Predicting the Depression of the South Korean Elderly using SMOTE and an Imbalanced Binary Dataset

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

***Abstract*—Since the number of healthy people is much more than that of ill people, it is highly likely that the problem of imbalanced data will occur when predicting the depression of the elderly living in the community using big data. When raw data are directly analyzed without using supplementary techniques such as a sample algorithm for datasets, which have imbalanced class ratios, it can decrease the performance of machine learning by causing prediction errors in the analysis process. Therefore, it is necessary to use a data sampling technique for overcoming this imbalanced data issue. As a result, this study tried to identify an effective way for processing imbalanced data to develop ensemble-based machine learning by comparing the performance of sampling methods using the depression data of the elderly living in South Korean communities, which had quite imbalanced class ratios. This study developed a model for predicting the depression of the elderly living in the community using a logistic regression model, gradient boosting machine (GBM), and random forest, and compared the accuracy, sensitivity, and specificity of them to evaluate the prediction performance of them. This study analyzed 4,085 elderly people (≥60 years old) living in the community. The depression data of the elderly in the community used in this study had an unbalance issue: the result of the depression screening test showed that 87.5% of subjects did not have depression, while 12.5% of them had depression. This study used oversampling, undersampling, and SMOTE methods to overcome the unbalance problem of the binary dataset, and the prediction performance (accuracy, sensitivity, and specificity) of each sampling method was compared. The results of this study confirmed that the SMOTE-based random forest algorithm showing the highest accuracy (a sensitivity ≥ 0.6 and a specificity ≥ 0.6) was best prediction performance among random forest, GBM, and logistic regression analysis. Further studies are needed to compare the accuracy of SMOTE, undersampling, and oversampling for imbalanced data with high dimensional y-variables.**

Keywords*—Random forests; gradient boosting machine; SMOTE; undersampling; imbalanced data; oversampling

I. Introduction

Depression is one of the important mood disorders at senescence. It is very important to diagnose and treat depression at an early stage because it is possible to treat and cure depression using medication or psychosocial therapy even after its onset [1]. Depressive symptoms in old age differ from those in young age. First, it is difficult to clearly distinguish depressive symptoms from dementia symptoms [2]. Pseudodementia, similar to dementia, shows a decline in cognitive ability in the dementia screening test, similar to the cognitive function test result of depression [3, 4]. In particular, the elderly accompanied by depression often express a subjectively recognized decrease in memory and cognitive function, which are not common with adolescents [5,6]. Moreover, the elderly with depression suffer from a decrease in memory and cognition more than the healthy elderly [5, 6].

Second, even though young patients complain about various physical symptoms, the key to diagnose depression, these physical symptoms are not very useful for diagnosing depression for elderly patients. For example, sleep disorder is a common symptom in adolescent depression, but elderly people frequently experience it regardless of depression [7,8]. Physical symptoms such as a normal decline in sexual function, constipation, and joint pain, associated with aging, are commonly found even in the elderly without depression [9]. Consequently, it is critical to accurately determine whether the depressive symptoms complained by the elderly are due to normal aging or depressive disorder.

Nevertheless, most of the studies that evaluated the depression of South Korean elderly were mainly regarding the factual survey for one city in terms of mental health, depression assessment, and the effectiveness of interventions for depression prevention and management. There are much fewer predictive model studies for identifying the factors associated with the depression of the elderly living in the community than patient-control group comparison studies. Previous studies [1,10,11,12,13] that evaluated the factors related to geriatric depression in South Korea local communities reported that health, socioeconomic status, education level, age, spouse, and social activities affected geriatric depression. Since regression analysis was mainly used as a modeling method to predict depression, they were efficient in identifying individual risk factors [14,15]. However, they were limited in identifying compound-risk factors (multivariate) such as sociodemographic variables and living habits [14,15]. Moreover, since regression analysis assumes independence, normality, and homoscedasticity, there is a possibility of producing biased results when the model is developed using data in violation of normality [16]. As a way to overcome the limitation of the regression model, big data-based analysis, called machine learning or data mining, has been widely used in various fields. Machine learning can analyze data accurately even if the data somewhat violate the assumption of normality such as nonlinear data in the estimation process [17]. Especially, it has been known that gradient boosting machine (GBM), which generates many classifiers and combines the predictions to derive more accurate results, and ensemble learning models such as random forest have much higher sensitivity and accuracy than a single decision tree [18,19]. Nonetheless, since the predictive performance of the ensemble learning model has been mainly tested using simulation data [20], it is necessary to conduct additional validation and verification for confirming the predictive performance of the ensemble learning model for using it for disease data, which are mostly imbalanced data [21].

Since the number of healthy people is much more than that of ill people, it is highly likely that the problem of imbalanced data will occur when predicting the depression of the elderly living in the community using big data [22]. When raw data are directly analyzed without using supplementary techniques such as a sample algorithm for datasets, which have imbalanced class ratios, it can decrease the performance of machine learning by causing prediction errors in the analysis process [23]. Therefore, it is necessary to use a data sampling technique for overcoming this imbalanced data issue [24]. As a result, this study tried to identify an effective way for processing imbalanced data to develop ensemble-based machine learning by comparing the performance of sampling methods using the depression data of the elderly living in South Korean communities, which had quite imbalanced class ratios.

II. Methods and Materials

A. Data Source

This study analyzed the raw data of the 2016 Seoul Panel Study (SEPANS) data. The SEPANS data was conducted from June 1 to August 31, 2016, for the purpose of estimating the welfare level of Seoul citizens and the actual status situation of socially vulnerable class. The population of this study was the households in Seoul at the time of the survey among the households subject to the 2005 Population and Housing Census. The stratified cluster sampling method was used for sampling households in 25 districts in Seoul. This study excluded foreigners and those admitted to retirement homes or nursing hospitals among the survey subjects. This study used the computer aided personal interview method that an interviewer visited the target households and entered the response to the structured questionnaire into a portable computer. This study analyzed 4,085 elderly people ($\geqslant$60 years old) living in the community.

B. Variable Measurement

Depression, the outcome variable, was defined according to the Korean version of Center for Epidemiologic Studies Depression Scale-Revised (K-CESD) [25]. K-CES-D is a self-administered depression scale composed of 20 items and it was developed by the National Institute of Mental Health. It is a primary screening tool for depression. The maximum score is 60, and a higher score indicates more severe depression. The cut-off score of K-CES-D, the threshold of depression, was defined as 25 points.

Explanatory variables were age, gender, educational level (elementary school graduate and below, middle school graduate, high school graduate, or college graduate or above), smoking (smokers or non-smokers), drinking (less than once a week or twice or more per week), economic activity (yes or no), social activities for the past month (yes or no), mean monthly household income (less than KRW 1.5 million, KRW 1.5-3 million, or KRW 3 million or more), spouse living together (living together, bereavement/separated, or single), disease/accident/addiction in the last two weeks (yes or no), subjective health status (good, fair, or bad), subjective stress (yes or no), days of walking for 30 minutes or more per day (less than 1 day per week or 2 days or more per week), the frequency of meetings neighbors (less than once a month or twice or more per month), and the frequency of meeting relatives (less than once a month or twice or more per month).

III. Analysis

A. Model Development and Evaluation

This study developed a model for predicting the depression of the elderly living in the community using a logistic regression model, GBM, and random forest, and compared the accuracy, sensitivity, and specificity of them to evaluate the prediction performance of them. To test the prediction performance of them, the data were randomly divided into train dataset (70%) and test dataset (30%). Prediction models were developed using the training dataset and the accuracy, sensitivity, and specificity of them were calculated by using the test dataset. Since GBM and Random forest have random characteristics, models were developed while the seed was fixed as 123456 for repeated measurement. The predictive performance of each model was evaluated by the area under the curve (AUC) of the ROC curve, and the accuracy, sensitivity, and specificity of each model were calculated as evaluation indices for the model performance. Accuracy means the percentage of successful predictions in all data. Sensitivity indicates the rate of a model predicting a senior with depression as depression. Specificity is a true negative rate, indicating how accurately a model predicts a senior without depression and not depression. This study defined the best predictive performance model as a model with the highest accuracy while sensitivity and specificity were 0.6 or higher, and the model was selected as the final model for predicting the depression of the elderly living in the community. . All analyses were performed using R version 4.0.2 (Foundation for Statistical Computing, Vienna, Austria) and Python version 3.8.0 (https://www.python.org).

B. Random Forest

Random forest is an ensemble model and it generates a number of decision trees to calculate predictions. The ensemble model is a method of integrating the classification results of multiple decision trees and using them for making the final decision. A number of studies [26, 27, 28] reported that the ensemble model had higher predictive power than single decision tree models. The ensemble model can be divided into bagging and boosting. Bagging is a way to predict by combining the results of each model through averaging or voting after generating multiple decision tree models by sampling raw data. It has the advantage of reducing the variance of predicted values [29]. Boosting is a machine learning method that enables better classification for the observation values that are difficult to classify by using more misclassified observations. It has the advantage of reducing the bias of predicted values [30]. The concepts of bagging and boosting are presented in Fig. 1.

These ensemble models supplement the poor performance of decision trees when handling data that are not divided well into horizontal or vertical division [31]. Random forest samples data to create multiple tree models and then vote or average the results of each tree. It is widely used in various fields because it can handle the multicollinearity problem of trees by randomly selecting variables as well as sampling data from each model [32, 33]. The concept of random forest is presented in Fig. 2.

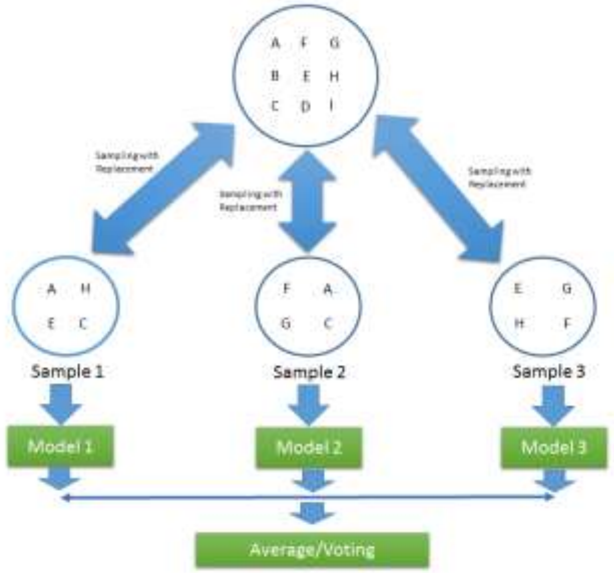


Fig. 1. The Concept of Bagging [34].

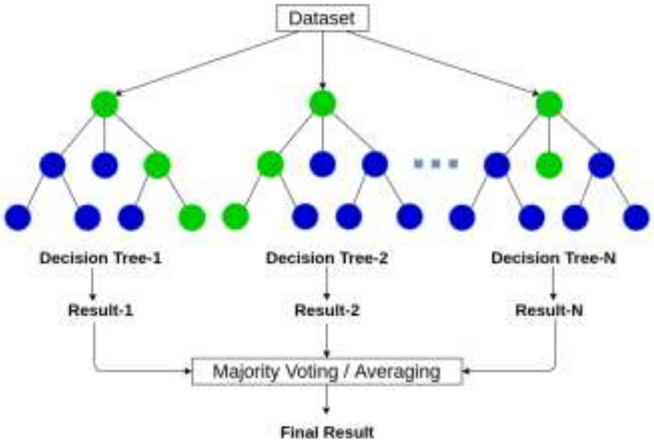
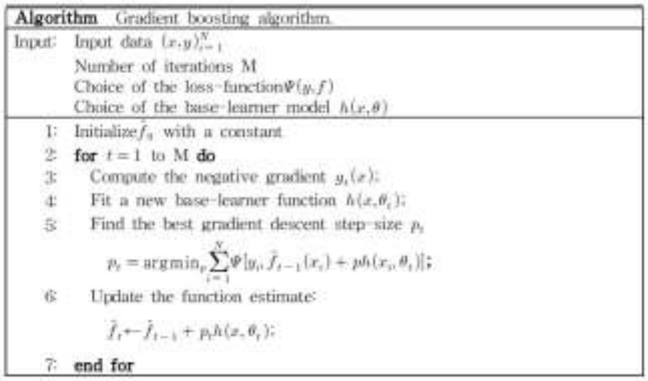


Fig. 2. Concept of Random Forest [35].

C. GBM

The GBM is a machine learning algorithm designed by Friedman (2001) [36] that generates a prediction model by combining weak learners of traditional decision trees using ensemble techniques. This model generalizes the model by generating models for each step and optimizing the loss function that can randomly differentiate, like other boosting methods. In machine learning, boosting refers to a method of generating strong learners by combining weak learners [36]. It generates a model even if the accuracy of it is low, and the error of this model is supplemented by the next model. A more accurate model is created through this process, and the basic principle of it is to increase accuracy by repeating this process. The prediction model learning is to find a parameter that minimizes the loss function. One of the ways to find the optimal parameter is gradient descent. When a slope is calculated by differentiating the loss function with parameters and moving the parameters in the direction of decreasing the value, it reaches the point where the loss function is minimized. In the gradient boosting process, this exploration process is carried out in the functional space. Therefore, it differentiates the loss function by the model function learned so far, instead of the parameter. GBM's algorithm is presented in Fig. 3.



```
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 $h(x,\theta)$
Initialize $\hat{f}_0$ with a constant
for $t=1$ to M do
   Compute the negative gradient $g_t(x)$;
   Fit a new base-learner function $h(x,\theta_t)$;
   Find the best gradient descent step-size $\rho_t$
      $\rho_t = \arg\min_{\rho} \sum_{i=1}^{N} \Psi[y_i, \hat{f}_{t-1}(x_i) + \rho h(x_i,\theta_t)]$;
   Update the function estimate:
      $\hat{f}_t \leftarrow \hat{f}_{t-1} + \rho_t h(x,\theta_t)$;
end for
```

Fig. 3. The Algorithm of Gradient Boosting Machine [36].

D. Sampling Techniques for Resolving Imbalanced Data

Disease data generally poses the problem of imbalanced classes because the number of people with a disease is smaller than those without a disease. The depression data of the elderly in the community used in this study also had an unbalance issue: the result of the depression screening test showed that 87.5% of subjects did not have depression, while 12.5% of them had depression. This study used oversampling [37], undersampling [38], and SMOTE [24] methods to overcome the unbalance problem of the binary dataset, and the prediction performance (accuracy, sensitivity, and specificity) of each sampling method was compared.

The undersampling method is a technique of randomly deleting data of multiple classes to match with the number of data in a class with small data. It is the fastest because it deletes data without conducting separate calculations, but the variation of performance is large because it deletes data randomly [38]. When a pair of data belonging to different classes and there is no data closer to each other, it is called

Tomek link. The Tomek link technique is a way to exclude data belonging to a class with more data. It has the effect of pushing the boundary line toward the class with many data. The edited nearest neighbors (ENN) technique is a technique that deletes the nearest k data out of a class with many data unless all or several of them belong to the class with many data. In other words, this technique deletes data of a class with more data that are around a class with fewer data. Since these traditional undersampling techniques delete data, they incur a loss of data and weaken the representativeness of data.

The oversampling technique is to use the data of a class with fewer data repetitively and randomly, which increases the weight. Like the random undersampling technique, it is the fastest because it copies data without conducting separate calculations, but the performance varies greatly because it copies data randomly.

The SMOTE technique finds n nearest neighbors of a class with small data regarding certain data belonging to the same class with a small data size, draws a straight line with the neighbor, and generates points until the random points have a balanced ratio. The concept of sampling types is presented in Fig. 4. Moreover, the algorithm of SMOTE is presented in Fig. 5. The Python code for executing SMOTE is presented in Fig. 6.

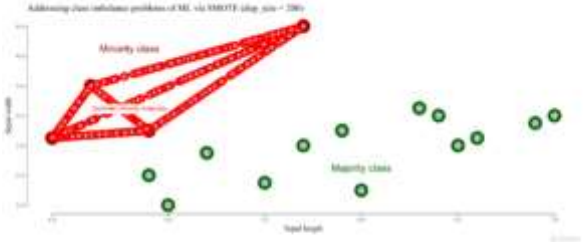
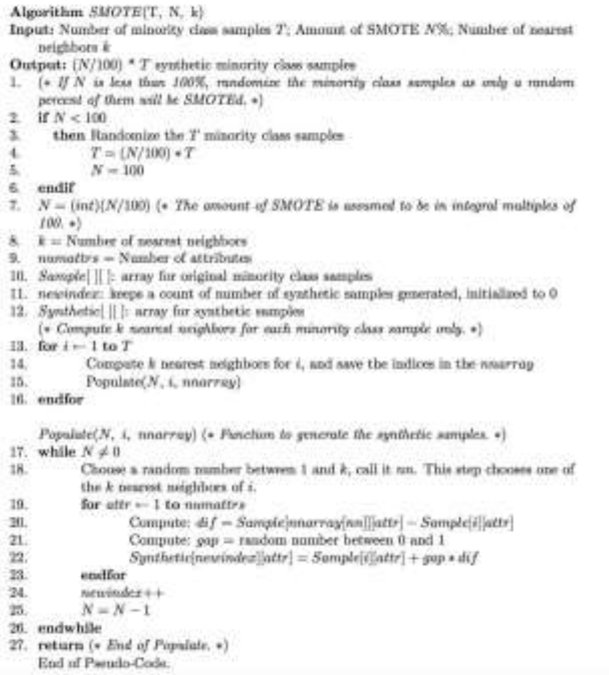


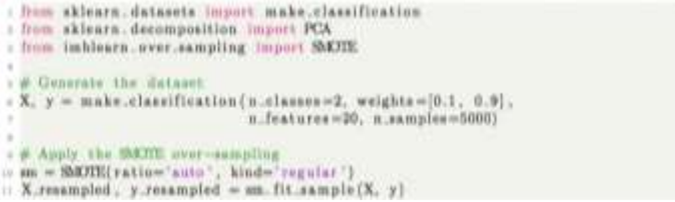
Fig. 4. Type of Sampling [39].



```
Algorithm SMOTE(T, N, k)
Input: Number of minority class samples T; Amount of SMOTE N%; Number of nearest
    neighbors k
Output: (N/100) * T synthetic minority class samples
(* If N is less than 100%, randomize the minority class samples as only a random
    percent of them will be SMOTEd. *)
if N < 100
    then Randomize the T minority class samples
        T = (N/100) * T
        N = 100
endif
N = (int)(N/100) (* The amount of SMOTE is assumed to be in integral multiples of
    100. *)
k = Number of nearest neighbors
numattrs = Number of attributes
Sample[ ][ ]: array for original minority class samples
newindex: keeps a count of number of synthetic samples generated, initialized to 0
Synthetic[ ][ ]: array for synthetic samples
(* Compute k nearest neighbors for each minority class sample only. *)
for i ← 1 to T
        Compute k nearest neighbors for i, and save the indices in the nnarray
        Populate(N, i, nnarray)
endfor

Populate(N, i, nnarray) (* Function to generate the synthetic samples. *)
while N ≠ 0
        Choose a random number between 1 and k, call it nn. This step chooses one of
        the k nearest neighbors of i.
        for attr ← 1 to numattrs
                Compute: dif = Sample[nnarray[nn]][attr] − Sample[i][attr]
                Compute: gap = random number between 0 and 1
                Synthetic[newindex][attr] = Sample[i][attr] + gap * dif
        endfor
        newindex++
        N = N − 1
endwhile
return (* End of Populate. *)
End of Pseudo-Code.
```

Fig. 5. The Algorithm of SMOTE.



```python
from sklearn.datasets import make_classification
from sklearn.decomposition import PCA
from imblearn.over_sampling import SMOTE

# Generate the dataset
X, y = make_classification(n_classes=2, weights=[0.1, 0.9],
                           n_features=20, n_samples=5000)

# Apply the SMOTE over-sampling
sm = SMOTE(ratio='auto', kind='regular')
X_resampled, y_resampled = sm.fit_sample(X, y)
```

Fig. 6. Code for Executing SMOTE in Python.

IV. Results

A. Comparing the Prediction Performance of the Model for Predicting Senile Depression

Table I shows the prediction performance (accuracy, sensitivity, and specificity) of oversampling, undersampling, and SMOTE. This study defined the final model with the best predictive performance as a model with the highest accuracy while sensitivity and specificity were 0.6 or higher. As a result, this study chose the SMOTE-based random forest algorithm, showing an accuracy of 0.68, a sensitivity of 0.83, and a specificity of 0.74, as the final model for predicting senile depression.

TABLE I. RESULTS OF THE PREDICTION PERFORMANCE (ACCURACY, SENSITIVITY, AND SPECIFICITY) OF OVERSAMPLING, UNDERSAMPLING, AND SMOTE

Type	Raw Data			Undersampling			Oversampling			SMOTE		
	Acc	Sen	Spe	Acc	Sen	Spe	Acc	Sen	Spe	Acc	Sen	Spe
LR	0.78	0.52	0.83	0.64	0.66	0.81	0.57	0.54	0.81	0.63	0.69	0.79
GBM	0.67	0.45	0.93	0.50	0.51	0.98	0.63	0.65	0.90	0.65	0.71	0.77
RF	0.73	0.65	0.88	0.61	0.75	0.92	0.78	0.64	0.91	0.68	0.83	0.74

Acc=accuracy; Sen=sensitivity; Spe=specificity; LR= Logistic regression; RF= Random forest

B. Major Predictors of Senile Depression

The model to predict the depression was developed through the GBM and the predictive power was compared with the results of random forest and logistic regression (Table II, Table III). Random forest 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 63.0% for logistic regression, 65.1% for GBM, and 68.3% for random forest. Table III shows the major predictors of senile depression according to the SMOTE algorithm.

TABLE II. NUMBER OF MAJOR DEPRESSION PREDICTORS BY THE ALGORITHM

Model	Factors
Logistic regression-raw data	8
GBM-raw data	10
Random forest-raw data	12
Logistic regression-undersampling	7
GBM-undersampling	10
Random forest-undersampling	12
Logistic regression-oversampling	6
GBM-oversampling	9
Random forest-oversampling	11
Logistic regression-SMOTE	8
GBM-SMOTE	10
Random forest-SMOTE	12

TABLE III. RESULTS OF MAJOR PREDICTORS OF SENILE DEPRESSION

Model	Characteristics
Random forest-SMOTE	Age, gender, educational level, economic activity, social activities for the past month, mean monthly household income, spouse living together, disease/accident/addiction in the last two weeks, subjective health status, subjective stress, the frequency of meetings neighbors, the frequency of meeting relatives.
Logistic regression-SMOTE	Age, gender, educational level, mean monthly household income, spouse living together, disease/accident/addiction in the last two weeks, subjective health status, subjective stress
GBM-SMOTE	Age, gender, educational level, social activities for the past month, mean monthly household income, spouse living together, disease/accident/addiction in the last two weeks, subjective health status, subjective stress, the frequency of meetings neighbors

V. CONCLUSION

This study compared the performance of ensemble-based machine learning sampling methods using the depression data of the elderly in the community, which had an imbalanced class ratio. The results of this study confirmed that the SMOTE-based random forest algorithm showing the highest accuracy (a sensitivity ⩾ 0.6 and a specificity ⩾ 0.6) was the final model with the best prediction performance among random forest, GBM, and logistic regression analysis. Since specificity and sensitivity have a trade-off relationship (when one value increases, the other value decreases), the ratio of specificity and sensitivity is selected according to the judgment of the researcher using a model. This study proposes to compare the performance of machine learning suitable for the study objective by considering accuracy, specificity, and sensitivity instead of considering only accuracy when future studies on prediction models will compare models and evaluate predictive performance.

This study compared the prediction performance of ensemble models built on imbalanced data by sampling method and found that SMOTE showed the best performance. Previous studies also reported that SMOTE had better predictive performance than undersampling and oversampling when analyzing imbalanced data [40]. The SMOTE technique has shown successful performance in various applied fields [41]. The ADASYN technique generates more realistic points deviated from the line by producing random points and adding random noise and it is a recently developed improved version of SMOTE. There have been continuous attempts to develop advanced algorithms that have better accuracy than SMOTE [42].

The results of this study suggest that using SMOTE as a sampling method to overcome the imbalance can be an efficient option when developing a prediction model using imbalanced binary data like disease data. SMOTE can alleviate the overfitting problem due to random oversampling and has the advantage of not losing useful data compared to undersampling or oversampling techniques [40]. However, it has also been reported that SMOTE may cause class overlapping, induce additional noise, and not be effective for treating imbalanced data with a high-dimensional y variable [42]. Therefore, although this study confirmed the effectiveness of SMOTE using an imbalanced binary dataset, the results cannot be generalized for all dimensions of data and the result should be interpreted with caution. Further studies are needed to compare the accuracy of SMOTE, undersampling, and oversampling for imbalanced data with high dimensional y-variables.

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] H. Byeon, Relationship between physical activity level and depression of elderly people living alone. International journal of environmental research and public health, vol. 16, no. 20, pp. 4051, 2019.

[2] H. Kang, F. Zhao, L. You, C. Giorgetta, D. Venkatest, S. Sarkhel, and R. Prakash, Pseudo-dementia: A neuropsychological review. Annals of Indian Academy of Neurology, vol. 17, no. 2, pp. 147-154, 2014.

[3] J. A. Sáez-Fonseca, J. A. Lee, and Z. Walker, Long-term outcome of depressive pseudodementia in the elderly. Journal of Affective Disorders, vol. 101, no. 1-3, pp. 123-129, 2006.

[4] M. H. Connors, L. Quinto, and H. Brodaty, Longitudinal outcomes of patients with pseudodementia: a systematic review. Psychological Medicine, vol. 49, no. 5, pp. 727-737, 2019.

[5] H. Byeon, Development of depression prediction models for caregivers of patients with dementia using decision tree learning algorithm. International Journal of Gerontology, vol. 13, no. 4, pp. 314-319, 2019.

[6] S. M. McClintock, M. M. Husain, T. L. Greer, and C. M. Cullum, Association between depression severity and neurocognitive function in major depressive disorder: a review and synthesis. Neuropsychology, vol. 24, no. 1, pp. 9-34, 2010.

[7] S. H. Kang, I. Y. Yoon, S. D. Lee, J. W. Han, T. H. Kim, and K. W. Kim, REM sleep behavior disorder in the Korean elderly population: prevalence and clinical characteristic. Sleep, vol. 36, no. 8, pp. 1147-1152, 2013.

[8] S. Lerche, A. Gutfreund, K. Brockmann, M. A. Hobert, I. Wurster, U. Sünkel, G. W. Eschweiler, F. G. Metzger, W. Maetzler, and D. Berg, Effect of physical activity on cognitive flexibility, depression and RBD in healthy elderly. Clinical Neurology and Neurosurgery, vol. 165, pp. 88-93, 2018.

[9] J. F. Gallegos-Orozco, A. E. Foxx-Orenstein, S. M. Sterler, and J. M Stoa, Chronic constipation in the elderly. American Journal of Gastroenterology, vol. 107, no. 1, pp. 18-25, 2012.

[10] S. K. Kahng, and B. K. Chung, Predictors of elderly depression using the andersen model. Korean Journal of Gerontological Social Welfare, vol. 49, no. 1, pp. 7-29, 2010.

[11] S. D. Chung, and M. J. Koo, Factors influencing depression: a comparison among babyboomers, the pre-elderly, and the elderly. Journal of Gerontological Social Welfare, vol. 52, no. 1, pp. 305-324, 2011.

[12] H. S. Jeon, and S. K. Kahng, Predictors of Depression trajectory among the elderly: using the Korean welfare panel data. Journal of the Korea Gerontological Society, vol. 29, no. 4, pp. 1611-1628, 2009.

[13] H. K. Han, and Y. R. Lee, A study on factors impacting on the mental health level of the elderly people living alone. Journal of the Korea Gerontological Society, vol. 29, no. 3, pp. 805-822, 2009.

[14] 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.

[15] 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.

[16] 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.

[17] A. Iqbal, S. Aftab, U. Ali, Z. Nawaz, L. Sana, M. Ahmad, and A. Husen, Performance analysis of machine learning techniques on software defect prediction using NASA datasets. International Journal of Advanced Computer Science and Applications, vol. 10, no. 5, pp. 300-308, 2019.

[18] J. Nobre, and 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.

[19] H. Nguyen, X. N. Bui, H. B. Bui, and 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, no. 2, pp. 477-490, 2019.

[20] L. Pourjafar, M. Sadeghzadeh, and M. Abdeyazdan, Combination of neural networks and fuzzy clustering algorithm to evalution training simulation-based training. International Journal of Advanced Computer Science and Applications, vol. 7, no. 7, pp. 31-38, 2016.

[21] H. Byeon, Application of machine learning technique to distinguish Parkinson’s disease dementia and Alzheimer’s dementia: predictive power of Parkinson’s disease-related non-motor symptoms and neuropsychological profile. Journal of Personalized Medicine, vol. 10, no. 2, pp. 31, 2020.

[22] H. Byeon, Is the random forest algorithm suitable for predicting Parkinson’s disease with mild cognitive impairment out of Parkinson’s disease with normal cognition?. International Journal of Environmental Research and Public Health, vol. 17. no. 7, pp. 2594, 2020.

[23] A. Fernández, S. García, F. Herrera, and 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.

[24] S. Shrivastava, P. M. Jeyanthi, and S. Singh, Failure prediction of Indian Banks using SMOTE, Lasso regression, bagging and boosting. Cogent Economics & Finance, vol. 8, no. 1, e-pub. 1729569, 2020.

[25] S. Lee, S. T. Oh, S. Y. Ryu, K. Lee, E, Lee, J. Y. Park, S. W. Yi, and W. J. Choi, Validation of the Korean version of Center for Epidemiologic Studies Depression Scale-Revised (K-CESD-R), Korean Journal of Psychosomatic Medicine, vol. 24, no. 1, pp. 83-93, 2016.

[26] M. Skurichina, and R. P. Duin, Bagging, boosting and the random subspace method for linear classifiers. Pattern Analysis and Applications, vol. 5, no.2, pp. 121-135, 2002.

[27] H. Byeon, Exploring the predictors of rapid eye movement sleep behavior disorder for Parkinson’s disease patients using classifier ensemble. In Healthcare, vol. 8, no. 2, pp. 121, 2020.

[28] B. Hyeon, Development of a depression in Parkinson's disease prediction model using machine learning. World Journal of Psychiatry, vol. 10, no. 10, pp. 234-244, 2020.

[29] Y. Grandvalet, Bagging equalizes influence. Machine Learning, vol. 55, no. 3, pp. 251-270, 2004.

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

[31] F. Tang, and 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.

[32] P. T. Noi, and 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, pp. 18, 2018.

[33] H. Byeon, Best early-onset Parkinson dementia predictor using ensemble learning among Parkinson's symptoms, rapid eye movement sleep disorder, and neuropsychological profile. World Journal of Psychiatry, vol. 10, no. 11, pp. 245-259, 2020.

[34] R. Kumar, Machine Learning quick reference: quick and essential machine learning hacks for training smart data models. Packt Publishing, Birmingham, 2019.

[35] A. Sharma, Decision Tree vs. Random Forest – Which Algorithm Should you Use?. Analytics Vidhya, Gurgaon, 2020.

[36] J. H. Friedman, Greedy function approximation: a gradient boosting machine. Annals of statistics, vol. 29, no. 5, pp. 1189-1232, 2001.

[37] L. Abdi, and 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, 2016.

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

[39] R. Kunert, Smote explained for noobs-synthetic minority over-sampling technique line by line. Rich Data, Berlin.2017.

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

[41] B. Santoso, H. Wijayanto, K. A. Notodiputro, and B. Sartono, Synthetic over sampling methods for handling class imbalanced problems: a review. IOP Conference Series: Earth and Environmental Science, vol. 58, no. 1, pp. 012031, 2017.

[42] M. Karajizadeh, M. Nasiri, M. Yadollahi, A. H. Zolfaghari, and A. Pakdam, Mortality prediction from hospital-acquired infections in trauma patients using an unbalanced dataset. Healthcare Informatics Research, vol. 26, no. 4, pp. 284-294, 2020.