# A Prediction Model for Mild Cognitive Impairment Using Random Forests

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

***Abstract*—Dementia is a geriatric disease which has emerged as a serious social and economic problem in an aging society and early diagnosis is very important for it. Especially, early diagnosis and early intervention of Mild Cognitive Impairment (MCI) which is the preliminary stage of dementia can reduce the onset rate of dementia. This study developed MCI prediction model for the Korean elderly in local communities and provides a basic material for the prevention of cognitive impairment. The subjects of this study were 3,240 elderly (1,502 males, 1,738 females) in local communities over the age of 65 who participated in the Korean Longitudinal Survey of Aging (close) conducted in 2012. The outcome was defined as having MCI and set as explanatory variables were gender, age, level of education, level of income, marital status, smoking, drinking habits, regular exercise more than once a week, monthly average hours of participation in social activities, subjective health, diabetes and high blood pressure. The random Forests algorithm was used to develop a prediction model and the result was compared with logistic regression model and decision tree model. As the result of this study, significant predictors of MCI were age, gender, level of education, level of income, subjective health, marital status, smoking, drinking, regular exercise and high blood pressure. In addition, Random Forests Model was more accurate than the logistic regression model and decision tree model. Based on these results, it is necessary to build monitoring system which can diagnose MCI at an early stage.**

***Keywords*—random forests; data mining; dementia; mild cognitive impairment; risk factors***

## I. INTRODUCTION

As worldwide aged population increases with the development of science, technology and medicine, number of geriatric diseases increases radically as well. Especially, dementia, a typical geriatric disease, is expected to experience an unprecedentedly rapid increase. According to World Alzheimer Report (2015), worldwide dementia population recorded 44 million in 2013 and will increase more than 3-fold to 135 million in 2050 [1].

In Korea, dementia also increases rapidly due to fast aging. According to a survey on prevalence rate of dementia conducted by Ministry of Health and Welfare, the number of dementia patients in Korea was 540,000 in 2012 and the number is expected rapidly increase to 840,000 in 2020, 1.27 million in 2030 and 2.71 million in 2050 [2]. In particular, as Korea shows the most rapid rate of increase in the world, it is urgent to take measures for geriatric cognitive impairment [3].

Although treatment methods for dementia have been developed globally over the last 20 years, no treatment method developed so far can provide full recovery. It is only possible to postpone cognitive decline of dementia when cognitive function is managed systematically by using drugs such as donepezil [4]. As medication of this kind can produce greater effect with earlier application, early diagnosis and intervention is crucial in dementia.

Especially, as dementia incurs tremendous socio-economical costs, systematic management is required through early intervention. According to a report to Korean National Assembly, social cost for dementia patients is estimated to be U$ 37.3 billion in 2050, which amounts to 1.5% of GDP [3]. Thus, reduction of prevalence rate through early discovery of dementia can decrease unnecessary social and economic costs [5].

Like this, as early diagnosis of dementia becomes important, Mild Cognitive Impairment (MCI) which is a previous stage of dementia is gaining attention. MCI is defined as intermediate stage between normal aging and dementia with its decline of cognitive function out of normal range but its severity still not reaching the state of dementia [6]. MCI, an earliest stage to discover dementia, is important as a primary target for dementia treatment since its early discovery and treatment can postpone the progress of dementia.

Along with dementia, MCI is also on the rapid increase. MCI in Korea has increased 4.3-fold from 24,000 in 2010 to 105,000 in 2014, attracting attention to its early discovery and prevention [2].

Over the last 20 years, numerous studies have reported that risk factors of MCI were gender, age, smoking, drinking, eating habits, exercise, diabetes and hypertension [5, 7, 8, 9]. And opinions exist that there are limitations to explain the outbreak of MCI with these individual risk factors and studies report different results on affecting risk factors [9]. In addition, necessity to consider mental health such as depression is recently being raised in the search for factors related to MCI [10]. In particular, as there are differences among races in outbreak pattern of cognitive impairment and risk factors, it is necessary to develop MCI prediction model reflecting the living patterns of the Korean elderly.

Meanwhile, as analysis on big data becomes possible with the development of computer, attention is being paid to data mining techniques in developing prediction models. Data

mining is a method of analysis to predict data based on already known attributes by using training data [11, 12]. Especially, Random Forests developed as one analysis method of data mining has high level of prediction capability as it creates multiple decision trees by implementing random sampling in an identical data set, combines them and finally predicts target variables [13, 14]. In addition, Random Forests is known to have an excellent prediction capability in finding out correlation between explanatory variables and a disease and prevent overfit when there are many kinds of explanatory variables applied to the model [15, 16].

This study developed a prediction model of Mild Cognitive Impairment for the elderly in Korean local communities based on random forests algorithm and compared it with logistic regression model and decision tree model to verify its results and accuracy.

Construction of this study is as follows; chapter II explains study subjects and analyzed variables and chapter III defines random forests and explains the procedure of model development. Chapter IV compares the results of developed prediction model with those of logistic regression model and decision tree model. Lastly, chapter V presents conclusion and direction for future studies.

## II. METHODS

### A. Sources of data

This study analyzed a total of 3,240 elderly people (1,502 males and 1,738 females) over the age of 65 who participated in 2012 Korean Longitudinal Survey of Aging (KLoSA).

KLoSA is supervised by Korea Labor Institute and TNS Korea conducted the survey on commission from July 7, 2012 through December 2012 [17]. Sampling frame was districts of Population and Housing Census 2005 and 261,237 districts were set as sampling units. In 2012 survey, 10,000 people was set as maximum valid sample size and considering that average population over the age of 45 was 1.67 per household in 2000 Population and Housing Census, 1,000 sampling districts were selected. The method of the survey was computer-assisted personal interviews using laptop computers.

### B. Measurements

Outcome was defined as prevalence of MCI. Explanatory variables were included as gender, age (65-75, 75+), level of education (middle school and lower, over high school), level of income, marital status (have spouse, divorced or separation, separation by death), smoking (non-smoking, past smoking, current smoking), drinking habits (non-drinker, past drinker, current drinker), regular exercise more than once a week (yes, no), monthly average hours of participation in social activities (less than 1 hour, over 1 hour), subjective health (good, fair, bad), diabetes (yes, no), and hypertension (yes, no).

## III. STATISTICAL ANALYSIS

### A. Development and evaluation of model

In order to develop MCI prediction model, this study divided data into 70% of training data and 30% of test data. Random forest algorithm was used to develop prediction model and results of developed prediction model were compared with those of decision tree based on multivariate logistic regression analysis and CART (classification and regression tree). Accuracies of developed models were evaluated with correct classification rate, and importance of variables and major risk factors drawn out were compared respectively.

### B. Random Forests

Random forest is a type of ensemble classifier which randomly learns multiple decision trees and is composed of training stage which construct many decision trees and test stage which classifies and predicts incoming input data [18] (Figure1).

Ensemble form of training data can be expressed in Forest F = {f1, … , fn} (Figure2).

Distributions earned from decision trees of each forest are averaged by T, the number of decision trees, and finally classified.

$$L(p) = \frac{1}{T}\sum_{t=1}^{T} P_t(b|I,p)$$

Figure 3 shows bagging algorithm which creates final model by conducting *n* times of random sampling on raw data and combining prediction variables coming out of modeling of each sample. For combining method of prediction variables of each sample, average was used when a target variable was a continuous variable while majority vote was used when a target variable was a categorical variable.

Although random forest is similar to bagging in that in enhances stability by combining decision trees created in multiple bootstrap samples with majority principle, it is different from bagging in that it uses only a few explanatory variables randomly chosen from bootstrap samples. In order to adjust correlation of combination model, random forest establishes decision tree by randomly extracting several explanatory variables from boot strap samples and establishes a model with as few pruning as possible.

Random forest has higher prediction capability than decision tree and strength that it can prevent overfitting [19]. This study established random forest model first and then compared it with the results drawn out from multivariate logistic regression analysis and decision tree and accuracy of model respectively.

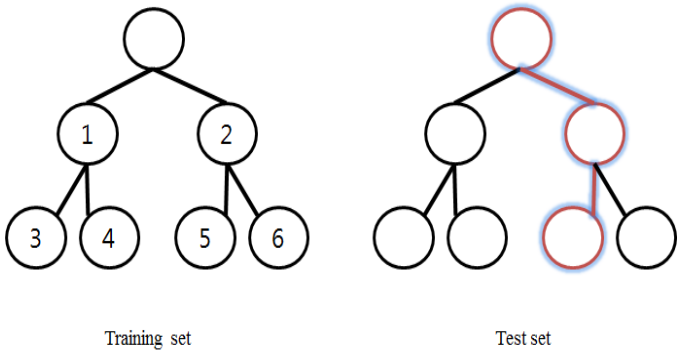


Fig. 1. Training and testing datasets

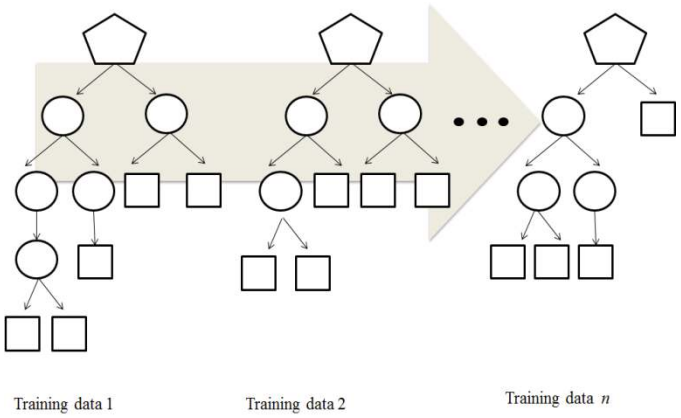


Fig. 2. Random forest: a classifier that combines many single decision trees

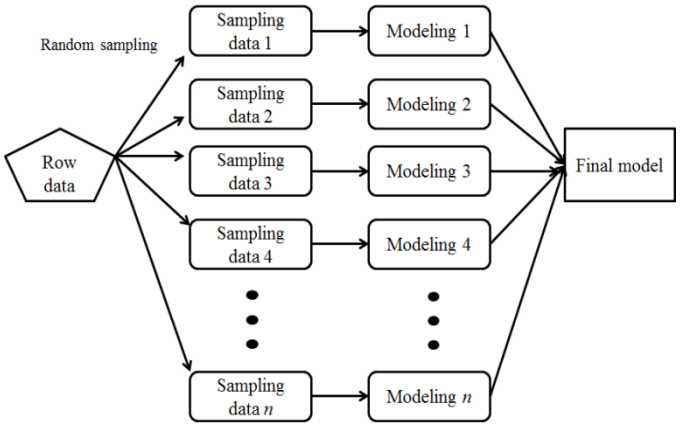


Fig. 3. Bagging algorithm

## IV. RESULTS

### A. Characteristics of subjects

Among total 4,134 subjects, males were 46.4% and females were 53.6%. Average age was 72 (range=65-99, standard deviation=5.9). As for level of education, high school graduates and over were 24.8% and those living with spouse were 71%. Current smokers were 13.1% and current drinkers were 28.8% while the elderly who exercise regularly more than once a week were 37.3%. 49.8% had hypertension and 20.9% had diabetes. Prevalence rate of MCI was 31.3%.

### B. Potential risk factors for Mild Cognitive Impairment (univariate analysis)

Potential risk factors for MCI are presented in Table 1. As the result of cross-tabulation analysis, there were significant differences between the normal elderly and those with MCI in gender, age, level of education, level of income, marital status, smoking, drinking habit, regular exercise of more than once per week, monthly average hour of participation in social activities, subjective health and hypertension ($p<0.05$).

Prevalence rate of MCI was high in the elderly over the age of 75 (43.7%), females (38.6%), middle school graduates and lower (36.4%), the bereaved of spouses (42.3%), nonsmokers (34.6%), nondrinkers (35.6%), elderly who do not exercise regularly (35.9%), elderly who participate in social activities less than average 1 hour per month (31.8%), elderly with poor subjective health (43.9%) and elderly with hypertension (34.3%).

TABLE I. GENERAL CHARACTERISTICS OF THE SUBJECTS BASED ON MCI (UNIVARIATE ANALYSIS), N (%)

| Characteristics | MCI | | p |
|---|---|---|---|
| | *Yes (n=1,015)* | *No (n=2,225)* | |
| Age | | | <0.001 |
| 65-75 | 475 (23.7) | 1,528 (76.3) | |
| 75+ | 540 (43.7) | 697 (56.3) | |
| Sex | | | <0.001 |
| Male | 345 (23.0) | 1,157 (77.0) | |
| Female | 670 (38.6) | 1,068 (61.4) | |
| Level of education | | | <0.001 |
| Middle school and lower | 887 (36.4) | 1,549 (63.6) | |
| Over high school | 128 (15.9) | 676 (84.1) | |
| Level of income (quartile) | | | <0.001 |
| First quartile | 478 (40.5) | 703 (59.5) | |
| Second quartile | 270 (29.1) | 659 (70.9) | |
| Third quartile | 174 (23.1) | 579 (76.9) | |
| Fourth quartile | 93 (24.7) | 284 (75.3) | |
| Marital status | | | <0.001 |
| Have spouse | 625 (27.2) | 1674 (72.8) | |
| Divorced/separation | 21 (30.9) | 47 (69.1) | |
| Separation by death | 369 (42.3) | 504 (57.7) | |
| Smoking | | | <0.001 |
| Non-smoking | 767 (34.6) | 1,449 (65.4) | |
| Past smoking | 141 (23.5) | 459 (76.5) | |
| Current smoking | 107 (25.2) | 317 (74.8) | |
| Drinking | | | <0.001 |
| Non- Drinking | 628 (35.6) | 1,138 (64.4) | |
| Past Drinking | 159 (29.4) | 382 (70.6) | |
| Current Drinking | 228 (24.4) | 705 (75.6) | |
| Regular exercise more than once a week | | | <0.001 |
| Yes | 287 (23.7) | 923 (76.3) | |
| No | 728 (35.9) | 1,302 (64.1) | |
| Monthly average hours of participation in social activities | | | 0.001 |
| Less than 1 hour | 996 (31.8) | 2,134 (68.2) | |
| Over 1 hour | 19 (17.3) | 91 (82.7) | |
| Subjective health | | | <0.001 |
| Good | 106 (17.4) | 502 (82.6) | |
| Fair | 392 (27.0) | 1,061 (73.0) | |
| Bad | 517 (43.9) | 662 (56.1) | |
| Hypertension | | | <0.001 |
| Yes | 554 (34.3) | 1,061 (65.7) | |
| No | 461 (28.4) | 1,164 (71.6) | |
| Diabetes | | | 0.185 |
| Yes | 226 (33.4) | 450 (66.6) | |
| No | 789 (30.8) | 1,775 (69.2) | |

### C. Accuracy comparison between models

Prediction model was developed by using random forests and its accuracy was compared with those of logistic regression model and decision tree (Table 2). As the result of analysis on training data, random forests showed very high accuracy of 72.5% (Figure 4, Figure 5). On the other hand, accuracy of decision tress was 71.2% and accuracy of logistic regression model was the lowest with 68.7%.

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

TABLE II. ACCURACY COMPARISON BETWEEN MODELS

| Data | Model | Accuracy (%) |
|---|---|---|
| Training data | Logistic regression | 68.7 |
| | Decision tree | 71.2 |
| | Random Forests | 72.5 |
| Test data | Logistic regression | 67.5 |
| | Decision tree | 70.8 |
| | Random Forests | 72.1 |

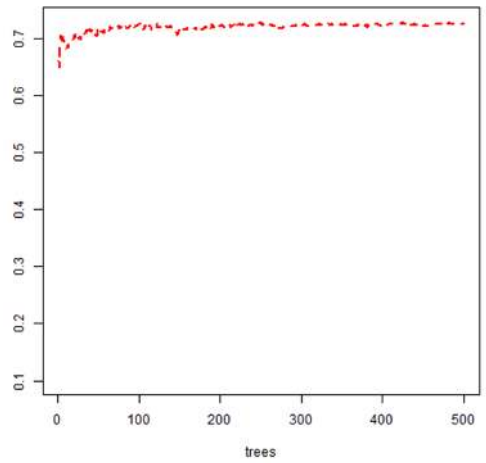


Fig. 4. Accuracy of Random Forests model

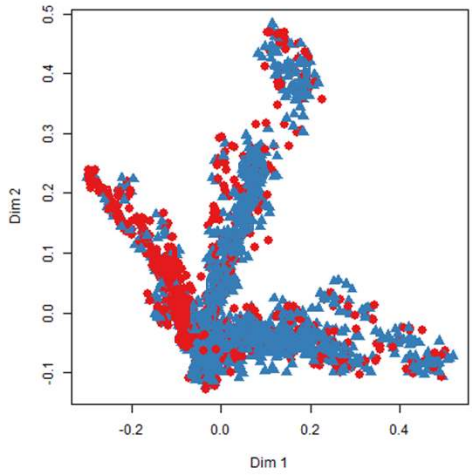


Fig. 5. Multidimensional scaling plot of proximities

### D. Comparison of risk factors based on prediction model

The results of prediction models established based on logistic regression model, decision tree and random forest by using a total of 12 explanatory variables to predict MCI are presented in Table 3. Among the prediction models used in this study, major risk factors for random forests model were presumed by using decrease of GINI coefficient.

In logistic regression model, risk factors for MCI were total of 7 variables, which were age, gender, level of income, hour of participation in social activities, subjective health and regular exercise and its accuracy was 67.5%.

Decision tree model predicted 8 variables as risk factors for MCI, which were age, gender, level of education, level of income, subjective health, marital status, smoking, regular exercise and its accuracy was 70.8%.

Random forests model predicted age, gender, level of education, level of income, subjective health, marital status, smoking, drinking, regular exercise and hypertension as risk factors for MCI, and its accuracy was 72.1%.

TABLE III. COMPARISON OF RISK FACTORS BASED ON PREDICTION MODEL

| Model | Number of factors | Risk factors |
|---|---|---|
| Logistic regression | 7 | age, gender, level of education, level of income, social activities, subjective health, regular exercise |
| Decision tree | 8 | age, gender, level of education, level of income, subjective health, marital status, smoking, regular exercise |
| Random Forests | 10 | age, gender, level of education, level of income, subjective health, marital status, smoking, drinking, regular exercise, hypertension |

## V. CONCLUSION

Early discovery of MCI is clinically important as it can postpone cognitive decline. This study developed MCI prediction model for the elderly in local communities by using random forest algorithm.

As the result of developing prediction model based on random forests, major risk factors for MCI were age, gender, level of education, level of income, subjective health, marital status, smoking, drinking, regular exercise and hypertension. Many studies have reported that socio-demographic factors such as old age and level of education and health risk behaviors such as smoking and drinking are risk factors of MCI [3, 5]. In particular, smoking was the most important variable except socio-demographic factors in the MCI prediction model of this study. Although smoking is a modifiable factor compared with socio-demographic factors like age and level of education, quitting smoking is difficult since the elderly have been exposed to smoking for a very long period of time and they do not have strong will to quit. Smoking rate of the elderly over the age of 65 in Korea is still very high of 23.3% [2]. Thus, in order to prevent MCI and maintain healthy cognitive function, quitting smoking is required more than anything else.

As the result of comparison of accuracy among random forests, logistic regression model and decision tree, random forests were the most accurate, which is speculated, because random forests is based on bagging algorithm which creates various decision trees from around 500 bootstrap samples.

As decision tree can compose a node even in the case of outlier, the influence of parameter deciding node is great, which creates risk of overfitting [12]. On the other hand, in the case of random forests which predict target variables through average or probability of each tree, as the bias of trees is maintained and variance decreases, its accuracy is higher than that of decision tree [16]. Therefore, when using data with many variables like disease examination data or establishing prediction model using distributed processing system such as big data, random forests is considered the most proper which extracts training data to create trees and predicts target variables. In order to further enhance accuracy of random forests, future studies are required to develop prediction model using weighted voting.

This study has the strength that it developed MCI prediction model by using examination data representing the whole population. It is necessary to establish a monitoring system which can diagnose old-age cognitive impairment in an early stage based on the MCI prediction model developed by this study.

## ACKNOWLEDGMENT

The author wishes to thank the Korea Labor Institute that provided the raw data for analysis.

## REFERENCES

[1] Alzheimer`s Disease International, World Alzheimer Report 2015. London, Alzheimer`s Disease International, 2015.

[2] Ministry of Health & Welfare, Nationwide Study on the Prevalence of Dementia in Korean Elders 2012. Sejong, Ministry of Health & Welfare, 2013.

[3] S. Kim, Analysis on Management Policies for the Dementia. Seoul, National Assembly Budget Office, 2014.

[4] 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.

[5] H. Byeon, Y. Lee, S. Y. Lee, K. S. Lee, S. Y. Moon, H. Kim, C. H. Hong, S. J. Son, and S. H. Choi, Association of alcohol drinking with verbal and visuospatial memory impairment in older adults: Clinical Research Center for Dementia of South Korea (CREDOS) study. International Psychogeriatrics, vol. 27, no. 3, pp. 455–461, 2015.

[6] H. A. Tuokko, and D. F. Hultsch, Mild cognitive impairment: International perspectives. New York, Psychology Press, 2013.

[7] G. Cheng, C. Huang, H. Deng, and H. Wang, Diabetes as a risk factor for dementia and mild cognitive impairment: a meta-analysis of longitudinal studies. Internal medicine journal, vol. 42, no. 5, pp. 484–491, 2012.

[8] M. Ganguli, B. Fu, B. E. Snitz, T. F. Hughes, and C. C. H. Chang, Mild cognitive impairment Incidence and vascular risk factors in a population-based cohort. Neurology, vol. 80, no. 23, pp. 2112–2120, 2013.

[9] T. Etgen, D. Sander, H. Bickel, and H. Förstl, Mild cognitive impairment and dementia: the importance of modifiable risk factors. Deutsches Ärzteblatt International, vol. 108, no. 44, p. 743, 2011.

[10] R. C. Petersen, B. Caracciolo, C. Brayne, S. Gauthier, V. Jelic, and L. Fratiglioni, Mild cognitive impairment: a concept in evolution. Journal of internal medicine, vol. 275, no. 3, pp. 214–228, 2014.

[11] H. Byeon, Development of prediction model for endocrine disorders in the Korean elderly using CART algorithm: results from a population-based study. International Journal of Advanced Computer Science and Applications, vol. 6, no. 9, 215–219. 2015.

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

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

[14] 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.

[15] C. Strobl, A. L. Boulesteix, A. Zeileis, and T. Hothorn, Bias in random forest variable importance measures: illustrations, sources and a solution. BMC bioinformatics, vol. 8, no. 1, p. 25, 2007.

[16] K. L. Lunetta, L. B. Hayward, J. Segal, and P. Van Eerdewegh, Screening large-scale association study data: exploiting interactions using random forests. BMC genetics, vol. 5, no. 1, p. 32, 2004.

[17] Korea Labor Institute, Korean Longitudinal Survey of Ageing 2011. Sejong, Korea Labor Institute, 2014.

[18] 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.

[19] 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.