting the most appropriate feature set is a crucial step prior to applying classification techniques, often affecting the model accuracy and consequently the success of the research outcomes. Incorporating additional features is sometimes costly and may even be against the goal of achieving optimality [52].

We performed a two-step approach called Maximum Information Gain Minimum Correlation to select an appropriate feature set. Our approach is based on the minimal-redundancy-maximal-relevance (MRMR) concept proposed by Peng et al. [53]. However, we did not apply the same procedure they proposed. First, we investigated the most influential features using information gain algorithm (IG) in WEKA. Information gain is a feature selection method used in gait analysis studies. It works by measuring the decrease in entropy in the presence and absence of the feature [28]. We used information gain evaluation on the feature set combined with ranker method to extract and rank the most influential features in classifying individuals into three associated groups. Then, we

took a step further and came up with a feature set in which features are maximally dissimilar to each other. This step was done by pair-wise Pearson correlation analysis. The threshold of 80% was considered to determine highly correlated features. Among all pairwise correlations, we picked one with the highest value and its associated features. If we had more than one pair of values with the highest correlation value, we randomly chose one. In order to decide which one of the features in the selected pair of features to remain in our feature set, we selected the one with higher discriminating power. To choose this feature, we performed ANOVA followed by Tukey post-hoc multiple comparison test. The feature with the highest absolute value of groups mean differences was selected. If both features showed the same discriminating power, then we calculated sum of the pairwise correlations each feature had with other features. The one with the lower value of accumulated correlation was selected to remain in our feature set. We repeated this process until we did not have any pairwise correlation value higher than 80%. Figure 3 illustrates MIGMC feature selection technique in details.

3.4. Classification models

Support Vector Machine (SVM) with non-linear kernel, naïve Bayes classifier, Random Forest, Bagging, and AdaBoost have been shown to be powerful supervised learning techniques for sensor-based gait classification [28], [29], [54]. We used Naïve Bayes, SVM, Random Forest, Bagging, and AdaBoost classifiers using a repeated 5-fold cross validation method in our analysis and compared the accuracy of each model across different feature sets and across standardized and non-standardized vectors of feature. Each classifier's accuracy was calculated using the measures of accuracy, precision, and recall.

3.5. Similarity Network

A similarity network for classification of individuals based on gait patterns is proposed by Rastegari et al., [42] in which correlation value between each pair of individuals is considered as the measure of similarity. In this network, vertices represent individuals and there is an edge between two individuals if the correlation value between their corresponding gait vectors is equal to or greater than a predefined threshold. Correlation analysis considers all of the selected parameters together and examine how similar two subjects are regarding their gait patterns. We performed correlation network analysis with the threshold value of 85% as an alternative and validation approach to show the effect of our proposed

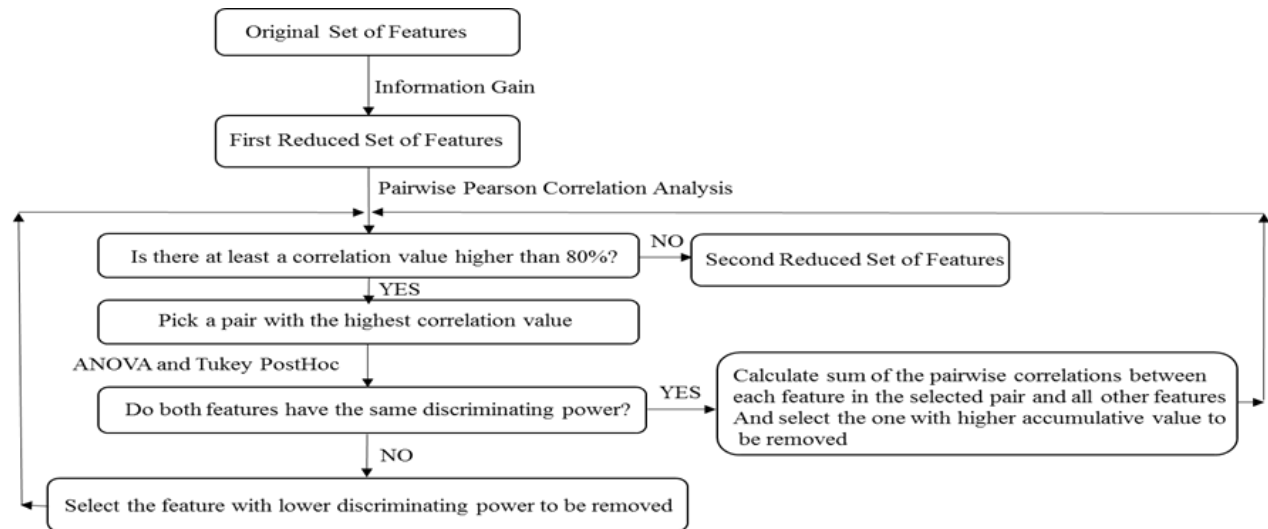


Figure 3. MIGMC feature selection technique

Table 3. Optimal set of features obtained using MIGMC

Feature's Name	Description	Feature category
Variability_StrideTime	Variability of stride time	Signal level
Variability_VM	Variability of vector magnitude	Signal level
Variability_RMSX	Variability of root mean square in the AP direction	Signal level
Variability_RMSZ	Variability of root mean square in the ML direction	Signal level
Velocity	Velocity	Signal level
Smoothness_X	Smoothness in the AP direction	Signal level
Smoothness_Z	Smoothness in the ML direction	Signal level
RMSZR	Root mean square relative to the mean value in the ML direction	Stride level

feature selection technique as well as feature standardization.

4. Analysis and Results

Before performing the classification, we have ranked all the features according to their information gain to obtain a set of features that are potentially significant in classifying three populations. We had 32 features in our original feature set, out of them, 22 were identified as the most influential features using the information gain method.

Then, we reduced the first reduced feature set to a set of features that are maximally dissimilar and obtained by applying Pearson correlation analysis and Post-hoc multiple comparison test. Table 3 shows the second reduced set of features along with a brief description of each feature. As it can be seen from Table 3, all selected features except RMSZ are signal level features, indicating that investigating the complete gait sequence can reveal more

discriminating information compared to stride level feature.

To make sure that the second reduced set of features has the minimum number of features contributing to the highest performance of the classification task, we tried to remove only one feature at a time, run all classification techniques using 7 remaining features and see whether we get at least the same accuracy we could get using 8 features identified by our feature selection technique. The last row in both Table 4 and Table 5 shows the results of classification when smoothness in the AP direction was removed from the feature set. Although Adaboost and Random Forest classifiers' performance stayed the same in some cases, overall, results indicated that for any set of features with less than the eight identified features, most of the classifiers could not perform at their best. Therefore, as it can be seen in Table 4, the second reduced set of features can be considered as the optimal feature set.

Table 4 and Table 5 show the performance of all classifiers regarding accuracy, precision, and recall

based on both standardized and non-standardized feature sets, and across different sets of features. One of the obvious observations is that standardization contributed to a higher accuracy. As it can be seen in both Table 4 and Table 5, AdaBoost showed the best

performance regardless of the feature set. Using eight features (optimal set) selected by MIGMC provided equal accuracy or even higher accuracy compared to other sets of features.

Table 4. Classification's performance based on standardized feature vectors- * Reduced Feature Set

Features	SVM			Random forest			AdaBoost			Bagging			Naïve Bayes		
	Acc	Pre	Rec	Acc	Pre	Rec	Acc	Pre	Rec	Acc	Pre	Rec	Acc	Pre	Rec
All (32)	83.6	85.6	83.3	86.6	86.9	86.7	96.7	97.0	96.7	96.7	97.0	96.7	90.0	92.3	90.0
RFS*1 (22)	80.4	81.5	80.0	93.3	93.6	93.3	100	100	100	96.7	97.0	96.7	90.0	92.3	90.0
RFS2 (8)	83.6	85.6	83.3	100	100	100	100	100	100	96.7	97.0	96.7	90.0	92.3	90.0
RFS3 (7)	69.9	81.1	73.3	83.3	83.5	83.3	73.3	74.7	73.3	70.5	71.1	70.0	83.1	83.7	83.3

Table 5. Classifications' performance based on non-standardized feature vectors

Features	SVM			Random forest			AdaBoost			Bagging			Naïve Bayes		
	Acc	Pre	Rec	Acc	Pre	Rec	Acc	Pre	Rec	Acc	Pre	Rec	Acc	Pre	Rec
All (32)	83.6	85.6	83.3	80.4	81.5	80.0	83.6	84.2	83.3	68.4	70.0	70.0	67.7	71.0	66.7
RFS*1 (22)	90.2	90.6	90.0	80.4	81.0	80.0	83.6	84.2	83.3	75.5	78.1	76.7	67.7	71.0	66.7
RFS2 (8)	93.3	93.6	93.3	80.4	81.0	80.0	83.6	84.2	83.3	79.8	79.9	80.0	70.5	71.1	70.0
RFS3 (7)	79.9	80.6	80.0	80.4	81.0	80.0	73.3	74.7	73.3	79.8	79.9	80.0	65.4	70.0	66.7

Results of Similarity Network Analysis is illustrated in Figure 4 and Figure 5. Red vertices represent healthy subjects, blue ones represent geriatrics, and green vertices are associated with PD patients. Figure 4 shows the results of Similarity Network analysis using all 32 identified features with and without feature standardization. We can see that in the case of modeling based on the standardized features, grouping is done far way better than the model made without feature standardization. This is consistent with the results from our classification with and without feature standardization.

In Figure 5, we can see three network models associated with the first, second, and third reduced feature sets including twenty-two, eight, and seven features, respectively. The network model using all 32

features is depicted in Figure 4-a. Comparing four network models indicates that using the first and second reduced sets of features improved the Similarity Network model's performance by increasing the discriminative power of the feature set. Thus, fewer number of healthy subjects are grouped together with PD patients in the network depicted in Figure 5-b compared to Figure 5-a and Figure 4-a networks. On the other hand, reducing the number of features further below the identified optimal set of features (second reduced set of features) reduced the performance of Similarity Network model as it is illustrated in Figure 5-c. In this figure, most of geriatrics (in blue) are still grouped together and differentiated from two other groups, indicating that their gait pattern is significantly different from others which is very likely due to the age difference.

However, there is no obvious differentiation between healthy subjects and patients with PD. Results of Similarity Network model using different feature sets is consistent with our classification results, validating

our optimal set of features as well as MIGMC feature selection technique. These results emphasize on the importance of choosing an optimal set of features for classification of individuals using their gait patterns.

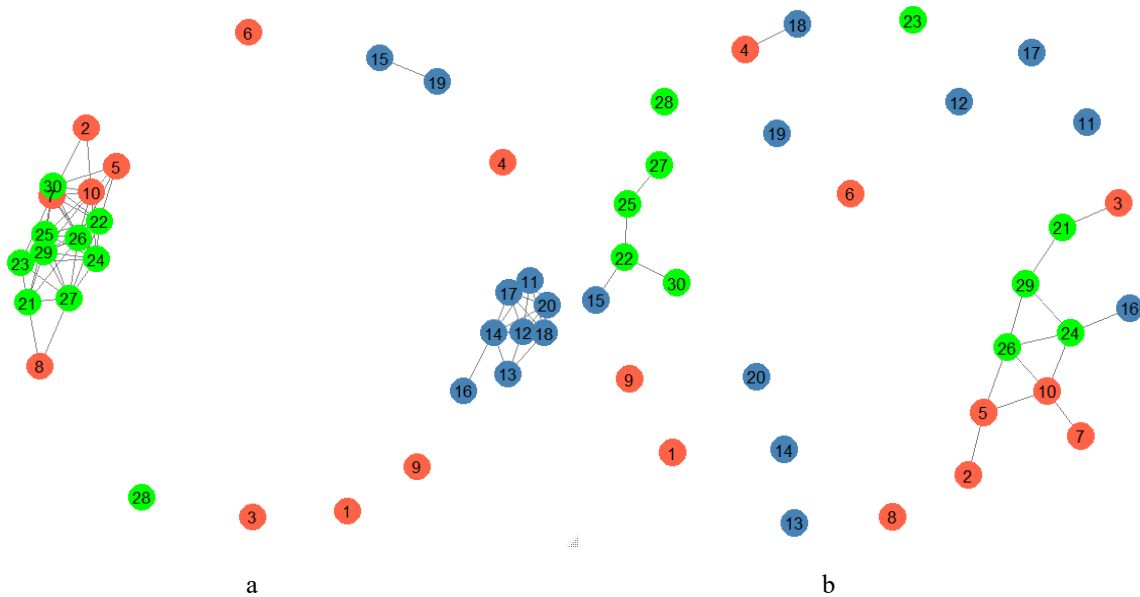


Figure 4. Similarity Network Model with (a) and without (b) feature standardization

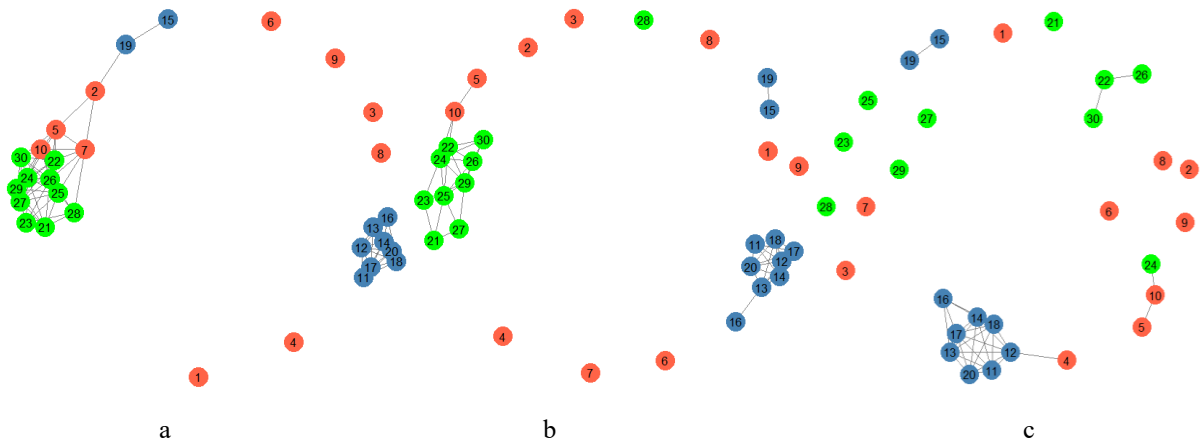


Figure 5. Similarity Network Model using first (a), second (b) and third reduced sets of features(c)

5. Discussion and Limitations

In this study, a set of gait features, several classifiers and a Similarity Network model were used to develop an automatic gait analysis system for diagnosis of PD. The developed model was able to discriminate between patients with PD, healthy elderly, and geriatrics using gait features extracted

from two accelerometers placed on both ankles. Geriatrics were considered in analysis since aging has significant impact on gait and similar gait deficiencies (e.g., variability, asymmetry) are reported in the literature for both patients with PD and geriatrics. To avoid the confusion of machine learning techniques and Similarity Network analysis in automatic diagnosis of PD, including other populations with

similar gait deficiencies in analysis seems necessary which is missed in the literature. A geriatric who is suffering from movement disorder should be classified neither in healthy nor in PD groups. Our results show that the developed model could be an objective tool for assessment of gait alterations in PD patients and early diagnosis of the disease without any need for extra data sources such as gyroscopes.

It is necessary to select an optimized feature set because a high dimension vector of features requires high computational cost and increases the risk of overfitting. To select an optimal set of features, we performed MIGMC feature selection technique. Although this feature selection technique and MRMR [53] share the same concept, we implemented MIGMC feature selection technique differently using Information Gain method, pairwise Pearson correlation analysis, ANOVA, and Tukey post-hoc.

Obtained results showed that standardization increased the performance of almost all classifiers and the Similarity Network model. Comparison of various feature sets revealed that the optimal feature set outperforms its counterparts. AdaBoost classifier showed the overall best performance, which is in agreement with the results of two studies by Barth et al. and Klucken et al. on the classification of PD patients and healthy individuals using gait parameters [27], [28]. Both Adaboost and Random Forest classifiers performed their best when we considered the optimal set of features obtained using MIGMC feature reduction technique. In the study by Barth et al., data from both accelerometers and gyroscopes were analyzed and 12 number of features were identified and contributed to the highest accuracy value of the classification model [27]. Kluchen et al. extracted 694 features associated with a combination of tasks performed by each subject, out of them 23 features were selected and contributed to the highest accuracy value of the classification [28]. Comparing our findings with the findings from these two studies illustrates that our classification models and the Similarity Network Model for diagnosis of PD outperform models developed by Barth et al. and Kluchen et al while a smaller feature set, including only 8 features, was utilized in our study. This set of features was extracted only from the accelerometer data and they were associated with a single task of walking, while Kluche et al. and Barth et al. both employed a multi-sensor system, including gyroscope and accelerometer. Furthermore, classification of subjects in the study by Kluchen et al. was based on features associated with a combination of tasks performed by subjects. Dealing with a lower number of data sources as well as a smaller set of features,

would lead to a less complex and computationally intensive task of continuous gait monitoring.

Comparing classifiers used in this study, we observed that ensemble methods and specifically boosting classifiers have better performances than other classifiers. These classifiers could distinguish not only between patients with PD and their healthy counterparts, but also between these two groups and geriatrics. This indicates that although gait analysis by itself may not be a solution to PD, gait features could provide information that are complementary to other sources of information.

Eight features identified by MIGMC technique provided the best classification performance which is consistent with our results obtained from Similarity Network analysis. Among these features, four of them represent variability regarding various gait parameters, including stride time, signal vector magnitude, and root mean square in both AP and ML directions. The remaining four features are velocity, body oscillation in ML direction, and smoothness in both AP and ML direction. The fact that four measures of variability are included in our optimal set of features illustrates that different measures of variability are not necessarily highly correlated and each of them can play a distinct role in discriminating groups under the study. All features in our optimal set of features except RMSZR are signal-level features, indicating that signal-level features play a more important role in discriminating different populations based on gait parameters.

Considering that our study was conducted on a small-size dataset, the danger of overfitting while training the classifiers is not deniable. However, Similarity Network Model is not prone to overfitting since it is not made using any training or test data but using a similarity measure between each pair of subjects. Similarity Network Models made using various feature sets confirm the fact that the optimal set of features introduced in our study contributes to creating models with a higher accuracy values compared to other feature sets introduced in our work and those introduced in the literature.

This study has obviously its own limitations. One of them is having a limited data set, which makes it impossible to draw a general conclusion about the results. However, this study is a step towards objective assessment and early diagnosis of Parkinson's disease which would benefit patients and healthcare providers. The other limitation of this study is the lack of having detailed information (e.g., UPDRS or H&Y score) about each individual. Having more detailed

information would have allowed us to have a more concise interpretation of our results such as investigating the classification errors with regards to age or UPDRS score. Although the results of classification and Similarity Network Model both confirm the validity of our proposed feature selection technique, one of the other limitations of this work is that we did not compare other feature selection techniques with MIGMC which remains the objective of our future works. Our future research plan is also to use another dataset including patients with more advanced PD to investigate the power of our approach in identifying various stages of PD. Moreover, using another dataset as an independent validation set might be useful to further confirm the results of this study and mitigate the effect of the small-size dataset.

6. Conclusion

The purpose of this study was to develop and validate an automated gait analysis system using lower-body motion data and pattern recognition algorithms to distinguish between three groups of people: healthy elderly, geriatrics, and patients with PD. Using a proposed feature selection technique based on the maximum information gain and minimum correlation among the features, an optimal set of gait features was obtained only from accelerometer data. The new set of features developed in our study together with machine learning techniques could distinguish PD patients from healthy elderly and geriatrics. A Similarity Network Model was used to validate the efficiency of the optimal set of gait features obtained using our proposed feature selection technique. In conclusion, our results support the potential benefit of accelerometers attached to the ankle as an objective tool for diagnostic purposes in PD.

References

- [1] S. D. Kim, N. E. Allen, C. G. Canning, and V. S. Fung, "Postural instability in patients with Parkinson's disease," *CNS Drugs*, vol. 27, no. 2, pp. 97–112, 2013.
- [2] J. Jankovic, "Parkinson's disease: clinical features and diagnosis," *J. Neurol. Neurosurg. Psychiatry*, vol. 79, no. 4, pp. 368–376, 2008.
- [3] K. R. Chaudhuri, D. G. Healy, and A. H. Schapira, "Non-motor symptoms of Parkinson's disease: diagnosis and management," *Lancet Neurol.*, vol. 5, no. 3, pp. 235–245, 2006.
- [4] J. M. Hausdorff, "Gait dynamics in Parkinson's disease: common and distinct behavior among stride length, gait variability, and fractal-like scaling," *Chaos Interdiscip. J. Nonlinear Sci.*, vol. 19, no. 2, p. 026113, 2009.
- [5] T. P. Andriacchi and E. J. Alexander, "Studies of human locomotion: past, present and future," *J. Biomech.*, vol. 33, no. 10, pp. 1217–1224, 2000.
- [6] A. Paraschiv-Ionescu, E. E. Buchser, B. Rutschmann, B. Najafi, and K. Aminian, "Ambulatory system for the quantitative and qualitative analysis of gait and posture in chronic pain patients treated with spinal cord stimulation," *Gait Posture*, vol. 20, no. 2, pp. 113–125, 2004.
- [7] A. Weiss, T. Herman, M. Plotnik, M. Brozgov, "Can an accelerometer enhance the utility of the Timed Up & Go Test when evaluating patients with Parkinson's disease?," *Med. Eng. Phys.*, vol. 32, no. 2, pp. 119–125, Mar. 2010.
- [8] B. R. Bloem, J. Haan, A. M. Lagaay, W. van Beek, A. R. Wintzen, and R. A. C. Roos, "Investigation of Gait in Elderly Subjects Over 88 Years of Age," *Top. Geriatr.*, vol. 5, no. 2, pp. 78–84, Apr. 1992.
- [9] Tripoliti, E.E., Tzallas, A.T., Tsipouras, M.G., Rigas, G., Bougia, P., Leontiou, M., Konitsiotis, S., Chondrogiorgi, M., Tsouli, S. and Fotiadis, D.I., "Comput. Methods Programs Biomed.", vol. 110, no. 1, pp. 12–26, 2013.
- [10] N. L. Keijsers, M. W. Horstink, and S. C. Gielen, "Automatic assessment of levodopa-induced dyskinesias in daily life by neural networks," *Mov. Disord.*, vol. 18, no. 1, pp. 70–80, 2003.
- [11] S. Joshi, D. Shenoy, P. L. Rrashmi, K. R. Venugopal, and L. M. Patnaik, "Classification of Alzheimer's disease and Parkinson's disease by using machine learning and neural network methods," in *Machine Learning and Computing (ICMLC)*, 2010 Second International Conference on, 2010, pp. 218–222.
- [12] S. Bind, AK Tiwari, AK Sahani, "A survey of machine learning based approaches for Parkinson disease prediction," *Int. J. Comput. Sci. Inf. Technol.*, vol. 6, no. 2, pp. 1648–1655, 2015.
- [13] I. Rustempasic and M. Can, "Diagnosis of parkinson's disease using fuzzy c-means clustering and pattern recognition," *Southeast Eur. J. Soft Comput.*, vol. 2, no. 1, 2013.
- [14] M. Shahbakhhi, D. T. Far, and E. Tahami, "Speech analysis for diagnosis of parkinson's disease using genetic algorithm and support vector machine," *J. Biomed. Sci. Eng.*, vol. 7, no. 4, pp. 147–156, 2014.
- [15] R. Das, "A comparison of multiple classification methods for diagnosis of Parkinson disease," *Expert Syst. Appl.*, vol. 37, no. 2, pp. 1568–1572, 2010.

- [16] A. Ozcift, "SVM feature selection based rotation forest ensemble classifiers to improve computer-aided diagnosis of Parkinson disease," *J. Med. Syst.*, vol. 36, no. 4, pp. 2141–2147, 2012.
- [17] H.-J. Huppertz, L. Möller, M. Südmeyer, "Differentiation of neurodegenerative parkinsonian syndromes by volumetric magnetic resonance imaging analysis and support vector machine classification," *Mov. Disord.*, vol. 31, no. 10, pp. 1506–1517, 2016.
- [18] G. Orru, W. Pettersson-Yeo, A. F. Marquand, G. Sartori, and A. Mechelli, "Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review," *Neurosci. Biobehav. Rev.*, vol. 36, no. 4, pp. 1140–1152, 2012.
- [19] K. Tejeswinee, G. J. Shomona, and R. Athilakshmi, "Feature Selection Techniques for Prediction of Neuro-Degenerative Disorders: A Case-Study with Alzheimer's And Parkinson's Disease," *Procedia Comput. Sci.*, vol. 115, pp. 188–194, 2017.
- [20] A. H. Chen, C. H. Lin, and C. H. Cheng, "New approaches to improve the performance of disease classification using nested-random forest and nested-support vector machine classifiers," *Res. Notes Inf. Sci. RNIS*, vol. 14, p. 102, 2013.
- [21] 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 Syst. Appl.*, vol. 36, no. 3, pp. 7033–7039, 2009.
- [22] M. V. Albert, S. Toledo, M. Shapiro, and K. Koerding, "Using mobile phones for activity recognition in Parkinson's patients," *Front. Neurol.*, vol. 3, p. 158, 2012.
- [23] S. Arora, V. Venkataraman, A. Zhan, S. Donohue, K.M. Biglan, E.R. Dorsey, "Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study," *Parkinsonism Relat. Disord.*, vol. 21, no. 6, pp. 650–653, Jun. 2015.
- [24] S. Patel, K. Lorincz, R. Hughes, N. Huggins, John Growdon, David Standaert, Metin Akay, Jennifer Dy, Math Welsh, Paolo Bonato, "Monitoring motor fluctuations in patients with Parkinson's disease using wearable sensors," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 6, pp. 864–873, 2009.
- [25] M. G. Tsipouras, A. T. Tzallas, G. Rigas, P. Bougia, D. I. Fotiadis, and S. Konitsiotis, "Automated Levodopa-induced dyskinesia assessment," in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, 2010, pp. 2411–2414.
- [26] G. Rigas et al., "Assessment of tremor activity in the Parkinson's disease using a set of wearable sensors," *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 3, pp. 478–487, 2012.
- [27] J. Barth, J. Klucken, P. Kugler, T. Kammerer, R. Steidl, J. Winkler, J. Horneegger, B. Eskofier, "Biometric and mobile gait analysis for early diagnosis and therapy monitoring in Parkinson's disease," in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, 2011, pp. 868–871.
- [28] J. Klucken, J. Barth, P. Kugler, J. Schlachetzki, T. Henze, F. Marxreiter, Z. Kohl, R. Steidl, J. Harnegger, B. Eskofier, J. Winkler, "Unbiased and Mobile Gait Analysis Detects Motor Impairment in Parkinson's Disease," *PLOS ONE*, vol. 8, no. 2, p. e56956, Feb. 2013.
- [29] J. Barth, Jens Barth, Michael Sünkel, Katharina Bergner, Gerald Schickhuber, Jürgen Winkler, Jochen Klucken, Björn Eskofier, "Combined analysis of sensor data from hand and gait motor function improves automatic recognition of Parkinson's disease," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2012, pp. 5122–5125.
- [30] D. Joshi, A. Khajuria, and P. Joshi, "An automatic non-invasive method for Parkinson's disease classification," *Comput. Methods Programs Biomed.*, vol. 145, pp. 135–145, 2017.
- [31] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," *Future Gener. Comput. Syst.*, vol. 29, no. 7, pp. 1645–1660, 2013.
- [32] M. Chan, D. Estève, C. Escriba, and E. Campo, "A review of smart homes—Present state and future challenges," *Comput. Methods Programs Biomed.*, vol. 91, no. 1, pp. 55–81, 2008.
- [33] H. Alemdar and C. Ersoy, "Wireless sensor networks for healthcare: A survey," *Comput. Netw.*, vol. 54, no. 15, pp. 2688–2710, 2010.
- [34] J. Bart, C. Oberndorfer, C. Pasluosta, S. Schüle, Heiko Gassner, Samuel Reinfelder, Patrick Kugler, Dominik Schuldhaus, Jürgen Winkler, Jochen Klucken, Björn M. Eskofier, "Stride Segmentation during Free Walk Movements Using Multi-Dimensional Subsequence Dynamic Time Warping on Inertial Sensor Data," *Sensors*, vol. 15, no. 3, pp. 6419–6440, Mar. 2015.
- [35] A. Burns, BR Greene, MJ McGrath, Terrance J. O'Shea, Benjamin Kuris, Steven M. Ayer, Florin Stroiescu, Victor Cionca, "SHIMMER: A Wireless Sensor Platform for Noninvasive Biomedical Research," *IEEE Sens. J.*, vol. 10, no. 9, pp. 1527–1534, Sep. 2010.
- [36] U. Lindemann, B. Najafi, W. Zijlstra, K. Hauer, R. Muche, C. Becker, K. Aminian, "Distance to achieve

steady state walking speed in frail elderly persons," *Gait Posture*, vol. 27, no. 1, pp. 91–96, Jan. 2008.

[37] D. Kobsar, C. Olson, R. Paranjape, T. Hadjistavropoulos, and J. M. Barden, "Evaluation of age-related differences in the stride-to-stride fluctuations, regularity and symmetry of gait using a waist-mounted tri-axial accelerometer," *Gait Posture*, vol. 39, no. 1, pp. 553–557, Jan. 2014.

[38] J. M. Hausdorff, M. E. Cudkowicz, R. Firtion, J. Y. Wei, and A. L. Goldberger, "Gait variability and basal ganglia disorders: Stride-to-stride variations of gait cycle timing in parkinson's disease and Huntington's disease," *Mov. Disord.*, vol. 13, no. 3, pp. 428–437, 1998.

[39] S. Okud, S. Takano, M. Ueno, Y. Hara, Yasushi Chida, Tomoko Ikkaku, Fumio Kanda, Tatsushi Toda, "Gait analysis of patients with Parkinson's disease using a portable triaxial accelerometer," *Neurol. Clin. Neurosci.*, vol. 4, no. 3, pp. 93–97, May 2016.

[40] M. Sekine, T. Tamura, M. Yoshida, Y. Suda, Y. Kimura, H. Miyoshi, Y. Kijima, Y. Higashi and T. Fujimoto, "A gait abnormality measure based on root mean square of trunk acceleration," *J. NeuroEngineering Rehabil.*, vol. 10, p. 118, Dec. 2013.

[41] P. Fazio, G. Granieri, I. Casetta, E. Cesnik, S. Mazzacane, P. Caliendo, F. Pedrielli, E. Granier, "Gait measures with a triaxial accelerometer among patients with neurological impairment," *Neurol. Sci.*, vol. 34, no. 4, pp. 435–440, Apr. 2013.

[42] E. Rastegari and H. Ali, "A Correlation Network Model Utilizing Gait Parameters for Evaluating Health Levels," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2017, pp. 568–574.

[43] L. A. Talbot, J. M. Gaines, T. N. Huynh, and E. J. Metter, "A Home-Based Pedometer-Driven Walking Program to Increase Physical Activity in Older Adults with Osteoarthritis of the Knee: A Preliminary Study," *J. Am. Geriatr. Soc.*, vol. 51, no. 3, pp. 387–392, Mar. 2003.

[44] J. M. Paquet, B. Auvinet, D. Chaleil, and E. Barrey, "[Analysis of gait disorders in Parkinson's disease assessed with an accelerometer]," *Rev. Neurol. (Paris)*, vol. 159, no. 8–9, pp. 786–789, Sep. 2003.

[45] E. Rastegari, V. Marmelat, L. Najjar, D. Bastola, and H. H. Ali, "Using gait parameters to recognize various stages of Parkinson," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2017, pp. 1647–1651.

[46] G. Yogeve, M. Plotnik, C. Peretz, N. Giladi, and J. M. Hausdorff, "Gait asymmetry in patients with Parkinson's disease and elderly fallers: when does the bilateral coordination of gait require attention?,"

Exp. Brain Res., vol. 177, no. 3, pp. 336–346, Mar. 2007.

[47] B. Galna, S. Lord, and L. Rochester, "Is gait variability reliable in older adults and Parkinson's disease? Towards an optimal testing protocol," *Gait Posture*, vol. 37, no. 4, pp. 580–585, Apr. 2013.

[48] R. Baltadjieva, N. Giladi, L. Gruendlinger, C. Peretz, and J. M. Hausdorff, "Marked alterations in the gait timing and rhythmicity of patients with de novo Parkinson's disease," *Eur. J. Neurosci.*, vol. 24, no. 6, pp. 1815–1820, 2006.

[49] M. Nallegowda, M., Singh, U., Handa, G., Khanna, M., Wadhwa, S., Yadav, S.L., Kumar, G. and Behari, M., "Role of Sensory Input and Muscle Strength in Maintenance of Balance, Gait, and Posture in Parkinson's Disease: A Pilot Study," *Am. J. Phys. Med. Rehabil.*, vol. 83, no. 12, p. 898, Dec. 2004.

[50] S. Volpato and J. M. Guralnik, "The Different Domains of the Comprehensive Geriatric Assessment," in *Comprehensive Geriatric Assessment*, Springer, 2018, pp. 11–25.

[51] N. M. Tahir and H. H. Manap, "Parkinson Disease Gait Classification based on Machine Learning Approach," *J. Appl. Sci.*, vol. 12, no. 2, pp. 180–185, 2012.

[52] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in *International Conference on Pervasive Computing*, 2004, pp. 1–17.

[53] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J. Bioinform. Comput. Biol.*, vol. 3, no. 02, pp. 185–205, 2005.

[54] A. K. Tiwari, "Machine learning based approaches for prediction of Parkinson disease," *Mach Learn Appl*, vol. 3, no. 2, pp. 33–39, 2016.