A Supervised Machine Learning Approach using Different Feature Selection Techniques on Voice Datasets for Prediction of Parkinson's Disease

Satyabrata Aich*, Hee-Cheol Kim*, Kim younga**, Kueh Lee Hui***, Ahmed Abdulhakim Al-Absi**** and Mangal Sain*****

*Department of Computer Engineering, Inje University, South Korea

*Department of Rehabilitation Science, Inje University, South Korea

***Department of Electrical Engineering, Dong-A University, South Korea

****Department of Computer Engineering, Kyungdong University- Global Campus, Gangwondo, South Korea

*****Department of Computer Engineering, Dongseo University, South Korea

satyabrataaich@gmail.com, heeki@inje.ac.kr, kya2664@hanmail.net, leehkueh@dau.ac.kr, absiahmed@kduniv.ac.kr, mangalsain1@gmail.com

Abstract—**Among the neurological diseases, parkinson's disease is the second most common disease, which affect the old age people over the age of 65 year. It is also mentioned that the number of people affected with Parkinson's disease will increase at a higher rate until 2050, and it will be a rising concern to many developed countries because the cost due to the healthcare service of these disease is really high. Parkinson's disease (PD) belongs to the group of neurological disorder, which directly affect the brain cells and the effect is shown in terms of movement, voice and other cognitive disabilities. Past few years researchers are working for detection and monitoring of the Parkinson's disease by using the speech analysis as well as the gait analysis data. Machine learning and artificial intelligence techniques are gaining popularity because these techniques are able to automate the pattern recognition process with high accuracy.**

However so far, no body has compared the performance metrics using different feature sets by applying nonlinear and linear classification approaches based on the voice data. So, in this paper we have proposed a new approach by comparing the performance metrics with different feature sets such as genetic algorithm-based feature sets as well as Principal Component Analysis based feature reduction technique for selecting the feature sets. We have used different classification approaches to compare the performance metrics. We have found an accuracy of 97.57% using SVM with RBF by using genetic algorithm-based feature sets. This analysis will help the clinicians to differentiate the PD group from healthy group based on the voice data.

Keywords— **Parkinson's disease, machine learning, feature selection, voice data, genetic algorithm**

I. Introduction

Past few years a lot of research has been going on the Parkinson's disease because the healthcare related cost due to this disease is keeping on increasing as the longevity of the population is increasing in the developed countries. Since this disease affect most of the old people, it is become necessary for the developed counties to detect the disease at the early stage. The early detection will help the developed country in economic perspective as well as social perspective because it can be assessed well. Parkinson's disease belongs to one of the category of neurodegenerative disease which directly as well as indirectly affects the brain cells that will affect the movement, speech and other cognitive parts [1, 2, and 3]. The Parkinson's disease is progressive in nature. As the disease progresses more than 90% of the patients has the speech disorder [4]. The symptom related to the vocal impairment of Parkinson's disease patients is called dysphonia. The clinicians measured some indicators related to dysphonia to assess the PD patients. The measures related to dysphonia

Manuscript received on January 2, 2018. This work was supported by Kyungdong University grant, 2018 and a follow-up the invited journal to the outstanding paper of the 20th International Conference on Advanced Communication Technology (ICACT 2018).

Satyabrata Aich is with the Department of Computer Engineering, Inje University, South Korea (e-mail: satyabrataaich@gmail.com)

Hee-Cheol Kim is with the Department of Computer Engineering, Inje University, South Korea (e-mail: heeki@inje.ac.kr)

Kim younga is with the Department of Rehabilitation Science, Inje University, South Korea (e-mail: kya2664@hanmail.net)

Kueh Lee Hui is with the Department of Electrical Engineering, Dong-A University, South Korea (e-mail: leehkueh@dau.ac.kr)

Ahmed Abdulhakim Al-Absi is with the Department of Computer Engineering, Kyungdong University- Global Campus, Gangwondo, South Korea (e-mail: absiahmed@kduniv.ac.kr)

Mangal Sain is with the Department of Computer Engineering, Dongseo University, South Korea. He is the corresponding author of this paper. (Corresponding author phone: +8251-320-2009; e-mail: mangalsain1@gmail.com).

could be treated as an important and most reliable tool to assess the voice related problem and monitor it at different stage [5, 6]. Usually the measures have lot of features which does not helpful for machine learning approaches, so feature selection method has been used for proper assessment. The feature selection method will help to evaluate the important contribution of the features in the assessment of the disease at different stage and also it helps to achieve good accuracy [7, 8].

The traditional diagnosis needs lot of observations related to the daily living activities, motor skills and other neurological parameters to assess the progression of PD, but this process is not suitable for the early detection of the PD. With respect to the past research it is found that artificial intelligence and machine learning techniques have good potential for the classification and it also found that the classification system helps to improve the accuracy and the reliability of the diagnosis and also minimize the errors as well as make the system more efficient [9]. improvement on the prediction of accuracy on the progression of PD is getting lot of attention these days [10, 11].

So in this paper an attempt has been made to check the improvement in the accuracy while classifying the PD group from the healthy control group by using different machine learning algorithm with different feature sets such as the genetic algorithm (GA) based feature set as well as the PCA based feature sets. Finally, a comparison has made in terms of performance metrics using different feature sets. The structure of the paper is organized as follows: Section 2 presents the past work related to classification model used for voice datasets. Section 3 describes about the methodologies used for this research work. Section 4 describes about the result of feature selection as well as the result of classification. Section 5 describes about the conclusion and future work.

II. RELATED WORK

Some of the Past research works are mentioned below to give overall ideas about the amount of work has been done in this field. Shahbaba and Neal used nonlinear based approach for classification of PD. They have used Dirichlet process mixtures and compared the results with other classification model such as decision trees, support vector machine and multinomial logit model and they found Dirichlet process-based method provides best classification approach of 87.7% compared to the other model [12]. Sakar and Kursun used feature selection method as well as machine learning based method for diagnosis of PD. They have used mutual information-based feature selection and support vector machine as the classification approach and they found their approach gives an accuracy of 92.75% [13].

Li et al used fuzzy based method to extend the classification related information and then they have used principal component analysis-based method for feature selections and the optimal features has been integrated with SVM based method provides a good accuracy of 93.47% [14]. Spadoto et al used evolutionary base techniques for feature selection and they have used Optimum-path Forest Classifier to detect the Parkinson’s disease and they found this approach provides a best accuracy of 84.01% while detecting the PD [15].

Luukka have proposed a feature selection method based on the fuzzy entropy measure and used similarity classifiers to classify PD. The best classification accuracy obtained by that method was 85.03% [16]. AStröm and Koker proposed a method that is used parallel feed-forward neural network-based approach to predict the PD. They have found the model is robust and the best classification accuracy obtained from that approach is 91.20% [17]. Nilashi et al proposed a method for the prediction of PD progression using clustering and prediction methods. They have applied Adaptive Neuro-Fuzzy Inference system (ANFIS) and Support Vector Regression for prediction of PD progression. They found this proposed method helps to improve the accuracy of the progression of PD [18].

Abdulhay et al proposed a method to investigate gait and tremor by using machine learning techniques based on the gait data. They have extracted various gait features using the peak detection and pulse duration and they found accuracy of 92.7% for diagnosis of parkinson’s disease [19]. Er et al proposed a method to distinguish dementia patients (AD) from the age-related cognitive decline (ARCD) by using machine learning techniques. They found that these techniques are able to distinguish ARCD and AD at a success rate of 100% based on neurocognitive tests [20]. Zeng et al proposed a method to classify the patients with PD. They have used deterministic learning method to distinguish PD from healthy control group. They found accuracy, sensitivity and specificity of 96.39%, 96.77% and 95.89% respectively [21]. Armañanzas et al proposed a method to find the non-motor related PD severity based on the machine learning approach. They have used Hoehn & Yahr index and clinical impression of severity index as the measure of the severity for the PD patients. They have used classification with feature selection using evolutionary algorithm and found accuracy ranging from 72-92% [22]. Kubota et al. discussed about the benefits of using wearable sensors for measuring the gait parameters and also mentioned about signal processing and machine learning approaches used for extracting meaningful information from data. This kind of information helps to understand the potentials of the technology on PD research and practices [23].

Abós et al proposed a method distinguishing the PD patients based on cognitive status using machine learning technique. They have used randomized logistic regression for feature selection and SVM for machine learning and found accuracy of 82.6%. They found connection-wise patterns of functional connectivity may be helpful for distinguishing the PD patients. [24]. Singh and Samavedham proposed a methodology using machine learning tools as statistical measure for determining most important features to distinguish different neurological disease(ND). They have found accuracy of 99.93% while distinguishing various ND such as HC, PD and SWEDD patients.

The above past works motivated us to try a different approach. In this paper we have tried different feature selection approach such as principal component analysis (PCA) approach and genetic algorithm-based approach and then the performance measures are compared with different machine learning classifiers.

III. PROPOSED TECHNIQUE

The flow chart of the proposed methodologies is shown in the Fig 1. In this paper we have used the dataset created by Max little University Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado, who recorded the speech signals [26]. The original data collected from the dataset composed of voice measurements from 31 people out of which 23 were diagnosed with PD. We have used Principal Component Analysis (PCA) algorithm on the original feature sets.

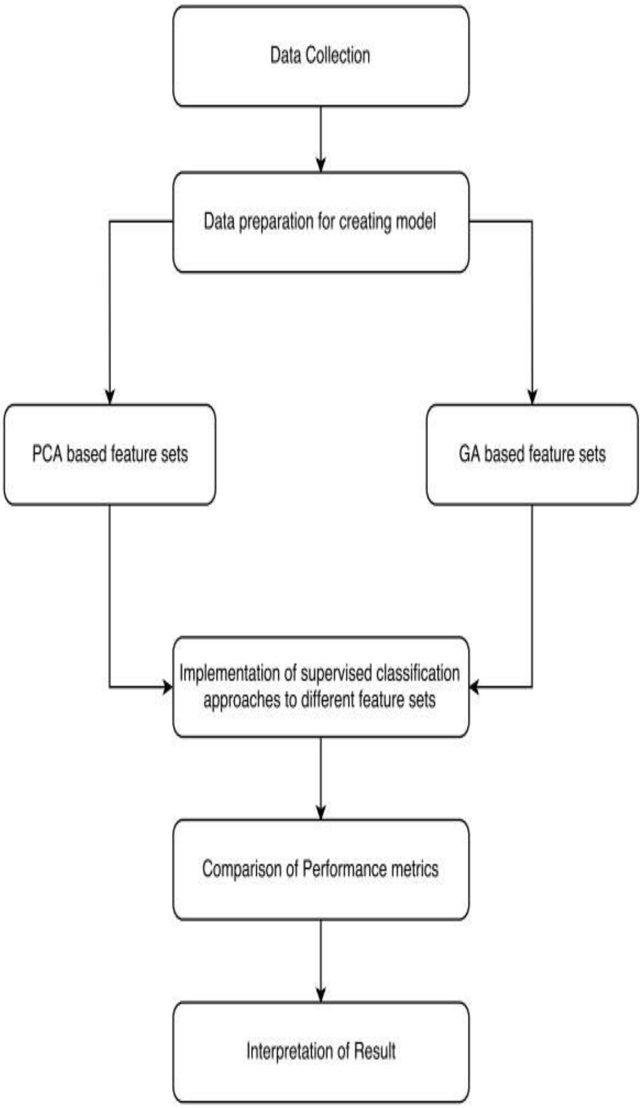
Data Collection
Data preparation for creating model
PCA based feature sets
GA based feature sets
Implementation of supervised classification approaches to different feature sets
Comparison of Performance metrics
Interpretation of Result

Fig. 1. Flowchart of the proposed method

The most widely applied technique for data reduction and feature selection is done using PCA. In PCA, the principal components or the latent variables are obtained from the variance of the data by maximizing it. The number of principal components is lesser than the regular variables. PCA reduces the dimensionality of the space so that the data can be visualized in the low dimensional space. The feature selection process is done by removing the redundant variables. [27] We have found 11 features after implementing the algorithm to the original feature sets. We have also used genetic algorithm (GA)-based feature sets for feature selection. Pledsoe first presented an adaptive optimization search methodology is called genetic algorithm and Holland mathematically presented the genetic algorithm-based approach by getting inspiration from Darwin's theory of evolution. A variable is mentioned as a gene. A chromosome is nothing but a sequence of gene. An initialization is done randomly by using population of chromosome. The quality of the chromosomes is evaluated according to a predefined fitness function. High performance chromosomes are used to produce the offspring. The genetic operators such as mutation and crossover are used to form the offspring. In this process the chromosomes are competing with each other and the fittest one survives at the end. The optimal solution comes after a series of iterative computations. [28, 29].

We have found 10 features using GA based feature sets. We have used different classification approach such as RPART, C4.5, PART, Bagging classification and Regression tree (Bagging CART), Random Forest, Boosted C5.0 and SVM.

Except SVM all other classifiers are belongs to nonlinear decision tree-based classifier. Decision tree-based classifiers are famous for giving good accuracy. SVM is most popular linear classification method. It uses different kernel function while doing classification. In this paper we have used radial basis function kernel function because it is recommended by many well-known data scientists. After implementing different classification approach, we have compared the performance measures such as accuracy, sensitivity, specificity, PPV and NPV.

Performance Metrics

The parameters used to compare the performance and validations of classifier are as follows: accuracy, sensitivity, specificity, positive predictive value (ppv), negative predictive value (npv). The sensitivity is defined as the ratio of true positives to the sum of true positives and false negatives. The specificity is defined as the ratio of true negatives to the sum of false positives and true negatives. In our research we have used the Positive predictive value and negative predictive value to check the present and absent of disease. So, the ppv is the probability that the disease is present given a positive test result and npv is the probability that the disease is absent given a negative test result [30]. Accuracy is defined as the ratio of number of correct predictions made to the total prediction made and the ratio is multiplied by 100 to make it in terms of percentage.

IV. Result and Discussions

We have used R programming language to write the code. We have reduced the original feature sets by feature selection techniques. We have chosen the PCA and GA as feature selection technique because these techniques are more widely used for feature selection without affecting the performance. As we have already discussed about the PCA method, after implementing it removes some redundant features and the reduced features are able to provide 99% variance, without affecting the performance. The GA method is different, which needs lot of iteration to reach the optimal result. With lot of iteration we have found optimized result with fewer features compared to the original features. We have trained each classifier based on the trained data and predict the power of classifier on the test data. The ratio of train data to test data is 70:30. In case of SVM we have used different Kernel function such as linear, polynomial, and radial basis function. So, each classifier able to show all the performance metrics based on the test data. We have plot the graphs shown below.

A. *Comparison of Accuracy*

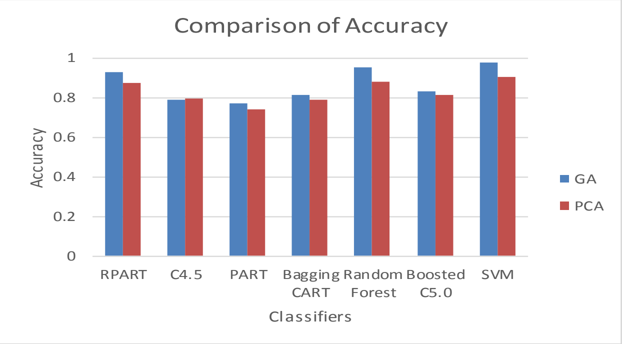

Fig. 2. Accuracy of different classifiers

Fig. 2 shows that SVM with RBF has highest accuracy of 97.57% with GA based feature sets followed by random forest and RPART classifiers with the same feature sets. Even though the other classifiers have less accuracy than SVM, but the difference in accuracy is not much.

B. *Comparison of Sensitivity*

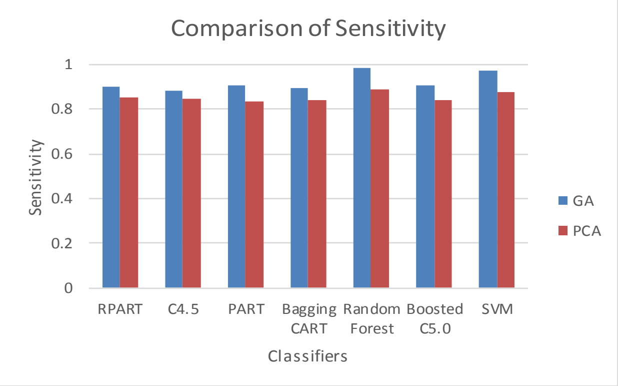

Fig. 3. Sensitivity of different classifiers

Fig .3 shows that random forest has highest sensitivity of 0.9985 with GA based feature sets followed by SVM with RBF classifiers with the same feature sets. The SVM classifier has a sensitivity of 0.9756.

C. *Comparison of Specificity*

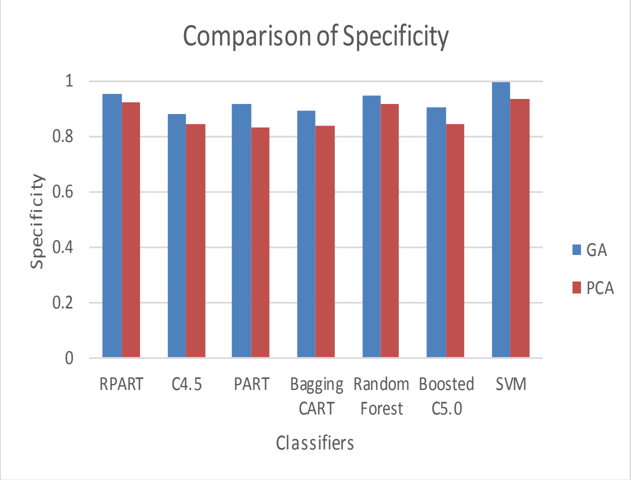

Fig. 4. Specificity of different classifiers

Fig. 4 shows that SVM with GA based feature set has highest specificity of 0.9987. RPART and random forest also follows SVM with good specificity with the same feature sets.

D. *Comparison of PPV*

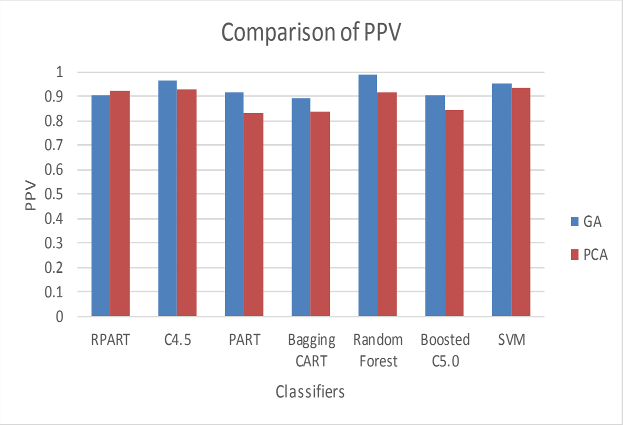

Fig. 5. PPV of different classifiers

Fig. 5 shows that the GA based feature set perform better compared to the PCA based feature set with random forest classifier. The maximum specificity achieved with this combination is 0.9934.

E. Comparison of NPV

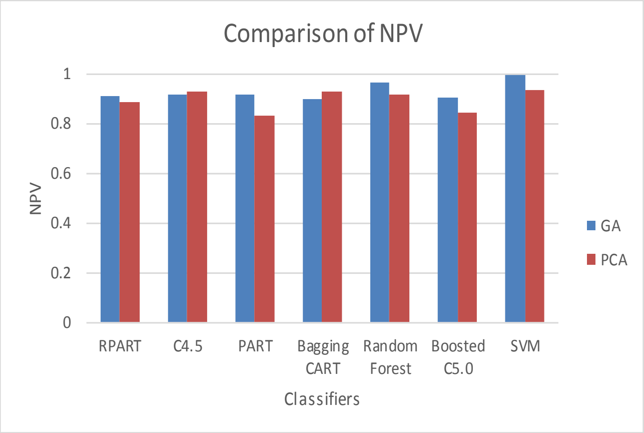

Fig. 6. NPV of different classifiers

Fig. 6 shows that SVM with GA based feature sets performs better compared to the PCA based feature sets. GA based feature sets shows maximum npv of 0.9995.

From the above plots we have seen that most of the classifiers are performing better while distinguishing between the Parkinson's patients to the control group. Basically, the feature selection processes help to save the time and space by removing the redundant sets, which has almost no impact on the performance. In our case we have seen the SVM and random forest are doing well in terms of the performance metric comparison with the GA based feature sets. We have seen that the GA based feature sets are performing well compared to PCA based feature sets. In theoretical case we may give a suggestion based on the accuracy while choosing the classifier, however in practical case lot of other performance measures is also come into picture. In our case we can suggest SVM is the best classifier and GA is best for feature selection, however it may provide different result with bigger dataset. This result actually gives an idea about the performance comparison and also gives an impression to analyze more deeply for implementing in practical life.

V. Conclusion and Future Work

This paper outlined some of the new feature selection technique as well as some of the supervised machine learning approach for distinguishing Parkinson's patient from the control group. We have seen the capability of the PCA based feature selection technique and GA based feature selection technique. We have also seen the performance of different classification approach when combined with different feature sets. We have found that GA based feature selection technique performs better when combined with SVM classifier in terms of all performance measure compared to the PCA based feature set. Mostly while looking for performance measure we usually concentrate on accuracy, in that respect our SVM classifiers provides highest accuracy of 97.57% with GA based feature sets. This type of analysis will save the time and efficiency while doing pattern classification comprising two groups such as PD and control groups. This will help the clinician to distinguish PD and control groups. This research can be further improved by adding more features and trying other classification techniques.

References

[1] S.Przedborski, M.Vila, AND V.Jackson-Lewis, "Series Introduction: Neurodegeneration: What is it and where are we?", *Journal of Clinical Investigation*, 111(1), pp. 3-10,2003.

[2] Y.Xu, X.Wei, X.Liu, J.Liao, J.Lin, C.Zhu ...andM.Cheng, "Low cerebral glucose metabolism: a potential predictor for the severity of vascular Parkinsonism and Parkinson's disease", *Aging and disease*, 6(6), pp. 426-436, 2015.

[3] K.Tjaden, "Speech and swallowing in Parkinson's disease.", *Topics in geriatric rehabilitation*, 24(2), pp. 115-126, 2008.

[4] A. K. Ho, R. Iansek, C. Marigliani, J. L. Bradshaw, and S. Gates, "Speech impairment in a large sample of patients with Parkinson's disease", *Behavioural Neurology*, 11(3), pp. 131–137,1998.

[5] Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., & Ramig, L. O. (2009). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE Transactions on Biomedical Engineering*, 56(4), 1015–1022.

[6] Rahn, D. A., Chou, M., Jiang, J. J., & Zhang, Y. (2007). Phonatory impairment in Parkinson's disease: evidence from nonlinear dynamic analysis and perturbation analysis. *Journal of Voice*, 21, 64–71.

[7] T. Hastie, R. Tibshirani, and J. H. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction: *With 200 Full-Color Illustrations. New York: Springer-Verlag*, 2001.

[8] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res*., vol. 3, pp. 1157–1182, 2003.

[9] D.Gil, and D.J.Manuel, "Diagnosing Parkinson by using artificial neural networks and support vector machines", *Global Journal of Computer Science and Technology*, 9(4), pp.63-71, 2009.

[10] W.Froelich, K.Wrobel, and P. Porwik, "Diagnosis of Parkinson's disease using speech samples and threshold-based classification", *Journal of Medical Imaging and Health Informatics*, 5(6), pp.1358-1363, 2015.

[11] M.Hariharan, K.Polat, and R.Sindhu, A new hybrid intelligent system for accurate detection of Parkinson's disease. *Computer methods and programs in biomedicine*, 113(3), pp.904-913, 2014.

[12] Shahbaba, B., & Neal, R. (2009). Nonlinear models using Dirichlet process mixtures. *The Journal of Machine Learning Research*, 10, 1829–1850..

[13] Sakar, C. O., & Kursun, O. (2010). Telediagnosis of Parkinson's disease using measurements of dysphonia. *Journal of Medical Systems*, 34, 1–9

[14] Li, D. C., Liu, C. W., & Hu, S. C. (2011). A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets. *Artificial Intelligence in Medicine*, 52, 45–52

[15] Spadoto, A. A., Guido, R. C., Carnevali, F. L., Pagnin, A. F., Falcao, A. X., & Papa, J. P. (2011). Improving Parkinson's disease identification through evolutionarybased feature selection. *In Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the* IEEE (pp. 7857–7860).

[16] Luukka, P. (2011). Feature selection using fuzzy entropy measures with similarity classifier. *Expert Systems with Applications*, 38, 4600–4607

[17] AStröm, F., & Koker, R. (2011). A parallel neural network approach to prediction of Parkinson's Disease. *Expert Systems with Applications*, 38, 12470–12474.

[18] M. Nilashi, O. Ibrahim, A. Ahani, "Accuracy Improvement for Predicting Parkinson's Disease Progression," *Scientific Reports*,vol.6,34181,2016

[19] Abdulhay, E., Arunkumar, N., Narasimhan, K., Vellaiappan, E., & Venkatraman, V. (2018). Gait and tremor investigation using machine learning techniques for the diagnosis of Parkinson disease. *Future Generation Computer Systems.*

[20] Er, F., Iscen, P., Sahin, S., Çinar, N., Karsidag, S., & Goularas, D. (2017). Distinguishing age-related cognitive decline from dementias: A study based on machine learning algorithms. *Journal of Clinical Neuroscience*, 42, 186-192.

[21] Zeng, W., Liu, F., Wang, Q., Wang, Y., Ma, L., & Zhang, Y. (2016). Parkinson's disease classification using gait analysis via deterministic learning. *Neuroscience letters*, 633, 268-278.

[22] Armañanzas, R., Bielza, C., Chaudhuri, K. R., Martinez-Martin, P., & Larrañaga, P. (2013). Unveiling relevant non-motor Parkinson's disease severity symptoms using a machine learning approach. *Artificial intelligence in medicine*, 58(3), 195-202.

[23] Kubota, K. J., Chen, J. A., & Little, M. A. (2016). Machine learning for large scale wearable sensor data in Parkinson's disease: Concepts, promises, pitfalls, and futures. *Movement disorders*, 31(9), 1314-1326.

[24] Abós, Alexandra, Hugo C. Baggio, Bàrbara Segura, Anna I. García-Díaz, Yaroslau Compta, Maria José Martí, Francesc Valldeoriola, and Carme Junqué. "Discriminating cognitive status in Parkinson's disease through functional connectomics and machine learning." *Scientific reports 7* (2017): 45347.

[25] Singh, G., & Samavedham, L. (2015). Unsupervised learning-based feature extraction for differential diagnosis of neurodegenerative diseases: a case study on early-stage diagnosis of Parkinson disease. *Journal of neuroscience methods*, 256, 30-40.

[26] M. A.Little, P.E. McSharry, E. J.Hunter, J.Spielman, and L. O.Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease", *IEEE transactions on biomedical engineering,* 2009, 56(4), pp.1015-1022.

[27] Guo, Q., Wu, W., Massart, D. L., Boucon, C., & De Jong, S. (2002). Feature selection in principal component analysis of analytical data. *Chemometrics and Intelligent Laboratory Systems*, 61(1-2), 123-132.

[28] Bledsoe, W. W. (1961). The use of biological concepts in the analytical study of systems. *In the ORSA-TIMS National Meeting*.

[29] Holland, J. H. (1992). Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. *MIT press.*

[30] H.B.Wong, G.H.Lim, Measures of diagnostic accuracy: sensitivity, specificity, PPV and NPV, *Proceedings of Singapore healthcare,* vol.20, no.4, pp.316-318, 2011.

Satyabrata Aich is working as a researcher in the field of computer engineering He has over four years of teaching, research and industry experience in India and abroad. He has published many research papers in journals and conferences in the realms of Supply Chain Management and data analytics. His research interests are natural language processing, Machine learning, supply chain management, data mining.

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. His primary concerns are u-Healthcare, smart home technology, e-learning and Human Computer Interaction (HCI). He is associate professor at Department of Computer Engineering, Inje University in Korea. He has published many research papers in journals and conferences in the realms of HCI and CSCW.

Young a Kim is a licensed Physical therapist and has received her Masters degree in Counseling and Psychology in 2008. She is working as a lecturer for Clinical Kinesiology and Functional Anatomy at Inje University. She is also a PhD candidate in the Department of Rehabilitation Science at the same University. Her research focuses on assessment tools and rehabilitation for the elderly, especially for the detection and treatment of age related conditions. She is a co-author of the book 'Community Occupational Therapy' and a frequent contributor to the journal of 'Physical Therapy Science' and a member in the Korean Ageing Friendly Industry Association as well as the Society of Occupational Therapy for the Aged and Dementia.

Kueh Lee Hui is working as an assistant professor at the department of Electrical Engineering, Dong-A University since 2012. She completed her PhD Degrees from Department of Electrical Engineering, Dong-A University, Korea. In 2009 she completed her BS degree in Electronic and Communication, Department of Electronic Engineering, University Malaysia of Sarawak, Malaysia. She also done MS in 2007 from Malaysia. Her research interests are image processing, face recognition, digital image forensic, intelligent control and control application, power system.

Ahmed Abdulhakim Al-Absi is an assistant professor in Department of Computer Engineering (Smart Computing) at Kyungdong University in South Korea. He earned a Ph.D. in computer science from Dongseo University in 2015. He received M.Sc. degree in information technology at University Utara Malaysia in 2011, and B.Sc. degree in computer applications at Bangalore University in 2008. His research interests include Big Data processing, Hadoop, Cloud computing, IoT, Distributed systems, Parallel computing, Bioinformatics, Security, and VANETs.

Mangal Sain received the M.Sc. degree in computer application from India in 2003 and the Ph.D. degree in computer science in 2011. Since 2012, he has been an Assistant Professor with the Department of Computer Engineering, Dongseo University, South Korea. His research interest includes wireless sensor network, cloud computing, Internet of Things, embedded systems, and middleware. He has authored over 50 international publications including journals and international conferences. He is a member of TIIS and a TPC member of more than ten international conferences.