Assessment for the Model Predicting of the Cognitive and Language Ability in the Mild Dementia by the Method of Data-Mining Technique

Haewon Byeon
Department of Speech Language
Pathology & Audiology
Nambu University
Gwangju, Republic of Korea

Dongwoo Lee
Department of Physical Therapy
Honam University
Gwangju, Republic of Korea

Sunghyoun Cho*
Department of Physical Therapy
Nambu University, Gwangju,
Republic of Korea

***Abstract*—Assessments of cognitive and verbal functions are widely used as screening tests to detect early dementia. This study developed an early dementia prediction model for Korean elderly based on random forest algorithm and compared its results and precision with those of logistic regression model and decision tree model. Subjects of the study were 418 elderly (135 males and 283 females) over the age of 60 in local communities. Outcome was defined as having dementia and explanatory variables included digit span forward, digit span backward, confrontational naming, Rey Complex Figure Test (RCFT) copy score, RCFT immediate recall, RCFT delayed recall, RCFT recognition true positive, RCFT recognition false positive, Seoul Verbal Learning Test (SVLT) immediate recall, SVLT delayed recall, SVLT recognition true positive, SVLT recognition false positive, Korean Color Word Stroop Test (K-CWST) color reading correct, and K-CWST color reading error. The Random Forests algorithm was used to develop prediction model and the result was compared with logistic regression model and decision tree based on chi-square automatic interaction detector (CHAID). As the result of the study, the tests with high level of predictive power in the detection of early dementia were verbal memory, visuospatial memory, naming, visuospatial functions, and executive functions. In addition, the random forests model was more accurate than logistic regression and CHIAD. In order to effectively detect early dementia, development of screening test programs is required which are composed of tests with high predictive power.**

Keywords—random forests; data mining; mild dementia; risk factors; neuropsychological test

I. Introduction

Dementia is rapidly increasing in line with worldwide aging. As of 2013, the global dementia population was over 44 million, and it is expected to increase by more than three times to 135 million by 2050 [1]. In particular, the dementia population in Korea is increasing the fastest in the world. That is, it was 610,000 as of 2014, and it is predicted to increase two-fold every 20 years, multiplying by more than four times and reaching 2.71 million by 2050 [2].

The increase of the dementia population is expected to lead to enormous social and economic costs by increasing medical costs and various supporting costs. According to a 2014 survey by the National Health Insurance Service, one of two (48.7%) recipients of long-term senior care insurance was a senior with dementia, and the annual medical cost for dementia per patient was reported to be US$ 2,650, which is more than that for cardiovascular diseases (US$ 1,130) and diabetes (US$ 505) [3]. In addition, the number of seniors who received treatment as outpatients increased from 8.2 persons per 100,000 in 1999 to 66.4 in 2010, which is around an eight-fold increase. Total supporting costs for dementia in Korea as of 2010 were estimated to be US$ 7.4 billion, and they are predicted to increase two-fold every 10 years and reach US$ 37.3 by 2050, exceeding 1.5% of GDP [4]. Measures must be taken, as the increase in the number of seniors with dementia leads to considerable losses, not only for the patients but also for supporting families, local communities and the country as a whole.

Dementia is known to be a disease associated with the gradual decline of cognitive functions for which full recovery is impossible. Reports suggest that cognitive decline in dementia can be postponed if cognitive functions are systematically managed with medicines, such as cholinesterase inhibitors, in the early stages of dementia [5]. Thus, the focus is now on the treatment and early detection of dementia. In particular, prolonging the onset of dementia, even for just two years, with early detection and treatment can lower its prevalence rate by 20% and decrease dementia patients' problem behaviours [6]. Thus, the early detection of dementia is crucial from a clinical perspective.

The early detection of dementia is performed based on interviews, standardised neuropsychological tests and neurological tests. Among them, neuropsychological tests composed of assessments of cognitive/verbal functions have been widely used as screening tests to detect early dementia [7]. In particular, as the usefulness of verbal ability for detecting dementia has been verified [8], verbal tests have been emphasised as effective screening tests for dementia. Nevertheless, few Korean studies have investigated the characteristics of the cognitive and verbal functions of the elderly using standardised neuropsychological test tools.

Meanwhile, as pattern analysis becomes possible on big data, data-mining analysis, which detects the possible onset of a disease by drawing out reliable conclusions based on data, is gaining attention in the healthcare area. In particular, random forest, which is a machine-learning algorithm using the bagging approach, has high accuracy and predictive power, because it predicts the final target variables after creating and combining multiple decision trees with random sampling [9, 10].

This study developed an early dementia prediction model for Korean seniors based on the random forest algorithm and compared its results and precision with those of a logistic regression model and decision tree model based on chi-square automatic interaction detection (CHAID).

This study is organised as follows: Section II describes the study participants and analysed variables, and Section III defines random forest and explains the model development procedure. Section IV compares the results of the developed prediction model with those of existing models. Lastly, Section V presents conclusions and suggestions for future studies.

II. Methods

A. Study participants

Data were collected from face-to-face interviews with voluntary participants aged 60–90 living in Seoul and Incheon. Subjects with depression and those taking medicines that hamper cognitive functions were excluded.

The seniors with cognitive impairment were selected as a group suspected of dementia by using the Korean-Mini Mental State Examination (K-MMSE) [11], and dementia was screened with the diagnostic standards for Alzheimer's dementia of the Diagnostic and Statistical Manual of Mental Disorders-Fourth Edition [12] and standards for probable Alzheimer's disease of NINCDS-ADRDA [13]. In this study, patients with mild dementia were defined as those scoring 0.5–1 point on the Clinical Dementia Rating Scale [14]. A total of 418 seniors (135 males, 283 females) were finally analysed.

B. Measurements

Cognitive and verbal ability was measured by Seoul Neuropsychological Screening Battery(SNSB)[15], which is composed of cognitive tests such as attention (digit span forward, digit span backward), verbal memory (Seoul Verbal Learning Test (SVLT) immediate recall, SVLT delayed recall, SVLT recognition true positive, SVLT recognition false positive), visuospatial memory (Rey Complex Figure Test (RCFT) copy score, RCFT immediate recall, RCFT delayed recall, RCFT recognition true positive, RCFT recognition false positive), language ability (confrontational naming), visuospatial functions (Rey Complex Figure Test(RCFT) copy score), and executive function (Korean Color Word Stroop Test (K-CWST) color reading correct, K-CWST color reading error).

Independent variables were including digit span forward, digit span backward, confrontational naming, RCFT copy score, RCFT Immediate recall, RCFT delayed recall, RCFT recognition true positive, RCFT recognition false positive, SVLT immediate recall, SVLT delayed recall, SVLT recognition true positive, SVLT recognition false positive, K-CWST color reading correct, K-CWST color reading error.

III. Statistical Analysis

A. Development of mild dementia prediction model

In order to develop the mild dementia prediction model, this study divided data into training data (70%) and test data (30%). The random forest algorithm was used to develop the prediction model, and the results of the developed prediction model were compared with those of a decision tree based on the CHAID algorithm. The accuracies of the developed models were evaluated with the correct classification rate, and the importance of variables and major factors drawn out were compared respectively.

B. Random forest model

The random forest model is a data-mining technique that combines multiple decision trees in an ensemble classifier [16]. Random forest is composed of a training stage, which constructs multiple decision trees, and a test stage, which makes classifications or predictions when there are input vectors [17] (Figure 1).

As random forest is based on decision trees, it has a fast learning speed and the ability to process a large amount of data [18]. In addition, random forest has a higher prediction capability than a decision tree, and it can prevent overfitting [19].

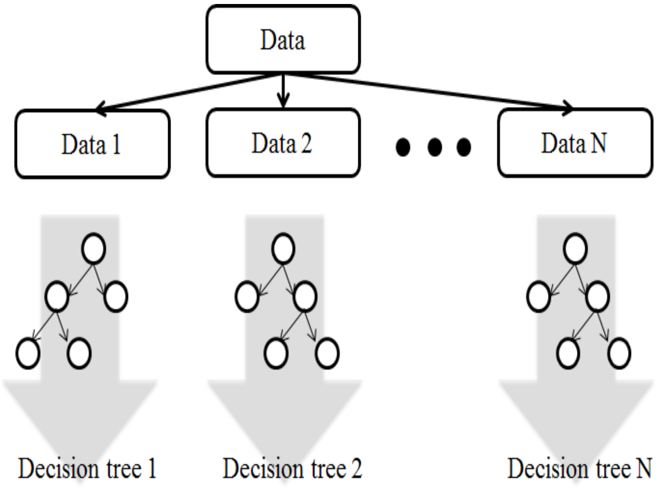

Fig. 1. Random Forests algorithm

IV. Results

A. General characteristics of participants

Among the 418 participants, 32.3% (n=135) were males and 67.7% (n=283) were females. The average age was 67.5 (standard deviation=4.3). Over 18.8% were high school graduates, and 76.5% were living with a spouse. Roughly 15.3% were current smokers, 26.5% were current drinkers and 33.8% exercised regularly (i.e. more than once a week). The prevalence rate of mild dementia was 8.4%.

B. *Results of neuropsychological test for healthy seniors and seniors with mild dementia*

The results of the neuropsychological test for healthy seniors and seniors with mild dementia are presented in Table 1. The results of the independent t-test revealed there were significant differences between healthy seniors and seniors with dementia for several factors. These included digit span forward, digit span backward, confrontational naming, RCFT copy score, RCFT immediate recall, RCFT delayed recall, RCFT recognition true positive, SVLT immediate recall, SVLT delayed recall, SVLT recognition true positive, SVLT recognition false positive, and K-CWST colour reading correct (p<0.05).

TABLE I. THE RESULTS OF NEUROPSYCHOLOGICAL TEST FOR HEALTHY ELDERLY AND ELDERLY WITH MILD DEMENTIA, MEAN±SD

Tests	Healthy elderly	Mild dementia
Attention		
Digit span forward*	5.34±1.78	4.35±1.13
Digit span backward*	3.83±1.25	2.65±1.11
Language function		
Confrontational naming*	40.83±12.18	31.52±8.84
Visuospatial function		
RCFT copy score*	26.73±9.01	21.88±10.81
Visuospatial memory		
RCFT immediate recall*	12.11±9.54	3.35±3.32
RCFT delayed recall*	10.88±8.35	2.45±2.86
RCFT recognition true positive*	9.12±2.53	7.38±3.15
RCFT recognition false positive	3.38±3.31	3.01±2.13
Verbal memory		
SVLT immediate recall*	17.35±6.31	10.08±3.56
SVLT delayed recall*	5.21±2.86	1.03±1.31
SVLT recognition true positive*	10.01±1.91	7.53±2.83
SVLT recognition false positive*	2.15±1.83	3.31±2.21
Executive functions		
K-CWST Color reading correct*	73.34±21.80	56.19±28.85
K-CWST Color reading error	4.89±6.31	9.36±10.38

*P<0.05

RCFT=Rey Complex Figure Test; SVLT=Seoul Verbal Learning Test; K-CWST=Korean Color Word Stroop Test

C. *Accuracy comparison among random forest, logistic regression model, and decision tree*

The prediction model was developed by using random forest, and its accuracy was compared with those developed using a logistic regression model and a decision tree (Table 2). The results of the analysis on the training data revealed that random forest showed very high accuracy of 72.5% (Figure 4, Figure 5). On the other hand, the accuracy of the decision tree was 71.2%, and the accuracy of the logistic regression model was the lowest with 68.7%.

In the test data, random forest showed the highest accuracy with 72.1%, while the logistic regression model had the lowest accuracy with 67.5%. Hence, random forest had the highest accuracy in both the training data and test data.

TABLE II. ACCURACY COMPARISON BETWEEN MODELS

Data	Model	Accuracy (%)
Training data	Logistic regression	67.9
	Decision tree	70.5
	Random Forests	73.7
Test data	Logistic regression	67.7
	Decision tree	70.8
	Random Forests	72.7

D. *Comparison of neuropsychological tests for prediction of dementia*

The results of the prediction models established based on a logistic regression model, a decision tree and random forest by using 14 neuropsychological tests to predict mild dementia are presented in Table 3.

In the logistic regression model, the prediction of mild dementia involved 12 tests, and its accuracy was 67.7%. These tests included digit span forward, digit span backward, confrontational naming, RCFT copy score, RCFT immediate recall, RCFT delayed recall, RCFT recognition true positive, SVLT immediate recall, SVLT delayed recall, SVLT recognition true positive, SVLT recognition false positive, and K-CWST colour reading correct.

The decision tree based on CHAID involved nine tests for the prediction of mild dementia, and its accuracy was 70.8%. These tests included digit span forward, digit span backward, confrontational naming, RCFT copy score, RCFT immediate recall, RCFT delayed recall, SVLT immediate recall, SVLT delayed recall, and K-CWST colour reading correct

Random forest involved 12 tests for the prediction of dementia, and its accuracy was 72.7%. These tests included digit span forward, digit span backward, confrontational naming, RCFT copy score, RCFT immediate recall, RCFT delayed recall, RCFT recognition true positive, SVLT immediate recall, SVLT delayed recall, SVLT recognition true positive, SVLT recognition false positive, and K-CWST colour reading correct.

TABLE III. COMPARISON OF NEUROPSYCHOLOGICAL TESTS FOR THE PREDICTION OF DEMENTIA

Model	Number of factors	Tests
Logistic regression	12	Digit span forward, digit span backward, confrontational naming, RCFT copy score, RCFT immediate recall, RCFT delayed recall, RCFT recognition true positive, SVLT immediate recall, SVLT delayed recall, SVLT recognition true positive, SVLT recognition false positive, K-CWST color reading correct

Decision tree	9	Digit span forward, digit span backward, confrontational naming, RCFT copy score, RCFT immediate recall, RCFT delayed recall, SVLT immediate recall, SVLT delayed recall, K-CWST color reading correct
Random Forests	12	Digit span forward, digit span backward, confrontational naming, RCFT copy score, RCFT immediate recall, RCFT delayed recall, RCFT recognition true positive, SVLT immediate recall, SVLT delayed recall, SVLT recognition true positive, SVLT recognition false positive, K-CWST color reading correct

V. Conclusion

The early diagnosis of dementia is important, because it not only reduces the number of cases that progress into dementia but also eases the individual and social burden of support for dementia patients.

As a result of the development of the early dementia prediction model for Korean seniors based on the random forest algorithm in this study, a number of factors were verified to be important indices in detecting mild dementia. These included digit span forward, digit span backward, confrontational naming, RCFT copy score, RCFT immediate recall, RCFT delayed recall, RCFT recognition true positive, SVLT immediate recall, SVLT delayed recall, SVLT recognition true positive, SVLT recognition false positive, and K-CWST colour reading correct.

Numerous studies have reported that verbal memory, visuospatial memory and naming are effective tests for distinguishing seniors with early dementia from healthy seniors [20, 21]. In particular, naming is known to be the most sensitive test for predicting the progress into dementia [8]. In addition, among the various neurological functions that decline with aging, delayed recall and selective attention have been reported to be the most sensitive items for predicting the onset of dementia from mild cognitive impairment [22].

Meanwhile, Artero et al. (2003) reported that the progress from mild cognitive impairment to dementia was best predicted when verbal memory and visuospatial ability were assessed together [23]. Moreover, in a cohort study on local communities, Dickerson et al. (2007) reported that the decline of not only verbal memory but also executive functions affects the progress into dementia [24]. These results imply that integrated assessment including verbal memory, visuospatial memory and performing ability is important in predicting cognitive decline and dementia in old age.

According to the results of the comparison of the accuracies of random forest, the logistic regression model and the decision tree, the accuracy of random forest was the highest. This is presumed to be because random forest is based on a bootstrap aggregating algorithm that creates various decision trees out of 500-odd bootstrap samples. While the decision tree has a risk of overfitting, random forest has higher accuracy than the decision tree, since it is based on a bootstrap aggregating algorithm that predicts target variables through means or probability [19, 25, 26]. Random forest is deemed to be more effective in conducting prediction analysis by using data with many variables to measure, since it draws out multiple training data, forms trees and predicts target variables.

The results of this study imply that verbal memory, visuospatial memory, naming, visuospatial functions and executive functions are cognitive domains that should be included before others in neuropsychological assessment to screen for mild dementia. In addition, in order to effectively detect early dementia, the development of screening test programmes composed of tests with high predictive power is required.

Acknowledgment

The authors wish to thank all of the subjects who participated in this study.

References

[1] World Health Organization and Alzheimer's Disease International, Dementia: A Public Health Priority 2013. Geneva, World Health Organization, 2013.

[2] Ministry of Health and Welfare, 2012 survey of dementia prevalence. Seoul, Ministry of Health and Welfare, 2013.

[3] National health insurance service, Organization of dementia care in 2014 basic education. Seoul, National Health Insurance Service, 2014.

[4] S. W. Kim, Status of dementia management. Seoul, National Assembly Budget Office, 2014.

[5] P. Anand, and B. Singh, A review on cholinesterase inhibitors for Alzheimer's disease. Archives of Pharmacal Research, vol. 36, no. 4, pp. 375–399, 2013.

[6] National Health Insurance Institute, Analysis of medical use of Alzheimer's disease 2013. Sejong, Ministry of Health and Welfare, 2013.

[7] S. Weintraub, D. Salmon, N. Mercaldo, S. Ferris, N. R. Graff-Radford, H. Chui, E. Peskind, W. Dietrich, D. L. Beekly, W. A. Kukull, and J. C. Morris, The Alzheimer's disease centers' uniform data set: The neuropsychological test battery. Alzheimer Disease & Associated Disorders, vol. 23, no. 2, pp. 91–101, 2009.

[8] F. Boller, J. T. Becker, A. L. Holland, M. M. Forbes, P. C. Hood, and K. L. McGonigle-Gibson, Predictors of decline in Alzheimer's disease. Cortex, vol. 27, no. 1, pp. 9–17, 1991.

[9] G. Biau, Analysis of a random forests model. The Journal of Machine Learning Research, vol. 13, no. 1, pp. 1063–1095, 2012.

[10] A. Shameem, and D. Manimeglai, Analysis of significant factors for dengue infection prognosis using the Random Forest Classifier. International Journal of Advanced Computer Science and Applications, vol. 6, no. 2, pp. 240–245, 2015.

[11] C. Han, S. A. Jo, I. Jo, E. Kim, M. H. Park, and Y. Kang, An adaptation of the Korean mini-mental state examination in elderly Koreans: demographic influence and population-based norms. Archives of gerontology and geriatrics, vol. 47, no. 3, pp. 302–310, 2008.

[12] American Psychiatric Association, Diagnostic and statistical manual of mental disorders American Psychiatric Association. Washington DC, American Psychiatric Association, 1994.

[13] B. Dubois, A. Slachevsky, I. Litvan, and B. Pillon, The FAB: a Frontal Assessment Battery atbedside. Neurology, vol. 55, no. 11, pp. 1621–1626, 2000.

[14] J. C. Morris, Clinical dementia rating: a reliable and valid diagnostic and staging measure for dementia of the Alzheimer type. International psychogeriatrics, vol. 9, no. 1, pp. 173–176, 1997.

[15] Y. Kang, and D. L. Na, Seoul neuropsychological screening battery. Incheon, Human brain research & consulting co, 2003.

[16] L. Breiman, Random forests. Machine learning, vol. 45, no. 1, pp. 5–32, 2001.

[17] S. N. Devi, and S. P. Rajagopalan, A study on feature selection techniques in bio-informatics. International Journal of Advanced Computer Science and Applications, vol. 2, no. 1, pp. 138–144, 2011.

[18] S. Hussain, and G. C. Hazarika, Educational data mining model using rattle. International Journal of Advanced Computer Science and Applications, vol. 5, no. 6, pp. 22–27, 2014.

[19] H. Byeon, A prediction model for mild cognitive impairment using random forests. International Journal of Advanced Computer Science and Applications, vol. 6, no. 12, pp. 8–12, 2015.

[20] M. C. Greenaway, L. H. Lacritz, D. Binegar, M. F. Weiner, A. Lipton, and C. M. Cullum, Patterns of verbal memory performance in mild cognitive impairment, Alzheimer disease, and normal aging. Cognitive and Behavioral Neurology, vol. 19, no. 2, pp. 79–84, 2006.

[21] M. D. Calero, M. L. Arnedo, E. Navarro, M. Ruiz-Pedrosa, and C. Carnero, Usefulness of a 15-item version of the Boston Naming Test in neuropsychological assessment of low-educational elders with dementia. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, vol. 57, no. 2, pp. 187–191, 2002.

[22] E. Mariani, R. Monastero, and P. Mecocci, Mild cognitive impairment: a systematic review. Journal of Alzheimer's Disease, vol. 12, no. 1, pp. 23-35, 2007.

[23] S. Artero, M.C. Tierney, J. Touchon, and K. Ritchie, Prediction of transition from cognitiveimpairment to senile dementia: a prospective, longitudinal study. Acta Psychiatrica Scandinavica, vol. 107, no. 5, pp. 390–393, 2003.

[24] B. C. Dickerson, R.A. Sperling, B.T. Hyman, M.S. Albert, and D. Blacker, Clinical prediction of Alzheimer disease dementia across the spectrum of mild cognitive impairment. Archives of General Psychiatry, vol. 64, no. 12, pp. 1443-1450, 2007.

[25] D. T. Larose, Discovering knowledge in data: an introduction to data mining. Hoboken, John Wiley & Sons, 2014.

[26] K. L. Lunetta, L. B. Hayward, J. Segal, and P. Van Eerdewegh, Screening large-scale association study data: exploiting interactions using random forests. BioMed Central genetics, vol. 5, no. 1, oi: 10.1186/1471-2156-5-32, 2004.