Development of a Physical Impairment Prediction Model for Korean Elderly People using Synthetic Minority Over-Sampling Technique and XGBoost

Haewon Byeon
Department of Medical Big Data
College of AI Convergence, Inje University
Gimhae 50834, Gyeonsangnamdo, South Korea

***Abstract*—The old people's 'physical functioning' is a key factor of active ageing as well as a major factor in determining the quality of life and the need for long-term care in old age. Previous studies that identified factors related to ADL mostly used regression analysis to predict groups of high physical impairment risk. Regression analysis is useful for confirming individual risk factors, but has limitations in grasping multiple risk factors. As methods for resolving this limitation of regression models, machine learning ensemble boosting models such as random forest and eXtreme Gradient Boosting (XGBoost) are widely used. Nonetheless, the prediction performances of XGBoost, such as accuracy and sensitivity, remain to be verified additionally by follow-up studies. This article proposes an effective method of dealing with imbalanced data for the development of ensemble-based machine learning, by comparing the performances of disease data sampling methods. This study analyzed 3,351 old people aged 65 or above who resided in local communities and completed the survey. As machine learning models to predict physical impairment in old age, this study compared the logistic regression model, XGBoost and random forest, with respect to the predictive performances of accuracy, sensitivity, and specificity. This study selected as the final model a model whose sensitivity and specificity were 0.6 or above and whose accuracy was highest. As a result, synthetic minority over-sampling technique (SMOTE)-based XGBoost whose accuracy, sensitivity, and specificity were 0.67, 0.81, and 0.75, respectively, was determined as the most excellent predictive performance. The results of this study suggest that in case of developing a predictive model using imbalanced data like disease data, it is efficient to use the SMOTE-based XGBoost model.**

***Keywords*—Random forest; XGBoost; GBM; gradient boosting machine; physical impairment prediction model**

I. Introduction

According as ageing progresses across the world, ageing-related new concepts such as 'healthy ageing' and 'successful ageing' have emerged [1]. There are several standards for successful ageing, but in general, successful ageing is defined as having the high levels of physical, psychological, and social functions and satisfaction with life, a step further from physically healthy ageing [2]. It has been reported that factors affecting the successful ageing include age, abstaining from smoking, disability, arthritis, and diabetes, and particularly that the better their subjective health, family support and physical activities, the higher their level of successful aging [1,3,4,5].

The World Health Organization (WHO) introduced the concept of active ageing in order to promote the development of policies to cope with the problem of ageing [6,7]. According to the definition of WHO [6], active ageing is the process of optimizing opportunities for health, participation and security in order to enhance quality of life as people age. That is, active ageing supports people with ADL functions so that they actively participate in social activities, and induces people with ADL dysfunction to actively perform daily life by enhancing their ADL functions with appropriate support [8].

On the other hand, old people's 'physical functioning' is a key factor of active ageing as well as a major factor in determining the quality of life and the need for long-term care in old age [9,10]. Old people's state of physical functioning is mostly assessed in terms of activities of daily living (ADL), with which it can be judged whether an elderly can lead an independent life or not [11,12]. For the assessment of ADL, Katz Index, Barthel Index, and MBI are usually used [13]; and in the Korea National Health and Nutrition Examination Survey, the Korean Activities of Daily Living scale(K-ADL), a standardized test tool for physical functioning developed by reflecting Korean old people's living environment and culture, was used [14]. Old age, low educational level, the beneficiary of medical benefits, non-subscriber of health insurance, stroke, urinary incontinence, diabetes, and lung cancer have been reported as risk factors affecting K-ADL [15-21].

Previous studies that identified factors related to ADL [15-21] mostly used regression analysis to predict groups of high physical impairment risk. Regression analysis is useful for confirming individual risk factors, but has limitations in grasping multiple risk factors [22,23]. In addition, regression models assume the independence and normality of variables; however, it is difficult to derive accurate results in case of data that violate the normality of distribution, as in disease [24]. As methods for resolving this limitation of regression models, machine learning ensemble boosting models such as random forest, gradient boosting machine (GBM), and eXtreme Gradient Boosting (XGBoost) are widely used [25,26]. Ensemble learning is a technique for deriving more accurate final prediction by generating several classifiers and combining their predictions, and some previous studies [27,28] have reported that XGBoost developed recently shows performance exceeding that of the existing random forest or gradient boosting. Nonetheless, the prediction performances of

XGBoost, such as accuracy and sensitivity, remain to be verified additionally by follow-up studies.

On the other hand, it is highly probable that the problem of imbalanced data will occur in the prediction of impairment using big data [29]. Particularly, in the case of disease data, data are highly probable to distribute unequally because generally the number of patients is very fewer than those without disease. These imbalanced data cause prediction error in the process of machine learning and deteriorate the performance of a model, and thus techniques for dealing with imbalanced data are required in order to resolve this problem [30].

Hence, first, this article prepares basic data for policy-making to respond to ageing by predicting and analyzing the tendencies of physical impairment risk among Korean old people in local communities, and second, this article proposes an effective method of dealing with imbalanced data for the development of ensemble-based machine learning, by comparing the performances of disease data sampling methods.

II. Research Methods

A. Sources of Data

This study used and analyzed the raw data of Seoul Panel Study Data (SEPANS), which was carried out with Seoul citizens by the Seoul Welfare Foundation from June 1, 2016 to August 31, 2016. The SEPANS was conducted to grasp the welfare levels of households residing in Seoul, find out the actual state of vulnerable groups, and estimate demand for welfare service. Its population was households in Seoul as of the survey period among households subjected to 2005 Population and Housing Census, excluding foreigners and those in nursing homes, the military, and prisons. As for the sampling method, the stratified cluster sampling was used. As for the survey method, the computer-aided personal interview was used in which an interviewer visited households to be surveyed and inputted their responses to a structured questionnaire into a portable computer. This study analyzed 3,351 old people aged 65 or above who resided in local communities and completed the survey.

B. Measurement of Variables

The outcome variable was defined as physical impairment of the elderly measured by means of K-ADL, a standardized test. According to Won (2002) [14], reliability and validity were high according as the reliability coefficient of K-ADL was 0.7 or higher at the stage of test development (standardization) and the inter-item consistency of the questionnaire was 0.937. K-ADL consisted of 7 items of the most basic physical functions in daily life, including dressing, washing the face, bathing, self-feeding, moving out of bed, using the toilet, and relieving oneself. In the event of answering with partial help or complete dependence to any item of K-ADL, the respondent was classified into the group of physical impairment, and in the event of answering with complete independence to all the items, the respondent was classified into the group of non-physical impairment.

Explanatory variables included sex, age, educational background (elementary school or below, middle school, high school, college graduate or above), being insured or not, stroke, diabetes, arthritis, monthly total household income (below KRW 2 million, KRW 2-4 million, KRW 4 million or above), the presence of a spouse (cohabiting with a spouse, having but not cohabiting with a spouse, no spouse), smoking (non-smoking, smoking in the past, smoking at present), and the presence of depressive symptoms (yes, no), which were reported to have associations with Korean old people's ADL.

C. Predictive Model

As machine learning models to predict physical impairment in old age, this study compared the logistic regression model, XGBoost and random forest, with respect to the predictive performances of accuracy, sensitivity, and specificity. In testing the predictive performances, data were randomly divided into train data and test data in the proportion of 7:3; a predictive model was generated from the train data, and the performance of the model was tested with the test data. Random forest and XGBoost are models that include randomness, and the models were developed with the seed being fixed to 1234. The value of predictive performance for each model was predicted by means of the Area Under the Curve (AUC) of ROC curve (Fig. 1). As model performance assessment indices, the accuracy, sensitivity, and specificity of each model were obtained. Accuracy is the ratio of successful predictions to all predictions. Sensitivity is the ratio of a predictive model's predicting accurately old people to whom actual impairment will occur. Specificity is the ratio of a predictive model's predicting accurately that impairment will not occur to healthy old people to whom impairment will not occur actually. This study defined as the model of the best predictive performance a model whose sensitivity and specificity were 0.6 or above and whose accuracy was found to be highest after comparison with other models; and selected it as the final model for the prediction of physical impairment in old age. In all the analyses, R version 4.0.2 (Foundation for Statistical Computing, Vienna, Austria) was used.

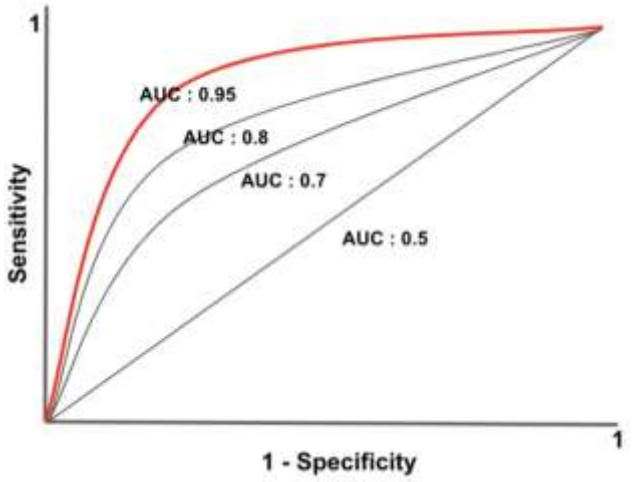

Fig. 1. Concepts of ROC Curve.

D. Ensemble Model

The ensemble model combines the prediction or classification results of several models, and use them in final decision-making, and a number of studies have shown that the model has better predictive performance than single decision tree models [31]. Ensemble methods are classified into boosting and bagging. Bagging is a method of generating several models through the sampling of source data and then making prediction by combining the outcomes of the models by voting or averaging; and reduces the variance of predicted values [32]. Boosting is a machine learning algorithm; is a method for better classifying observation values difficult to classify, by using more misclassified observation values; and reduces the bias of predictive values [33]. The concepts of boosting and bagging are presented in Fig. 2.

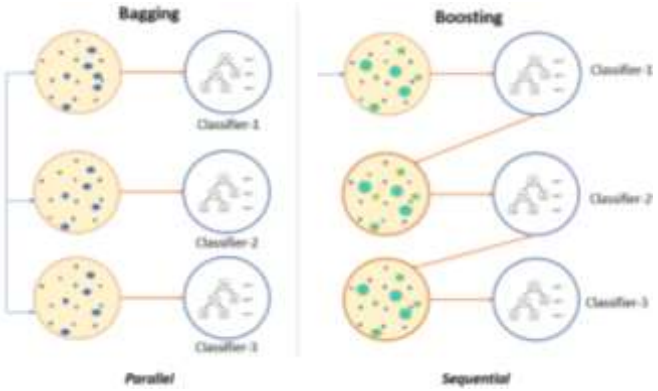

Fig. 2. Concepts of Bagging and Bossting Algorithm [34].

E. Random Forest

The forest ensemble is an ensemble form of decision trees. The decision tree is a model that divides the scope of variables in each branching, and can be used regardless of continuous/categorical target variables. It has the advantage of being capable of explaining a model easily, but its performance drops in the event that data has a structure not easily divided with horizontal partitioning or vertical partitioning [35]. A method developed to remedy this shortcoming is the random forest. The random forest samples data, generates several tree models, and then votes or averages the outcomes of the trees. It is similar to bagging, but is different from bagging in that it supplements the problem of multicollinearity by random selection of variables as well as sampling of data [36]. The concept of random forest is presented in Fig. 3.

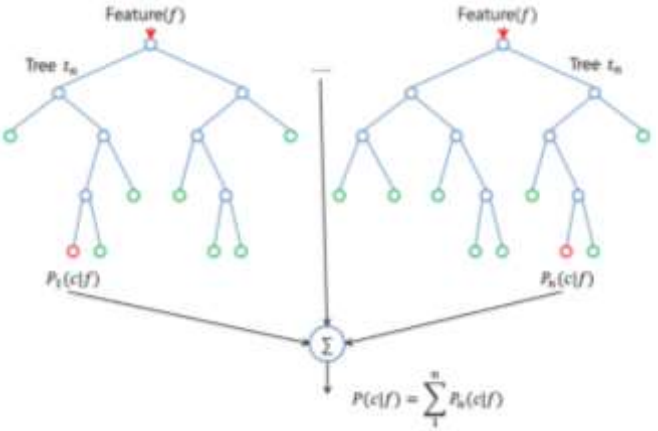

Fig. 3. Concepts of Random Forest Algorithm [37].

F. XGBoost

XGBoost is one of boosting methods, and uses a misclassified observation value more in the next model when generating a tree [38]. That is, it is a boosting algorithm that trains with a method for improving performance as to misclassified observation values. XGBoost has the advantage of speedy calculation process owing to parallel computing that uses all CPU cores in learning, and is very useful because it supports various programming languages including Python and R [38]. The concept of XGBoost is presented in Fig. 4.

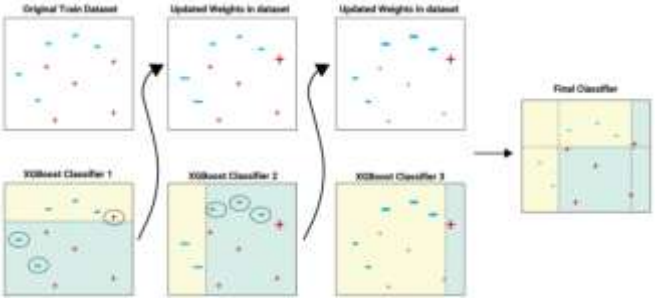

Fig. 4. Concepts of XGBoost Algorithm [39].

G. Sampling

Disease data generally have the problem of imbalance because the number of patients is fewer than healthy people. In the case of data used in this study, the ratio of normal old people was found to be 92%, and the ratio of old people with physical impairment only 8.0%, respectively, as a result of ADL assessment, which shows the problem of imbalance. To resolve the problem of imbalance, this study used the algorithms of under-sampling [40], over-sampling [41], and synthetic minority over-sampling technique (SMOTE) [30].

The under-sampling is a method of resolving the problem of data imbalance by randomly removing the major class among classes of response variables. The technique of under-sampling can reduce the speed of model construction by removing data amount, but has the shortcoming of information loss.

The over-sampling is a method of resolving the problem of imbalance by randomly copying the minor class among classes of response variables. The over-sampling technique may cause the problem of overfitting because the speed of model construction increases due to the increase in data amount and it copies a small number of categories.

SMOTE (Synthetic Minority Over-sampling Technique) is a method for supplementing overfitting, the shortcoming of over-sampling. One of minor classes among classes of response variables is randomly chosen, and then k nearest neighbors of this data is found. And the difference between this chosen sample and k neighbors is obtained, and the difference multiplied by any value between 0 and 1 is added to the existing sample, and then the resulting value is added to the training data. Lastly, this process is repeated. The SMOTE

algorithm is similar to over-sampling in that it increases data of a minor class of few categories, but it is known that it supplements overfitting, the shortcoming of over-sampling, through creating a new sample by properly combining the existing data, not copying the same data. The concepts of sampling types are presented in Fig. 5.

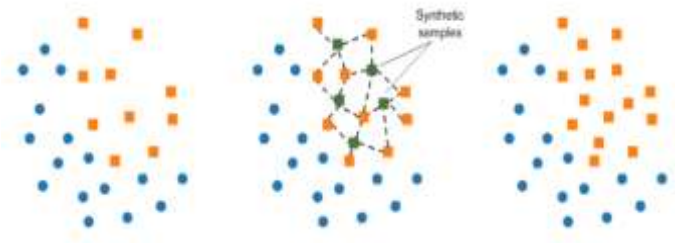

Fig. 5. Concepts of SMOTE Algorithm [42].

III. Results

A. Accuracy of Predictive Models

The accuracy, sensitivity, and specificity of predictive models to which sampling methods were applied are presented in Table I. This study selected as the final model a model whose sensitivity and specificity were 0.6 or above and whose accuracy was highest. As a result, SMOTE-based XGBoost whose accuracy, sensitivity, and specificity were 0.67, 0.81, and 0.75, respectively, was determined as the final predictive model.

B. Results of XGBoost Model Development

The model to predict the physical impairment was developed through the Xgboost and the predictive power was compared with the results of random forest and logistic regression (Table II). Xgboost had higher classification accuracy than other predictive model in both training and test data. The analysis results of test data showed that the classification accuracy was 67.2% for Xgboost, 65.0% for logistic regression, and 62.1% for random forest.

TABLE I. Performance (accuracy, sensitivity, and specificity) of predictive models to which sampling methods were applied

Model		Random Forest	Logistic regression	XGBoost
Raw data	Accuracy	0.73	0.65	0.75
	Sensitivity	0.50	0.43	0.63
	Specificity	0.85	0.94	0.90
Under-sampling	Accuracy	0.63	0.49	0.60
	Sensitivity	0.63	0.48	0.74
	Specificity	0.79	1.00	0.93
Over-sampling	Accuracy	0.54	0.62	0.77
	Sensitivity	0.52	0.63	0.65
	Specificity	0.80	0.91	0.93
SMOTE	Accuracy	0.62	0.65	0.67
	Sensitivity	0.68	0.70	0.81
	Specificity	0.81	0.78	0.75

TABLE II. Results of model to predict the physical impairment

Model	Factors	Characteristics
Random forest	9	sex, age, educational background, being insured or not, stroke, diabetes, arthritis, the presence of a spouse, the presence of depressive symptoms
Logistic regression	6	sex, age, educational background, stroke, diabetes, arthritis
XGBoost	7	sex, age, educational background, being insured or not, stroke, diabetes, arthritis

IV. Conclusion

Sensitivity and specificity are in the relationship of trade-off. Therefore, the proportions of sensitivity and specificity are selected by the judgment of a researcher who uses a model. In this article, among random forest, logistic regression and XGBoost, the SMOTE-based XGBoost model, which showed the sensitivity and specificity of 0.6 or above and the highest accuracy, was derived as the final model of the most excellent predictive performance.

Similarly to the results of this study, previous studies also reported that XGBoost is more excellent than other ensemble models, such as GBM, in terms of accuracy [27,28]. It is presumed that XGBoost displayed excellent predictive performance in the areas of classification and regression because although it, one of tree-based ensemble learning algorithms, is based on GBM, it is equipped with its own functions of overfitting regularization and early stopping [27,28]. Further, previous studies [43,44] reported that XGBoost shows faster execution time than GBM. Therefore, the results of this study suggest that in case of developing a predictive model using imbalanced data like disease data, it is efficient to use the SMOTE-based XGBoost model.

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).

References

[1] S. H. Kim, S. Park, A meta-analysis of the correlates of successful aging in older adults. Research on Aging, vol. 39, no. 5, pp. 657-677, 2017.

[2] J. Woo, J. Leung, T. Zhang, Successful aging and frailty: opposite sides of the same coin?. Journal of the American Medical Directors Association, vol. 17, no. 9, pp. 797-801, 2016.

[3] J. Mana, O. Bezdicek, Cognition in successful aging: Systematic review and future directions. Clinical Gerontologist, pp. 1-9, 2020.

[4] B. Teater, J. M. Chonody, How do older adults define successful aging? A scoping review. The International Journal of Aging and Human Development, vol. 91, no. 4, pp. 599-625, 2020.

[5] J. H. Cho, The Influence of Self-Efficacy, Self-Esteem, Aging Anxiety on Successful Aging in Middle-Aged Women. Medico Legal Update, vol. 20, no. 1, pp. 2265-2270, 2020.

[6] A. I. Hijas-Gómez, A. Ayala, M. P. Rodríguez-García, C. Rodríguez-Blázquez, V. Rodríguez-Rodríguez, F. Rojo-Pérez, G. Fernández-Mayoralas, A. Rodríguez-Laso, A. Calderón-Larrañaga, M. J. Forjaz, The WHO active ageing pillars and its association with survival: Findings from a population-based study in Spain. Archives of Gerontology and Geriatrics, vol. 90, 104114, 2020.

[7] E. Thalassinos, M. Cristea, G. G. Noja, Measuring active ageing within the European Union: implications on economic development.

Equilibrium. Quarterly Journal of Economics and Economic Policy, vol. 14, no. 4, pp. 591-609, 2019.

[8] D. Sánchez-González, F. Rojo-Pérez, V. Rodríguez-Rodríguez, G. Fernández-Mayoralas, Environmental and psychosocial interventions in age-friendly communities and active ageing: a systematic review. International Journal of Environmental Research and Public Health, vol. 17, no. 22, 8305, 2020.

[9] L. Dipietro, W. W. Campbell, D. M. Buchner, K. I. Erickson, K. E. Powell, B. Bloodgood, T. Hughes, K. R. Day, K. L. Piercy, A. Vaux-Bjerke, R. D. Olson, 2018 Physical Activity Guidelines Advisory Committee, Physical Activity, Injurious Falls, and Physical Function in Aging: An Umbrella Review. Medicine and Science in Sports and Exercise, vol. 51, no. 6, pp. 1303-1313, 2019.

[10] D. O'Neill, D. E. Forman, The importance of physical function as a clinical outcome: Assessment and enhancement. Clinical Cardiology, vol. 43, no. 2, pp. 108-117, 2020.

[11] M. A. Sharkawi, S. M. Zulfarina, S. M. Z. Aqilah-SN, N. M. Isa, A. M. Sabarul, A. S. Nazrun, Systematic review on the functional status of elderly hip fracture patients using Katz Index of Activity of Daily Living (Katz ADL) score. IIUM Medical Journal Malaysia, vol. 15, no. 2, E-pub, doi:10.31436/imjm.v15i2.397, 2016.

[12] F. Sharifi, M. Alizadeh-Khoei, H. Saghebi, L. Angooti-Oshnari, S. Fadaee, S. Hormozi, F. Taati, M. Haghi, H. Fakhrzadeh, Validation study of ADL-Katz scale in the Iranian elderly nursing homes. Ageing International, vol. 43, no. 4, pp. 508-523, 2018.

[13] L. T. Y. D. Silveira, J. M. D. Silva, J. M. P. Soler, C. Y. L. Sun, C. Tanaka, C. Fu, Assessing functional status after intensive care unit stay: the Barthel Index and the Katz Index. International Journal for Quality in Health Care, vol. 30, no. 4, pp. 265-270, 2018.

[14] C. W. Won, K. Y. Yang, Y. G. Rho, S. Y. Kim, E. J. Lee, J. L. Yoon, K. H. Cho, H. C. Shin, B. R. Cho, J. R. Oh, D. K. Yoon, H. S. Lee, Y. S. Lee, The development of Korean activities of daily living (K-ADL) and Korean instrumental activities of daily living (K-IADL) scale. Journal of the Korean Geriatrics Society, vol. 6, no. 2, pp. 107-120, 2002.

[15] S. Y. Sohn, Factors Related to the Health Related Quality. of Life in Elderly Women. Korean Journal of Women Health Nursing, vol. 15, no. 2, pp. 99-107, 2009.

[16] S. M. Bang, J. O. Lee, Y. J. Kim, K. W. Lee, S. Lim, J. H. Kim, Y. J. Park, H. J. Chin, K. W. Kim, H. C. Jang, J. S. Lee, Anemia and activities of daily living in the Korean urban elderly population: results from the Korean Longitudinal Study on Health and Aging (KLoSHA). Annals of Hematology, vol. 92, no. 1, pp. 59-65, 2013.

[17] J. W. Noh, K. B. Kim, J. H. Lee, M. H. Kim, Y. D. Kwon, Relationship of health, sociodemographic, and economic factors and life satisfaction in young-old and old-old elderly: a cross-sectional analysis of data from the Korean longitudinal study of aging. Journal of Physical Therapy Science, vol. 29, no. 9, pp. 1483-1489, 2017.

[18] H. M. Ku, J. H. Kim, H. S. Lee, H. J. Ko, E. J. Kwon, S. Jo, D. K. Kim, A study on the reliability and validity of Seoul-Activities of Daily Living (S-ADL). Journal of the Korean Geriatrics Society, vol. 8, no. 4, pp. 206-214, 2004.

[19] H. Byeon, H. W. Koh, The relationship between communication activities of daily living and quality of life among the elderly suffering from stroke. Journal of Physical Therapy Science, vol. 28, no. 5, pp. 1450-1453, 2016.

[20] H. Byeon, S. Cho, Association between Drinking Behavior and Activities of Daily Living in Community-dwelling Older Adults. International Journal of Bio-Science and Bio-Technology, vol. 7, no. 4, pp. 135-144, 2015.

[21] K. S. Ahn, S. K. Park, Y. C. Cho, Physical Function (ADL, IADL) and Related Factors in the Elderly People Institutionalized in Long-term Care Facilities. Journal of the Korea Academia-Industrial cooperation Society, vol. 17, no. 3, pp. 480-488, 2016.

[22] H. Byeon, Developing a model to predict the occurrence of the cardio-cerebrovascular disease for the Korean elderly using the random forests algorithm. International Journal of Advanced Computer Science and Applications, vol. 9, no. 9, pp. 494-499, 2018.

[23] H. Byeon, S. Cha, K. Lim, Exploring Factors Associated with Voucher Program for Speech Language Therapy for the Preschoolers of Parents with Communication Disorder using Weighted Random Forests. International Journal of Advanced Computer Science and Applications, vol. 10, no. 5, pp. 12-17, 2019.

[24] H. Byeon, Model development for predicting the occurrence of benign laryngeal lesions using support vector machine: focusing on South Korean adults living in local communities. International Journal of Advanced Computer Science and Applications, vol. 9, no. 10, pp. 222-227, 2018.

[25] J. Nobre, R. F. Neves, Combining principal component analysis, discrete wavelet transform and XGBoost to trade in the financial markets. Expert Systems with Applications, vol. 125, pp. 181-194, 2019.

[26] H. Nguyen, X. N. Bui, H. B. Bui, D. T. Cuong, Developing an XGBoost model to predict blast-induced peak particle velocity in an open-pit mine: a case study. Acta Geophysica, vol. 67, pp. 477-490, 2019.

[27] E. K. Sahin, Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. SN Applied Sciences, vol. 2, 1308, 2020.

[28] T. Hengl, J. G. Leenaars, K. D. Shepherd, M. G. Walsh, G. B. Heuvelink, T. Mamo, H. Tilahun, E. Berkhout, M. Cooper, E. Fegraus, I. Wheeler, N. A. Kwabena, Soil nutrient maps of Sub-Saharan Africa: assessment of soil nutrient content at 250 m spatial resolution using machine learning. Nutrient Cycling in Agroecosystems, vol. 109, pp. 77-102, 2017.

[29] A. Fernández, S. García, F. Herrera, N. V. Chawla, SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. Journal of Artificial Intelligence Research, vol. 61, pp. 863-905, 2018.

[30] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.

[31] M. Skurichina, R. P. Duin, Bagging, boosting and the random subspace method for linear classifiers. Pattern Analysis and Applications, vol. 5, pp. 121-135, 2002.

[32] Y. Grandvalet, Bagging equalizes influence. Machine Learning, vol. 55, pp. 251-270, 2004.

[33] A. Mayr, B. Hofner, E. Waldmann, T. Hepp, S. Meyer, O. Gefeller, An update on statistical boosting in biomedicine. Computational and Mathematical Methods in Medicine, vol. 2017, E-pub, 6083072, doi:10.1155/2017/6083072, 2017.

[34] https://www.pluralsight.com/guides/ensemble-methods:-bagging-versus-boosting.

[35] F. Tang, H. Ishwaran, Random forest missing data algorithms. Statistical Analysis and Data Mining: The ASA Data Science Journal, vol. 10, no. 6, pp. 363-377, 2017.

[36] P. Thanh Noi, M. Kappas, Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery. Sensors, vol. 18, no. 1, 18, 2018.

[37] https://builtin.com/data-science/random-forest-algorithm.

[38] R. Mitchell, E. Frank, Accelerating the XGBoost algorithm using GPU computing. PeerJ Computer Science, vol. 3, E-pub, e127, doi:10.7717/peerj-cs.127, 2017.

[39] https://blog.thinknewfound.com/2020/05/defensive-equity-with-machine-learning/xg-boost-final-01/.

[40] S. J. Yen, Y. S. Lee, Cluster-based under-sampling approaches for imbalanced data distributions. Expert Systems with Applications, vol. 36, no. 3, pp. 5718-5727, 2009.

[41] L. Abdi, S. Hashemi, To combat multi-class imbalanced problems by means of over-sampling techniques. IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 1, pp. 238-251, 2015.

[42] https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets.

[43] Z. Ding, H. Nguyen, X. N. Bui, J. Zhou, H. Moayedi, Computational intelligence model for estimating intensity of blast-induced ground vibration in a mine based on imperialist competitive and extreme gradient boosting algorithms. Natural Resources Research, vol. 29, pp. 751-769, 2020.

[44] A. Kadiyala, A. Kumar, Applications of python to evaluate the performance of decision tree - based boosting algorithms. Environmental Progress and Sustainable Energy, vol. 37, pp. 618-623, 2018.