

Improvisation of classification performance based on feature optimization for differentiation of Parkinson's disease from other neurological diseases using gait characteristics

Satyabrata Aich¹, Moon-il Joo², Hee-Cheol Kim³, Jinse Park⁴

^{1,2,3}Department of Computer Engineering/Institute of Digital Anti-aging healthcare, Inje University, Republic of Korea

⁴Department of Neurology, Haeundae-Paik Hospital, Inje University College of Medicine, Republic of Korea

Article Info

Article history:

Received Apr 25, 2019

Revised Jul 9, 2019

Accepted Jul 17, 2019

Keywords:

Gait characteristics

Machine learning

Feature optimization

Parkinson's disease

Neurological disorders

ABSTRACT

Most neurological disorders that include Parkinson's disease (PD) as well as other neurological diseases such as Amyotrophic Lateral Sclerosis (ALS) and Huntington's disease (HD) have some common abnormalities regarding the movement, vocal, and cognitive behaviors of sufferers. Variations in the manifestation of these types of abnormality help distinguish one disorder from another. In this study, differentiation was performed based on the gait characteristics of patients afflicted by different neurological disorders. In the recent past, many researchers have applied different machine learning and feature selection techniques to the classification of different groups of patients based on common abnormalities. However, in an era of modernization where the focus is on timely low-cost automatization and pattern recognition, such techniques require improvisation to provide high performance. We attempted to improve the performance of such techniques using different feature optimization methods, such as a genetic algorithm (GA) and principal component analysis (PCA), and applying different classification approaches, i.e., linear, nonlinear, and probabilistic classifiers. In this study, gait dynamics data of patients suffering with PD, ALS, and HD were collated from a public database, and a binary classification approach was used by taking PD as one group and adopting ALS+HD as another group. Performance comparison was achieved using different classification techniques that incorporated optimized feature sets obtained from GA and PCA. In comparison with other classifiers using different feature sets, the highest accuracy (97.87%) was obtained using random forest combined with GA-based feature sets. The results provide evidence that could assist medical practitioners in differentiating PD from other neurological diseases using gait characteristics.

Copyright © 2019 Institute of Advanced Engineering and Science.
All rights reserved.

Corresponding Author:

Jinse Park,

Departement of Neurology, Haeundae-Paik Hospital,

Inje University College of Medicine,

875, Haeun-daero, Haeundae-gu, Busan, 612-896, Republic of Korea.

Email: jinsepark@gmail.com

1. INTRODUCTION

Disorders related to Parkinson's disease (PD), which usually affect multiple parts of the body, are characterized by motor symptoms such as tremors, abnormalities in gait, and postural rigidity and by non-motor symptoms such as cognitive disorders that affect the brain cells. The progressive nature of this type of disease affects the quality of life of those affected [1]. The difficulties associated with this disease limit the ability of sufferers to participate in ordinary daily, social, and sporting activities and enforce their dependence on others and healthcare service systems [2]. Patients affected by other neurological diseases

such as Amyotrophic Lateral Sclerosis (ALS) and Huntington's disease (HD) can also experience similar abnormalities. Gait abnormalities are most common amongst those affected by one of the various neurological disorders, and evaluation of gait characteristics is often conducted for early diagnosis as well as for prediction of different neurological diseases [3]. It has been found that gait analysis and PD have a strong relationship [4]. In the recent past, several researchers have used machine learning techniques as well as feature selection techniques to distinguish the various neurological diseases of patients. However, they have tended to use either only one feature selection technique (or none at all) and only a few machine learning techniques to attempt to distinguish neurological diseases such as PD or ALS [5-7]. No comprehensive study has been undertaken to compare the performance of such an approach considering all three neurological diseases (PD, ALS, and HD) when using various feature optimization and machine learning techniques. This study compared the improvement in the accuracy of distinguishing PD, ALS, and HD patients using popular feature selection algorithms such as a genetic algorithm (GA) and principal component analysis (PCA) together with a range of classifiers, i.e., linear, nonlinear, and probabilistic classifiers.

2. BACKGROUND

In 2017, Kotsavasiloglou *et al.* [8] studied the differences in the movement of the hand and coordination of the muscles between PD patients and healthy subjects when using a pen and tablet device. They used different classification techniques such as Naïve Bayes, AdaBoost, logistic regression, J48, support vector machine (SVM), and random forest (RF) for the classification tasks. They found average accuracy of 91%, sensitivity of 88%, and specificity of 95% when differentiating the two groups. Eskidere *et al.* [9] analyzed the performance of the least square SVM (LS-SVM), multilayer perceptron neural network, and general regression neural network regression methods to track remotely the progression of PD. They found LS-SVM produced the best result in comparison with the other three methods. Gharehchopogh *et al.* [10] used a method based on an artificial neural network for the diagnosis of PD. They used the multilayer perceptron (MLP) approach with a back-propagation algorithm and a radial basis function to distinguish patients with PD. They found the MLP method performed well with accuracy of 93.22%. Nilashi *et al.* [11] proposed a hybrid intelligent method for the prediction of PD progression using noise removal, clustering, and prediction methods. To overcome the multicollinearity problem, they used PCA and an expectation maximization method. They also applied an adaptive neuro-fuzzy inference system and support vector regression for the prediction of PD progression. They found this proposed method improved the accuracy for the prediction of PD progression.

Muniz *et al.* [12] proposed a method based on a comparison of three different classifiers (i.e., logistic regression, probabilistic neural network, and SVM) for distinguishing normal subjects from PD subjects by assessing the effects of deep brain stimulation on ground reaction force both with and without medication. They observed that classifiers indicating deep brain stimulation only were more effective than those indicating medication only. Polat [13] used fuzzy c-means clustering feature weighting (FCMFW) and the k-NN classification technique for the detection of PD. The FCMFW-based PD datasets were used as input to the k-NN-based classification system. The best k value was obtained after several iterations using the FCMFW dataset. It was found that the FCMFW-based dataset combined with the k-NN classification system offered reasonable performance for the detection of PD. Nilashi *et al.* [14] proposed a knowledge-based system for disease prediction using clustering, noise removal, and prediction techniques. They used expectation maximization, PCA, classification and regression tree (CART), and a fuzzy-based method for the prediction of disease. They found a combined system that included a fuzzy-based system and noise removal using a CART together with clustering techniques was effective in disease prediction using real-world medical datasets.

Parisi *et al.* [15] used dysphonia measures and clinical scores as input data for different machine learning techniques that included a combination of multilayer perceptron and Lagrangian SVM (LSVM), a combination of MLP and SVM, and both MLP and SVM alone for the diagnosis of early stage PD. They used four performance metrics (i.e., accuracy, sensitivity, specificity, and area under the curve) for comparison of the different machine learning techniques as well as the hybrid machine learning techniques. They found the MLP-LSVM hybrid technique provided accuracy of 100% and that it converged at a faster rate than the other techniques. They suggested the MLP-LSVM hybrid technique could have potential for use by clinicians for early detection of PD. Aich *et al.* [16] used spatiotemporal gait parameters as input to different machine learning approaches to distinguish PD from Alzheimer's disease. Two different feature sets were used for comparison of the performance of the different machine learning techniques. The machine learning techniques used in their approach were linear discriminant analysis, Naïve Bayes, recursive partitioning and regression tree, C4.5, pruning rule-based classification tree, bagging CART, RF, Boosted C 5.0, and SVM. They found classification accuracy of 92.59% with the SVM classifier combined with

the original feature sets. Kim proposed a new technique. Yi and Yu [17] used accelerometer and gyroscope data to develop a fall detection algorithm with 100% sensitivity and 98.75% specificity. Aich *et al.* [18] used gait signals those were quantified using 3D motion captures as an input to the machine learning algorithms to distinguish Parkinson's disease patients with shuffling gait from older adults. Aich *et al.* [19] proposed a multicriteria decision modelling approach to compare the classification performance based on different feature selection techniques for the gait analysis of the PD patients.

3. METHODOLOGY

Figure 1 presents a flowchart of the methodology used in this study. The study focused on the gait dynamics of 48 individuals from different neurodegenerative groups: 15 PD patients, 20 HD patients, and 13 ALS patients [20]. The 12 parameters of gait dynamics obtained from the datasets were as follows: left stride interval (LSTRI), right stride interval (RSTRI), left swing interval (LSWIT), right swing interval (RSWIT), left swing interval_% of stride (LSWIPOS), right swing interval_% of stride (RSWIPOS), left stance interval (LSTNI), right stance interval (RSTNI), left stance interval_% of stride (LSTNIPOS), right stance interval_% of stride (RSTNIPOS), double support interval (DSI), and double support interval_% of stride (DSIPOS). We used two feature selection methods: GA and PCA. The GA method is an adaptive optimization search methodology that employs iterative computation to obtain the optimal solution. Figure 2 presents a flowchart of the working principle of GA [21]. PCA was used to compress the data and to extract information [22].

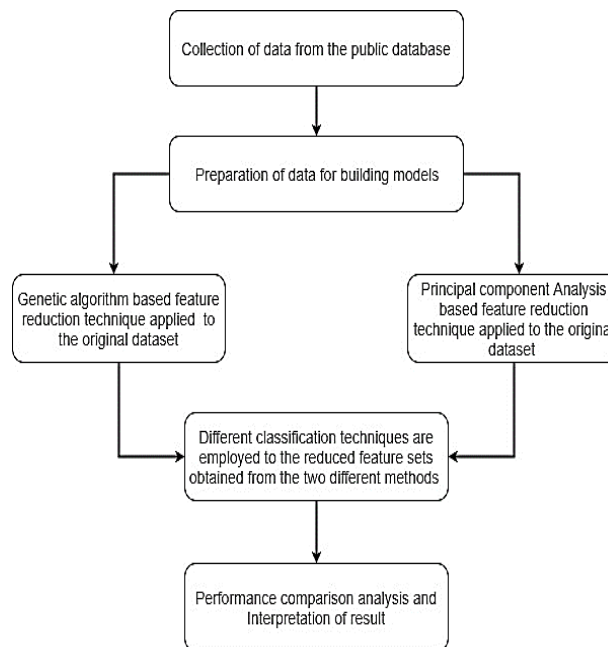


Figure 1. Flowchart of proposed methodology

3.1. Classification methods

We used recursive partitioning and regression tree, multilayer perceptron, pruning rule-based classification tree, bagging CART, and RF as nonlinear classifiers, the Naïve Bayes method as a probabilistic classifier, and SVM and a radial basis function as linear classifiers.

3.2. Performance measure metrics

The parameters used to assess and validate the classifiers were as follows: accuracy, sensitivity, specificity, PPV, and NPV. Sensitivity is defined as the ratio of true positives to the sum of true positives and false negatives. Specificity is the ratio of true negatives to the sum of false positives and true negatives. In the current study, we used PPV and NPV to verify the presence or absence of PD. Therefore, PPV represents the probability that PD is present and given a positive test result, and NPV is the probability that PD is absent and given a negative test result [23]. Accuracy is defined as the ratio of the number of correct predictions to the total number of predictions, multiplied by 100 to convert to a percentage.

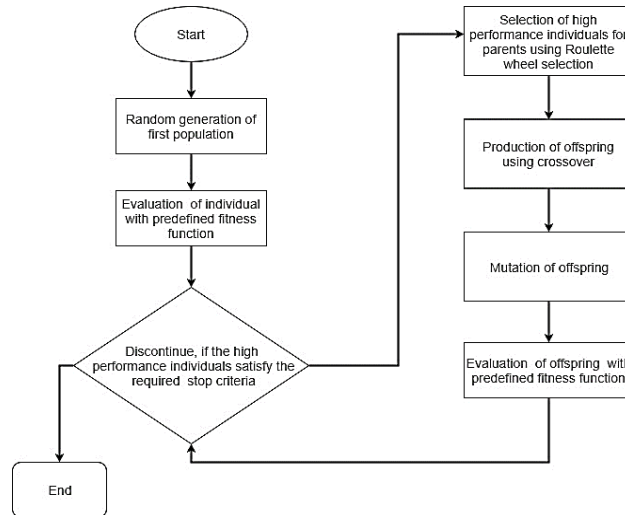


Figure 2. Flowchart of the working principle of the genetic algorithm

4. RESULT AND DISCUSSIONS

The original feature sets consisted of 12 features. The features obtained based on the GA and PCA methods are detailed in Table 1. For the GA method, the features were selected based on the survival of the fittest method, whereby the high performers survived a repetitive iteration process and they were able to achieve the desired criteria specified as the fitness function. The GA is a dimension reduction technique and the principal components are the directions in which variance is greatest, i.e., greater dispersion. Figure 3 shows the variance of the selected variables and their contributions to feature selection; seven variables accounted for 99.1% of the variance. Figure 4 shows the cumulative variance of the principal components. Both feature selection techniques were used to identify a small number of important features and these features were employed as input to the various classifiers to analyze the performance of the classifiers based on the performance metrics. Figure 5 shows the resultant principal components. Figure 6 shows the relative importance of the features selected by PCA.

Table 1. Feature selection using different methods

Method	Feature
GA	Left stride interval, right stride interval, left stance interval, left stance interval_% of stride, left swing interval, right swing interval, right swing interval_% of stride, double support interval,
PCA	Left stride interval, left stance interval, left stance interval_% of stride, left swing interval, right swing interval, double support interval, double support interval_% of stride

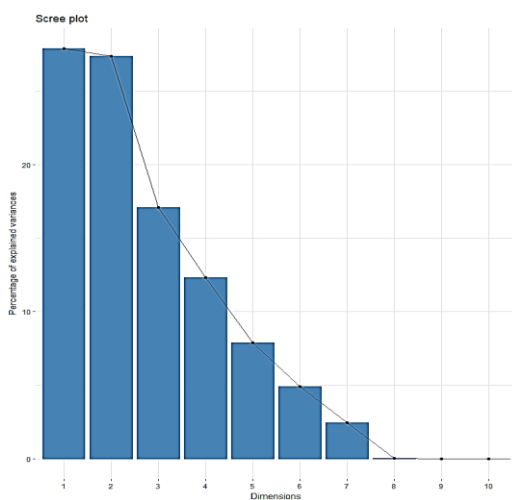


Figure 3. Variances of the principal components

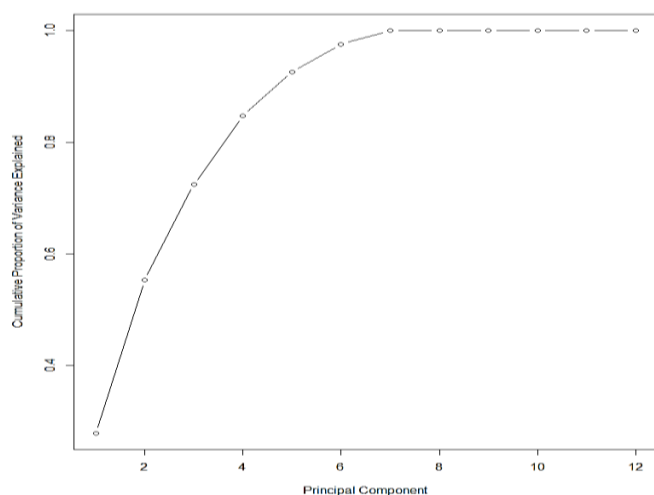


Figure 4. Cumulative variances of principal components

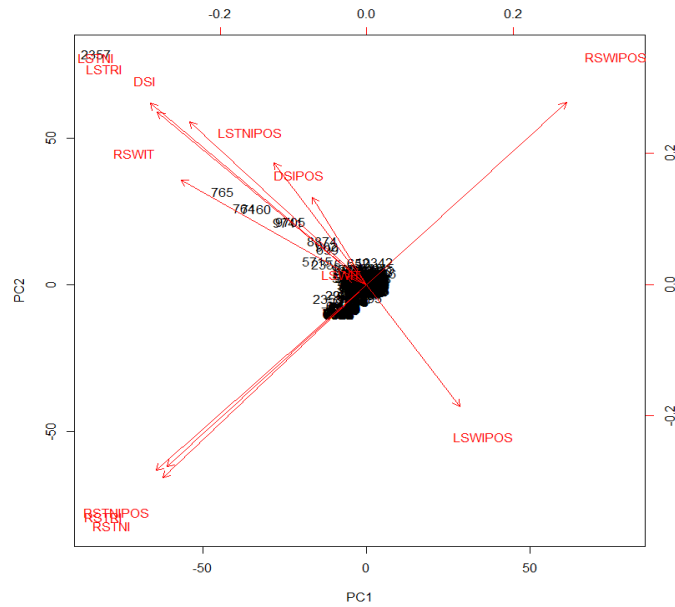


Figure 5. Resultant principal components

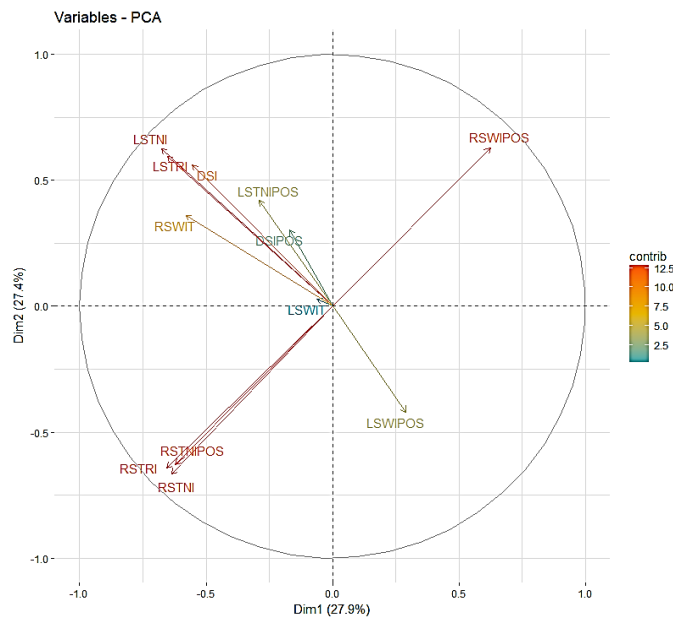


Figure 6. Relative importance of the features selected by principal component analysis (PCA)

After feature selection, we applied classification techniques to the selected features. We divided the datasets into two: one set was used for training and the other was used for testing. The results of the performance measures for different feature sets are shown in Figure 7 to Figure 11. Figure 7 illustrates the accuracy of the classifiers with different feature sets. It can be observed that features obtained using the GA sets and the RF classifier show the highest accuracy (97.87%) followed by the SVM classifier with the same GA sets (accuracy: 92.89%). Figure 8 presents the sensitivity of the classifiers with different feature sets. It can be seen that features obtained using the GA sets and the RF classifier show the highest sensitivity (0.9883) followed by the bagging CART classifier with the same GA sets (sensitivity: 0.9428). Figure 9 shows the specificity of the classifiers with different feature sets. It can be observed that features obtained using the GA sets and the RF classifier show the highest specificity (0.9835) followed by the SVM classifier with the same GA sets (specificity: 0.9489). Figure 10 presents the PPV of the classifiers with different

feature sets. It can be seen that features obtained using the GA sets and the RF classifier show the highest PPV (0.9775) followed by SVM classifier with the same GA sets (PPV: 0.9565). Figure 11 presents the NPV of the classifiers with different feature sets. It can be observed that features obtained using the GA sets and the RF classifier show the highest NPV (0.9825) followed by the Naïve Bayes classifier with the same GA sets (NPV: 0.933).

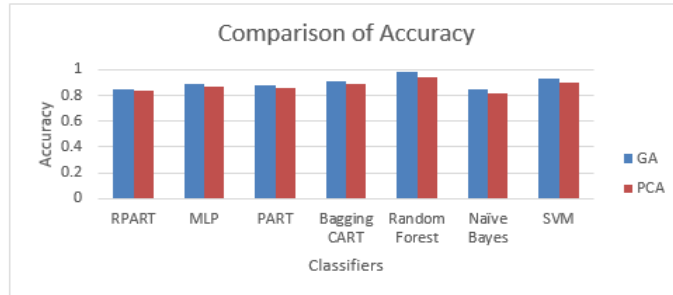


Figure 7. Comparison of the genetic algorithm (GA) and principal component analysis (PCA) accuracy

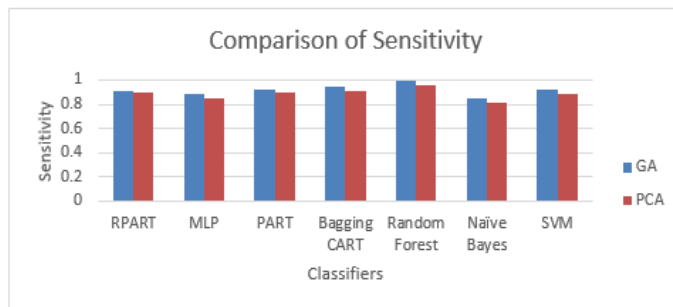


Figure 8. Comparison of the genetic algorithm (GA) and principal component analysis (PCA) sensitivity

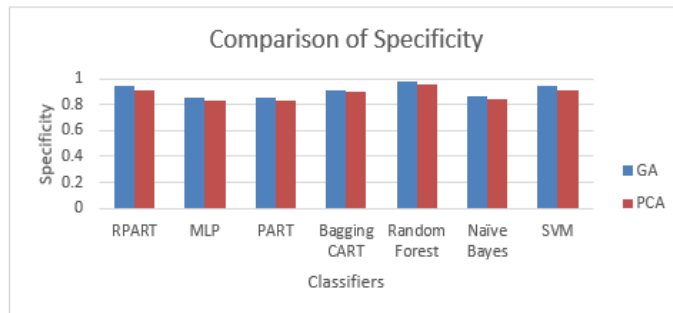


Figure 9. Comparison of the genetic algorithm (GA) and principal component analysis (PCA) specificity

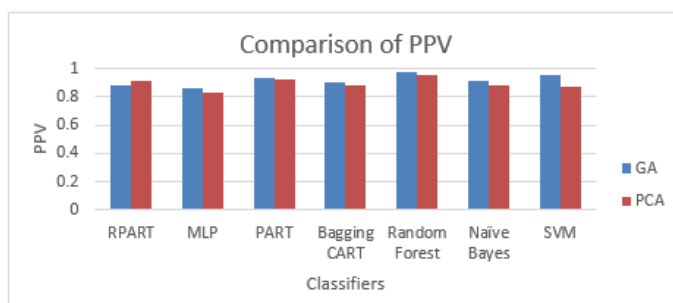


Figure 10. Comparison of the genetic algorithm (GA) and principal component analysis (PCA) positive predictive value (PPV)

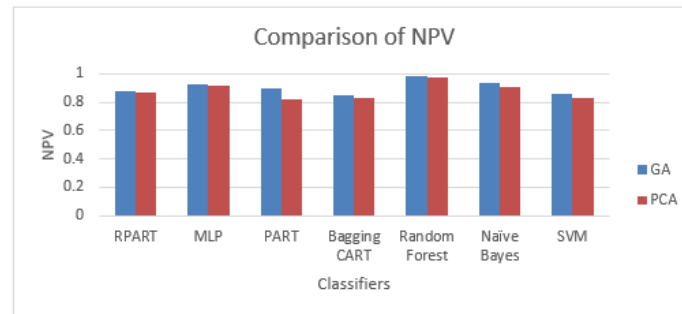


Figure 11. Comparison of the genetic algorithm (GA) and principal component analysis (PCA) negative predictive value (NPV)

The results demonstrate that the RF classifier with the GA feature sets performed better in terms of accuracy in comparison with other classifiers with the same GA feature sets as well as the same group of classifiers with PCA feature sets. Furthermore, all classifiers were found to perform better with the GA feature sets than with the PCA feature sets in terms of the other performance measures, i.e., sensitivity, specificity, PPV, and NPV. In our analysis, we prioritized accuracy over the other performance measures; thus, we consider the RF classifier with GA sets more important than other combinations. The previous study [24] related to the classification of Parkinson's disease with FoG and No FoG also observed classification accuracy of 91.42%. Another study [25] that used voice parameters and using various feature selection technique also observed classification accuracy of 97.57%. The results obtained in this proposed approach are in accordance with previous reports.

5. CONCLUSION

This study used linear, nonlinear, and probabilistic classifiers to distinguish PD from other neurological diseases (i.e., ALS and HD) by considering two different feature selection methods using gait dynamics data. In a comparison of various classifiers and different feature sets, we found the highest accuracy (97.87%) when using the RF classifier in combination with GA-based feature sets. This analysis helps us explore different feature selection methods as well as the importance of the features for improvement in classification performance. The results demonstrated could help clinicians and practitioners distinguish PD from ALS and HD using gait dynamics data.

ACKNOWLEDGEMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2017R1D1A3B04032905). This paper is a revised and expanded version of a paper entitled "A Machine Learning Approach to Discriminate the Parkinson's disease from Other Neurological Diseases using Different Feature Selection Techniques Based on Gait Parameters" presented at "International Conference on Future Information & Communication Engineering (ICFICE), Pattaya, Thailand, 26 June-1 July 2018.

REFERENCES

- [1] C.T.R. Suratos, *et al.*, "Quality of life and Parkinson's disease: Philippine translation and validation of the Parkinson's disease questionnaire," *Journal of Clinical Neuroscience*, vol. 54, pp. 156-160, 2018.
- [2] L.J. Findley, "The economic impact of Parkinson's disease," *Parkinsonism & related disorders*, vol. 13, pp. S8-S12, 2007.
- [3] A. Klomsae, *et al.*, "String Grammar Unsupervised Possibilistic Fuzzy C-Medians for Gait Pattern Classification in Patients with Neurodegenerative Diseases," *Computational Intelligence and Neuroscience*, vol. 2018, no. 3, pp. 1-10, 2018.
- [4] S.Aich, *et al.*, "A text mining approach to identify the relationship between gait-Parkinson's disease (PD) from PD based research articles," *In Inventive Computing and Informatics (ICICI), International Conference on, IEEE*, pp. 481-485, 2017.
- [5] Y. Wu and L. Shi, "Analysis of altered gait cycle duration in amyotrophic lateral sclerosis based on nonparametric probability density function estimation," *Medical Engineering & Physics*, vol. 33, no. 3, pp. 347-355, 2011.

- [6] M. R. Daliri, "Chi-square distance kernel of the gaits for the diagnosis of Parkinson's disease," *Biomedical Signal Processing and Control*, vol. 8, pp. 66–70, 2013.
- [7] Y. Wu and S. Krishnan, "Computer-aided analysis of gait rhythm fluctuations in amyotrophic lateral sclerosis," *Medical & Biological Engineering & Computing*, vol. 47, pp. 1165–1171, 2009.
- [8] C. Kotsavasiloglou, *et al.*, "Machine learning-based classification of simple drawing movements in Parkinson's disease," *Biomedical Signal Processing and Control*, vol. 31, pp. 174–180, 2017.
- [9] Ö. Eskidere, *et al.*, "A comparison of regression methods for remote tracking of Parkinson's disease progression," *Expert Systems with Applications*, vol. 39, pp. 5523–5528, 2012.
- [10] F.S.Gharehchopogh, P.Mohammadi, "A Case Study of Parkinson's Disease Diagnosis using Artificial Neural Networks," *International Journal of Computer Applications*, vol. 73, pp. 1–6, 2013.
- [11] M. Nilashi, *et al.*, "Accuracy Improvement for Predicting Parkinson's disease Progression," *Scientific Reports*, vol. 6, no. 34181, 2016.
- [12] A.M.S. Muniz, *et al.*, "Comparison among probabilistic neural network, support vector machine and logistic regression for evaluating the effect of subthalamic stimulation in Parkinson disease on ground reaction force during gait," *Journal of Biomechanics*, vol. 43, pp. 720–726, 2010.
- [13] K.Polat, "Classification of Parkinson's disease using feature weighting method on the basis of fuzzy C-means clustering," *International Journal of Systems Science*, vol. 43, pp. 597–609, 2012.
- [14] M.Nilashi, *et al.*, "An Analytical Method for Diseases Prediction Using Machine Learning Techniques," *Computers & Chemical Engineering*, vol. 106, pp. 212–223, 2017.
- [15] L. Parisi, *et al.*, "Feature-driven Machine Learning to Improve Early Diagnosis of Parkinson's Disease," *Expert Systems with Applications*, vol. 110, pp. 182–190, 2018.
- [16] S.Aich, *et al.*, "A Performance Comparison Based on Machine Learning Approaches to Distinguish Parkinson's Disease from Alzheimer Disease Using Spatiotemporal Gait signals," *Advanced Science Letters*, vol. 24, pp. 2058–2062, 2018.
- [17] Y.J. Yi and Y.S. Yu, "Emergency monitoring system based on a newly-developed fall detection algorithm," *Journal of information and communication convergence engineering*, vol. 11, pp. 199–206, 2013.
- [18] S.Aich, *et al.*, "A machine learning approach to distinguish Parkinson's disease (PD) patient's with shuffling gait from older adults based on gait signals using 3D motion analysis," *International Journal of Engineering & Technology*, vol. 7, no. 3.29, pp. 153–156, 2018.
- [19] S.Aich, *et al.*, "A Multi Criteria Decision Modelling Approach for Gait Analysis of Parkinson's Disease Using Wearable Sensors to Compare the Classification Performance Based on the Different Feature Selection Methods," *In International Conference on Frontier Computing*, Springer, Singapore, pp. 528–534, 2018.
- [20] A. L. Goldberger, *et al.*, PhysioBank, PhysioToolkit, and Physio Net, *Circulation*, vol. 101, pp. 215–220, 2000.
- [21] C.C Kang, *et al.*, "A genetic algorithm-based Boolean delay model of intracellular signal transduction in inflammation," *BMC bioinformatics*, vol. 12, pp. S17, 2011.
- [22] E.R. Hruschka, N.F. Ebecken, "Extracting rules from multilayer perceptrons in classification problems: A clustering-based approach," *Neurocomputing*, vol. 70, pp. 384–397, 2006.
- [23] H.B. Wong, G.H. Lim, Measures of diagnostic accuracy: sensitivity, specificity, PPV and NPV, *Proceedings of Singapore healthcare*, vol. 20, pp. 316–318, 2011.
- [24] S.Aich, *et al.*, "A validation study of freezing of gait (FoG) detection and machine-learning-based FoG prediction using estimated gait characteristics with a wearable accelerometer," *Sensors*, vol. 18, no. 10, pp. 3287, 2018.
- [25] S.Aich, *et al.*, "A Supervised Machine Learning Approach using Different Feature Selection Techniques on Voice Datasets for Prediction of Parkinson's Disease," *In 2019 21st International Conference on Advanced Communication Technology (ICACT) IEEE*, pp. 1116–1121, 2019.

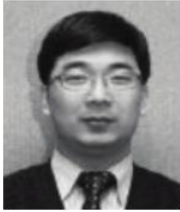
BIOGRAPHIES OF AUTHORS



Satyabrata Aich, is working as a researcher in the field of computer engineering and digital healthcare. He has published many research papers in journals and conferences in the realms of Machine learning, Text mining, and Supply Chain Management. His research interests are natural language processing, Machine learning, supply chain management, Text mining, Bio informatics.



Moon-II Joo received the PhD degree in computer engineering from Inje University in 2018. He is currently working as a research professor at the Institute of digital anti-aging healthcare, Inje University, Korea. His research interests are in software engineering, human computer interaction, smart phone programming, and component based development.



Hee-Cheol Kim, received his BSc at Department of Mathematics, MSc at Department of Computer Science in SoGang University in Korea, and PhD at Numerical Analysis and Computing Science, Stockholm University in Sweden in 2001. He is Professor at Department of Computer Engineering and Head of the Institute of Digital Anti-aging Healthcare, Inje University in Korea. His research interests include Machine learning, Text mining, and Bio Informatics



Jinse Park is associate professor who work in Haeundae Paik Hospital, Inje University, Busan, Korea. He is expert of movement disorder including Parkinson's disease. He graduated from college of medicine in Hanyang University in 2003. He got doctor degree from Hanyang University in 2016. His research field is gait analysis in Parkinson's disease. He wrote many reports including SCI and SCOPUS journals and patents about gait analysis in Parkinson's disease. He got research fund from 2016 Inje University specialization project for this research. He was got best paper award from Korean society of clinical neurophysiology in 2018.