Developing A Model to Predict the Occurrence of the Cardio-Cerebrovascular Disease for the Korean Elderly using the Random Forests Algorithm

Haewon Byeon
Department of Speech Language Pathology
Honam University
Gwangju, Republic of Korea

***Abstract*—This study aimed to develop a model for predicting the cardio-cerebrovascular disease of the South Korean elderly using the random forests technique. This study analyzed 2,111 respondents (879 males and 1,232 females), who were age 60 or older, out of total 7,761 respondents, who completed the Seoul Welfare Panel Study. The result variable was defined as the cardio-cerebrovascular disease (e.g., hypertension, cerebral infarction, hyperlipidemia, cardiac infarction, and angina). As a result of developing a random forest-based model, the major determinants of the cardio-cerebrovascular diseases of the South Korean elderly were mean monthly household income, the highest level of education, subjective health condition, subjective friendship, subjective family relationship, smoking, regular exercise, age, marital status, gender, depression experience, economic activity, and high-risk drinking. Among them, mean monthly household income was the most important predictor of the cardio-cerebrovascular disease. Based on the developed prediction model, it is needed to develop a systematic program for preventing the cardio-cerebrovascular disease of the Korean elderly.**

***Keywords*—Prediction model; data mining; random forest; risk factors; cardio-cerebrovascular disease; stroke**

I. Introduction

The cardio-cerebrovascular diseases include cerebrovascular diseases (e.g., cerebral hemorrhage and cerebral infarction), cardiac disorders (e.g., cardiac insufficiency, angina, and cardiac infarction), and vascular abnormalities (e.g., hypertension, diabetes, hyperlipidemia, and arteriosclerosis). As of 2013, the mortality due to cardio-cerebrovascular diseases accounts for more than 25% of the national mortality. The annual death toll is 50.3 people for cardiovascular diseases and 50.2 people for cerebrovascular per 100,000 population [1]. The cardio-cerebrovascular disease is the second leading cause of death in South Korea [1]. The cardio-cerebrovascular disease has increased by 1.35 times over the past decade and has become a critical health problem in South Korea [1].

Particularly, the cardio-cerebrovascular disease is a representative chronic disease of the elderly. It is known that the mortality rate increases rapidly with age. Especially, previous studies reported that it increased abruptly in the elderly over 70 years old [1]. Additionally, the cardio-cerebrovascular disease of the elderly is often accompanied by severe disability even if surgical treatment is successful. Therefore, they tend to have a hard time to return to the society even after recovery [2]. Consequently, it is essential to identify factors associated with the cardio-cerebrovascular disease and prevent them for achieving the successful aging.

As more people die from cardio-cerebrovascular diseases, there is a growing interest in managing and preventing the diseases. In the past 20 years, a number of studies have attempted to evaluate various risk factors for cardio-cerebrovascular diseases such as sociodemographic factors, lifestyle, and family history [3-6]. The results of these studies have identified various risk factors encompassing those that cannot be controlled (e.g., age and gender) and those that can be controlled (e.g., eat habits and physical activities) [3-6].

However, it has been pointed out that these individual risk factors have limitations in explaining the onset of a cardio-cerebrovascular disease [7]. Moreover, studies have indicated different factors as the most important risk factor. Additionally, although the cardio-cerebrovascular disease is known as a complex disease due to the interactions of multiple factors including sociodemographic factors (e.g., age and gender), environmental factors (e.g., lifestyle), and causative disease factors (e.g., hypertension and hyperlipidemia) [8-9], recent studies reported that psychological factors such as depression were major risk factors as well [10-11].

Moreover, the occurrence patterns and the risk factors of the cardio-cerebrovascular disease vary greatly among different ethnic groups. Therefore, it is difficult to establish a prevention and management strategy based on the results of previous studies conducted for different ethnic groups. Additionally, the lifestyle, an important factor in deciding the health, is determined by cultural influences as well as personal characteristics. Therefore, it is necessary to develop a model for predicting the cardio-cerebrovascular disease with reflecting the characteristics of the elderly living in the local communities in South Korea using big data.

The random forests technique has been used more frequently as a data mining algorithms for predicting the risk factors of target variables such as a disease or a disability [12-14]. The random forests technique is a method of combining multiple decision trees based on the ensemble technique in order to minimize the over-fitting, which is a shortfall of the

decision tree. The technique shows a good prediction ability, which is an advantage of this technique. This study aimed to develop a model for predicting the cardio-cerebrovascular disease of the South Korean elderly using the random forests technique.

Construction of this study is as follows. chapter II explains data source and materials and chapter III defines random forests and explains the procedure of final model development. Chapter IV compares the results of developed final prediction model. Lastly, chapter V presents discussion and direction for future studies.

II. Materials and Methods

A. Data Source

This study analyzed a portion of the raw data of the Seoul Welfare Panel Study, which was conducted by Seoul Welfare Foundation to survey Seoul citizens from Jun 1 to August 31, 2010. Seoul Welfare Panel Study was approved (#20113) by Statistics Korea in 2009 and it has been conducted to identify the welfare level of households residing in Seoul, understand the status of the welfare vulnerable class, and estimate the demand for welfare services [15]. This study targeted households in Seoul as of "the 2005 Population and Housing Census" and sampled using the stratified cluster sampling method for 25 districts in Seoul. The main survey items were income, economic level, health, living conditions, and the demand for welfare services. The survey was conducted by using the computer-assisted personal interviewing method: the interviewer visited the surveyed households and inputted responses according to a structured questionnaire into a portable computer. This study analyzed 2,111 respondents (879 males and 1,232 females), who were age 60 or older, out of total 7,761 respondents, who completed the survey.

B. Measurements and Definitions of Variables

The result variable was defined as the cardio-cerebrovascular disease (e.g., hypertension, cerebral infarction, hyperlipidemia, cardiac infarction, and angina). The explanatory variables included age (60years or older and younger than 70years or 70years or older), gender (male or female), the highest level of education (below elementary school, junior high school, high school, and college graduation and over), economic activity (yes or no), mean monthly household income (less than 2 million KRW, 2-4 million KRW, and more than 4 million KRW), marital status (living with a spouse, married but not living with a spouse, or single), High-risk drinking (yes or no), Smoking (non-smoker, past smoker, current smoker), subjective health condition (good, normal, or poor), subjective family relationship (good, average, or bad), subjective friendship (good, average, or poor), regular exercise (no or yes), and the depression symptom in the past one month (no or yes).

III. Analysis Methods

A. Exploring Potential Factors of the Cardio-Cerebrovascular Disease in Old Age

The prevalence of the cardio-cerebrovascular disease between groups was analyzed by using the chi-square test. When the significance level of an explanatory variable was 0.1 or below, it was considered as a potential factor of the cardio-cerebrovascular disease and it was included in the random forest model.

B. Random Forests Algorithm

The random forests technique [16] is an algorithm that creates various sample datasets using bootstrap. This method has an advantage of increasing the diversity of the decision tree because it repeats the process of randomly selecting several variables [17]. Unlike the decision trees, which present each node with the partition showing the most optimum results by using all variables, the random forests select explanatory variables randomly and use the method showing the most optimum results among the selected explanatory variable groups (Figure 1). The process of random forests is shown in Eq. (1).

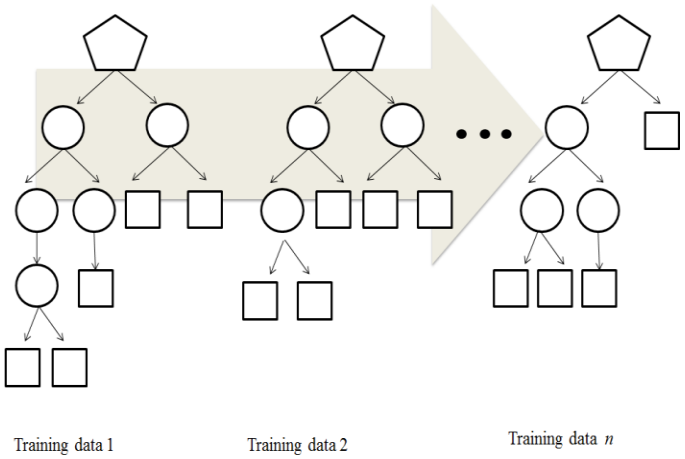

Fig. 1. Random forest classifier: source is Byeon [12]

1: Generate B bootstrap samples $L_1, ..., L_B$ from the original training data set L

2: Grow a random forest tree using a random feature selection from bootstrapped data.

:randomly select $\sqrt{p}$ predictors at each node and split the data using the best predictors.

3: Construct train classifiers $C_b(\mathbf{x}), b = 1, ...B$ from each of L_b samples

4: Aggregate the B train classifiers.

Let N_j be the number of times that classified j

$$N_j = \sum_{b=1}^{B} I[C_b\mathbf{x} = j], \text{ for } j = 1, ...J$$

5: final classification: $arg\,max_j N_j$

(1)

Another advantage of the random forests is to reduce the variance compared to the bagging method because it decreases the correlation between trees. Moreover, it presents more accurate results than other algorithms and it is useful to find an important variable in big data because it utilizes thousands of independent variables without eliminating variables [18]. Especially, when there are many input variables, it often shows similar or better prediction power than bagging or boosting. The input source of the R program for performing random forests analysis is shown in Fig 2.

```
X_train, X_test, y_train, y_test = train_test_split(

cvd.data, cvd.target, random_state=0)

forest = RandomForestClassifier(n_estimators=500, random_state=0)

forest.fit(X_train, y_train)



print("train: {:.3f}".format(forest.score(X_train, y_train)))

print("test: {:.3f}".format(forest.score(X_test, y_test)))
```

Fig. 2. Input source of the R program for performing random forests

In this study, the number of trees in the model was set to 500. The analysis was conducted using R version 3.4.2 and Waikato Environment for Knowledge Analysis (WEKA) version 3.6.0 [19].

IV. Results

A. General Characteristics of Subjects

The characteristics the data (n=2,111) were analyzed and the results showed that the 53.5% of study subjects were between 60 and 69 years old and the 58.4% of them were women. The majority of the subjects lived with their spouses (67.2%), were elementary school graduation or below (43.3%), had the mean monthly income less than 2 million KRW (64.7%), were not economically active (83.2%), did not exercise regularly (55.4%), had poor subjective health (39.4%), had good subjective family relationship (59.1%), had average subjective friendship, and did not experience a depression symptom in the past one month (74.4%). The prevalence of the cardio-cerebrovascular disease was 42.3 %.

B. Potential Factors of Cardio-Cerebrovascular Disease in Old Age

Table 1 shows the general characteristics and potential factors of subjects according to the prevalence of cardio-cerebrovascular diseases. The prevalence of cardio-cerebrovascular diseases, which indicated the proportion of subjects suffering hypertension, cerebral infarction, hyperlipidemia, cardiac infarction, and angina, was 42.3% (n=894). The results of chi-square test showed that the elderly with cardio-cerebrovascular diseases and those without cardio-cerebrovascular diseases were there were significant ($p<0.05$) different in age, marital status, economic activity, smoking, the depression symptom in the past one month, subjective health condition, and subjective family relationship. The prevalence of cardio-cerebrovascular diseases was significantly higher for the elderly equal to or older than 70 years (50.8%), not living with a spouse (47.8%), not economically active (44.1%), former smoker (43.8%), depression symptom experience in the past one month (49.7%), poor subjective health (51.1%), and average family relationship (46.4%).

TABLE I. General Characteristics Of The Subjects By Cardio-Cerebrovascular Disease, N (%)

Characteristics	cardio-cerebrovascular disease		p
	Yes (n=894)	No (n=1,217)	
Age			<0.001
60-69	396 (35.0)	734 (65.0)	
70+	498 (50.8)	483 (49.2)	
Gender			0.315
Male	316 (41.1)	518 (58.9)	
Female	533 (43.3)	699 (56.7)	
Marital Status			0.002
Living with a spouse	563 (39.7)	856 (60.3)	
Married but not living with a spouse	20 (47.6)	22 (52.4)	
Single	311 (47.8)	339 (52.2)	
The highest level of education			0.071
Below elementary school	406 (44.4)	508 (55.6)	
Junior high school	160 (42.9)	213 (57.1)	
High school	185 (37.3)	311 (62.7)	
College graduation and over	143 (43.6)	185 (56.4)	
Mean monthly household income			0.056
Less Than 2 million KRW	601 (44.0)	765 (56.0)	
2-4 million KRW	194 (39.9)	292 (60.1)	
More than 4 million KRW	31 (33.3)	62 (66.7)	
Economic activity			<0.001
Yes	120 (33.9)	234 (66.1)	
No	774 (44.1)	983 (55.9)	
Smoking			0.035
Non-smoker	619 (43.1)	817 (56.9)	
Past smoker	202 (43.8)	259 (56.2)	
Current smoker	73 (34.1)	141 (65.9)	
High-risk drinking			0.299
No	122 (38.4)	196 (61.6)	
Yes	61 (33.7)	120 (66.3)	
Regular exercise			0.689
No	500 (42.7)	670 (57.3)	
Yes	394 (41.9)	547 (58.1)	
Depression symptom in the past one month			<0.001
No	625 (39.8)	945 (60.2)	
Yes	269 (49.7)	272 (50.3)	
Subjective health condition			<0.001
Good	158 (27.1)	426 (72.9)	
Normal	312 (44.7)	386 (55.3)	
Poor	424 (51.1)	405 (48.9)	
Subjective family relationship			0.014
Good	479 (39.5)	735 (60.5)	
Average	307 (46.4)	355 (53.6)	
Bad	78 (43.6)	101 (56.4)	
Subjective friendship			0.146
Good	277 (39.4)	426 (60.6)	
Average	481 (44.0)	611 (56.0)	
Bad	136 (43.0)	180 (57.0)	

C. Predict Occurrence of Cardio-Cerebrovascular Disease for Korean elderly

The importance of variables (the decrement of node impurity) based on random forests is shown in Table 2 and Figure 3. The results showed that the major determinants of the

cardio-cerebrovascular diseases of the South Korean elderly were mean monthly household income, the highest level of education, subjective health condition, subjective friendship, subjective family relationship, smoking, regular exercise, age, marital status, gender, depression experience, economic activity, and high-risk drinking. Among them, mean monthly household income was the most important predictor of the cardio-cerebrovascular disease.

TABLE II. IMPORTANCE OF VARIABLES: THE DECREMENT OF NODE IMPURITY

Importance of variables	Decrement of node impurity
Household income	53.351
Highest level of education	39.502
Ubjective health condition	33.180
Subjective friendship	27.991
Subjective family relationship	25.104
Smoking	23.234
Regular exercise	16.904
Age	16.635
Marital status	16.501
Gender	14.559
Depression experience	12.522
Economic activity	11.815
High-risk drinking	8.330

Figure 4 shows the error rate graphs for each prediction model for each of the extracted 500 bootstrap samples. The error rate of the developed random forests was 0.24 and the prediction rate was 76.5%.

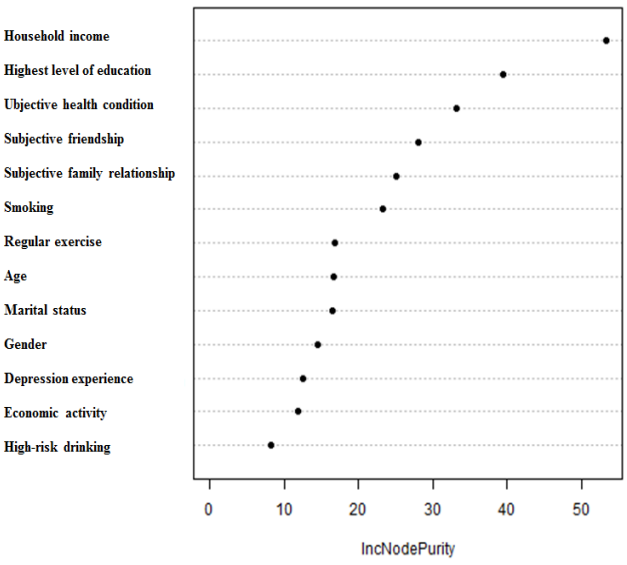

Fig. 3. Variable importance plot

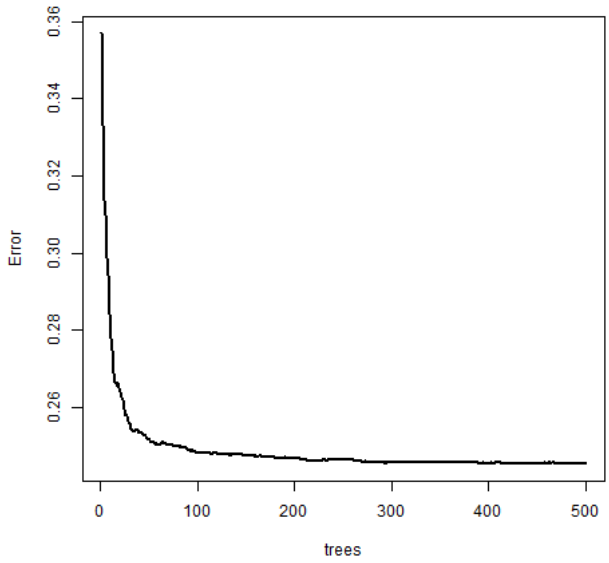

Fig. 4. Error rate graphs for each prediction model

V. DISCUSSION

This study developed a model for predicting the cardio-cerebrovascular disease of the elderly living in the community using the random forests technique, which is a data minding algorithm based on the classification learning. This study constructed a prediction model of cardio-cerebrovascular disease considering multiple risk factors. The results of the constructed model showed that household income was the most important factor followed by the education level, which indicated that socio-economic factors were major risk factors.

Many previous studies have reported that the prevalence of the cardio-cerebrovascular disease is affected by the socioeconomic levels [20-22]. The mechanism of the socioeconomic factors can be explained by the changes in hemodynamics due to the increase of stress and the lack of health life practice, HDL-cholesterol, insulin resistance, and blood coagulation-related factors [20]. Therefore, the low-income and poorly educated groups should be sufficiently considered when establishing prevention programs.

The results of this study confirmed that depression was a major predictor of the cardio-cerebrovascular disease. Previous studies reported that the depression and the cardio-cerebrovascular disease were highly associated. Particularly, the cardio-cerebrovascular disorder was identified as a risk factor of depression [6, 23], and those with depression were at risk for the cardiovascular disease and their mortality risk was twice than others without it [24]. Morris et al. (1993) also found that cerebrovascular disorders were frequently accompanied by depression and people with depression had an 8-fold higher risk of death from cerebral infarction than others without it [25]. Byeon (2015) identified the cardio-cerebrovascular disease risk groups using the QUEST algorithm and also predicted that the elderly who experienced depression would have a higher risk of the cardio-cerebrovascular disease [6]. The results of this study indicated that the depression in the old age was a risk factor of the cardio-cerebrovascular disease. Therefore, it will be necessary

to develop the cardio-cerebrovascular disease prevention program for the elderly with depression for preventing the cardio-cerebrovascular disease of the elderly in the local community.

This study developed a depression prediction model for children from multicultural families by using CHAID algorithm and found that the experience of social discrimination is the most critical factor affecting depression. Although it is hard to compare the results of this study directly, the previous studies evaluating the relationship between social discrimination and mental health reported that the economic discrimination and the discrimination against a specific group (e.g., the elderly group) were significant predictor variables negatively influencing mental health [18]. Therefore, it is necessary to establish a legal system and pay social level interests to overcome the discrimination and prejudice against adolescents from multicultural families based on the results of this study.

The results of this study showed that the cardio-cerebrovascular disease prediction model based on the random forests technique had stronger prediction power than the previously developed cardio-cerebrovascular disease prediction model based on QUEST algorithm [6]. The random forests showed superior prediction performance than the decision tree and it produced more stable results because it made decisions by integrating the prediction results of multiple decision trees using the bootstrap sample [26]. Therefore, it was believed that using the random forests model would be more effective than using the decision tree model when estimating the importance of variables in the development of disease prediction models. It will be necessary to compare the predictive performance of the logistic regression model, the decision tree model, and the random forests in the future.

VI. Conclusion

Based on the developed prediction model, it is needed to develop a systematic program for preventing the cardio-cerebrovascular disease of the Korean elderly.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (NRF-2018R1D1A1B07041091).

References

[1] Statistical Office, Cause of death statistics, National Statistical Office, Daejeon, 2015.

[2] P. B. Gorelick, R. L. Sacco, D. B. Smith, M. Alberts, L. M. Alexander, D. Rader, J. L. Ross, E. Raps, M. N. Ozer, L. M. Brass, M. E. Malone, S. Goldberg, J. Booss, D. F. Hanley, J. F. Toole, N. L. Greengold and D. C. Rhew, Prevention of a first stroke-a review of guidelines and a multidisciplinary consensus statement fromt The National Stroke Association. Journal of the American Medical Association, vol. 281, no. 12, pp. 1112-1120, 1999.

[3] N. Y. Han, E. A. Ko and S. Y. Hwang, Knowledge of stroke symptoms and risk factors among older adults. Korean Journal of Adult Nursing, vol. 21, no. 3, pp. 314-323, 2009.

[4] D. Ettehad, C. A. Emdin, A. Kiran, S. G. Anderson, T. Callender, J. Emberson, J. Chalmers, A. Rodgers and K. Rahimi, Blood pressure lowering for prevention of cardiovascular disease and death: a systematic review and meta-analysis. The Lancet, vol. 387, no. 10022, pp. 957-967, 2016.

[5] S. Mottillo, K. B. Filion, J. Genest, L. Joseph, L. Pilote, P. Poirier, S. Rinfret, E. L. Schiffrin and M. J. Eisenberg, The metabolic syndrome and cardiovascular risk: a systematic review and meta-analysis. Journal of the American College of Cardiology, vol. 56, no. 14, pp. 1113-1132, 2010.

[6] S. Cho and H. Byeon, Predictive modeling of hypertension in Korean old adults using QUEST algorithm. Asia Life Sciences, pp. 21-30, 2015.

[7] T. W. Smith and J. M. Ruiz, Psychosocial influences on the development and course of coronary heart disease: current status and implications for research and practice. Journal of Consulting and Clinical Psychology, vol. 70, pp.5 48-568, 2002.

[8] R. M. Conroy, K. Pyörälä, A. P. Fitzgerald, S. Sans, A. Menotti, G. De Backer, P. De Bacquer, P. Ducimetière, P. U. Jousilahti, I. Keil, R. G. Njølstad, T. Oganov, H. Thomsen, A. Tunstall-Pedoe, H. Tverdal, P. Wedel and I. M. Wilhelmsen, Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. European Heart Journal, vol. 24, no. 11, pp. 987-1003, 2003.

[9] J. D. Berry, A. Dyer, X. Cai, D. B. Garside, H. Ning, A, Thomas, A. Thomas, P. Greenland, L. V. Horn, R. P. Tracy and D. M. Lloyd-Jones, Lifetime risks of cardiovascular disease. New England Journal of Medicine, vol. 366, no. 4, pp. 321-329, 2012.

[10] A. Rozanski, J. A. Blumenthal and J. Kaplan, Impact of psychological factors on the pathogenesis of cardiovascular disease and implications for therapy. Circulation, vol. 99, pp. 2192-2217, 1999.

[11] B. W. Penninx, Depression and cardiovascular disease: epidemiological evidence on their linking mechanisms. Neuroscience & Biobehavioral Reviews, vol. 74, pp. 277-286, 2017.

[12] 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.

[13] K. R. Gray, P. Aljabar, R. A. Heckemann, A. Hammers, D. Rueckert and Alzheimer's Disease Neuroimaging Initiative. Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. NeuroImage, vol. 65, pp. 167-175, 2013.

[14] H. Byeon, H. Jin, and S. Cho, Development of parkinson's disease dementia prediction model based on verbal memory, visuospatial memory, and executive function. Journal of Medical Imaging and Health Informatics, vol. 7, no. 7, pp. 1517-1521, 2017.

[15] Seoul Welfare Foundation, Seoul Welfare Panel Study 2010, Seoul Welfare Foundation, Seoul, 2010.

[16] L. Breiman, Random forests. Machine learning, vol. 45 no. 1, pp. 5-32, 2010.

[17] A. Singh, M. N. Halgamuge and R. Lakshmiganthan, Impact of different data types on classifier performance of random forest, naive bayes, and k-nearest neighbors algorithms. International Journal of Advanced Computer Science and Applications, vol. 8, no. 12, pp. 1-10, 2017.

[18] H. Byeon, D. Lee, and S. Cho, Assessment for the model predicting of the cognitive and language ability in the mild dementia by the method of data-mining technique. International Journal of Advanced Computer Science and Applications, vol. 7, no. 6, 2016.

[19] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, Burlington, 2016.

[20] J. F. Meschia, C. Bushnell, B. Boden-Albala, L. T. Braun, D. M. Bravata, S. Chaturvedi, and L. B. Goldstein, Guidelines for the primary prevention of stroke: a statement for healthcare professionals from the American Heart Association/American Stroke Association. Stroke, vol. 45, no. 12, pp. 3754-3832, 2014.

[21] K. Backholer, S. A. Peters, S. H. Bots, A. Peeters, R. R. Huxley, R. R. and M. Woodward, M, Sex differences in the relationship between socioeconomic status and cardiovascular disease: a systematic review and meta-analysis. Journal of Epidemiology and Community Health, jech-2016, 2016.

[22] C. T. January, L. Samuel Wann, J. S. Alpert, H. Calkins, J. E. Cigarroa,, J. C. Cleveland Jr, J. B. Conti, P. T. Ellinor, M. D. Ezekowitz, M. E. Field, K. T. Murray, R. L. Sacco, W. G. Stevenson, P. J. Tchou, C. M Tracy and C. W. Yancy, 2014 AHA/ACC/HRS guideline for the

management of patients with atrial fibrillation: executive summary: A report of the American College of Cardiology/American Heart Association Task Force on practice guidelines and the Heart Rhythm Society. Journal of the American College of Cardiology, 64(21), 2246-2280.

[23] B. E. Cohen, D. Edmondson and I. M. Kronish, State of the art review: depression, stress, anxiety, and cardiovascular disease. American Journal of Hypertension, vol. 28 no. 11, pp. 1295-1302, 2015.

[24] J. C. Barefoot and M. Schroll, Symptoms of depression, acute myocardial infarction, and total mortality in a community sample. Circulation. vol. 93, no. 11, pp. 1976-1980, 1996.

[25] P. L. Morris, R. G. Robinson and J. Samuels, Depression, introversion and mortality following stroke. Australian & New Zealand Journal of Psychiatry, vol. 27, no.3, pp. 443-449, 1993.

[26] J. Maroco, D. Silva, A. Rodrigues, M. Guerreiro, I. Santana and A. de Mendonça, Data mining methods in the prediction of dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BMC research notes, vol. 4, no. 1, 299, 2011.