Predicting Neurodegenerative Diseases Using a Novel Blood Biomarkers-based Model by Machine Learning

Shu-I Chiu [a]
sichiu@ntu.edu.tw

Chin-Hsien Lin [b]
chlin@ntu.edu.tw

Wee Shin Lim [a]
leolim3092@csie.ntu.edu.tw

Ming-Jang Chiu [b]
mjchiu@ntu.edu.tw

Ta-Fu Chen [b]
tfchen@ntu.edu.tw

Jyh-Shing Roger Jang [a]
jang@csie.ntu.edu.tw

[a] Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
[b] Department of Neurology, National Taiwan University Hospital, College of Medicine, National Taiwan University, Taipei, Taiwan

Abstract—**This paper presents machine learning based framework to the analysis and modeling of several neurodegenerative diseases by using features from blood-based biomarkers. The proposed approaches can be employed for early detection of Alzheimer's disease (AD) or Parkinson's disease (PD). In particular, we applied LDA (linear discriminant analysis) for visualizing the dataset as 2D or 3D scatter plots. Moreover, we constructed various classifiers for several different tasks of classification, and explore the accuracy of these classifiers. Based on our experiments, random forests are in general a very good choice of these tasks considering both the computing time (during modeling and prediction) and the accuracy.**

Keywords— Linear discriminant analysis; Classification; Multivariate imputation by chained equations; Neurodegenerative disease; Biomarkers

I. INTRODUCTION

People worldwide are living longer. Today, 125 million people are aged 80 years or older. According to World Health Organization (WHO), the number of people over 60 years is expected to rise from 900 million in 2015 to more than 2 billion in 2050. Similarly, approximately 47 million people are living with dementia worldwide, and this number is expected to increase to 82 million in 2030 and 150 million by 2050. The pace of population aging around the world is also increasing dramatically. As populations get older, age-related neurodegenerative diseases such as Alzheimer's disease (AD) and Parkinson's disease (PD) have become more common [1].

AD is the most important cause of dementia in the elder population and the pathological hallmarks are intraneuronal tau accumulations as neurofibrillary tangles and amyloid plaques depositions. Elderly people would progress to mild cognitive impairment (MCI) and then to the extent of clinically significant cognitive decline of AD presentation. PD is the second most common neurodegenerative disorder. Pathologically, PD is characterized by formation of intracellular α-synuclein containing Lewy bodies in the dopaminergic neurons of substantia nigra. Notably, PD patients deteriorate not only in their motor aspects but also in cognitive function, which is defined as PD with dementia (PDD). Mild cognitive impairment in PD (PD-MCI) refers to the stage between normal cognitive (PD-NC) functioning and PDD. In addition, the second most common dementia, frontotemporal dementia (FTD), which is characterized by intraneuronal phosphorylated tau depositions, is often clinically difficult to differentiate with AD or PD.

Given the likely entry of several classes of mechanism-targeted therapies for mitigating neurodegeneration in AD or PD into early human clinical trials, the identification of easily accessible biomarkers that could reflect disease severity is urgently needed. These neuropathology-related proteins are present in human body fluids including cerebrospinal fluid (CSF) and blood plasma [2, 3], which are good candidates for surrogate biomarker for disease severity in AD and PDD.

Machine learning algorithms are broadly applied to support healthcare systems such as early diagnosing, precision medicine, and genetic screening [4, 5, 6]. In this paper, we use these plasma biomarkers as features to classify neurodegenerative diseases based on machine learning.

Two major contributions of our work are:

- We use linear discriminant analysis (LDA) algorithm to obtain individual differences in various neurodegenerative diseases for diagnostic verification. The result can distinguish degree of deterioration between neurodegenerative diseases.
- Our model constructed by machine learning can effectively differentiates the disease groups and reflect the disease severity in either AD or PD group.

We will briefly review the related works in Section 2, describe the proposed methods in Section 3, depict experiments in Section 4, and finally conclude this paper in Section 5.

978-1-7281-4666-9/19/$31.00 ©2019 IEEE

II. Related Work

We previously have collaborated with MagQu company, which developed an ultra-sensitive immunoassay utilizing immuno-magnetic reduction (IMR) method [7, 8, 9] that could quantitatively detect biomolecules at ultra-low concentrations with a fg/ml limit of detection. IMR is a method to assay target molecules via measuring the reduction in the mixed frequency magnetic susceptibility of magnetic reagent due to binding of the target proteins to magnetic nanoparticles. We already have a longitudinal follow-up cohort of AD and PD patients. By using the IMR method, we have established the platform to detect plasma levels of disease-related proteins, including amyloid beta 42 (Aβ42), amyloid beta 40 (Aβ40), total tau, phosphorylated tau (p-tau181) and α-synuclein [7, 8, 9, 10]. However, differentiating different neurodegenerative disorders is difficult, especially in the early disease stages. We therefore aim to develop a machine learning-based model using plasma biomarker data collected from more than 400 participants of normal aging, AD or PD spectrum, and FTD, to predict and differentiate different neurodegenerative disorders.

Machine learning algorithms are broadly applied to support healthcare systems such as early diagnosing, precision medicine, and genetic screening [4, 5, 6]. With the advancing of computing power, medical images such as magnetic resonance imaging (MRI) and positron emission tomography (PET-CT) are widely used in machine learning for neurodegenerative disorder studies. These studies include diagnosing of AD, PD, Huntington's disease, and other types of neurodegenerative diseases [11, 12, 13, 14, 15, 16, 17].

Most of the tasks, if not applied with machine learning are considered time-consuming and required highly experience individual to perform the diagnose [18]. Machine learning supported diagnose not only increase the effectiveness but also decreases the chance of misdiagnosing. However, the acquisition of fMRI and CT images is expensive. Furthermore, diagnosing neurodegenerative disease depend only with images data is inadequate and sometimes misleading. Other types of medical data such as gene, Electroencephalography (EEG) and medical history records are also used in diagnosing neurodegenerative disease [19, 20, 21]. Several studies are using next-generation sequencing and machine learning to screen for candidate miRNA, which can be used as diagnosing tools for specific type of neurodegenerative disease. Most of these tasks are screening for the change of miRNA expression level in blood or other body fluid by comparing patient cases and controls. Statistical analysis is used in these studies, and large-scale dataset is handled with machine learning approaches [22, 23, 24].

In [25], Kruthika et al. disclose a multistage classifier, including Naive Bayes (NB) classifier, support vector machine (SVM) and k-Nearest Neighbor (kNN) to classify MRI data for detecting AD. Ahmed et al. apply machine learning to create a model for MRI data, MCI and AD before symptoms occur during 2015 to 2018 [26]. In [26], MRI of multi-type dementia is used to reduce the dimensionality of extracted key features using PCA. In contrast, this paper applies LDA to reduce the dimensionality of blood biomarkers features. In addition, we distinguish degree of deterioration between neurodegenerative diseases using machine learning algorithms.

III. Method

Dataset Descriptions: This study uses a dataset of 377 plasma samples from patients with or without neurodegenerative diseases, who visited National Taiwan University Hospital (which is a tertiary referral center in Taiwan) from 2012 to mid 2019. The dataset is divided into 7 classes as the outputs (or ground truth) to be predicted, including healthy individuals serving as controls ("Normal"), and patients with PD AD, MCI, PDD, PD-MCI, PD-NC, and FTD. The size of each output is shown in Table 1. Moreover, there are 7 features (inputs) of the dataset, including basic information (age and gender) and blood biomarkers (tau, p-Tau181, Aβ40, Aβ42, and α- synuclein).

TABLE 1 The size of each class

Class	*Normal*	*AD*	*MCI*	*PDD*	*PD-MCI*	*PD-NC*	*FTD*
size	97	35	41	87	29	57	31

There are some missing values in the dataset due to incomplete measurements. Since the dataset is not big and each entry is valuable, we use MICE (multivariate imputation by chained equations) [27] to do data imputation. In particular, we use CART (classification and regression trees) [28] as the model to predict each of the missing values. Moreover, we perform the following two operations of data adjustment to make the dataset more compliant for machine learning:

- The values of α-synuclein are too small, so we put them through the logarithm function.
- To make each feature have a similar range, we put each feature into a linear min-max normalization such that each feature has a minimum value of 0 and a maximum value of 1.

The workflow of the above data preprocessing is illustrated in Fig. 1.

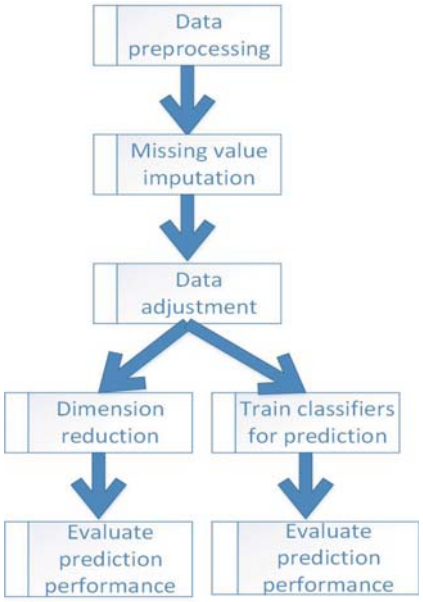


Fig. 1. Data processing

Dimensionality reduction: In statistics and machine learning, dimensionality reduction is the process of reducing the number of features such that the characteristics of the reduced dataset can be retained as much as possible. Approaches of dimensionality reduction can be divided into feature selection and feature extraction. In this paper, we employ LDA (linear discriminant analysis) to perform visualization of scatter plots in 2D or 3D. Such reduction can sometimes lead to a better classification accuracy since it can avoid the effects of the curse of dimensionality.

IV. Experiments

First, we perform dimensionality reduction by using LDA to create scatter plots in 2D or 3D. As suggested by physicians, for ease of analysis and interpretation, we define 3 different classification tasks, as shown in Table 2.

TABLE 2 The three tasks

Name of classification tasks	***Classes of the tasks***			
Individual dementia syndromes	Normal	AD	FTD	PDD
AD spectrum	Normal	MCI	AD	
PD spectrum	Normal	PD-NC	PD-MCI	PDD

For the task of individual dementia syndromes, LDA reduces the dimensionality to 3. Fig. 2 shows the 3D scatter plot of the samples.

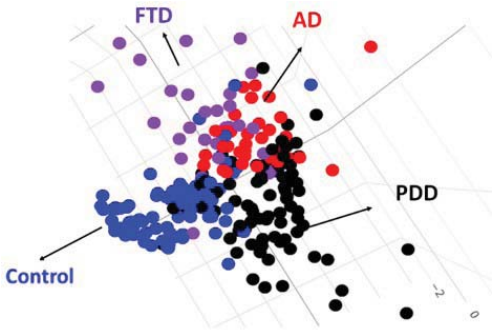


Fig. 2. 3D cluster plot for individual dementia syndromes

For the task of AD spectrum, LDA reduces the dimensionality to 2. Fig. 3 is the scatter plot, which clearly demonstrate that sample points of different classes are more or less separated due to LDA which takes class labels into consideration for feature extraction.

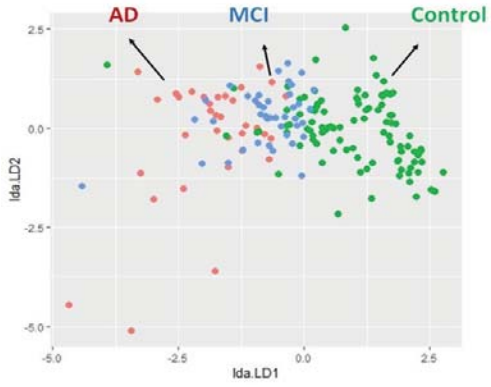


Fig. 3. 2D cluster plot for AD spectrum

For the task of PD spectrum, LDA reduces the dimensionality to 3. Fig. 4 shows the 3D scatter plot. These 3 best selected (transformed) features places the samples in a 3D space that are easier for further classification.

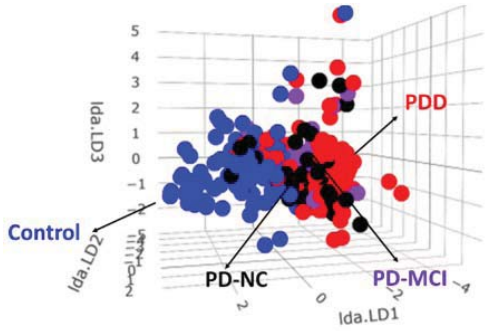


Fig. 4. 3D cluster plot for PD spectrum

Regarding classification algorithms, we use those top classifiers use in [29], including NB [30, 31], kNN [32, 33], SVM [34], C4.5 decision tree (C4.5) [35], and CART. Previous studies [36, 37] also use these classification algorithms for

prediction. For completeness, we use two more classifiers of random forests and logistic regression. Random forests (RFs) are a popular ensemble method that can build models for classification and regression efficiently [38]. Logistic regression (LogReg) can be used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables. We use these 7 classifiers (i.e. SVM, CART, C4.5, NB, LogReg, kNN, and RF) and compare their performance in terms of accuracy for multiclass classification, or AUROC (area under the receiver operating characteristic curve) for binary classification.

Furthermore, we use leave-one-out cross-validation (LOOCV) to have an objective estimate of the performance of our model construction procedure. This is particularly important since our dataset is not too big for LOOCV, which is the most unbiased version of cross validation. LOOCV is essentially an estimate of the generalization performance of a model trained on n−1 samples of data, which is generally a slightly pessimistic estimate of the performance of a model trained on all n samples.

We compare the accuracy of the three classification tasks (i.e. individual dementia syndromes, AD spectrum, and PD spectrum) using LDA algorithm for every category in Fig. 5. It shows that the normal category performs better than others; the accuracy is higher than 0.9. The second is PDD; the accuracy is close to 0.8. PD-NC and PD-MCI would not perform well. AD is sensitive to these tasks. For AD spectrum, the accuracy is 0.57; for individual dementia syndromes, the accuracy is 0.37.

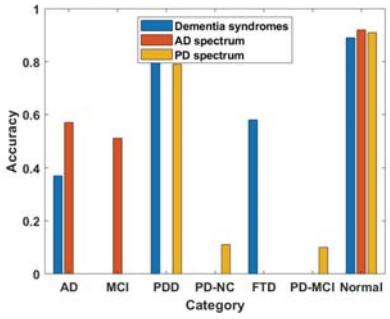


Fig. 5. Accuracy for the three tasks

We can use LDA to perform linear transformation on the features to better classify neurodegenerative diseases. Fig. 6 shows accuracies of the tasks of individual dementia syndromes. In general, for this task, RF is the best classifier and the accuracy with 3 transformed features is over 0.76.

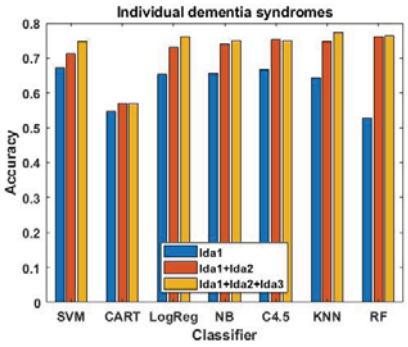


Fig. 6. The accuracy of individual dementia syndromes

For the tasks of AD spectrum, the accuracy is showed in Fig. 7. Again, RF is the best classifiers and the accuracy with 2 (transformed) features is over 0.83.

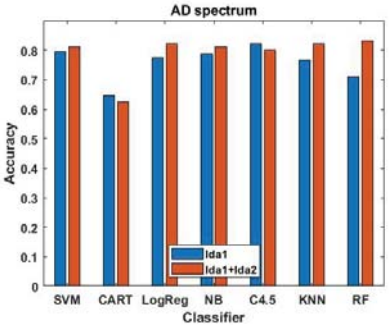


Fig. 7. The accuracy of AD spectrum

For the task of PD spectrum, the accuracy is showed in Fig. 8. LDA is the best classification algorithm and the accuracy with 2 linear discriminant variables is 0.68. However, the accuracy with 3 linear discriminant variables is lower than the accuracy with 2 ones. LDA classification algorithm with only 2 linear discriminant variables has the best performance.

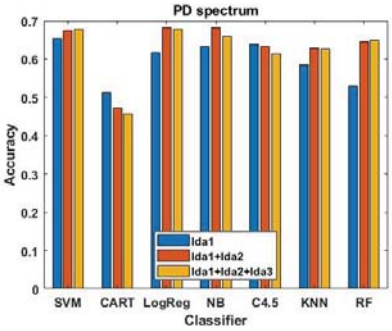
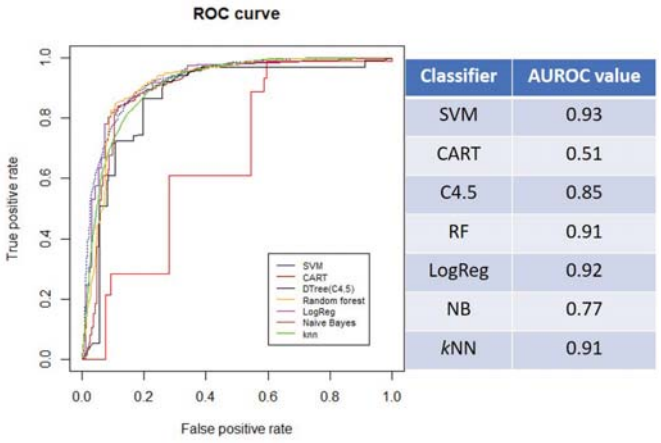


Fig. 8. The accuracy of PD spectrum

In addition, we define another set of classification tasks, from easy to hard ones, as shown in Table 3, to classify neurodegenerative diseases. The task set includes task 1 of two classes, task 2 of 4 classes, and task 3 of 7 classes. First of all, task 1 has two classes, including healthy samples (i.e. normal category) and plasma samples from patients with AD, PDD, FTD, MCI, PD-NC, or PD-MCI. Fig. 9 shows the ROC curve for the binary classes. SVM, logistic regression, random forest

and *k*NN algorithms all perform quite well, with AUROC values over 0.9. The best classifier is SVM while the worst classifier is CART. Tree-based algorithms such as CART and decision tree do not perform well except random forest.



Classifier	AUROC value
SVM	0.93
CART	0.51
C4.5	0.85
RF	0.91
LogReg	0.92
NB	0.77
*k*NN	0.91

Fig. 9. ROC curve for binary classes

TABLE 3 THE THREE CLASSIFICATION TASKS

Tasks	***Classes***						
Task 1	Normal	AD, PDD, FTD, MCI, PD-NC, and PD-MCI					
Task 2	Normal	AD group (i.e. AD and MCI)		PD group (i.e. PDD, PD-NC, and PD-MCI)			FTD
Task 3	Normal	AD	MCI	PD-NC	PD-MCI	PDD	FTD

Fig. 10 shows the accuracy of tasks 1, 2, and 3, respectively. In general, the best classifier is RF; the worst is CART. For task 1, the accuracy of each classifier is over 0.75. The best classifiers are RF and SVM with accuracy up to 0.89. For task 2, all classifiers yield a 0.7 accuracy excluding CART. For task 3, the best classifier is RF with an accuracy up to 0.6. But some classifiers do not perform well, with accuracy close to 0.5. For the task 3, each class has fewer data, which could lead to poor performance of this task.

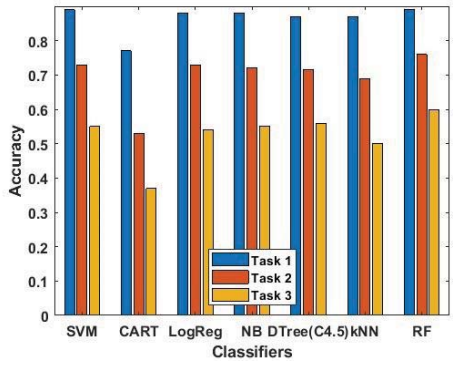


Fig. 10. Accuracy for the three tasks

V. CONCLUSIONS

This paper presents several machine learning based approaches to the visualization and classification of neurodegenerative diseases. As the first step, we use MICE for imputing missing values to avoid the loss of valuable data. Then we use LDA for dimensionality reduction such that the samples can be shown as 2D or 3D scatter plots for better visualization. Moreover, we defined several important classification tasks which are meaningful and interpretable to physicians, and employed various classifications to achieve different levels of performance. Several better performed classifiers are suggested for such classification tasks. Potential future directions involve the use of feature selection which can prioritize features for classification. Moreover, in-depth error analysis should be performed extensively in order to boost the accuracy to the next level.

ACKNOWLEDGMENT

The work presented in this paper was supported in part by the Ministry of Science and Technology, R.O.C., under grant number 107-2634-F-002-015. We are grateful to the Department of Neurology, National Taiwan University Hospital for the support. The authors acknowledge the support.

REFERENCES

[1] I. P. Johnson, "Age-related neurodegenerative disease research needs aging models," Frontiers in aging neuroscience7: 168, 2015.

[2] M. J. Chiu, et al., "Combined plasma biomarkers for diagnosing mild cognition impairment and Alzheimer's disease," ACS chemical neuroscience, vol. 4, no. 12, pp. 1530-1536, 2013.

[3] M. J. Chiu, et al., "Plasma α-synuclein predicts cognitive decline in Parkinson's disease," J Neurol Neurosurg Psychiatry, vol. 88(10), pp. 818-824.

[4] S.S. Johnston, et al., "Using Machine Learning Applied to Real-World Healthcare Data for Predictive Analytics: An Applied Example in Bariatric Surgery," Value Health, vol. 22, no. 5, pp. 580-586, 2019.

[5] T. Hulsen, et al., "From Big Data to Precision Medicine," Front Med (Lausanne), vol. 6, pp. 34, 2019.

[6] P. Hunter, "The advent of AI and deep learning in diagnostics and imaging: Machine learning systems have potential to improve diagnostics in healthcare and imaging systems in research," EMBO Rep, vol. 20, no. 7, p. e48559, 2019.

[7] C. C. Yang, et al., "Biofunctionalized magnetic nanoparticles for specifically detecting biomarkers of Alzheimer's disease in vitro," ACS chemical neuroscience, vol. 2, no. 9, pp. 500-505, 2011.

[8] M. J. Chiu, M. et al., "Plasma tau as a window to the brain—negative associations with brain volume and memory function in mild cognitive impairment and early alzheimer's disease," Human brain mapping, vol. 35, no. 7, pp. 3132-3142, 2014.

[9] M. J. Chiu, et al., "New assay for old markers-plasma beta amyloid of mild cognitive impairment and Alzheimer's disease," Current Alzheimer Research , vol. 9, no. 10, pp. 1142-1148, 2012.

[10] C. C. Lin, et al., "Plasma α-synuclein predicts cognitive decline in Parkinson's disease," J Neurol Neurosurg Psychiatry, vol. 88, no. 10, pp. 818-824, 2017.

[11] T.A. Shaikh and Ali R, "Automated atrophy assessment for Alzheimer's disease diagnosis from brain MRI images," Magn Reson Imaging, vol. 62, pp. 167-173, 2019.

[12] S. Klöppel, et al., "Automatic detection of preclinical neurodegeneration: presymptomatic Huntington disease," Neurology, vol. 72, no. 5, pp. 426-431, 2009.

[13] E. Moradi, A. Pepe, C. Gaser, H. Huttunen, J. Tohka, and Alzheimer's Disease Neuroimaging Initiative, "Machine learning framework for

early MRI-based Alzheimer's conversion prediction in MCI subjects. Neuroimage," vol. 104, pp. 398-412, 2015.

[14] J. Samper-Gonzalez, et al., "Reproducible evaluation of classification methods in Alzheimer's disease: Framework and application to MRI and PET data," NeuroImage, vol. 183, pp. 504-521, 2018.

[15] C. Salvatore, et al., "Machine learning on brain MRI data for differential diagnosis of Parkinson's disease and Progressive Supranuclear Palsy," Journal of Neuroscience Methods, vol. 222, pp. 230-237, 2014.

[16] M. Signaevsky, et al., "Artificial intelligence in neuropathology: deep learning-based assessment of tauopathy," Laboratory Investigation, vol. 99, pp. 1019-1029, 2019.

[17] A. Kautzky, R. Seiger, A. Hahn, P. Fischer,W. Krampla, S. Kasper, G. G. Kovacs, R. Lanzenberger, "Prediction of Autopsy Verified Neuropathological Change of Alzheimer's Disease Using Machine Learning and MRI," Frontiers in Aging Neuroscience, vol. 10, 2018.

[18] S. Klöppel, et al., "Automatic classification of MR scans in Alzheimer's disease," Brain vol. 131, no. 3, pp. 681-689, 2008.

[19] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," IEEE transactions on biomedical engineering, vol. 59, no. 5, pp. 1264-1271, 2012.

[20] B. M. Eskofier, et al., "Recent machine learning advancements in sensor-based mobility analysis: Deep learning for Parkinson's disease assessment," Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 655-658, 2016.

[21] A. Hall, et al., "Prediction models for dementia and neuropathology in the oldest old: the Vantaa 85+ cohort study," Alzheimer's research & therapy, vol. 11(1): 11, 2019.

[22] D. Santos, Marcia Cristina T., et al., "miRNA-based signatures in cerebrospinal fluid as potential diagnostic tools for early stage Parkinson's disease," Oncotarget, vol. 9, no. 25: 17455, 2018.

[23] A. Keller, et al., "Validating Alzheimer's disease micro RNAs using next-generation sequencing," Alzheimer's & Dementia vol. 12, no .5 pp. 565-576, 2016.

[24] P. Leidinger, et al., "A blood based 12-miRNA signature of Alzheimer disease patients," Genome biology vol. 14, no. 7, R78, 2013.

[25] K. R. Kruthika, H. D. Maheshappa, and Alzheimer's Disease Neuroimaging Initiative, "Multistage classifier-based approach for Alzheimer's disease prediction and retrieval," Informatics in Medicine Unlocked, vol. 14 pp. 34-42, 2019.

[26] M. R. Ahmed, Y. Zhang, Z. Feng, B. Lo, O. T. Inan, and H. Liao, "Neuroimaging and Machine Learning for Dementia Diagnosis: Recent Advancements and Future Prospects," IEEE reviews in biomedical engineering, vol. 12, pp. 19-33, 2018.

[27] S. V. Buuren, and K. Groothuis-Oudshoorn, "MICE: Multivariate imputation by chained equations in R," Journal of statistical software, pp. 1-68.

[28] L. Breiman, J. Friedman, R. A. Olshen, Stone CJ, Classification and regression trees, Wadsworth, Belmont, 1984.

[29] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," Knowl. Inf. Syst., vol. 14, no. 1, pp. 1-37, 2007.

[30] D. J. Hand and K. Yu, "Idiot's Bayes—not so stupid after all?," International Statistical Review vol. 69, no. 3, pp. 385-398, 2001.

[31] H. Zhang, "The optimality of naive Bayes," Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference, AAAI, pp. 562-567, 2004.

[32] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," IEEE Trans. Inf. Theor., vol. 13, no. 1, pp. 21-27, 2006.

[33] A. Mucherino, P. J. Papajorgji, and P. M. Pardalos, k-nearest neighbor classification, Data Mining in Agriculture, Volume 34 of the series Springer Optimization and Its Applications. Springer, New York, NY, pp. 83-106, 2009.

[34] I. Steinwart and A. Christmann, Support Vector Machines (1st ed.), Springer Publishing Company, Incorporated, 2008.

[35] J. R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, Inc., 1993.

[36] S. I. Chiu and K. W. Hsu, "Predicting Political Tendency of Posts on Facebook," Proceedings of the 2018 7th International Conference on Software and Computer Applications, pp. 110-114, 2018.

[37] C. C. Chang, S. I. Chiu, and K. W. Hsu, "Predicting political affiliation of posts on Facebook," Proceedings of the 11th International Conference on Ubiquitous Information Management and Communication, p. 57, 2017.

[38] L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, pp. 5-32, 2001.