

Exploring Factors Associated with Subjective Health of Older-Old using ReLU Deep Neural Network and Machine Learning

Haewon Byeon

Department of Medical Big Data, College of AI Convergence
Inje University, Gimhae 50834, Republic of Korea

Abstract—Resolving the health issues of the elderly has emerged as an important task in the current society. This study developed models that could predict the subjective health of the older-old based on gradient boosting machine (GBM), naive Bayes model, classification and regression trees (CART), deep neural network, and random forest by using the health survey data of the elderly and compared their prediction performance (i.e., accuracy, sensitivity, specificity) the models. This study analyzed 851 older-old people (≥ 75 years old) who resided in the community. This study compared the accuracy, sensitivity, and specificity of the developed models to evaluate their prediction performance. This study conducted 5-fold cross-validation to validate the developed models. The results of this study showed that the deep neural network with an accuracy of 0.75, a sensitivity of 0.73, and a specificity of 0.81 was the model with the best prediction performance. The normalized importance of variables derived from deep neural network analysis showed that depression, subjective stress recognition, the number of accompanying chronic diseases, subjective oral conditions, and the number of days walking more than 30 minutes were major predictors for the subjective health of the older-old. Further studies are needed to identify factors associated with the subjective health of the older-old with considering the age-period-cohort effects.

Keywords—Gradient boosting machine; classification and regression trees; Naive Bayes model; deep learning; subjective health

I. INTRODUCTION

The elderly population is increasing rapidly in South Korea. As of 2018, South Korea has entered the aging society because the elderly population (≥ 65 years old) exceeded 14% of the total population [1,2]. Moreover, it is expected that South Korea will become a super-aged society, where the elderly population is at least 20% of the total population [1,2]. If this trend continues, it is estimated that the elderly population will be 16.16 million people, 3.7 folds of that in 2005 [1]. In particular, as the average life expectancy is extended, the older-old (75 years or older) is also expected to increase by 8.6 times compared to 2005 [1].

The increased elderly population due to aging caused health problems such as an increase in geriatric diseases and social problems such as an increase in the burden of medical expenses [3,4]. Resolving the health issues of the elderly has emerged as an important task in the current society [5]. In particular, as the baby boomer generation is rapidly turning

into an elderly population, South Korean society needs to prepare measures for faces more serious and complex elderly issues to be resolved [6,7,8]. Consequently, the subjective health of the elderly has been studied extensively [6,7,8]. However, these previous studies have limitations. First, most previous studies [9, 10] that identified factors affecting the subjective health status of the elderly usually treated the elderly, ≥ 65 years old, as a homogeneous group and developed a model for predicting a subjective health condition without categorizing them into different age groups. Second, it is difficult to generalize and extrapolate the results of previous studies [11] to the whole elderly population in South Korea because they mostly targeted a small elderly group in a specific area. Third, previous studies [12] targeting the elderly in South Korea mainly used tools that were developed in South Korea to measure subjective health or quality of life, and as a result, it is difficult to compare the results of these studies with those published in other countries that do not speak Korean. In particular, as researchers began to recognize issues associated with defining the elderly, ≥ 65 years old, as a homogenous group [13], some studies [14,15] subdivided the elderly into the young-old (< 75 years old) and the older-old (≥ 75 years old).

As South Korea entered the aging society in 2017, it is necessary to explicitly identify the subjective health-related factors of the older-old whose daily living ability has declined. Previous studies [16,17,18,19,20,21,22] evaluated factors related to the subjective health of the older-old and reported that the number of accompanying chronic diseases, depression, socioeconomic status, gender, and social network with friends and relatives as the predictors of subjective health. However, since these previous studies used regression analysis to develop a prediction model, it was efficient to identify individual risk factors but the model is limited in identifying multiple risk factors such as lifestyle and socioeconomic level [23]. In addition, the regression model assumes independence, normality, and homoscedasticity [24]. If the model is developed using data that violate the normality assumption, it may produce biased results [24].

Recently, machine learning has been widely used in various fields as a way to solve the limitations of this regression model. Machine learning can accurately analyze even data that somewhat violate the normality assumption and nonlinear data in the estimation process, which are advantages of machine learning [25]. This study developed models that could predict

the subjective health of the older-old based on gradient boosting machine (GBM), naive Bayes model, classification and regression trees (CART), deep neural network, and random forest by using the health survey data of the elderly and compared their prediction performance (i.e., accuracy, sensitivity, specificity) the models.

II. METHODS AND MATERIALS

A. Study Subjects

This study analyzed the raw data of the 2016 Seoul Panel Study Data (SPS-data). The SPS-data was carried out from June 1 to August 31, 2016, for understanding the status of the welfare vulnerable class living in Seoul and estimating the welfare level of citizens. The population was the households living in Seoul at the time of the 2005 Population and Housing Census, a complete enumeration census. The sample was extracted from the census data using the randomized stratified sampling method for 25 districts in Seoul. This study excluded those who were imprisoned, were admitted to a nursing hospital, or moved to a silver town, and foreigners. As for the survey method, the interviewer visited the target household and conducted a computer-aided personal interview. This study analyzed 851 older-old people who resided in the community and completed the survey.

B. Measurements and Definitions of Variables

Subjective health status, outcome variable, was measured on a 5-point Likert scale (very good, good, moderate, bad, and very bad). Explanatory variables included age, sex, smoking (smoker or non-smoker), drinking (once a week or less or more than twice a week), economic activity (yes or no), mean monthly household income (less than KRW 1.5 million, KRW 1.5-3 million, or KRW 3 million or more), educational level (elementary school graduation and below, middle school graduation, high school graduation, or college graduation or above), social activities in the past month (yes or no), living with a spouse (living together, bereavement/separation, or single), disease/accident/addition experience in the past two weeks (yes or no), subjective stress recognition (yes or no), number of accompanying chronic diseases, body mass index, number of days walking for 30 minutes or more (1 or less than day per week or 2 days per week or more), depression, subjective oral conditions, self-perceived diet, frequency of meeting with neighbors (once a month or less or more than twice a month), and meeting frequency with relatives (once a month or less or more than twice a month).

The number of chronic diseases (e.g., diabetes, hypertension, and hyperlipidemia) was classified into none, 1, 2, and 3 or more. The body mass index (BMI) was classified as “underweight” when it was less than 18.5, “normal” when it was 18.5 or more and less than 25, “overweight” when it was 25 or more and less than 30, and “obesity” when it was 30 or more. Subjective stress was defined as a “high-stress group” when the respondents said they felt a lot of stress and a “low-stress group” when they answered that they felt a little stress. Depressive symptoms were classified as “depressed” when feeling depressed for more than two weeks in a row and “not depressed” when not feeling depressed for two weeks or more in a row. Subjective oral conditions were defined as “good”,

“normal”, and “bad”. The self-perceived dietary life was classified into “sufficient diet life” when the respondents answered “ate enough quantity of food and a variety of foods in the last year” or “ate enough quantity of food and not various foods” and “insufficient diet life” when they answered “insufficient food from time to time due to economic difficulties” or “frequently insufficient food due to economic difficulties”.

C. Development of Prediction Models: GBM

GBM is a machine learning algorithm that creates a strong learner by combining weak learners of a decision tree by using an ensemble technique [26]. This model generalizes the model by generating a model for each step like other boosting methods and optimizing a loss function that can be arbitrarily differentiated. A model is created, even if the accuracy is low, and the errors of the generated model are compensated by the following model. Through this process, the accuracy of the current model becomes better than that of the previous model: a more accurate model (or a stronger learner). The basic principle is to increase accuracy by repeating this process. GBM's algorithm is presented in Fig. 1.

Algorithm	Gradient boosting algorithm.
Input:	Input data $(x, y)_{i=1}^N$ Number of iterations M Choice of the loss-function $\Psi(y, f)$ Choice of the base-learner model $h(x, \theta)$
1:	Initialize \hat{f}_0 with a constant
2:	for $t = 1$ to M do
3:	Compute the negative gradient $g_t(x)$;
4:	Fit a new base-learner function $h(x, \theta_t)$;
5:	Find the best gradient descent step-size p_t
	$p_t = \operatorname{argmin}_p \sum_{i=1}^N \Psi[y_i, \hat{f}_{t-1}(x_i) + ph(x_i, \theta_t)];$
6:	Update the function estimate:
	$\hat{f}_t \leftarrow \hat{f}_{t-1} + p_t h(x, \theta_t);$
7:	end for

Fig. 1. Algorithm of GBM [26].

D. Naïve Bayes Classification

Naïve Bayes classification is a method of classifying observations into different groups using Bayes theory (Fig. 2)[27]. Bayes theory refers to a way of calculating the posterior probability by using an observation when there is a prior probability.

E. CART

CART is one of the analysis algorithms of the statistical decision classification model, which measures impurity using the Gini Index [28]. It is an algorithm based on a binary split, which generates only two child nodes from a parent node [28]. The advantages of CART are that it is easy to interpret the generated rules and can analyze both continuous and categorical variables. For continuous variables, a separation rule is created in the form of “ $X \leq C$?” or “ $X \geq C$?” [28]. For categorical variables, a rule of binary separation is created in the form of “ $X \in \{A, B\}$ ” [28]. In the model of this study, the criterion for separating and merging decision rules for CART was set to 0.05. The number of child nodes was limited to 100, the number of parent nodes was limited to 200, and the number of branch branches was limited to 5.

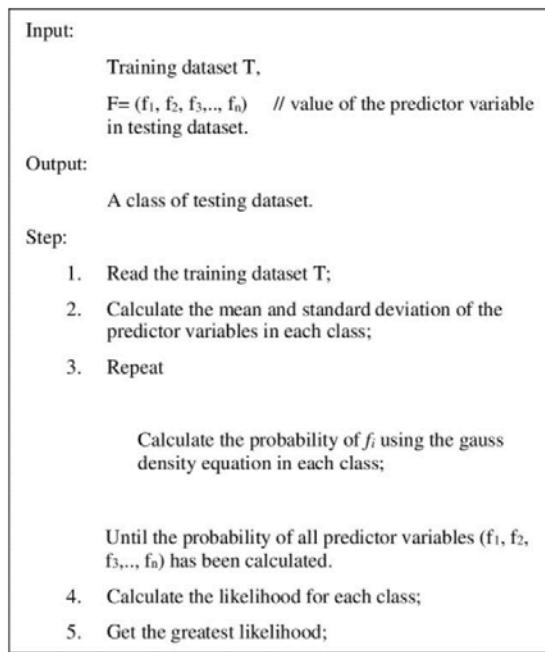


Fig. 2. Concept of Naïve Bayes Classification [27].

F. Deep Neural Network

Deep neural network is an algorithm made up of an input layer composed of independent variables, an output layer composed of dependent variables, and two or more hidden layers between the input and output layers [29]. Each layer has independent nodes and a node is connected to other nodes in other layers by weighted neurons (connecting lines) (Fig. 3).

This study used H2O Deep Learning among various deep learning types. H2O's Deep Learning is based on a multi-layer feedforward artificial neural network that is trained with stochastic gradient descent using back-propagation.

In this study, the number of hidden layers was set to 2, nodes of each layer were set to 10 (total of 20), and epochs (number of repetitions) were set to 5. In this study, a model was developed using the Rectifier Linear Unit (ReLU) designated as a default as the active function of deep learning (Fig. 4).

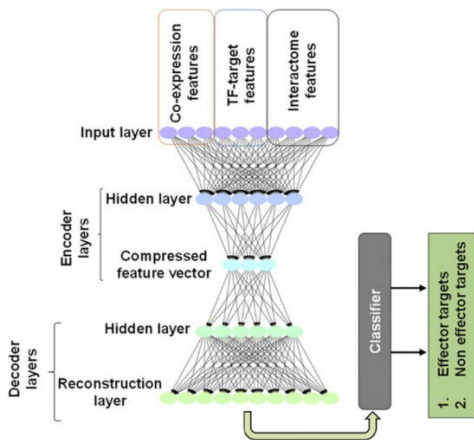


Fig. 3. The Concept of Deep Neural Network [30].

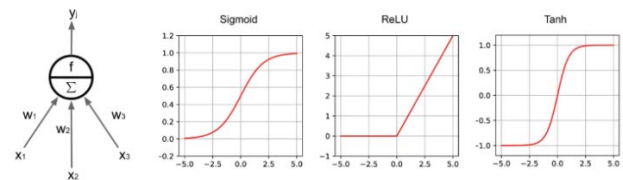


Fig. 4. The Widely used Nonlinear Activation Function of Deep Neural Networks [31].

G. Random Forest

Random Forest is an ensemble method that learns an independent decision tree for each sample after generating a large number of random samples from training data using bootstrap (conduct random restoration sampling of the same sample size from a given data) and decides a final model by synthesizing the results (Fig. 5).

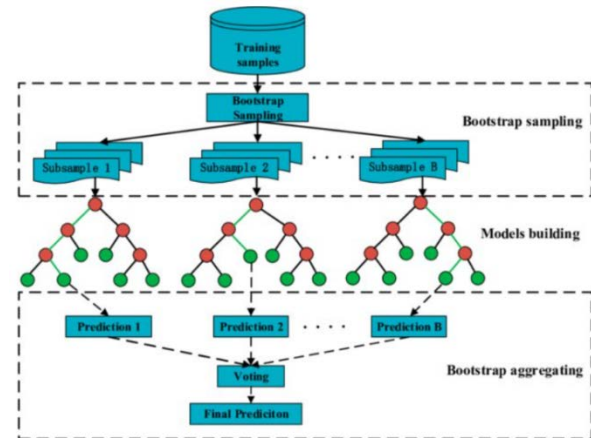


Fig. 5. The Concept of the Random Forest Model [32].

H. Validation of the Prediction Models

This study developed models for predicting the subjective health of the older-old using GBM, naive Bayes model, cart, deep neural network, and random forest. This study compared the accuracy, sensitivity, and specificity of the developed models to evaluate their prediction performance. This study conducted 5-fold cross-validation to validate the developed models.

In the analysis stage, models containing randomness, such as a random forest, were developed while fixing the seed number to "9876543". This study defined the model with the highest accuracy as the best prediction model (best prediction performance) after comparing prediction performance. If two models had the same accuracy, a model with a higher sensitivity was defined as a better prediction model. All analyzes were performed using R version 3.6.2 (Foundation for Statistical Computing, Vienna, Austria).

III. RESULTS

The accuracy of five models (GBM, naive Bayes model, cart, deep neural network, and random forest) for predicting the subjective health of the older-old is presented in Fig. 6. The results of this study showed that the deep neural network with an accuracy of 0.75, a sensitivity of 0.73, and a specificity of 0.81 was the model with the best prediction performance.

The normalized importance of variables derived from deep neural network analysis showed that depression, subjective stress recognition, the number of accompanying chronic diseases, subjective oral conditions, and the number of days walking more than 30 minutes were major predictors for the subjective health of the older-old (Fig. 7). Among them, depression had the highest importance (Fig. 7).

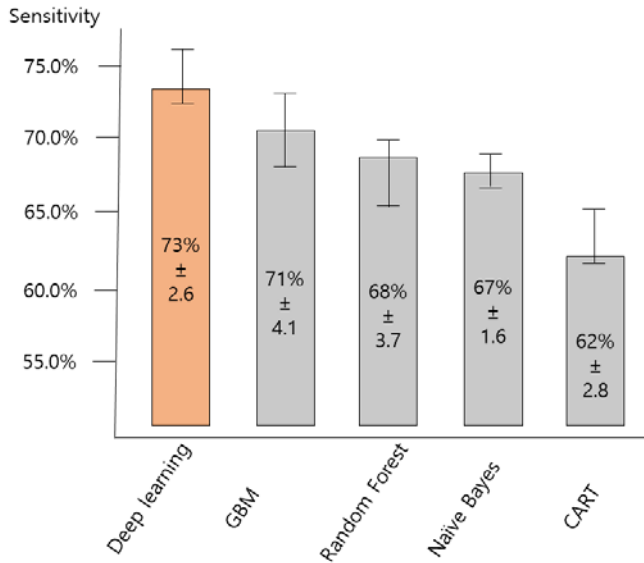


Fig. 6. Comparing the Sensitivity of a Machine Learning Model and a Deep Learning Model for Predicting the Subjective Health of the Older-Old.

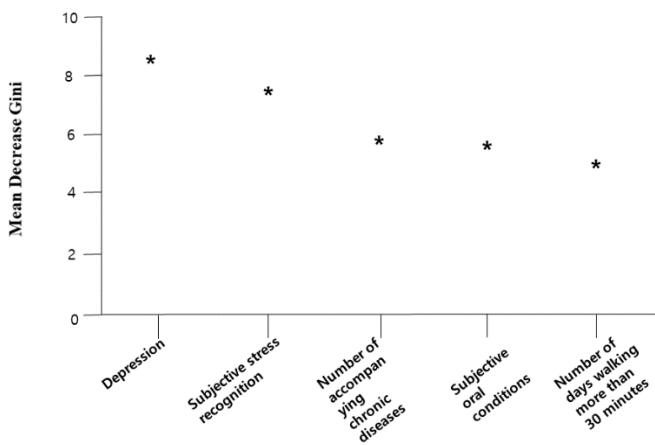


Fig. 7. The Importance of Variables in the Prediction Model for the Subjective Health of the Older-Old (Only the Top 5 Variables are Presented) Conclusions.

IV. CONCLUSIONS

This study compared the accuracy of prediction models for the subjective health of the older-old in South Korea and confirmed that the deep neural network-based prediction model had the best prediction performance among GBM, naive Bayes model, cart, deep neural network, and random forest. These results are consistent with the results of da Silva et al. [33], which predicted the health condition of the elderly and showed that the performance of deep learning was superior to that of

machine learning methods such as naive Bayes, J48 Decision Tree, and SVM.

Another finding of this study is that depression, subjective stress recognition, number of accompanying chronic diseases, subjective oral conditions, and the number of days of walking more than 30 minutes were identified as the main predictors for the subjective health of the older-old. The result was different from the previous studies [34,35,36] that identified the subjective health of the elderly using regression analysis and reported that social network (e.g., the presence of a spouse), education level, job status, and household income level were the main factors for subjective health. It is believed that these results reflected the interactions of the aging effect, period effect, and cohort effect, indicating the similar socioeconomic level of the older-old in South Korea, where unhealthy elderly people have already passed away. However, Kim et al. [37] evaluated the subjective health of the elderly and reported that the higher the household income level improved the satisfaction of subjective health. Further studies are needed to identify factors associated with the subjective health of the older-old with considering the age-period-cohort effects. Furthermore, additional studies are needed to compare prediction performance such as accuracy, sensitivity, and specificity using data of various diseases in order to prove the effectiveness of deep learning in epidemiological data.

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, NRF-2019S1A5A8034211, NRF-2021S1A5A8062526).

REFERENCES

- [1] Korea National Statistical Office. Population projections for Korea: 2005-2050 based on the 2005 census. Korea National Statistical Office, Daejeon, 2006.
- [2] M. H. Oh, Making use of older people's human resources in an aged Korea. Health and Welfare Policy Forum, vol. 254, pp. 50-66, 2017.
- [3] J. S. Kim, and Y. J. Han, The effect of household type on the medical burden of the elderly living in a local government that has entered a super-aged society. The Journal of the Korea Contents Association, vol. 17, no. 7, pp. 610-621, 2017.
- [4] M. De Nardi, E. French, and J. B. Jones, Why do the elderly save? the role of medical expenses. Journal of Political Economy, vol. 118, no. 1, pp. 39-75, 2010. doi: 10.5392/JKCA.2017.17.07.610.
- [5] T. Suzuki, Health status of older adults living in the community in Japan: recent changes and significance in the super-aged society. Geriatrics & Gerontology International, vol. 18, no. 5, pp. 667-677, 2018. doi: 10.1111/ggi.13266.
- [6] H. S. Gweon, The effect of social participation on the life satisfaction of the elderly -focusing on the mediating effects of depression and self-reported health. Korean Journal of Human Ecology, vol. 18, no. 5, pp. 995-1008, 2009. doi: 10.5934/KJHE.2009.18.5.995.
- [7] Y. Kwak, and Y. Kim, Quality of life and subjective health status according to handgrip strength in the elderly: a cross-sectional study. Aging and Mental Health, vol. 23, no. 1, pp. 107-112, 2019. doi: 10.1080/13607863.2017.1387766.
- [8] S. Y. Sohn, W. Joo, W. J. Kim, S. J. Kim, Y. Youm, H. C. kim, Y. R. Park, and E. Lee, Social network types among older Korean adults: associations with subjective health. Social Science & Medicine, vol. 173, pp. 88-95. doi: 10.1016/j.socscimed.2016.11.042.

- [9] M. S. Lee, and H. J. Lim, Factors related to health promoting behaviors of young-old and old-old elderly in rural areas. *Journal of Agricultural Medicine and Community Health*, vol. 35, no. 4, pp. 370-382, 2010.
- [10] S. H. Kim, Health literacy and functional health status in Korean older adults. *Journal of Clinical Nursing*, vol. 18, no. 16, pp. 2337-2343, 2009. doi: 10.5393/JAMCH.2010.35.4.370.
- [11] K. S. You, H. O. Lee, J. J. Fitzpatrick, S. Kim, E. Marui, J. S. Lee, and P. Cook, Spirituality, depression, living alone, and perceived health among Korean older adults in the community. *Archives of Psychiatric Nursing*, vol. 23, no. 4, pp. 309-322, 2009. doi: 10.1016/j.apnu.2008.07.003.
- [12] J. Y. Kim, S. G. Lee, and S. K. Lee, The relationship between health behaviors, health status, activities of daily living and health-related Quality of Life in the Elderly. *Journal of the Korean Gerontological Society*, vol. 30, no. 2, pp. 471-484, 2010. doi: 10.21215/kjfp.2018.8.2.220.
- [13] N. C. Davis, and D. Friedrich, Knowledge of aging and life satisfaction among older adults. *The International Journal of Aging and Human Development*, vol. 59, no. 1, pp. 43-61, 2004. doi: 10.2190/U9WD-M79K-9HB8-G9JY.
- [14] A. T. F. Beekman, D. M. W. Kriegsman, D. J. H. Deeg, and W. Van Tilburg, The association of physical health and depressive symptoms in the older population: age and sex differences. *Social Psychiatry and Psychiatric Epidemiology*, vol. 30, no. 1, pp. 32-38, 1995. doi: 10.1007/bf00784432.
- [15] R. Choi, and B. D. Hwang, Health care utilization of age group in the elderly on the Korean health panel. *The Korean Journal of Health Service Management*, vol. 8, no. 3, pp. 49-61, 2014. doi: 10.12811/kshsm.2014.8.3.049.
- [16] S. Suh, H. Choi, C. Lee, M. Cha, and I. Jo, Association between knowledge and attitude about aging and life satisfaction among older Koreans. *Asian Nursing Research*, vol. 6, no. 3, pp. 96-101, 2012. doi: 10.1016/j.anr.2012.07.002.
- [17] S. Shiovitz-Ezra, and H. Litwin, Social network type and health-related behaviors: evidence from an American national survey. *Social Science & Medicine*, vol. 75, no. 5, pp. 901-904, 2012. doi: 10.1016/j.socscimed.2012.04.031.
- [18] G. Low, A. E. Molzahn, and D. Schopflocher, Attitudes to aging mediate the relationship between older peoples' subjective health and quality of life in 20 countries. *Health and Quality of Life Outcomes*, vol. 11, no. 1, pp. 1-10, 2013. doi: 10.1186/1477-7525-11-146.
- [19] M. Luo, D. Ding, A. Bauman, J. Negin, and P. Phongsavan, Social engagement pattern, health behaviors and subjective well-being of older adults: an international perspective using WHO-SAGE survey data. *BMC Public Health*, vol. 20, no. 1, p. 99, 2020. doi: 10.1186/s12889-019-7841-7.
- [20] F. Desmyter, and R. De Raedt, The relationship between time perspective and subjective well-being of older adults. *Psychologica Belgica*, vol. 52, no. 1, pp. 19-38, 2012. doi: 10.5334/pb-52-1-19.
- [21] H. Litwin, and K. J. Stoeckel, Social networks and subjective wellbeing among older Europeans: does age make a difference?. *Ageing and Society*, vol. 33, no. 7, pp. 1263-1281, 2013. doi: 10.1017/S0144686X12000645.
- [22] H. Galenkamp, A. W. Braam, M. Huisman, and D. J. Deeg, Somatic multimorbidity and self-rated health in the older population. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, vol. 66, no. 3, pp. 380-386, 2011. doi: 10.1093/geronb/gbr032.
- [23] H. Byeon, A laryngeal disorders prediction model based on cluster analysis and regression analysis. *Medicine*, vol. 98, no. 31, pp. e16686, 2019. doi: 10.1097/md.00000000000016686.
- [24] H. Byeon, Development of a physical impairment prediction model for Korean elderly people using synthetic minority over-sampling technique and XGBoost. *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 1, pp. 36-41, 2021. doi: 10.14569/IJACSA.2021.0120105.
- [25] H. Byeon, Predicting the anxiety of patients with Alzheimer's dementia using boosting algorithm and data-level approach. *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 3, pp. 107-113, 2021. doi: 10.14569/IJACSA.2021.0120313.
- [26] J. H. Friedman, Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, vol. 29, no. 5, pp. 1189-1232, 2001. doi: 0.1214/aos/1013203451.
- [27] M. F. A. Saputra, T. Widiyaningtyas, and A. P. Wibawa, Illiteracy classification using K means-Naïve Bayes algorithm. *JOIV: International Journal on Informatics Visualization*, vol. 2, no. 3, pp. 153-158, 2018. doi: 10.30630/joiv.2.3.129.
- [28] L. Rutkowski, M. Jaworski, L. Pietruczuk, and P. Duda, The CART decision tree for mining data streams. *Information Sciences*, vol. 266, pp. 1-15, 2014. doi: 10.1016/j.ins.2013.12.060.
- [29] B. Belean, M. Streza, S. Crisan, and S. Emerich, Dorsal hand vein pattern analysis and neural networks for biometric authentication. *Studies in Informatics and Control*, vol. 26, no. 3, pp. 305-314, 2017. doi: 10.24846/v26i3y201706.
- [30] B. Mishra, N. Kumar, and M. S. Mukhtar, Systems biology and machine learning in plant-pathogen interactions. *Molecular Plant-Microbe Interactions*, vol. 32, no. 1, pp. 45-55, 2019. doi: 10.1094/MPMI-08-18-0221-FI.
- [31] C. Su, Z. Xu, J. Pathak, and F. Wang, Deep learning in mental health outcome research: a scoping review. *Translational Psychiatry*, vol. 10, no.1, pp. 1-26, 2020. doi: 10.1038/s41398-020-0780-3.
- [32] H. Wang, M. Lei, Y. Chen, M. Li, and L. Zou, Intelligent identification of maceral components of coal based on image segmentation and classification. *Applied Sciences*, vol. 9, no. 16, pp. 3245, 2019. doi: 10.3390/app9163245.
- [33] D. B. da Silva, D. Schmidt, C. A. da Costa, R. da Rosa Righi, and B. Eskofier, DeepSigns: a predictive model based on deep learning for the early detection of patient health deterioration. *Expert Systems with Applications*, vol. 165, p. 113905, 2021. doi: 10.1016/j.eswa.2020.113905.
- [34] S. Y. Im, The association of social network and health status among the elderly in Korea. master's thesis, Public administration The Graduated School of Ewha Womans University, Seoul, 2008.
- [35] K. H. Kim, and J. H. Kim, The effects of self-esteem on the relationship between the elderly depression and life satisfaction. *Family and Culture*, vol. 20, no. 4, pp. 95-116, 2008. doi: 10.21478/family.20.4.200812.004.
- [36] T. Lampert, and J. Hoebel, Socioeconomic inequalities in health in later life. *Zeitschrift für Gerontologie und Geriatrie*, vol. 52, no. Suppl 1, pp. 91-99, 2018. doi: 10.1007/s00391-018-01487-y.
- [37] S. K. Kim, The socioeconomic status and the self-reported health of the aged. *International Journal of Welfare for the Aged*, vol. 28, 2005.